UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE

EVALUACIÓN DE ESTIMADORES BASADOS EN MODELOS PARA EL CÁLCULO DEL RIESGO DE CRÉDITO BANCARIO EN ENTIDADES FINANCIERAS

Tesis doctoral presentada por:

Marta Vaca Lamata

Programa de Doctorado en
Gestión en Recursos Humanos, Trabajo y Organizaciones

Dirigida por:

Dr. Agustín Pérez Martín
Departamento de Estudios Económicos y Financieros
Universidad Miguel Hernández de Elche

Dr. Alejandro Rabasa Dolado
Departamento de Estadística Matemáticas e Informática
Universidad Miguel Hernández de Elche

UNIVERSITAS
Miguel
Hernández

UNIVERSITAS
Miguel
Hernández

UNIVERSITAS
Miguel
Hernández

EVALUACIÓN DE ESTIMADORES BASADOS EN MODELOS PARA EL CÁLCULO DEL RIESGO DE CRÉDITO BANCARIO EN ENTIDADES FINANCIERAS

Marta Vaca Lamata

Memoria presentada por
Marta Vaca Lamata
para optar al grado de doctor por la
Universidad Miguel Hernández de Elche.

Elche, septiembre 2017.

Directores:

Dr. Agustín Pérez Martín

Dr. Alejandro Rabasa Dolado

UNIVERSITAS
Miguel
Hernández

D. Agustín Pérez Martín, profesor Doctor del Departamento de Economía Financiera y Contabilidad de la Universidad Miguel Hernández de Elche y D. Alejandro Rabasa Dolado, profesor Doctor del Departamento de Estadística Matemáticas e Informática de la Universidad Miguel Hernández de Elche,

CERTIFICAN:

Que la memoria de investigación titulado "**Evaluación de estimadores basados en modelos para el cálculo del riesgo de crédito bancario en entidades financieras**", realizado por Dña. Marta Vaca Lamata ha sido llevado a cabo bajo nuestra dirección, y se encuentra en condiciones de ser leído y defendido como Tesis Doctoral en la Universidad Miguel Hernández de Elche.

Lo que firmamos para los efectos oportunos en Elche, julio de 2017.

D. Agustín Pérez Martín

D. Alejandro Rabasa Dolado

UNIVERSITAS
Miguel
Hernández

D. Juan Carlos Marzo Campos, director del Departamento de Psicología de la Salud de la Universidad Miguel Hernández de Elche.

AUTORIZA:

La presentación y defensa de la Tesis Doctoral titulada "**Evaluación de estimadores basados en modelos para el cálculo del riesgo de crédito bancario en entidades financieras**", realizado por Dña. Marta Vaca Lamata, bajo la dirección y supervisión de los doctores D. Agustín Pérez Martín, profesor del Departamento de Economía Financiera y Contabilidad de la Universidad Miguel Hernández de Elche y de D. Alejandro Rabasa Dolado, profesor del Departamento de Estadística Matemáticas e Informática de la Universidad Miguel Hernández de Elche.

Lo que firmo en Elche, julio de 2017

Fdo. Dr. Juan Carlos Marzo Campos
Director del Departamento de Psicología de la Salud

UNIVERSITAS
Miguel
Hernández

A mi marido e hijos

A mis padres, hermanos y sobrinos

UNIVERSITAS
Miguel
Hernández

Llibertat per triar,
il·lusió per arribar,
companyia per caminar,
constància per millorar
i esforç per superar.

UNIVERSITAS
Miguel
Hernández

Índice general

UNIVERSITAS
Miguel
Hernández

UNIVERSITAS
Miguel
Hernández

Agradecimientos

Voldria expressar el més sincer agraïment a totes les persones que d'una forma o un altra han permés dur a terme aquest treball, així com les què han contribuït a fer-me qui sóc. Cadascuna ha ajudat en moments puntuals o al llarg de la meua vida, sempre han aportat, han sumat. A totes elles, gràcies per les vostres aportacions, suggeriments, ànims, cafés i sonmriures. I a les què no han confiat, gràcies per donar-me l'empenta necessaria per a abastir les metes.

En primer lloc, m'agradaria dedicar aquest treball als meus pares. Tot el què sóc és un deute amb ells. Demetrio i Marta donaren tot sense demanar res a canvi, hui no estaria ací sense tot l'esforç que han realitzat al llarg de les seues vides, els seus sacrificis, els seus ensenyaments, la seua dedicació als fills, el seu interés per *"traure profit de la seua filla"*. Sempre és molt prompte per a què les persones més estimades deixen de ser la teua guia, però estic segura que hi ha una llumeneta al cel que els deixa veure'm i em dóna llum cada minut de la meua existència.

Ara sí que puc passar al terrenal, el més sincer agraïment al Director d'aquesta tesi, el Dr. D. Agustín Pérez Martín. *Perdón Agustín, en castellano te suena mejor aunque como te he dicho alguna vez hay palabras que son más bonitas dichas en valenciano, muchas gracias.* Gràcies per la teua dedicació, ànim i tenacitat perquè el treball isquera endavant. El teu recolzament, empenta, insistència, els teus valuosos consells i aportacions (tant personals com laborals). Vas ser la primera persona que vaig conèixer en arribar a aquesta Univesitat, i puc dir que des de el principi m'he sentit acompanyada i aconsellada. He rebut més del que t' he donat, però espere que tot siga qüestió de temps. Gràcies per tots els comentaris crítics, constructius, sense els quals aquest treball no fós possible. També al director Dr. Alejandro Rabasa Dolado, gràcies per les teues aportacions, sinceritat i contribució a aquesta tesi. I com no, a l'estudiant incansable, Agustín Pérez Torregrosa, una persona clau per la finalització d'aquesta tesi. Les seues aportacions han sigut molt valuoses i sempre dispost a fer un nou suggeriment per la millora de la mateixa fins al darrer minut.

Gràcies a Domingo Morales i Lidia Ortiz Henarejos, per oferir-me la oportunitat de refrescar part de l'estadística oblidada i ensenyar-me'n molt més, permetent-me la assistència a les seues classes.

Gràcies a Javier Toledo, per la seua ajuda matemàtica.

Gràcies a César Pérez López, de l'Institut d'Estudis Fiscals, Ministeri d'Hisenda i Funció Pública per la cessió de la base de dades.

Als meus companys de l'Àrea d'Economia Financera i Comptabilitat per l'ajuda prestada al llarg d'aquests anys. Molt especialment al Catedràtic Dr. D. José Francisco González Carbonell,

pel seu recolzament i facilitats perquè aquest treball arribés a bon termini. També als meus companys de Departament d'Estudis económics i Financers.

A Ángel Solanes Puchol com a responsable del Programa de Doctorat en Gestió de Recursos Humans, Treball i Organiztacions per donar-me l'oportunitat de finalitzar la tesi amb ells, i a l'equip directiu i administratiu del Departament de Psicologia de la Salut per facilitar-me els tràmits adients.

També m'agradaria dedicar aquest treball als meus excompanys de treball i amics de l' I.U. Biodiversitat CIBIO de la Universitat d'Alacant i l'Excm. Ajuntament de la Vila d'Ibi. Gràcies a tots ells pel seu recoltzament constant, tant quan compartíem hores de treball com després. Gràcies per la comprensió que tinguéreu quan vaig triar fer aquest pas, els ànims que em donàreu, el *"fins a sempre"* que em dedicàreu i el temps que m'heu regalat.

A la meua germana Matilde, "Titi", el meu gran suport incondicional, sempre present quan se la necessita i segurament la persona que menys ha demanat i més ha donat. Poques coses podria fer sense ella. Gràcies per la teua dedicació a la meua familia i la teua preocupació per mi. També un agraïment pels meus germans, nebots i la meua sogra, M^{a}Carmen, qui amb la seua alegria i bon humor trau un somriure a qualsevol.

A totes les amigues i amics, que d'una manera o altra han participat en aquest treball, se n'han interessat i sobretot han desitjat veure'l finalitzat. *"No patiu, ho celebrarem prompte!"*.

De forma molt especial vull expresar la gratitud més gran a la meua familia. El meu marit Jordi i els meus fills Elies i Arnau. Aquest treball no tindria cap sentit ni valor sense tots tres. Són la font de la meua inspiració, els qu em donem la força per continuar, pels què val la pena tot l'esforç i sacrifici que suposa dedicar el teu temps a aquesta tesi. La vostra estima, la vostra espera, el vostre ànim, per estar sempre al meu costat, per comprendre'm, confiar en mi, pel vostre recolzament incondicional i les vostres paraules d'alé. Gràcies per la teua paciència Jordi.

Per últim, agrair el suport económic rebut de les següents institucions i la cessió de la base de dades, per aquesta investigació:

* La Universitat Miguel Hernández d'Elx amb les ajudes per Grups d'Investigació emergents de "Projectes d'Investigació en Humanitats i Ciències Socials" de l'any 2013.

* Ajudes per Grups d'Investigació emergents de la Conselleria d'Educació de la Generalitat Valenciana (GVA/2016/053).

* L'Institut d'Estudis Fiscals, Ministeri d'Hisenda i Funció Pública.

Prólogo

Introducción

El crédito bancario es una pieza clave de la economía moderna y necesario para el desarrollo económico de cualquier país. Las entidades financieras son las encargadas de hacer llegar los flujos de efectivo a los agentes económicos mediante la concesión de préstamos. El objetivo de cualquier entidad financiera es maximizar su beneficio derivado de la concesión de un volumen máximo de crédito, pero al mismo tiempo están sujetas a normas emitidas por los Bancos Centrales y recomendaciones de organismos internacionales.

La concesión de un préstamo implica un riesgo para la entidad financiera, el riesgo viene dado por la posibilidad de que el prestatario no cumpla con la devolución del mismo, es decir la posibilidad de que exista morosidad. Las entidades financieras se enfrentan a un problema de decisión, conceder o no conceder el préstamo al cliente, que implica una valoración de la probabilidad que tiene cada solicitante de presentar problemas de morosidad. Por tanto, trata de calcular un suceso incierto con datos del pasado; es decir, es un problema de estimación o predicción denominado **credit scoring**.

El credit scoring trata de estimar las probabilidades de incumplimiento por parte del cliente con el préstamo en base a las características personales del individuo y del crédito que solicita, utilizando como información inicial el comportamiento de otros individuos que han recibido previamente un crédito en condiciones similares y del propio historial crediticio del cliente.

El crédito está vinculado al crecimiento económico, pero las políticas monetarias de los países o de la Unión Europea limitan y controlan la cantidad de dinero inyectada en el sistema. Por tanto, existen limitaciones legales en cuanto al dinero en circulación. Al mismo tiempo, todo préstamo conlleva un riesgo de incumplimiento. Las entidades financieras necesitan disponer de un buen historial de crédito y métodos para medir el riesgo. Cuanto más eficientes y eficaces sean los métodos de predicción mejor asignación de los recursos, menores costes y mayor facilidad de acceso a los créditos por parte del consumidor.

Desde comienzos de los años 60 la demanda de créditos bancarios ha venido experimentando un crecimiento exponencial, tanto en número como en cuantía. En sus inicios el método utilizado por las entidades financieras se basó en el juicio personal del analista. Conforme aumentaban las solicitudes de crédito para las entidades financieras era inviable analizar todas y cada una de ellas de forma personal. Se necesitaba disponer de un método estadístico para conocer la exposición al riesgo de insolvencia, un sistema experto. Se comenzó a aplicar técnicas de puntaje

de crédito, pero hasta los años 80 las técnicas de calificación crediticia no se extendieron a los préstamos. Su uso se generalizó en los 90, debido tanto a la mejora de los recursos informáticos como estadísticos. Algunas entidades financieras utilizaron ténicas estadísticas para optimizar la diferenciación entre préstamos con probabilidad de éxito o fracaso (default). Estas técnicas de evaluación se denominan técnicas de credit scoring e incluyen todo sistema de evaluación crediticia que permita valorar de forma automática el riesgo asociado a cada solicitud de crédito. El riesgo está en función de la solvencia del deudor, del tipo de crédito, de los plazos y de otras características propias del cliente y de la operación. Se trata de sistemas objetivos en el que la aprobación del crédito solicitado no depende de la discrecionalidad del analista y, además al ser un sistema automático va a permitir reducir costes y tiempo de tramitación. En este punto cabe destacar 2 tipos, por un lado los sistemas expertos, que se encargan automáticamente de tomar la decisión sobre el crédito, y por otro lado los sistemas de apoyo a la decisión, que después de valorar la operación dejan la decisión final en manos de la persona que la entidad financiera disponga para la concesión.

Las entidades crediticias están sometidas a lo que se llama "política prudencial", que se entiende por el número de recursos propios que estas deben mantener para asegurar su buen funcionamiento y cubrir los diferentes riesgos a los que están expuestos, entre ellos el riesgo de crédito. A finales de los años 80, el Comité de Basilea adoptó una serie de recomendaciones para la supervisión de las entidades de crédito con actividad internacional agrupadas en el documento conocido como "Basel Core Principles" que desarrolla los principios básicos de supervisión bancaria. Más tarde, en 1988, aparecen nuevas recomendaciones del Comité de Basilea, conocido como Basilea I que se centraban en la supervisión del riesgo de crédito.

En el año 2004 aparecieron las recomendaciones del Comité de Basilea (llamado Basilea II) sobre la supervisión bancaria. Desde entonces, el uso de métodos avanzados de puntuación de crédito se ha convertido en un requisito regulatorio para los bancos y las instituciones financieras con el fin de mejorar la eficiencia en la asignación de capital. Con la crisis financiera global, aparece un nuevo documento, Basilea III. Este documento pretende dotar al sistema financiero de una mayor estabilidad introduciendo cambios más exigentes en el control del capital prestado. La dificultad está en distinguir los solicitantes solventes y los que entrarán en mora, convirtiéndose en un problema de predicción. En consecuencia, la mejora en la precisión de la evaluación del riesgo de crédito, aunque sea pequeña, supone un potencial beneficio para las instituciones financieras.

Aunque los acuerdos de Basilea III no sean vinculantes para los países, los Bancos Centrales sí que han procedido a regular y endurecer significativamente los requerimientos de capital, aumentando las reservas en relación con la exposición al riesgo y gestión del mismo. Como consecuencia de dicha regulación, las técnicas de clasificación que se utilizan para desarrollar modelos precisos en la evaluación del credit scoring toman una gran importancia. Por tanto, la exactitud en la evaluación del riesgo de crédito además de ser un beneficio potencial para las instituciones financieras, se convierte en una exigencia por parte de las autoridades monetarias de los países.

Cabe añadir un problema no abordado hasta el momento, que tiene que ver con el volumen de las bases de datos de las entidades financieras. En los últimos años se ha producido un incremento de las operaciones financieras y por tanto también ha aumentado de modo exponencial el volumen de las bases de datos que las entidades manejan, no sólo para la gestión propia del negocio bancario, sino también de la que se nutren para el cálculo de probabilidades de riesgo en base a su experiencia. De este tipo de aumentos en el volumen de los datos nace el "**BigData**". "BigData" es el término que se utiliza para acuñar este nuevo problema al que se enfrentan actualmente todo tipo de empresas, y como no podía ser menos también en las entidades financieras. BigData viene a ayudar en cómo analizar todo ese volumen de datos para extraer su valor y tomar así mejores decisiones, sin que la componente tiempo de ejecución suponga un coste tan elevado que haga inabordable el problema. Por tanto, el coste computacional que requiere el procesamiento de los datos es un factor clave a tener en cuenta. Diferentes estudios revelan que grandes conjuntos de datos necesitan técnicas y algoritmos BigData, mediante computación en la nube, que produzcan no sólo estimadores precisos, si no que también sean eficientes computacionalmente, es decir, que sean capaces de dar resultados en tiempos razonables.

Si las entidades financieras son capaces de disponer de un método **eficiente** y **eficaz** para predecir el éxito o fracaso del pago en la concesión de un préstamo hipotecario, obtendrán mayores beneficios y más agentes tendrán acceso a los mismos. Si se aplica un método eficiente, se reducirá el tiempo de ejecución y, en consecuencia, el cliente tendrá la respuesta más rápida y, se podrán analizar mayor número de posibles clientes en menos tiempo.

La investigación presentada en esta memoria de tesis se centra en calcular las probabilidades de incumplimiento en los préstamos. En primer lugar, se realiza con experimentos computacionales de simulación Monte Carlo con distintos algoritmos y técnicas. El objetivo es medir la eficacia y eficiencia de cada uno de los métodos estadísticos y de minería de datos. Una vez comprobado en muestras sintéticas se realiza los mismos experimentos con una muestra semi-real y se comparan los resultados obtenidos. Por último se comparan los resultados con dos muestras reales.

Objetivos

En esta subsección se procede a enumerar los objetivos específicos que se pretenden alcanzar que posteriormente serán objeto de comprobación en cada uno de los diferentes capítulos.

Como principal objetivo, se pretende introducir la estimación del riesgo de crédito mediante modelos lineales mixtos (LMM), que hasta la fecha nunca se habían empleado en esta rama de las finanzas, y que en otras áreas científicas han dado muy buenos resultados. Posteriormente a la introducción se realizará una comparación con otros métodos y técnicas comúnmente empleados en la estimación del riesgo de crédito.

De manera más detallada se plantean los objetivos específicos siguientes:

1. Explorar en la literatura del tema cuales son los métodos comúnmente aceptados tanto

por entidades financieras, como en el ámbito académico.

2. Estudiar la robustez de los modelos lineales generalizados para la estimación del riesgo de crédito, así como el método de ajuste más adecuado dentro de estos.

3. Estudiar diferentes métodos de estimación del riesgo de crédito para su mutua comparación, atendiendo tanto a eficacia en la estimación como a eficiencia computacional, y que sirva para poder discriminar entre ellos bajo el efecto de presencia de grandes volúmenes de datos, tanto en datos sintéticos como la comprobación en su aplicación a casos reales.

4. Comprobar si los métodos de selección de variables consiguen una reducción en la dimensión de los ficheros de datos masivos y por tanto en la eficiencia computacional, sin que se produzca un significativo retroceso en la eficacia.

Contenido de la memoria

A lo largo de esta memoria de tesis doctoral, se realiza un estudio comparativo de diversos modelos estadísticos para la estimación de la probabilidad de éxito o fracaso en atender los pagos de un préstamo. Una vez comprobado el modelo a utilizar se estudian distintos métodos para la predicción del posible impago de un cliente de un préstamo.

La investigación pretende aportar conocimiento sobre el método o métodos más eficaces y eficientes para la predicción del impago del préstamo, mediante la comparación bajo distintas condiciones de estos, todo ello bajo la presencia del inconveniente de disponer de grandes volúmenes de datos. El objetivo de una entidad financiera es conseguir una mejor redistribución de los recursos. Si se reducen los impagos se puede disponer de más recursos, por tanto se puede aumentar el volumen de préstamos. Al mismo tiempo disminuyen los costes y se obtienen mayores beneficios.

La presente memoria consta de siete capítulos y un apéndice distribuidos del siguiente modo:

El **Capítulo 1**. Antecedentes y estado actual del tema. Se realiza un estudio de la evolución y estado actual de los distintos métodos que se han aplicado para la predicción del riesgo de crédito.

El **Capítulo 2**. Métodos estadísticos para la estimación del riesgo de crédito. Se analizan cinco métodos estadísticos para su aplicación. Los métodos analizados son paramétricos y no paramétricos como, el análisis discriminante, árboles de decisión y clasificiación CART; métodos estadísticos como, modelo lineal mixto y modelo lineal generalizado; y técnicas de inteligencia artificial como máquina de vectores soporte.

El **Capítulo 3**. Evaluación de la robustez en modelos lineales. Se compara el modelo lineal con efectos fijos y el modelo mixto, con efectos fijos y aleatorios, ajustados por máxima verosimilitud y máxima verosimilitud residual. También se comprueba el efecto que tienen en los distintos modelos la homocedasticidad y heterocedasticidad. Se comparan los mismos modelos

bajo los supuestos de normalidad de los errores, con errores distribuidos Weibull y Gamma. Se realizan 3 experimentos de simulación Monte Carlo para ver las bondades de cada uno de ellos.

El **Capítulo 4**. Evaluación de métodos estadísticos para la estimación del riesgo de crédito. Con bases de datos simuladas y bajo el modelo que resulta óptimo en los experimentos anteriores se aplican 5 métodos distintos de predicción para calcular el método más eficaz y eficiente. Se analizará la eficacia a través de la tasa de acierto y el error cuadrático medio. Los métodos que se analizan son los estudiados en el capítulo 2.

El **Capítulo 5**. Aplicación a datos semi-reales y reales. En este capítulo se aplica la misma metodología anterior a una base de datos semi-real y dos bases de datos reales, *German Credit* y *Australian Credit*. El objetivo es comprobar si las conclusiones obtenidas con las bases de datos sintéticas son trasladables a una base de datos semi-real y real para su aplicación en entidades financieras.

El **Capítulo 6**. Selección de variables. Una vez obtenidos los métodos más eficientes y eficaces se realizan pruebas para seleccionar variables con el objetivo de conseguir minimizar el número de variables que intervienen en el modelo sin perder eficacia en el método y ganar eficiencia computacional.

El **Capítulo 7**. Algunas conclusiones de carácter general y planteamiento de futuras líneas de investigación.

El **Apéndice A**. Se presentan tablas con valores numéricos correspondientes a la realización de los experimentos realizados a lo largo de la tesis.

UNIVERSITAS
Miguel
Hernández

1

Antecedentes y estado actual del tema

1.1. Introducción

En un mundo globalizado que se cimienta en el comercio, existen millones de transacciones entre actores económicos. En las economías desarrolladas, la actividad de las empresas, la inversión y el consumo de las familias se sustentan sobre la base del crédito, siendo muy relevante el papel de las entidades de crédito en este proceso. El crédito surge de forma natural como consecuencia de la actividad económica y se consolida como una condición necesaria para el crecimiento de la misma.

Aunque sería deseable que el resultado de la actividad económica se conociera con certeza, la incertidumbre está siempre presente y, más aún, en el negocio financiero. Por tanto la actividad bancaria se ve afectada muy seriamente por esa componente de incertidumbre: el riesgo.

Existen diversos tipos de riesgos asociados a la actividad bancaria (Crédito, Mercado, Liquidez, Operacional, País, Reputacional, de Cambio, Regulatorio, Comercial ...). Sin duda el Riesgo de Crédito es uno de los riesgos financieros soportados más elevados (más de la mitad del total de riesgos asumidos por cualquier entidad financiera) y ello es debido principalmente a la esencia de la industria bancaria.

El Riesgo de Crédito está especialmente vinculado a la operativa de los intermediarios financieros y presente en la mayor parte de sus operaciones de activo. El Riesgo de Crédito surge cuando sobreviene, de la actividad crediticia, la posibilidad de sufrir una pérdida, como consecuencia del incumplimiento de la otra parte de no asumir el pago o pagos acordados, bien por incapacidad o por falta de disposición. La existencia del riesgo crediticio depende pues de la solvencia y compromiso del deudor, pero a la vez la magnitud de esta pérdida está asociada al volumen de la operación.

El uso de métodos que permitan calcular todo esto es de vital importancia en la industria financiera y claramente afecta a la economía global, tal y como se demostró en los años 2007 y 2008 con la caída de los grandes bancos en Estados Unidos. Ello llevó a una reacción en cadena que afectó a la economía del país, y a los países desarrollados del mundo, llegando incluso algunos a la quiebra o siendo rescatados por ayudas de otros (Unión Europea). La titulización de activos inmobiliarios de dudoso origen y por tanto de dudosa calidad crediticia, cuyo subyacente eran préstamos con garantía hipotecaria de la que no se sabía su riesgo de crédito, llevó a aquellos en primer lugar perdiesen su valor, para después caer en impago por quiebra de la entidad emisora. Todo esto pese a que la industria financiera ha estado sometida a una política prudencial materializada en regulación por cada país que asumiese las recomendaciones de Basilea I, II y III.

Debido a este efecto en la economía de los países, su influencia en las personas y que la regulación del sistema no ha sido suficiente, determinados organismos a nivel nacional y supranacional han tenido que emitir nuevas normas que garanticen estabilidad a las cuentas de la industria bancaria. Entre ellos la IASB (International Accounting Standards Board) con los estándares contables internacionales, y más concretamente la NIIF 9, IFRS 9, (DOUE, 2016), cuya aplicación en la UE será obligatoria a partir del 1 de enero de 2018, y que dejará sin efecto la actual NIC 39. O la actual Circular 4/2016 de 4 de abril del Banco de España (BOE, 2016) por la que se modifican la Circular 4/2004, de 22 de diciembre, a entidades de crédito, sobre normas de información financiera pública y reservada y modelos de estados financieros (BOE, 2004) y la Circular 1/2013, de 24 de mayo, sobre la Central de Información de Riesgos (BOE, 2013).

La medición del riesgo de crédito por una entidad bancaria es una tarea que presenta gran dificultad, pero a pesar de ello, en los últimos años han sido numerosos los bancos, cajas de ahorro y cooperativas de crédito que se han interesado por herramientas de medición para establecer controles que minimicen el daño que este tipo de riesgo ocasiona a la cuenta de resultados. Es más, determinados requisitos regulatorios de solvencia (acuerdos del Comité de Basilea, normas internacionales como la NIIF 9 y las anteriores Circulares del Banco de España) hacen necesario pasar por el uso de herramientas que permitan dimensionar este riesgo.

Las entidades financieras necesitan dotarse de herramientas que les permitan la medición y el diagnóstico posterior de futuras operaciones. A todo sistema de evaluación crediticia que permite valorar de forma automática el riesgo asociado a cada solicitud de crédito se le denomina **credit scoring**. El riesgo estará en función de la solvencia del deudor, del tipo de crédito, de los plazos y de otras características propias del cliente, de la operación, de otras muchas magnitudes inherentes a la operación financiera, y todo ello contextualizado en el momento en el que se produce.

En los últimos años, se ha producido un incremento de las operaciones financieras y por tanto también ha aumentado de modo exponencial el volumen de las bases de datos que las entidades financieras manejan, no sólo para la gestión propia del negocio bancario, sino también de la que se nutren para el cálculo de probabilidades de riesgo en base a su experiencia. De

este tipo de aumentos en el volumen de los datos nace el "BigData". "BigData", "Grandes datos", o "datos masivos", es el término que se utiliza para acuñar este nuevo problema al que se enfrentan actualmente todo tipo de empresas, y como no podía ser menos también las entidades financieras. BigData viene a ayudar en cómo analizar todo ese volumen de datos que se dispone para extraer su valor y tomar así mejores decisiones, sin que la componente tiempo de ejecución suponga un coste tan elevado que haga inabordable el problema. Tal cantidad de datos suelen ser computacionalmente difíciles de tratar, por tanto, la computación en la nube (cloud computing), con este tipo de técnicas BigData, se hace imprescindible, a la par que permite acceder a cualquier interesado en la organización a los resultados que de este proceso se deriven.

Pese a su corta historia, son muy numerosas las aportaciones que se han hecho al riesgo de crédito. Desde los cercanos años 40 que se introdujo por primera vez, hasta la actualidad. Además como se podrá ver a lo largo de esta memoria, y más concretamente en este capítulo, no es un tema cerrado en cuanto a investigación científica y tampoco en cuanto a su aplicabilidad en el sector financiero, a tenor de lo sucedido en el sector inmobiliario, en la titulización de activos procedentes de este sector, y en la economía en general como consecuencia de la caída de éstos.
Los primeros sistemas para la decisión de concesión de crédito eran completamente rudimentarios y para nada apoyados científicamente. Por poner un ejemplo JP Morgan comenzó a conceder créditos a las personas que ya disponían de una tarjeta de crédito, pues entendían que ya habían sido evaluadas crediticiamente para la concesión de la tarjeta.
En este capítulo se enumeran, ubican y contextualizan las aportaciones científicas que se han producido en la línea de investigación.

1.2. Antecedentes

El riesgo de crédito ha ido tomando importancia desde mediados del siglo XX (principio de los años 40) hasta nuestros días. En sus comienzos era un simple método que se realizaba en la misma oficina bancaria a juicio personal del analista. Según señalan Abrahams and Zhang (2008), en los comienzos del análisis del riesgo de crédito, las decisiones se tomaban en función de las 3 Cs, 4 Cs o 5 Cs (carácter, capital, garantía, capacidad y condición). Pero este método se descartó porque supuso altos costes de capacitación, decisiones incorrectas e inconsistentes.

El primer método estadístico que se empezó a utilizar fue el análisis discriminante, basado en las aportaciones de Fisher (1936) por Durand (1941), que diferenció entre buenos y malos créditos. Fisher (1936) fue el primero en diferenciar estadísticamente dos grupos de población para realizar estudios taxonómicos, el objetivo del autor no era analizar el riesgo de crédito.

Con el boom económico de los años 60 en EEUU se produce una demanda masiva de tarjetas de crédito. Según Thomas (2000) y Myers and Forgy (1963) es el inicio de la aplicación de técnicas de credit scoring.

En la década de los 80 coinciden varios factores que influyen en el avance de las técnicas de credit scoring. Se produce un gran avance computacional, el aumento del capital prestado por las entidades financieras y que ya se habían probado las técnicas de credit scoring en la concesión de tarjetas de crédito con éxito. Según Marqués et al. (2013), el primero en aplicar credit scoring para los préstamos (tanto hipotecarios, como de consumo y a pequeñas empresas) fue Altman (1968).

A finales de los 90, en la economía mundial se produjo un gran aumento de la demanda del crédito y con ello la necesidad de aplicar técnicas de detección del riesgo de crédito más precisas. Al mismo tiempo, se produce una crisis del Sistema Monetario Europeo en 1992, junto con una crisis bancaria en México que tiene repercusión en el resto de países. Como consecuencia de los acontecimientos negativos en la economía bancaria aparece el primer documento de recomendaciones de supervisión de riesgo de crédito emitido por el Comité de Basilea de Supervisión Bancaria denominado Basilea I.

A partir del año 2004, y como continuación a la supervisión iniciada a nivel europeo, se amplían las recomendaciones emitidas por el Comité de Basilea de Supervisión Bancaria en el documento Basilea II. Se sugiere la utilización de métodos avanzados de credit scoring, se convierte en un requisito regulatorio para los bancos y entidades financieras con el fin de mejorar la eficiencia en la asignación de capital.

Poco más tarde, debido a la crisis financiera a nivel mundial del año 2007, la evaluación del riesgo de crédito se convierte en una cuestión sumamente importante en la gestión de las entidades financieras. Aparece una nueva regulación del Comité de Basilea de Supervisión Bancaria, Basilea III, que introduce nuevos cambios y mayores exigencias de control sobre el capital prestado por las entidades financieras y su consiguiente aumento de reservas en función de sus riesgos. Basilea III cumple el objetivo de dotar al sistema financiero de una mayor estabilidad. La dificultad está en distinguir los solicitantes solventes y los que entrarán en mora, convirtiéndose en un problema de predicción. En consecuencia, la mejora en la precisión de la evaluación del riesgo de crédito, aunque sea pequeña, supone un potencial beneficio para las instituciones financieras. Aunque los acuerdos de Basilea I, II y III no sean vinculantes para los países, los Bancos Centrales entre otras medidas aplicadas sí han procedido a regular y endurecer significativamente los requerimientos de capital y gestión del riesgo. Como consecuencia de dicha regulación, del incremento de las operaciones financieras y por tanto el aumento de sus bases de datos, las técnicas de clasificación que se utilizan para desarrollar modelos precisos en la evaluación del credit scoring toman mucha importancia. Al mismo tiempo, las entidades financieras se ven en la necesidad de aplicar métodos más precisos y eficaces.

Las aportaciones científicas a esta línea de investigación son muy numerosas, tal y como se puede observar en el capítulo de referencias bibliográficas. Es por este motivo, por el que en este capítulo de antecedentes y estado actual del tema, se ha optado por agrupar las investigaciones previas por la metodología empleada para el cálculo del riesgo de crédito o como en la mayoría de los casos, la probabilidad de mora o también llamada probabilidad de default.

Con este fin, a continuación se detallan dos artículos que realizan una exhaustiva revisión metodológica englobando técnicas y métodos en grandes grupos.

Thomas (2000) engloba la metodología existente hasta la fecha en dos grupos:

1. Métodos paramétricos, de investigación estadística u operacional.
    1.1. Análisis discriminante (Myers and Forgy, 1963).
    1.2. Regresión lineal (Orgler, 1971).
    1.3. Regresión logística (Grablowsky and Talley, 1981; Wiginton, 1980).
    1.4. Árboles de decisión o algoritmos de particionamiento recursivo (Breiman et al., 1984).
2. Métodos no paramétricos, de inteligencia artificial.
    2.1. Redes neuronales (Tam and Kiang, 1992).
    2.2. Sistemas expertos (Leonard, 1993).
    2.3. Algoritmos genéticos (Davis et al., 1992).
    2.4. Método de los K vecinos más cercanos (Henley and Hand, 1996).

Lean Yu et al. (2008) clasifican las técnicas de credit scoring en cuatro grupos:

1. Estadísticas:
    1.1. Análisis discriminante (Altman, 1968; Fisher, 1936).
    1.2. Regresión logística (Steenackers and Goovaerts, 1989; Wiginton, 1980).
    1.3. Regresión probit (Grablowsky and Talley, 1981).
    1.4. K vecinos más cercanos (Henley and Hand, 1996; Hand and Henley, 1997).
    1.5. Árboles de decisión (Carter and Catlett, 1987; Coffman, 1986; Makowski, 1985).
2. Investigación operativa
    2.1. Programación lineal (Hand, 1981).
    2.2. Programación entera (Kolesar and Showers, 1985; Showers and Chakrin, 1981).
3. Inteligencia Artificial
    3.1. Redes neuronales (Baesens et al., 2003a; Desai et al., 1996; Khashman, 2009; Malhotra and Malhotra, 2002; West, 2000; Yobas and Ross, 2000).
    3.2. Máquinas de vectores soporte (Baesens et al., 2003b; Schebesch and Stecking, 2005; Van Gestel et al., 2003; Yu et al., 2009a; Zhou et al., 2009).
    3.3. Algoritmo genético (Chen and Huang, 2003; Ong et al., 2005; Varetto, 1998).
    3.4. Conjunto aproximado (Beynon, 2001).

3.5. Razonamiento basado en casos (Li and Sun, 2008; Li and Sun, 2009b; Li and Sun, 2009c; Li and Sun, 2010; Li et al., 2009).

4. Híbrido, combinado y enfoque de conjunto.

4.1. Sistema difuso y la red neuronal artificial (Malhotra and Malhotra, 2002; Piramuthu, 1999).

4.2. Conjunto aproximado y red neuronal artificial (Ahn et al., 2000).

4.3. Conjunto aproximado y máquinas de vectores (Yu et al., 2008).

4.4. Sistema difuso y máquinas de vectores (Wang et al., 2005).

4.5. Casos basados en razonamiento y máquinas de vectores (Li and Sun, 2009a).

4.6. Conjunto de redes neuronales (Yu and Lai, 2008).

4.7. Conjunto de máquinas de vectores (Yu et al., 2010; Zhou et al., 2010).

4.8. Grupo de apoyo en toma de decisiones (Sun and Li, 2009; Yu et al., 2009b).

Tal y como se indica en Galindo and Tamayo (2000), la aplicación de la construcción de modelos para el cálculo del riesgo de crédito se ve obstaculizada por la disponibilidad y la calidad de los datos. En muchos casos, el proceso de recopilación de los datos no es exacto o completo y los datos a menudo contienen inconsistencias y valores perdidos o datos anómalos. Al mismo tiempo, las relaciones existentes entre las variables utilizadas pueden ser complejas, no lineales y no reflejan los cambios estructurales, como por ejemplo las tendencias demográficas o de mercado. Los autores argumentan que hasta cierto punto, cada conjunto de datos es idiosincrásico y único en el espacio y el tiempo. Al mismo tiempo, los datos financieros son dinámicos, por tanto en el proceso de construcción del modelo se tiene que tener en cuenta el modelado continuo.

Motivados por este hecho y comprobando la amplía variedad de técnicas y métodos empleados en la literatura de la materia, nace el interés por la investigación llevada a cabo en esta memoria, evaluar estos métodos bajo diferentes condiciones de carga, tanto en la vertiente de eficacia como de la eficiencia, esta última apenas evaluada hasta la fecha.

Tal y como señalan autores como Abdou and El Masry (2008), Bailey (2004) y Sullivan (1981), a pesar de las críticas que puedan tener las técnicas de credit scoring, estas han sido una de las principales herramientas en la evaluación del riesgo de crédito para los préstamos y por tanto pueden considerarse las de más éxito en el campo de los negocios y las finanzas.

A continuación se pasan a detallar en las siguientes secciones, las contribuciones científicas de los principales métodos que luego serán de profundo estudio en esta memoria.

1.3. Analisis Discriminante

El primer trabajo vinculado con riesgo crediticio que utilizó análisis discriminante fue propuesto por Durand (1941). El autor afirma que es un buen método para analizar las predicciones

de devolución de crédito. Los primeros trabajos en el ámbito corporativo fueron los estudios de Altman (1968), Orgler (1970), Deakin (1972) y Blum (1974). Otros autores más recientes que emplean el análisis discriminante son Reichert et al. (1983), Artís et al. (1994), Boj et al. (2009b), Bonilla et al. (2003), Trias et al. (2005) y Trias et al. (2008). También existen detractores del análisis discriminante aplicado a las finanzas y economía en general. Puede señalarse como críticos a Eisenbeis (1977), Eisenbeis (1978) y Rosenberg and Gleit (1994).

Hasta ahora se ha utilizado esta técnica como un punto de referencia para comparar la precisión de las predicciones con otras técnicas alternativas, Coats and Fant (1992), Back et al. (1996) y Laitinen and Kankaanpää (1999). Otros autores realizan estudios comparativos de análisis discriminante con otros métodos como los autores Myers and Forgy (1963), Lane (1972), Apilado et al. (1974) y Moses and Liao (1987) que comparan el análisis discriminante con análisis de regresión. Otro estudio es el de los autores Grablowsky and Talley (1981) que comparan el análisis discriminante lineal y el análisis probit mediante el uso de datos de una gran cadena de tiendas en el medio oeste de EE.UU.

En el trabajo de Altman et al. (1994) se realiza un estudio comparativo aplicando análisis discriminante lineal y redes neuronales (MLP). Concluyen que la red neuronal no es una técnica claramente dominante.

En otro sentido, está el trabajo de Coats and Fant (1993), en el que se compara la redes neuronal (MLP) con el análisis discriminante lineal, obteniendo que para la base de datos analizada, la red neuronal obtiene mejores predicciones que el análisis discriminante lineal.

En el trabajo de Lacher et al. (1995) y utlizando los mismos datos que Coats and Fant (1993), y una medida proporcionada por Altman (1968), concluyen que la red neuronal predice con mayor precisión.

El estudio de Boj et al. (2009a) aplica el análisis discriminante basado en distancias (ADBD) en dos conjuntos de datos reales de créditos de instituciones financieras (*German Credit* y *Australian Credit*) con distintos métodos. Realiza el estudio comparativo con los métodos siguientes:

1. Métodos no paramétricos
    1.1. Redes neuronales MOE (Mixture of experts)
    1.2. Redes neuronales RDF (Radial basis function)
    1.3. Redes neuronales MLP (Multi-layer perceptron)
    1.4. Redes neuronales LVQ (Learning vector quantization)
    1.5. Redes neuronales FAR (Fuzzy adaptive resonance)
    1.6. K vecinos más próximos
    1.7. Estimación núcleo de densidad
    1.8. Árbol de decisión CART
2. Métodos paramétricos

2.1. Análisis discriminante lineal

2.2. Regresión logística

El criterio de selección elegido se realiza en base a dos criterios:

1. Probabilidad de mala clasificación (falso negativo junto con falsos positivos). Para ello se calcula la matriz de confusión.

2. Coste del error, que es el coste de conceder un crédito a un candidato con mal riesgo (falso positivo) junto con el coste de denegar un crédito a un buen candidato (falso negativo).

Las conclusiones de los autores son las siguientes:

1. Para los datos de *German Credit*,

   1.1. Según el criterio de probabilidad de mala clasificación la metodología que obtiene una menor probabilidad de clasificar malos riesgos es el análisis discriminante basado en distancias, seguido del análisis discriminante lineal. Si el criterio es la probabilidad global de la mala clasificación, el método con el que se obtiene un menor error de clasificación es la regresión logística, seguido de la red neuronal MOE. Los autores destacan que se tiene que tener en cuenta que las técnicas con las que se obtiene una menor probabilidad global de mala clasificación coincide con las que se obtiene mayor probabilidad de clasificar mal los malos riesgos, es decir, disminuye la probabilidad global de clasificar mal, a costa de clasificar de modo incorrecto los buenos riesgos.

   1.2. Según el criterio de costes estimados. Se analiza los costes estimados con dos probabilidades a priori de estimar malos riesgos. Los autores consideran una buena estimación de estos la propuesta por West (2000), que propone la estimación a priori de los malos riesgos entre dos cotas $[0{,}144; 0{,}249]$. Con un escenario en el que la probabilidad a priori de los malos riesgos es 0,144 , la metodología con menor coste es la regresión logística. Cambiando el escenario de la probabilidad a priori de los malos riesgos a 0,249 la metodología con menor coste es el análisis discriminante lineal, seguida del análisis discriminante basado en distancias. Los autores destacan que los costes en el análisis discriminante basado en distancias ofrecen una variación menor en general al cambiar la probabilidad a priori.

2. Para los datos de *Australian Credit* ,

   2.1. Según el criterio de probabilidad de mala clasificación, la metodología que obtiene una menor probabilidad de clasificar malos riesgos es el árbol de decisión CART, seguido de la red neuronal MOE. Si el criterio es la probabilidad global de la mala clasificación, el método con el que se obtiene un menor error de clasificación es la regresión logística, seguido del análisis discriminante basado en distancias.

2.2. Según el criterio de costes estimados. Con un escenario en el que la probabilidad a priori de los malos riesgos es 0,144 , la metodología con menor coste es la red neuronal MOE igualada con la regresión logística. Cambiando el escenario de la probabilidad a priori de los malos riesgos a 0,249 la metodología con menor coste es la red neuronal MLP. Al igual que con los datos *German Credit*, los autores destacan que los costes en el análisis discriminante basado en distancias ofrecen una variación menor en general al cambiar la probabilidad a priori.

Los autores concluyen que no existe un método óptimo, cada metodología tiene sus ventajas e inconvenientes. Al mismo tiempo, se puede comprobar con los resultados obtenidos en el estudio, que dependiendo de los datos objeto de estudio, el método óptimo difiere.

En el capítulo 5.3 de esta memoria se procede a evaluar los dos conjuntos de datos estudiados por Boj et al. (2009a).

1.4. Árboles de decisión

Los árboles de decisión son métodos de particionamiento recursivo. Los referentes más importantes son los trabajos de Breiman et al. (1984) y Safavian and Landgrebe (1991). Distintos autores como Makowski (1985), Carter and Catlett (1987) y Mehta (1968), describen las aplicaciones de los métodos recursivos en el credit scoring. En el trabajo de Coffman (1986) se compara los árboles de decisión con el análisis discriminante, llegando a la conclusión que los árboles de decisión son mejores cuando hay una interacción entre las variables y el análisis discriminante es mejor cuando las variables están correlacionadas.

Distintos autores estudian la aplicación al credit scoring de árboles de decisión comparándolos con otros métodos. El autor Frydman et al. (1985) aplicó un particionamiento recursivo para clasificar y predecir compañías en quiebra, comparándolo con análisis discriminante. Llega a la conclusión que la aplicación del árbol de decisión es más precisa. En el trabajo de Boyle et al. (1992) se compara el particionamiento recursivo con el análisis discriminante. Concluyen que el mejor método aplicado al credit scoring son los modelos lineales, como es el caso del análisis discriminante.

Los autores Davis et al. (1992), compararon un método basado en redes neuronales (MLP) con árboles de decisión. Sus resultados se basan en una sola partición de datos y un único ensayo de red neuronal. Concluyen que la red neuronal (MLP) y el modelo del árbol de decisión tienen un nivel comparable de precisión.

El trabajo de los autores Galindo and Tamayo (2000) basan su investigación en la aplicación de cuatro algoritmos o métodos en una base de datos de 4 000 préstamos hipotecarios de un Banco Mexicano entre 1995 y 1996. Los autores realizan el estudio aplicando a cada una de las bases de datos la distribución siguiente: $\frac{1}{2}$ conjunto de datos se clasifica como datos de entrenamiento, $\frac{1}{4}$ datos para testear y $\frac{1}{4}$ datos para evaluar. Los métodos aplicados son el árbol de decisión CART, los k vecinos más cercanos, redes neuronales y probit. En la regresión probit

se establece como mejor valor de probabilidad a priori $\pi = 0{,}7$. El árbol de decisión se prueba con distintos nodos, siendo el óptimo 120. La red neuronal se prueba con distintos nodos e iteraciones, resultando óptimo 16 nodos con 80 iteraciones. Los k vecinos más cercanos se prueba con distintos k, resultando el óptimo 24 vecinos. Se concluye que el mejor modelo general es el árbol de decisión CART con 120 nodos, se obtiene una tasa de error menor, aunque para alcanzar un modelo predictivo óptimo se necesitaría una muestra de 21 675 registros. El siguiente método con mejores resultados es la red neuronal con 16 nodos y 80 iteraciones. En este caso para alcanzar el modelo predictivo óptimo se necesitaría 18 165 registros. En tercer lugar estaría los k vecinos más cercanos con $k = 24$, en este caso indican los autores que con un número mayor de registros disponibles se obtendrían mejores resultados. En último lugar estaría el modelo probit, aunque indican que a pesar de obtener mayor tasa de error en el caso de tener pequeñas muestras sería competitivo porque el tamaño óptimo de muestra se sitúa en 1 804, mucho menor a los anteriores. En todos los modelos estudiados, coincide que los errores son más altos para el caso del grupo moroso (no cumple).

Contrariamente a estos resultados, los trabajos de Laitinen and Kankaanpää (1999), West (2000) y Shin and Han (2001) indican que las técnicas de árboles de decisión obtuvieron peores resultados que las restantes técnicas utilizadas.

El autor Wang and Ma (2011) concluye que para una de las bases de datos de empresas chinas, la técnica que obtiene mejores resultados es la de árboles de decisión, aunque tan sólo levemente por encima. Sin embargo para la otra base de datos, esta misma técnica obtiene los peores resultados.

Otras investigaciones más recientes proponen un enfoque de minería de datos híbrido. En esta línea podemos señalar a Hsieh (2005) que propone un sistema híbrido basado en la agrupación junto a técnicas de redes neuronales. Los autores Lee and Chen (2005) proponen un procedimiento híbrido de dos etapas con redes neuronales y splines de regresión multivariante de adaptación. En el estudio de Lee et al. (2002) se integran backpropagation de redes neuronales con un enfoque tradicional de análisis discriminante. En el artículo de Chen and Huang (2003) se aplican redes neuronales y técnicas de algoritmos genéticos.

1.5. Modelos lineales

Autores como Lachenbruch (1975), Orgler (1970), Orgler (1971), Lucas (1992) y Henley (1995) utilizaron métodos lineales con el fin de predecir distintos comportamientos. Orgler (1970) emplea en su estudio el método lineal multivariable para compararlo con el método tradicional empleado hasta ese momento de revisión por parte del personal de la oficina bancaria. Las conclusiones a la que llega es que es un sistema sencillo y aplicable a cualquier entidad financiera, con ahorro de tiempo considerable y simplicidad. Indicando que los modelos a desarrollar deben de ir en la línea de predecir el futuro comportamiento de un cliente.

Las primeras aplicaciones de logit (o regresión logística) en el ámbito del riesgo de crédito se encuentran en Martín (1977), Orgler (1980) y Hammer (1983). El trabajo de Martín (1977), es-

tudia la posibilidad de predicción del default bancario aplicando el método de regresión logística, análisis discriminante lineal y cuadrático. Concluye que si los modelos se comparan en términos de clasificación en lugar de estimación de probabilidad, tanto el método logit como discriminante son similares, situándose el análisis discriminante lineal un poco mejor ya que minimiza los cálculos requeridos. Sin embargo, si los resultados de interés son en términos de probabilidad es preferible el método de análisis de regresión logística. Por tanto dependerá del objetivo del estudio.

Orgler (1980) aplica la regresión logística en su estudio argumentando que comparativamente con el análisis discriminante tiene ventajas como que no es necesario conocer las probabilidades a priori, no es necesario conocer la distribución de los predictores y el resultado obtenido en el análisis discriminante es simplemente una puntuación poco interpretable. Al mismo tiempo, aplicando la regresión logística la significación estadística de los diferentes predictores se obtiene mediante la teoría asintótica (con una muestra grande). Las conclusiones que alcanza son dos: en primer lugar, el poder de predicción de cualquier modelo depende de la información disponible. Y en segundo lugar, existe un poder predictivo en las transformaciones lineales de un vector de relaciones robusto, a través de la estimación de los procedimientos, siempre y cuando se tenga una muestra grande, por tanto se mejoraría significativamente la predicción con predictores adicionales.

El estudio de Hammer (1983), es descriptivo y compara los modelos logit, análisis discriminante lineal y cuadrático. Concluye que en general tanto los modelos logit como análisis discriminante lineal predicen al menos tan bien como el cuadrático, para los primeros tres años de estudio, sin embargo para el cuarto año la predicción no es buena en ningún método.

El estudio realizado por los autores Tam and Kiang (1992), utilizan variantes de los k vecinos 1-NN y 3-NN y lo comparan con el análisis discriminante, regresión logística, árbol de decisión ID3 y dos redes neuronales (MLP). El objetivo del estudio realizado es predecir la quiebra bancaria a uno y dos años antes del suceso de incumplimiento de un Banco de Texas. Sugieren que la red neuronal (MLP) es más precisa, seguida del análisis discriminante lineal, regresión logística, árbol de decisión y k vecinos más cercanos.

En el artículo de Salchenberger and Lash (1992) se compara la red neuronal (MLP) con el modelo de regresión logística. Concluyen que ambos son igual de buenos para los datos utilizados en el estudio.

En el artículo de Henley (1995), se realiza un estudio comparativo de la regresión logística frente a la regresión lineal. El autor concluye que la regresión logística no es mejor debido a la existencia, en la muestra utilizada para el estudio, de muchos solicitantes con puntajes asociados a probabilidades estimadas de ser buenos riesgos que oscilaban entre 0,2 y 0,8 , y esto produce que la curva logística se aproxime a la recta.

Otros autores como Masters (1995), Zekic-Susac et al. (2004) han realizado investigaciones en modelos de credit scoring utilizando el método de redes neuronales probabilísticas comparándolo con regresión logística y árboles de decisión CART, llegando a la conclusión que las redes neuronales son mejores en cuanto a representación de datos se refiere.

En el artículo de Desai et al. (1996) se realiza una comparación de técnicas estadísticas convencionales o métodos paramétricos y no paramétricos, fundamentalmente redes neuronales:

1. Métodos no paramétricos

   1.1. Redes neuronales perceptrón multicapa (MLP)

   1.2. Redes neuronales modulares (MNN)

2. Métodos paramétricos

   2.1. Análisis discriminante lineal (LDA)

   2.3. Regresión logística (LR)

Los autores realizan el estudio con la base de datos utilizada por Overstreet et al. (1992) y Overstreet and Bradley (1994). La base de datos ha sido construida a partir de los archivos de préstamos de tres cooperativas de crédito del sudeste de Estados Unidos para el período de 1988-91. Las cooperativas de crédito son de tres colectivos distintos, la cooperativa de crédito "L" está compuesta principalmente por profesores, la cooperativa de crédito "M" predominantemente son empleados de una compañía telefónica y la cooperativa de crédito "N" representa una muestra diversa. Los datos recogidos para el estudio, después de eliminar datos incompletos, son: 505 observaciones para la cooperativa "L", 762 para la cooperativa "M" y 695 observaciones para la cooperativa "N". Se seleccionan 18 variables explicativas de la base de datos, y la variable objetivo es dicotómica, se establece como "malo" si en los últimos 48 meses no se atiende el pago y "bueno" el caso contrario. Se realiza el estudio con 0,50 como punto de corte. La base de datos la particionan en 2/3 corresponde a la muestra de entrenamiento y 1/3 a la muestra de test. Las observaciones se asignan de forma aleatoria a la formación de los subconjuntos de datos y se crean diez pares de subconjuntos. Se realizan dos modelos, el personalizado para cada una de las cooperativas de crédito y el genérico, un modelo para el conjunto de todos los datos de las cooperativas.

Los autores Desai et al. (1996) llegan a las conclusiones siguientes:

1. Según el criterio de los préstamos correctamente clasificados:

   1.1 Modelos personalizados. Los modelos de redes neuronales superan al análisis discriminante lineal, pero son apenas un poco mejor que los modelos de regresión logística.

   1.2. Modelos genéricos, la regresión logística predice mejor que el análisis discriminante y en último lugar las redes neuronales.

2. Según el criterio de identificación de préstamos incobrables:

   2.1. Modelos personalizados. Los modelos de redes neuronales clasifican con mayor precisión que el análisis discriminante lineal o regresión logística

2.2. Modelos genéricos. Las redes neuronales son las que mejor clasifican, seguidas de regresión logística y por último análisis discriminante.

En general, los autores concluyen que las redes neuronales son un buen modelo predictivo, siendo la regresión logística también buena medida.

El trabajo de Wiginton (1980) fue uno de los primeros estudios en el que se aplicó la regresión logística en credit scoring comparando con el análisis discriminante. Concluyó que el enfoque logístico clasifica mejor, pero que ninguno de los dos métodos es suficientemente bueno desde el punto de vista del coste efectivo.

Según la investigación realizada por West (2000), teniendo en cuenta la investigación disponible sobre la predicción de dificultades financieras, sugiere que los modelos de redes neuronales muestran buen potencial pero carecen de las ventajas de las técnicas estadísticas clásicas. El autor sugiere emplear mayor número de repeticiones para la formación de redes neuronales para obtener mayor predicción, dada la naturaleza estocástica del proceso. El autor realiza una comparación de métodos no paramétricos, fundamentalmente redes neuronales, y métodos paramétricos :

1. Métodos no paramétricos
   1.1. Redes neuronales MOE (Mixture of experts)
   1.2. Redes neuronales RDF (Radial basis function)
   1.3. Redes neuronales MLP (Multi-layer perceptron)
   1.4. Redes neuronales LVQ (Learning vector quantization)
   1.5. Redes neuronales FAR (Fuzzy adaptive resonance)
   1.6. K vecinos más próximos
   1.7. Estimación núcleo de densidad o densidad de kernel.
   1.8. Árbol de decisión CART
2. Métodos paramétricos
   2.1. Análisis discriminante lineal
   2.2. Regresión logística

En la investigación se utilizan dos conjuntos de datos reales con particiones en datos de entrenamiento y de prueba, con 10 veces la validación cruzada. Se realizan diez repeticiones de cada ensayo de redes neuronales y por último aplica a los modelos el test de McNemar Chi Square, que según Dietterich (1998) ha demostrado ser el test más eficiente para los algoritmos de aprendizaje supervisado.

Analizando los resultados obtenidos, West (2000) llega a la conclusión que se obtienen resultados muy similares en ambas bases de datos:

1. Según el criterio de probabilidad de mala clasificación:

   1.1 El método que menor error global obtiene es la regresión logística, seguida de la red neuronal MOE y la red neuronal RBF.

   1.2. El análisis discriminante lineal obtiene un error global aceptable, pero su fortaleza radica en que obtiene el error más pequeño en la clasificación de los falsos positivos.

   1.3. El método menos preciso es la red neuronal FAR y la red neuronal LVQ, seguido de los K vecinos más cercanos, árbol de decisión CART y densidad del núcleo Kernel. Siguiendo a Dietterich, West refuerza las conclusiones anteriores aplicando la prueba de McNemar, prueba todos los modelos de credit scoring con el modelo más preciso. Aquel modelo que no es significativamente diferente con el modelo más preciso se etiqueta como superior. Obtiene que los modelos de redes neuronales MOE, RBF y MLP son los más precisos para ambos conjuntos de datos junto con la regresión logística. Para el conjunto de datos *Australian Credit*, el modelo de análisis discriminante lineal y k vecinos más cercanos son superiores. Los modelos de redes neuronales LVQ y FAR, nucleo de densidad de Kernel y árbol de decisión CART son modelos inferiores para ambos conjuntos de datos.

2. Según el criterio de coste estimado de los errores: Se analiza los costes estimados con dos probabilidades a priori de estimar malos riesgos (π_2). West (2000) propone como probabilidad a priori de malos riesgos entre dos cotas ($\pi_2[0{,}144, 0{,}249]$), calculadas a partir de las tasas de incumplimiento. Para los datos *German Credit*, en ambos escenarios de probabilidad a priori, la metodología con menor coste es el análisis discriminante lineal, seguido de la red neuronal MOE, RBF y la regresión logística. Para los datos *Australian Credit* los costes de los modelos de red neuronal MOE, RBF y MLP son casi idénticos en ambos niveles de probabilidad. Los costes de análisis discriminante lineal y regresión logística son comparables a los modelos de redes neuronales pero se vuelven menos eficientes cuando el valor de la probabilidad a priori aumenta.

En el estudio de Abdou and El Masry (2008) se realiza una comparación de métodos no paramétricos (fundamentalmente redes neuronales) y métodos paramétricos:

1. Métodos no paramétricos

   1.1. Redes neuronales probabilísticas (PNN)

   1.2. Redes neuronales multicapa feed forward (MLFN)

   1.3. La mejor red neuronal (BNS)

2. Métodos paramétricos

   2.1. Análisis discriminante lineal

   2.2. Análisis probit

2.3. Regresión logística

Los autores realizan el estudio con una base de datos de una banca comercial egipcia, con un total de 581 datos. De las 20 variables explicativas de la base de datos, seleccionan 12 para el estudio. Realizan el estudio con dos puntos de corte, 0,50 y 0,60, finalmente eligen para las conclusiones como punto de corte 0,50. La base de datos la particionan en 80 % muestra de entrenamiento y 20 % muestra de test. Tanto en el análisis discriminante, el análisis probit como en la regresión logística realizan dos estudios en cada uno de ellos, con todas las variables (las 12 seleccionadas) y el estudio por etapas, añadiendo las variables 1 a 1 que minimice lambda Willks' global (9 variables). En cuanto a las redes neuronales, el experimento se repite 20 veces para diferentes muestras de entrenamiento y test aleatorias. Una vez obtenidos los resultados seleccionan las 20 mejores redes neuronales. El caso de la mejor red neuronal, plantean las redes neuronales multicapa feed forward desde 2 nodos a 6 nodos y la red neuronal probabilística obtendrá seis modelos. Con todos los resultados anteriores selecciona de las redes neuronales las 5 mejores de cada experimento para compararlas con las dos obtenidas en análisis discriminante, análisis probit y regresión logística. Los autores concluyen que existen resultados muy similares en ambas bases de datos:

1. Según el criterio de probabilidad de mala clasificación:

   1.1 La regresión logística es la mejor técnica convencional.

   1.2. Todos los modelos predicen mejor el crédito bueno que el malo, a excepción del análisis discriminante.

   1.3. En general, las técnicas de redes neuronales son mejores que las técnicas convencionales.

2. Según el criterio de coste estimado de los errores, siguiendo a West (2000) en su definición del coste estimado, error tipo I, créditos buenos clasificados como malos o falsos negativos (FN), y error tipo II, créditos malos clasificados como buenos o falsos positivos (FP), y coste total estimado:

   2.1. En general los errores de clasificación tipo II son más altos que los errores de clasificación tipo I.

   2.2. Se cumple la generalidad anterior para las técnicas convencionales excepto para el análisis discriminante. Se predicen mejor los malos créditos.

   2.3. Para los modelos de redes neuronales, el error tipo II es más alto que el error tipo I.

Las conclusiones del trabajo de investigación plantean que si el criterio de elección es la tasa de clasificación correcta el mejor modelo será el de redes neuronales probabilísticas. Si el criterio es el coste estimado de errores de clasificación el mejor modelo será el de redes neuronales multicapa feedforward (MLFN).

Actualmente, logit es considerado un buen modelo de predicción y clasificación en el ámbito crediticio, y suele utilizarse para comparaciones frente a técnicas alternativas. Existen diversos estudios que concluyen la efectividad de los modelos logit frente a otras técnicas como, entre otros, Srinivasan and Kim (1987a), Leonard (1993), Back et al. (1996), Laitinen and Kankaanpää (1999), Huang et al. (2004), Baesens et al. (2003b).

Otras investigaciones más recientes proponen un enfoque de minería de datos híbrido. En esta línea podemos señalar a Hsieh (2005) que propuso un sistema híbrido basado en la agrupación junto a técnicas de redes neuronales. Los autores Lee and Chen (2005) proponen un procedimiento híbrido de dos etapas con redes neuronales y splines de regresión multivariante de adaptación. En el estudio de Lee et al. (2002) se integran backpropagation de redes neuronales con un enfoque tradicional de análisis discriminante. En el artículo de Chen and Huang (2003) se aplican redes neuronales y técnicas de algoritmos genéticos.

1.6. Máquina de vectores soporte (SVM)

Las máquinas de vectores soporte fueron sugeridas por primera vez por Vapnik (1995). En el ámbito financiero se empiezan a utilizar sobre el año 2000, fundamentalmente en dos áreas de conocimiento financiero:

1. Para la predicción de precios y volatilidad bursátil (Van Gestel et al., 2001; Tay and Cao, 2002; Cao, 2002; Huang et al., 2005).

2. En aplicaciones de riesgo crediticio y detección de fraude (Fan and Palaniswami, 2000; Baesens et al., 2003b).

En el trabajo de Fan and Palaniswami (2000) se predicen situaciones de incumplimiento en compañías. Los autores utilizan una muestra de 174 empresas australianas (86 con incumplimiento), y aplican modelos predictivos utilizados por los autores Altman (1968), Orgler (1980), Lincoln (1982) y uno propio (en el que utilizan las variables de los trabajos de estos tres autores más otras 5 adicionales). Los autores realizan la investigación comparando cuatro técnicas para la predicción: SVM, análisis discriminante, redes neuronales multiperceptrón y aprendizaje de cuantificación vectorial. Concluyen que la mejor técnica predictiva es SVM.

En los trabajos de Härdle et al. (2004) y Härdle et al. (2005) se utiliza la técnica SVM comparándola con análisis discriminante para predecir la quiebra de 84 compañías. Llegan a la conclusión que las diferencias obtenidas aplicando las dos técnicas no es estadísticamente significativa, por tanto no pueden concluir que SVM es mejor clasificador que el análisis discriminante.

En el artículo Baesens et al. (2003b) se desarrolla un estudio comparativo de diferentes técnicas de clasificación en ocho conjuntos de datos reales de créditos de instituciones financieras (Benelux1, Benelux2, UK1, UK2, UK3, UK4, Alemania y Australia). Realiza el estudio comparativo con los métodos siguientes:

1. Regresión logística (LR).

2. Análisis Discriminante lineal (LDA).
3. Análisis Discriminante cuadrático (QDA).
4. Programación lineal (LP).
5. Máquina de vectores con núcleo de base lineal (Lin SVM).
6. Máquina de vectores con núcleo de base radial (RBF SVM).
7. Máquina de vectores rectificada con núcleo de base lineal (Lin LS-SVM).
8. Máquina de vectores rectificada con núcleo de base radial (RBF LS-SVM).
9. Redes neuronales (NN).
10. Árbol clasificador de Bayes (NB).
11. Árbol clasificador de Bayes aumentado (TAN).
12. Árbol de decisión (C4.5).
13. Árbol de decisión y reglas (C4.5rules).
14. Árbol de decisión (C4.5dis).
15. Árbol de decisión y reglas (C4.5rules dis).
16. K vecinos más cercanos (KNN10).
17. K vecinos más cercanos (KNN100).

El criterio elegido de selección es por:

1. Porcentaje de los correctamente clasificados. Para ello calculan la matriz de confusión.
2. Curva ROC o área AUC que medirá la sensibilidad como la proporción de positivos que se prevé sean positivos y la especificidad que medirá la proporción de negativos que se espera sean negativos.

Los autores llegan a la conclusión que, tanto con el criterio del porcentaje de los correctamente clasificados como en términos de la curva AUC, los métodos que obtienen mejor resultados son los de máquinas de vectores y redes neuronales, seguidas de los métodos lineales de regresión logística y análisis discriminante lineal. La mayoría de técnicas obtuvieron buenos resultados, tan sólo obtuvieron un peor desempeño el análisis discriminante cuadrático, clasificador de Bayes y los árboles de decisión y reglas (C4.5rules).

Los autores Huang and Wang (2007) desarrollan un estudio comparativo de diferentes técnicas de clasificación en dos conjuntos de datos reales de créditos de instituciones financieras (*German Credit* y *Australian Credit*). Primero realizan el estudio comparativo con los métodos siguientes:

1. Máquina de vectores soporte + Grid search (SVM + Grid search).
2. Máquina de vectores soporte + Grid search + F-score (SVM + Grid search + F-score).
3. Máquina de vectores soporte + GA (SVM + GA).
4. Redes neuronales (NN).
5. Programación genética.
6. Árbol de decisión C4.5.

Los autores realizan el estudio aplicando los tres métodos indicados de máquinas de vectores soporte a los mismos conjunto de datos de entrenamiento y testeo. Eligen como punto de corte para realizar el estudio 0,5, mayor a ese punto es buen cliente, menor mal cliente. Al mismo tiempo que calculan el porcentaje de correctamente clasificados, para cada uno de los métodos y bases de datos, calculan la precisión en la clasificación a través del test de Friedman. Para las dos bases de datos coinciden los resultados obtenidos, tanto el porcentaje de correctamente clasificados como la precisión de la clasificación. Señalan como mejor método la máquina de vectores soporte + GA (SVM + GA). En cuanto a la comparación del método anteriormente seleccionado con redes neuronales, programación genética y árbol de decisión C4.5, para la base de datos *Australian Credit* no hay diferencias significativas entre ellos, pero en la base de datos *German Credit*, el modelo C4.5 fue significativamente inferior.

Los autores Yu et al. (2011) desarrollan un estudio comparativo de diferentes técnicas de clasificación en dos conjuntos de datos reales de créditos de instituciones financieras (*German Credit* y *Australian Credit*). El objetivo del trabajo es estudiar el método de máquina de vectores soporte ponderado con la función Kernel RBF por mínimos cuadrados con un algoritmo DOE (Zhou et al., 2009), LSSVMDOE, comparándolo con distintos métodos de evaluación de riesgo. Para el estudio comparativo utilizan los 17 métodos empleados por Baesens et al. (2003b) en su estudio y las 5 redes neuronales empleadas por West (2000) en su estudio. Por tanto el estudio comparativo se realiza en primer lugar con los métodos siguientes:

1. Regresión logística (LR)
2. Análisis Discriminante lineal (LDA)
3. Análisis Discriminante cuadrático (QDA)
4. Programación lineal (LP)
5. Máquina de vectores con núcleo de base lineal (Lin SVM)
6. Máquina de vectores con núcleo de base radial (RBF SVM)
7. Máquina de vectores rectificada con núcleo de base lineal (Lin LS-SVM)

8. Máquina de vectores rectificada con núcleo de base radial (RBF LS-SVM)
9. Red neuronal (NN)
10. Árbol clasificador de Bayes (NB)
11. Árbol clasificador de Bayes aumentado (TAN)
12. Árbol de decisión (C4.5)
13. Árbol de decisión y reglas (C4.5rules)
14. Árbol de decisión (C4.5dis)
15. Árbol de decisión y reglas (C4.5rules dis)
16. K vecinos más cercanos (KNN10)
17. K vecinos más cercanos (KNN100)
18. Redes neuronales MOE(Mixture of experts)
19. Redes neuronales RDF (Radial basis function)
20. Redes neuronales MLP (Multi-layer perceptron)
21. Redes neuronales LVQ (Learning vector quantization)
22. Redes neuronales FAR (Fuzzy adaptive resonance)
23. Máquina de vectores soporte ponderado con la función Kernel RBF por mínimos cuadrados con un algoritmo DOE (LSSVMDOE)

Los autores utilizan tres criterios de evaluación:

1 Precisión tipo I o especificidad. Definido como la proporción del total de observaciones malas las que son realmente malas.

2 Precisión tipo II o sensibilidad. Definido como la proporción del total de observaciones buenas las que son realmente buenas.

3 Precisión total. Definido como la proporción del total de observaciones de la muestra las que están correctamente clasificadas.

Los autores realizan el estudio aplicando a cada una de las bases de datos la distribución de $\frac{2}{3}$ conjunto de entrenamiento y $\frac{1}{3}$ datos para comprobación. Yu et al. llegan a la conclusión que en cuanto a precisión total el mejor método es máquina de vectores soporte ponderado con la función Kernel RBF por mínimos cuadrados con un algoritmo DOE, seguido de las técnicas

más simples como regresión logístia y análisis discriminante lineal para ambas bases de datos. En cuanto a la precisión tipo I los mejores resultados para ambas bases de datos se obtienen con máquina de vectores soporte ponderado con la función Kernel RBF por mínimos cuadrados con un algoritmo DOE. Los peores resultados se obtienen para los k vecinos más cercanos y el clasificador de Bayes. En cuanto a la precisión tipo II los mejores resultados se obtienen en la base de datos *German Credit* la técnica de clasificación de Bayes y en la base de datos *Australian Credit* la técnica de los k vecinos más cercanos. Una posible explicación es que las bases de datos estén sesgadas, por tanto predicen mejor la opción mayoritaria. Los autores concluyen en esta primera fase que el clasificador óptimo es máquina de vectores soporte ponderado con la función Kernel RBF por mínimos cuadrados con un algoritmo DOE ya que obtiene los mejores resultados en cuanto a precisión total y precisión tipo I, considerando que es más difícil de clasificar los malos clientes que los buenos.

Una vez realizado el estudio anterior los autores concluyen como mejor método la máquina de vectores soporte. Pero además de comprobar que es el mejor método de clasificación comprueban que el algoritmo utilizado en el estudio (DOE) es el que mejor selecciona los parámetros. Para realizarlo comparan diversos algortimos de máquinas de vectores. Los algoritmos utilizados son los siguientes:

1. Máquina de vectores soporte ponderado con la función Kernel RBF por mínimos cuadrados con un algoritmo DOE (LSSVMDOE).

2. Máquina de vectores soporte ponderado con la función Kernel RBF por mínimos cuadrados con un algoritmo genético (LSSVMGA).

3. Máquina de vectores soporte ponderado con la función Kernel RBF por mínimos cuadrados con un algoritmo "grid search"(LSSVMGS).

4. Máquina de vectores soporte ponderado con la función Kernel RBF por mínimos cuadrados con un algoritmo "direct search"(LSSVMDS).

En cuanto a la comparación del método máquina de vectores soporte ponderado con la función Kernel RBF por mínimos cuadrados con diversos algoritmos, los autores concluyen para ambos conjuntos de datos, que el algoritmo DOE obtiene mejores resultados en precisión tipo I. El algoritmo GS obtiene mejores resultados en términos de precisión tipo II y en precisión total se obtienen mejores resultados con el algoritmo DS, pero desde el punto de vista de eficiencia computacional se obtienen mejores resultados para el algoritmo DOE. Por tanto, concluyen que el mejor método para resolver el credit scoring es máquina de vectores soporte ponderado con la función Kernel RBF por mínimos cuadrados con un algoritmo DOE por ser el más preciso a la vez que el más eficiente computacionalmente (a tener en cuenta que las empresas financieras necesitan decisiones rápidas).

1.7. ¿Existe un método mejor?

En general, y como se puede comprobar a lo largo de las investigaciones realizadas por los distintos autores, no se puede concluir que exista un método mejor para medir el credit scoring. Dependiendo de la estructura de los datos, las características utilizadas, la posibilidad de separar las clases mediante el uso de esas características y el objetivo de la clasificación de la estructura de datos, será más apropiado la utilización de un método u otro (Hand and Henley, 1997; Yu and Lai, 2008).

En la investigación realizada por Henley and Hand (1996) se señala como método superior el de k vecinos más cercanos con adaptación métrica, pero al mismo tiempo advierten que la investigación se ha realizado con un conjunto de datos pequeño (5 000) y con una tasa de aceptación del 70 %. Los mismos autores conluyen que los métodos de clasificación más fáciles de entender como la regresión, k vecinos más cercanos y árboles de decisión son mucho más atractivos tanto para los usuarios como para los clientes que las redes neuronales que son cajas negras, difíciles de entender e interpretar.

Galindo and Tamayo (2000) señalan que ningún método o algoritmo es perfecto, por lo que el investigador debe de ser consciente de las limitaciones y fortalezas de cada uno.

Thomas (2000) realiza un estudio de las diferentes técnicas de clasificación de credit scoring utilizadas por diversos autores en distintos trabajos de investigación: Henley (1995), Boyle et al. (1992), Srinivasan and Kim (1987a), Srinivasan and Kim (1987b), Yobas et al. (1997), Desai et al. (1997). Se analizan tanto los distintos métodos de credit scoring aplicados como los resultados obtenidos en cuanto al método o métodos óptimos. El primer problema que destaca es que la información, en cuanto a la disposición de las bases de datos reales de préstamos, es limitada para los investigadores. Los estudios existentes coinciden en no señalar ningún método como óptimo, concluyen que dependiendo qué métodos se comparen, qué variables se utilicen, de qué país se extraen los datos, ... cada trabajo llega a unas conclusiones diferentes. Pero, en lo que sí coinciden todos los autores es que es mejor utilizar un método de credit scoring que no utilizar ninguno. Cada método muestra unas ventajas y unos inconvenientes. Thomas (2000), señala la importancia de incluir dentro del credit scoring el comportamiento del scoring (información conocida del prestatario de un período anterior), las condiciones macroeconómicas existentes en cada momento y el beneficio del scoring.

En el trabajo de Baesens et al. (2003b) se señala que las técnicas de clasificación más simples, como el análisis discriminante lineal y la regresión logística obtienen unos buenos resultados.

Huang et al. (2004) señalan que las técnicas de inteligencia artificial son superiores a las técnicas estadísticas para medir el credit scoring, pero indican que dentro de las técnicas de inteligencia artificial no se puede afirmar una como mejor. Otros autores como Huang and Wang (2007) señalan que las técnicas de minería de datos o inteligencia artificial, desarrolladas más recientemente como las redes neuronales, programación genética y máquinas de vectores soporte, pueden realizar la tarea de clasificación sin la limitación de las técnicas convencionales sobre la precisión del credit scoring. Al mismo tiempo estos métodos logran un mejor rendimiento.

En la misma línea que el estudio de Thomas (2000), Crook et al. (2007) realizan un estudio comparativo de los distintos métodos aplicados en las investigaciones de distintos autores como: Srinivasan and Kim (1987b), Boyle et al. (1992), Henley (1995), Yobas and Ross (2000) , West (2000), Lee et al. (2002), Malhotra and Malhotra (2003), Baesens et al. (2003b) y Ong et al. (2005). Analizan el porcentaje de los créditos correctamente clasificados comprobando que obtienen resultados distintos cada uno de los autores, por tanto no hay un método óptimo. En el artículo de Crook et al. (2007) llegan a la conclusión que el método más utilizado en el credit scoring es la regresión logística, aunque en los últimos años se han desarrollado más métodos y la evidencia sugiere que el más exacto probablemente sea el método de máquina de vectores soporte, aunque falta investigación sobre el mismo para poder concluir esta afirmación.

Otros autores como Abdou and El Masry (2008) señalan que en la mayoría de casos estudiados, en cuanto a comparación de métodos de credit scoring, las técnicas estadísticas más avanzadas como redes neuronales y algoritmos difusos son mejores que las tradicionales. Pero al mismo tiempo, no hay ninguna diferencia entre las diferentes técnicas estadísticas en términos de porcentaje de tasa media de clasificación correcta. Tal y como indican Desai et al. (1996), Blochlinger and Leippold (2006), Hoffmann et al. (2007), en algunos casos dependerá del grupo original que se utiliza para calcular la clasificación correcta de buenos y malos créditos.

En los trabajos de Huang et al. (2009), Lean Yu et al. (2008) se introduce en las técnicas de inteligencia artificial la posibilidad de estudiar métodos de aprendizaje conjuntos obteniendo mejores resultados que las técnicas individuales de inteligencia artificial. En esta línea el trabajo de Wang and Ma (2011) introducen tres métodos de aprendizaje conjunto (bagging, boosting y stacking) para distintos métodos de clasificación comparándolo con el método clásico de una muestra única, los resultados revelan que la aplicación del aprendizaje conjunto puede aportar una mejora sustancial para todos los métodos de clasificación, sin definirse por ningún método de clasificación como el mejor.

En cualquier caso, tal y como señalan Marqués et al. (2013) comparando los métodos subjetivos de credit scoring con cualquiera de los métodos automáticos de computación, estos últimos presentan una serie de ventajas interesantes (Rosenberg and Gleit, 1994; Thomas and Edelman, 2002; Blochlinger and Leippold, 2006):

1. La reducción en el coste del proceso de evaluación de crédito y el riesgo esperado de ser un mal préstamo

2. El ahorro de tiempo y esfuerzo

3. Recomendaciones coherentes basadas en información objetiva, lo que elimina los prejuicios humanos

4. Facilidad para incorporar cambios en el sistema

5. El proceso puede ser supervisado, rastreado y ajustado en cualquier momento.

Marqués et al. (2013) realizan una revisión de la literatura sobre la aplicación de la computación evolutiva para el credit scoring limitado a los artículos publicados en el período 2000-2012. Los autores, coincidiendo con Lucas (2001), destacan que la aplicación de las técnicas de computación evolutiva en el credit scoring son buenas para la predicción, la selección de variables y optimización de parámetros, pero sus resultados difícilmente interpretables por los analistas. Se admite que los algoritmos genéticos y programación genética son métodos eficientes y flexibles para encontrar soluciones óptimas. Pero no es posible concluir que los algoritmos genéticos y la programación genética son mejores o peores que otros modelos, simplemente es una alternativa a los métodos convencionales. Señalan como solución óptima la combinación de métodos de computación evolutiva con convencionales. Al mismo tiempo, en el estudio realizado, ponen de manifiesto debilidades y limitaciones de las investigaciones realizadas hasta el período de análisis a tener en cuenta en investigaciones futuras:

1. Falta de bases de datos de crédito disponibles públicamente y las pocas existentes no tienen un número suficientemente grande de datos. Este hecho limita tanto para generalizar conclusiones como para la realización de comparaciones de métodos y experimentos.

2. En los artículos referidos existen variedad de criterios de evaluación del desempeño, pero algunos no son los más adecuados debido a los diferentes costes de clasificación erróneo y/o distribuciones de clase desequilibradas. Menos del 20 % de los artículos utilizan los errores de tipo I y tipo II, siendo un dato a tener en cuenta en las aplicaciones reales ya que existe bastante consenso en las entidades financieras que los costes asociados a los errores de tipo II, malos clientes clasificados como buenos, son mucho mayores que los errores de tipo I, buenos clientes clasificados como malos (West, 2000; Baesens et al., 2003b).

3. Las conclusiones de la mayoría de los autores no han sido apoyadas por pruebas estadísticas para demostrar la importancia de los resultados experimentales y, en otros casos la estadística aplicada es inadecuada debido a las suposiciones incorrectas que han realizado.

4. En cuanto al pre-procesamiento de datos se refiere, los autores observan que en general, los conjuntos de datos de crédito son muy desequilibrados (número de casos por defecto bajo, datos solo de los que se les concedió en el pasado el crédito, datos incompletos). Los autores han observado que la selección de variables es el único problema de pre-procesamiento de datos que, en gran medida, ha sido abordado con técnicas evolutivas, pero no en todos los estudios.

A lo largo del capítulo se han enumerado distintos métodos utilizados para analizar el riesgo de crédito en las entidades financieras. Se realiza una revisión de los métodos utilizados desde mediados del siglo XX (principio de los años 40), hasta nuestros días. Las aportaciones científicas al análisis del riesgo de crédito, a lo largo de estos años, han sido importantes debido, en gran

parte, a los avances tecnológicos. Las técnicas de credit scoring incluyen tanto métodos estadísticos, como econométricos, de inteligencia artificial, de investigacin operativa e híbridas. Con este capítulo se ha dado respuesta al objetivo número 1 de la sección “Objetivos” del “Prólogo”.

2

Métodos estadísticos para el estimación del riesgo de crédito

Tal y como se ha indicado en el capítulo 1, si se cuestiona qué método es el mejor para predecir la probabilidad de éxito o por el contrario, la probabilidad de incumplimiento en un préstamo, no se puede concluir que exista ninguno mejor o peor, pues en primer lugar habrá que definir qué es mejor, y en segundo lugar, en la literatura de la materia no existe un consenso. Dependiendo de la estructura de los datos, las características utilizadas, la posibilidad de separar las clases mediante el uso de esas características y el objetivo de la clasificación de la estructura de datos, será más apropiado la utilización de un método u otro (Hand and Henley, 1997; Yu and Lai, 2008).

En el presente capítulo, se exponen los cinco métodos que por su habitual uso, se han considerado en esta tesis para su estudio y aplicación. Entre todas la técnicas aplicadas y contrastadas en la literatura de credit scoring, se han seleccionado aquellas que han obtenido buenos y contrastados resultados dentro de las técnicas paramétricas, técnicas no paramétricas, técnicas estadísticas y de inteligencia artificial. Dentro de las técnicas paramétricas y no paramétricas, se han seleccionado el análisis discriminante y árboles de decisión y clasificación CART; como técnicas estadísticas, se han seleccionado el modelo lineal mixto y modelo lineal generalizado para datos binarios; como técnicas de inteligencia artificial, se ha seleccionado las máquinas de vectores soporte. En este capítulo se va a plantear brevemente la metodología de cada una de las técnicas utilizadas.

En ningún momento se va a entrar a desarrollar la obtención de los estimadores, pues es un hecho que se puede encontrar en la literatura.

2.1. Método de Análisis Discriminante

El análisis discriminante es una técnica paramétrica enumerada por Fisher (1936). La finalidad de esta técnica es ver si existen diferencias entre los dos grupos de la variable objetivo para las que se miden p variables explicativas o también llamadas discriminantes. Se asume que las variables independientes $\boldsymbol{X}$ de la muestra provienen de una distribución normal multivariada. La función de densidad condicional, $f_k(\boldsymbol{x})$, utilizada en análisis discriminante es una estimación de la probabilidad a posteriori $\mathbb{P}(G = G_k|\boldsymbol{X} = \boldsymbol{x})$. La función $f_k(\boldsymbol{x})$ es la siguiente:

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\mu_k)^T\Sigma_k^{-1}(\boldsymbol{x}-\mu_k)\right\}, \quad para \quad \boldsymbol{x} \in \mathbb{R}^p$$

Para el uso de esta técnica se ha de definir una función discriminante, por ejemplo la distancia de Mahalanobis (Mahalanobis, 1936). Esta compara la distancia entre $\boldsymbol{x}$ y el centro de la clase k^{th} y viene expresada por:

$$\delta(\boldsymbol{x}, \mu_k) = [(\boldsymbol{x}-\mu_k)^T\Sigma_k^{-1}(\boldsymbol{x}-\mu_k)]^{1/2}$$

Suponiendo que se clasifica en dos clases 0 y 1, se compara la razón de ambas probabilidades a posteriori. La expresión matemática (Hastie et al., 2008) sería:

$$log\frac{\mathbb{P}(G = G_0|\boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}(G = G_1|\boldsymbol{X} = \boldsymbol{x})} = log\frac{f_0(\boldsymbol{x})}{f_1(\boldsymbol{x})} + log\frac{\pi_{(0)}}{\pi_{(1)}} = \tag{2.1}$$

$$log\frac{\pi_{(0)}}{\pi_{(1)}} + log\frac{|\Sigma_0|}{|\Sigma_1|} - \frac{1}{2}(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1) \quad +$$
$$\boldsymbol{x}^T(\Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1) - \frac{1}{2}\boldsymbol{x}^T(\Sigma_0^{-1} - \Sigma_1^{-1})\boldsymbol{x}$$

Si la matriz de varianzas Σ_k es igual para todas las clases k, el método discriminante es el análisis discriminante lineal (homocedasticidad) y si, por el contrario, no son iguales (heterocedasticidad), es análisis discriminante cuadrático.

Análisis discriminante lineal

El análisis discriminante hace referencia al problema de clasificar en distintos grupos un conjunto de observaciones vectoriales. A partir de una muestra conocida (training o muestra de entrenamiento) en la que se conocen las características (variables exógenas $\boldsymbol{x}$) de cada individuo, se clasifica su respuesta (variable endógena $\boldsymbol{y}$) en función de esas características. Se obtendrá una relación que separe, atendiendo a sus características, cada observación en uno de los dos grupos respuesta. Al mismo tiempo, para cada nueva observación se podrá predecir a qué grupo pertenecerá.

Se considera análisis discriminante lineal si la función de separación aplicada de las variables es lineal, y se emplea esa función para predecir la pertenencia de una nueva observación a

alguno de los grupos. El análisis discriminante lineal resulta adecuado cuando las variables provienen de una distribución normal multivariada con igual varianza dentro de cada grupo (homocedasticidad), esto requiere que las variables independientes tienen que ser cuantitativas, pero sus resultados pueden no ser válidos ante la presencia de pocos valores extremos (Khattree and Naik, 2000).

Existe una serie de supuestos que se deben de cumplir para la aplicación del análisis discriminante lineal:

1. Se ha de tener una variable categórica como variable respuesta y un conjunto de variables explicativas, independientes que pueden ser continuas o discretas ($\boldsymbol{x}$). La variable respuesta ha de cumplir que, $\boldsymbol{Y}_k = 1 \quad si \quad G_k = k, \quad \boldsymbol{Y}_k = 0 \quad si \quad G_k \neq k$

2. Es necesario que como mínimo existan dos clases (k y m), y para cada una de ellas como mínimo dos o más observaciones.

3. El número de variables discriminantes (las que se utilizarán como características en el modelo para poder discriminar por clases) debe ser menor que el número de variables totales (p) menos dos.

4. Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes.

5. El número máximo de funciones discriminantes (δ_k) es igual al mínimo entre el número de variables (p) y el número de grupos (K) menos 1.

6. La matriz de covarianzas dentro de cada clase debe ser aproximadamente igual. ($\Sigma = \Sigma_k \quad \forall \in 1, \ldots, K$).

7. Las variables continuas deben seguir una distribución normal multivariante.

Suponiendo que se clasifica en dos clases k y m, se compara la razón de ambas probabilidades a posteriori. La probabilidad a priori de la clase k es conocida y su valor se expresa como $\pi_{(k)}$, cumpliéndose que $\sum_{k=1}^{K} \pi_{(k)} = 1$, al mismo tiempo se cumple que la matriz de varianzas Σ_k es igual para todas las clases k, por tanto, $\Sigma = \Sigma_k$. La expresión matemática (Hastie et al., 2008) de la comparación de las probabilidades a posteriori es la siguiente:

La función discriminante lineal viene dada por la siguiente expresión:

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log \pi_k$$

donde el número de funciones discriminantes k será,

$$k = min(p, k-1) \quad y \quad corr(\delta_i, \delta_j) = 0 \quad \forall i \neq j$$

.

En el caso de una variable respuesta dicotómica la regla de decisión será igual a 1. La regla de decisión es $\hat{G(x)} = argmax_k \mathbb{P}(G = G_k | \boldsymbol{X} = \boldsymbol{x})$. El límite de decisión entre las clases k y m será cuando se cumpla que $\boldsymbol{x} : \delta_k(\boldsymbol{x}) = \delta_m(\boldsymbol{x})$ lo que equivale a que la expresión matemática 2.1 se iguale a cero.

Las funciones $\delta_1, \delta_2, \ldots, \delta_k$ denominadas funciones discriminantes canónicas, se construyen de modo que:

1. δ_1 sea la combinación lineal de $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ que mejor diferencia las k clases.

2. δ_2 sea la combinación lineal de $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ que mejor diferencia las k clases después de δ_1 tal que $corr(\delta_2, \delta_1) = 0$.

En general, δ_i es la combinación lineal de $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ que proporciona la mayor discriminación posible entre las clases después de δ_{i-1} y tal que $corr(\delta_i, \delta_j) = 0$ para $j = 1, \ldots, (i-1)$.

Por tanto, conociendo las p variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ sobre un nuevo individuo, es posible clasificarlo en uno de las K clases a partir de las funciones discriminantes $\delta_1, \delta_2, \ldots, \delta_k$.

En la práctica, los parámetros de la distribución gaussiana multivariable se estiman a partir de las observaciones $(\boldsymbol{x}_i, g_i)$ dentro del conjunto de entrenamiento, donde:

$$\hat{\pi_{(k)}} = \frac{N_k}{N};$$

$$\hat{\mu_k} = \sum_{g_i = k} \frac{\boldsymbol{x}_i}{N_k},$$

$$\hat{\Sigma} = \sum_{k=1}^{K} \frac{\sum_{g_i = k} (\boldsymbol{x}_i - \hat{\mu_k})^T}{N - K},$$

Donde, $N_k = \sum_{i=1}^{N} 1(g_i = G_k)$ es el número de observaciones de la clase k, N es el número de observaciones totales. Por tanto, se va a necesitar la diferencia entre las funciones discriminantes $\delta_k(\boldsymbol{x}) - \delta_K(\boldsymbol{x})$, cada una requiere $(p+1)$ parámetros, por tanto tendremos $(k-1)(p+1)$ parámetros en el ajuste con análisis discriminante lineal.

Se calculan los valores para esas k funciones en el registro nuevo y luego la distancia a cada uno de los vectores de las k funciones valorizadas en los promedios de K clases. El nuevo registro se le asigna a la clase cuyo promedio se encuentra a menor distancia. Si además se conoce la probabilidad a priori que un registro pertenezca a cada una de las K clases, puede usarse para mejorar la clasificación.

Para el cálculo del riesgo de crédito o la clasificación de riesgos mediante análisis discriminante se utilizará el paquete MASS de R, desarrollado por Venables and Ripley (2002).

2.2. Árboles de decisión

Los árboles de decisión se basan en la aplicación de un conjunto de reglas, del tipo If-Then. Utilizan funciones lógicas que llevan a tomar decisiones en un sentido u otro. Las funciones If-Then toman valores discretos. Los árboles de decisión permiten analizar un mapa de posibilidades y cuantificar cada una de ellas en términos de probabilidad. Cuanta mayor información se tenga más eficiente será el cálculo.

Para la elaboración del árbol se parte de un nodo inicial denominado raíz. Las distintas alternativas If-Then existentes son los nodos, es decir, se parte la base de datos en conjuntos disjuntos por cada nodo. Este corte vendrá dado por una condición en una de las variables explicativas. El particionamiento será recursivo y se detiene en los nodos terminales cuando no exista discriminación posible de la variable respuesta para cada una de las explicativas. A cada nodo final se le asignará uno de los estados de la variable respuesta, a estos se les denominará hojas.

Para cada nueva observación, el estado de la variable respuesta se predice por el estado del nodo terminal al que dicha observación pertenece.

El criterio que determina el corte en cada nodo, el mecanismo de segmentación y el criterio de parada (que permite decidir si un nodo es terminal o no) ha dado lugar a distintos algoritmos de segmentación: CHAID, desarrollado por Kass (1980)); CART, desarrollado por Breiman et al. (1984); ID3, desarrollado por Quinlan (1983); C4.5, version superior al ID3 desarrollada por Quinlan (1993).

Algoritmo CART

Breiman et al. (1984), desarrolló el algoritmo Classification and Regression Trees (CART) que es un tipo específico de árbol de decisión en el que las ramas representan conjuntos de decisiones y cada decisión genera reglas sucesivas para continuar la clasificación (partición) formando así grupos homogéneos respecto a la variable que se desea discriminar. CART es un método no paramétrico de segmentación binaria.

CART trabaja con variables tanto discretas como continuas. El corte con cada nodo viene dado por reglas de tipo binario. Es un algoritmo de partición recursiva. Se parte de una muestra original y se divide en submuestras utilizando unas reglas univariantes If-Then que buscarán aquella variable independiente que permita discriminar mejor la decisión. El proceso finaliza cuando resulte imposible realizar una nueva división que mejore la homogeneidad existente. El riesgo total del árbol se calcula sumando los riesgos correspondientes a cada uno de los nodos terminales.

El criterio para seleccionar la mejor división de cada nodo es el de reducción de la impureza del nodo, definida mediante la siguiente expresión.

$$i(t) = -\sum_{j} p(j/t) \quad log[p(j/t)]$$

Siendo $p(j/t)$ la proporción de la clase j en el nodo t.

Como medida de la homogeneidad o impureza se utiliza una extensión del índice de Gini para respuestas categóricas. El algoritmo optará por aquella división que mejore la impureza, mejora que se mide comparando la que presenta el nodo de procedencia con la correspondiente a las dos regiones obtenidas en la partición.

Este procedimiento, tal y como señala Friedman (1977) presenta el problema del sobreaprendizaje. Para evitar este problema una vez construido el árbol saturado, se inicia el proceso de poda, es decir, eliminar las divisiones que supongan un mayor coste de complejidad (Death and Fabricius, 2000) hasta encontrar el tamaño óptimo que será aquel que minimice ese coste. La calidad de un árbol viene determinada por la calidad de sus nodos terminales, que tienen que ser homogéneos.

Entre otras, las ventajas del método CART son su robustez a outliers, su firme estructura a transformaciones de las variables independientes, y su interpretabilidad, es un método de fácil comprensión de los resultados obtenidos.

Para la clasificación de riesgos mediante el algoritmo CART se utilizará el paquete rpart de R, desarrollado por Therneau et al. (2014).

2.3. Modelo lineal mixto

Los modelos lineales mixtos (LMM) tienen una estructura jerárquica y multinivel más compleja que los modelos lineales (LM). Las observaciones de distintos niveles o clusters son independientes, pero las observaciones dentro de un mismo cluster son dependientes ya que comparten un efecto aleatorio asociado a la subpoblación. En el contexto de los modelos mixtos se suele hablar de dos tipos de variabilidad: entre e intra clusters. La modelización de estos tipos de variabilidad es lo que da aplicabilidad a los LMM. Los LMM manejan conjuntos de datos en los que las observaciones no son independientes. Estos modelos permiten modelizar correlaciones entre efectos aleatorios, lo cual los hace muy aptos para el tratamiento estadístico de datos. Los LMM le dan un carácter aleatorio a determinadas variables categóricas, estimando su varianza y prediciendo su valor. Mediante los efectos aleatorios, el investigador puede hacer inferencias sobre poblaciones con una estructura más compleja que en el caso de usar LM. Razón por la cual uno de los objetivos del análisis de los modelos mixtos es el de estimar las varianzas de cada uno de los efectos aleatorios, así como sus covarianzas. Se pueden citar los siguientes textos: Jiang (2007), Demidenko (2004), Pérez-Martín (2008).

2.3.1. Modelos lineales mixtos con un factor aleatorio

La expresión matricial de un modelo lineal mixto es:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}, \tag{2.2}$$

donde $\boldsymbol{y}_{n\times 1}$ es el vector de observaciones, $\boldsymbol{\beta}_{p\times 1}$ es el vector de efectos fijos, $\boldsymbol{u}_{m\times 1}$ es el vector de efectos aleatorios, $\boldsymbol{X}_{n\times p}$ y $\boldsymbol{Z}_{n\times m}$ son las matrices de incidencia correspondientes y $\boldsymbol{e}_{n\times 1}$ es el vector de perturbaciones aleatorias. Se supone que el vector de efectos aleatorios $\boldsymbol{u}$ y el vector de perturbaciones $\boldsymbol{e}$ son independientes, se distribuyen como una normal, sus medias nulas y sus matrices de varianzas-covarianzas conocidas,

$$\boldsymbol{e} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{V}_0),$$

$$\boldsymbol{u} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{V}_1),$$

$$V[\boldsymbol{e}] = E[\boldsymbol{e}\boldsymbol{e}^t] = \boldsymbol{V}_0, \quad \textit{siendo} \quad \boldsymbol{V}_0 = \sigma_0^2 \boldsymbol{I}_n$$

$$V[\boldsymbol{u}] = E[\boldsymbol{u}\boldsymbol{u}^t] = \boldsymbol{V}_1, \quad \textit{siendo} \quad \boldsymbol{V}_1 = \sigma_u^2 \boldsymbol{I}_m$$

que dependen de unos parámetros $\boldsymbol{\theta}$ llamados componentes de la varianza. De (2.2) se obtiene:

$$\boldsymbol{V} = V[\boldsymbol{y}] = V[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}] = V[\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}] = \boldsymbol{Z}\boldsymbol{V}_1\boldsymbol{Z}^t + \boldsymbol{V}_0 = \boldsymbol{Z}\sigma_u^2\boldsymbol{I}_m\boldsymbol{Z}^t + \sigma_0^2\boldsymbol{I}_n = \boldsymbol{V},$$

siendo $\boldsymbol{V} = \sigma_u^2 \boldsymbol{Z}\boldsymbol{Z}^t + \sigma_0^2 \boldsymbol{I}_n$.
Se supone que $\boldsymbol{V}$ es no singular.
El *estimador de máxima verosimilitud* (ml) de $\boldsymbol{\beta}$ es,

$$\widehat{\boldsymbol{\beta}} = argmax_{\boldsymbol{\beta}} \left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right).$$

El uso de un modelo lineal mixto (LMM) cuando la distribución condicionada de la variable respuesta es binomial,

$$\boldsymbol{y} \sim Bin(n, p)$$

está plenamente justificada gracias al teorema de Moivre-Laplace (de Moivre, 1756) en el que se aproxima una distribución binomial a una normal bajo ciertas condiciones. Estas condiciones se cumplen sobradamente bajo un escenario BigData. La elección de este modelo, en lugar de un modelo lineal mixto generalizado para respuesta binaria, es que la expresión de verosimilitud de un modelo mixto es una integral sobre el espacio de efectos aleatorios y puede ser evaluada exactamente por máxima verosimilitud restringida, mientras que para un modelo mixto generalizado esta debe ser aproximada.

Para el cálculo del riesgo de crédito mediante un modelo lineal mixto y con método de ajuste reml se usará el paquete lme4 de R descrito por Bates et al. (2015b) y Bates et al. (2015a).

2.4. Modelo Lineal Generalizado para datos binarios

El Modelo lineal generalizado (GLM) tuvo mucha difusión a partir del libro de McCullagh and Nelder (1989).

En el GLM tenemos variables respuesta asociadas a covariables, pero a diferencia del modelo lineal, no se tiene por qué cumplir principios fundamentales que se cumplen en el modelo lineal como:

1. Aditividad de los efectos de las covariables.

2. Normalidad de la respuesta.

3. Homocedasticidad.

Se supone que tenemos $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ variables objetivo aleatorias univariantes e independientes, relacionadas con las covariables o variables explicativas $\boldsymbol{x}_{ji}$, desde $j = 1, \ldots, p$ en el i-ésimo elemento muestral, y $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ son parámetros desconocidos. Se supone que las distribuciones de las variables $\boldsymbol{Y}_i$, desde $i = 1, \ldots, n$ coinciden en la forma (perteneciente a la familia exponencial) pero no necesariamente en los parámetros. Es decir supongamos que la función de densidad de $\boldsymbol{Y}_i$ es:

$$f(\boldsymbol{y}_i, \boldsymbol{\theta}_i, \phi) = \exp \frac{\boldsymbol{y}_i \boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{a_i(\phi)} + c(\boldsymbol{y}_i, \phi) \tag{2.3}$$

Al plantear un modelo lineal generalizado se considera un conjunto de parámetros reducido $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$, siendo $p < n$, de modo que una transformación de $\eta_i = E[\boldsymbol{Y}_i]$ es una combinación lineal de ellos

$$g(E[\boldsymbol{Y}_i]) = \eta_i = \sum_{j=1}^{p} \boldsymbol{x}_{ij} \boldsymbol{\beta}_j = \boldsymbol{x}_i^t \boldsymbol{\beta}, \quad i = 1, \ldots, n.$$

donde,

1. g es una función monótona y diferenciable llamada nexo o link.

2. $\boldsymbol{x}_i^t = \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{ip}$ son los valores que toman las variables explicativas (covariables), $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, en el individuo i. Si la variable explicativa es continua, se utiliza una única variable $\boldsymbol{X}$. Si la variable explicativa es discreta (factor), se utilizan tantas variables $\boldsymbol{X}$ como niveles tenga el factor menos uno. Las variables artificiales toman los valores 1 y 0, es decir, que si el individuo i-ésimo pertenece al nivel k del factor, entonces la k-ésima variable artificial toma el valor 1 y las restantes el valor 0.

3. El vector de parámetros, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ es de dimensión p x 1. Las covariables tienen un único parámetro asociado, mientras que los factores tienen tantos parámetros asociados como niveles.

Por tanto, las componentes del modelo son:

1. Componente aleatoria. La variable respuesta $\boldsymbol{Y}$ tiene distribución exponencial donde $\boldsymbol{\theta}$ es el parámetro canónico, ϕ es el parámetro de ruido y las funciones $a()$, $b()$ y $c()$ son conocidas. Además se cumple que:

   a. $\boldsymbol{\mu} = \boldsymbol{E}(\boldsymbol{y}) = \boldsymbol{b}'(\boldsymbol{\theta})$

   b. $\boldsymbol{V}(\boldsymbol{Y}) = a(\phi)\boldsymbol{b}''(\boldsymbol{\theta})$

2. Componente sistemática. El vector de covaribles $\tilde{\boldsymbol{x}} = (\boldsymbol{x}_j, \ldots, \boldsymbol{x}_p)$ que da origen al predictor lineal

$$g(E[\boldsymbol{Y}_i]) = \eta = \sum_{i=1}^{p} \boldsymbol{\beta}_j \boldsymbol{X}_j i, \quad i = 1, \ldots, n,$$

   y $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ son parámetros deconocidos a estimar. Se va a tener, por tanto, un predictor lineal basado en una combinación lineal de variables explicativas. Las variables pueden ser continuas, categóricas o una mezcla de las dos.

3. Función nexo o link. Relaciona las dos componentes, la esperanza de la respuesta $\boldsymbol{\mu}$ y el predictor lineal η. La función $g(\boldsymbol{\mu})$ es continua y diferenciable, tal que $\eta_i = g(\boldsymbol{\mu}_i) = \boldsymbol{x}_i^t \boldsymbol{\beta}$, con $\boldsymbol{\mu}_i = E[\boldsymbol{Y}_i]$.
   Los links para los GLM se muestran en la tabla 2.1.

Familia	Normal	Poisson	Binomial	Gamma	Inversa
Link o nexo	$\eta = \boldsymbol{\mu}$	$\eta = log(\boldsymbol{\mu})$	$\eta = log(\boldsymbol{\mu}/(1-\boldsymbol{\mu})$	$\eta = \boldsymbol{\mu}^{-1}$	$\eta = \boldsymbol{\mu}^{-2}$
Función de varianza	1	$\boldsymbol{\mu}$	$\boldsymbol{\mu}(1-\boldsymbol{\mu})$	$\boldsymbol{\mu}^2$	$\boldsymbol{\mu}^3$

Tabla 2.1: Links posibles en un GLM

Los parámetros $\boldsymbol{\beta}$ se estiman utilizando el método de máxima verosimilitud. Dada las observaciones independientes $\boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n$, con funciones de densidad $f(\boldsymbol{y}_i, \boldsymbol{\theta}_i, \phi)$ (2.3), la función de verosimilitud es:

$$L(\boldsymbol{y}, \boldsymbol{\theta}, \phi) = \prod_{i=1}^{n} f(\boldsymbol{y}_i, \boldsymbol{\theta}_i, \phi) = exp\left[\sum_{i=1}^{n} \frac{\boldsymbol{y}_i\boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(\boldsymbol{y}_i, \phi)\right]$$

Donde $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, y $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$, es el parámetro canónico. La función de log-verosimilitud es:

$$l(\boldsymbol{y}, \boldsymbol{\theta}, \phi) = \sum_{i=1}^{n} \frac{\boldsymbol{y}_i\boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(\boldsymbol{y}_i, \phi) \doteq \sum_{i=1}^{n} l(\boldsymbol{y}_i, \boldsymbol{\theta}_i, \phi)$$

Suponiendo que el parámetro de escala, ϕ, es conocido. Se desea estimar $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$ por máxima verosimilitud. El estimador máximo verosimil es,

$$\hat{\boldsymbol{\beta}} = argmax_{\boldsymbol{\beta}} \quad l((\boldsymbol{y}, \boldsymbol{\theta}, \phi),$$

que se obtiene resolviendo las ecuaciones de verosimilitud:

$$U_j = \frac{\partial l(\boldsymbol{y}, \boldsymbol{\theta}, \phi)}{\partial \boldsymbol{\beta}_j} = 0, \quad j = 1, \ldots, p$$

Pero no siempre las ecuaciones de verosimilitud proporcionan soluciones explícitas para $\boldsymbol{\beta}_j, \quad j = 1, \ldots, p$. Es por ello que se debe de recurrir a métodos numéricos para la resolución, como:

1. El método de Newton-Raphson utiliza la ecuación recurrente:

$$\hat{\boldsymbol{\beta}}^{(r)} = \hat{\boldsymbol{\beta}}^{(r-1)} - H^{-1}(\hat{\boldsymbol{\beta}}^{(r-1)})U(\hat{\boldsymbol{\beta}}^{(r-1)}),$$

   Donde $U = \left(\frac{\partial l}{\partial \boldsymbol{\beta}_1}, \ldots, \frac{\partial l}{\partial \boldsymbol{\beta}_p}\right)^T$, $\hat{\boldsymbol{\beta}}^{(r)} = \left(\hat{\boldsymbol{\beta}_1}^{(r)}, \ldots, \hat{\boldsymbol{\beta}_p}^{(r)}\right)^T$, $H = \left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k}\right)_{j,k=1,\ldots,p}$

2. El método de las puntuaciones de Fisher (método de scoring) utiliza la ecuación recurrente:

$$\hat{\boldsymbol{\beta}}^{(r)} = \hat{\boldsymbol{\beta}}^{(r-1)} + \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}^{(r-1)})U(\hat{\boldsymbol{\beta}}^{(r-1)}), \tag{2.4}$$

   donde,

$$\boldsymbol{I} = -\boldsymbol{E}[H] = \sum_{i=1}^{n} \left(\boldsymbol{E}\left[-\frac{\partial^2 l_i}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_k}\right]\right)_{j,k=1,\ldots,p}$$

   Nótese que este método consiste en tomar esperanzas en el método Newton-Raphson pero cambiadas de signo.

3. Mínimos cuadrados ponderados iterados. Se considera el modelo lineal, con $\boldsymbol{e} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{V})$. Suponiendo que $\boldsymbol{V}_{n\times n}$ es una matriz conocida, definida positiva y simétrica, entonces existe una matriz invertible $\boldsymbol{K}_{n\times n}$ tal que $\boldsymbol{V} = \boldsymbol{K}\boldsymbol{K}^T$. Para obtener el estimador de máxima verosimilitud de $\boldsymbol{\beta}$ es suficiente con transformar el modelo de la siguiente forma:

$$\left.\begin{array}{c} \varepsilon = \boldsymbol{K}^{-1}\boldsymbol{Y} \\ \boldsymbol{M} = \boldsymbol{K}^{-1}\boldsymbol{X} \\ \epsilon = \boldsymbol{K}^{-1}\boldsymbol{e} \end{array}\right\} \Longrightarrow \varepsilon = \boldsymbol{M}\boldsymbol{\beta} + \epsilon,$$

   Donde, $\boldsymbol{e} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I})$.

   El estimador de máxima verosimilitud, $\hat{\boldsymbol{\beta}} = (\boldsymbol{M}^T\boldsymbol{M}^{-1})\boldsymbol{M}^T\varepsilon$, coincide con el estimador por mínimos cuadrados ponderados, es decir,

$$\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}}(\varepsilon - \boldsymbol{M}\boldsymbol{\beta})^T(\varepsilon - \boldsymbol{M}\boldsymbol{\beta}).$$

   Además,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}.$$

La función glm del software en R utiliza el ajuste por mínimos cuadrados ponderados iterados. Para el cálculo del riesgo de crédito mediante un modelo lineal generalizado se usará el paquete stats de R. Como función link se usará el nexo logit, R Core Team (2015).

2.5. Método de Máquinas de vectores soporte

La teoría de las máquinas de vectores soporte (SVM) fue desarrollada inicialmente por Vapnik (1995). El SVM es un método de clasificación y regresión basado en el principio de minimizacion de riesgo estructural. Se basa en la búsqueda de un hiperplano óptimo que separe las clases. Es una técnica robusta para la clasificación y regresión aplicada a grandes conjuntos de datos complejos con ruido. Se puede utilizar para resolver tanto problemas lineales como no lineales. Tal y como indica González-Abril (2003b) el problema consiste en buscar, para una tarea de aprendizaje dada con, una cantidad finita de datos, una adecuada función que permita llevar a cabo una buena generalización que sea resultado de una adecuada relación entre la precisión alcanzada con un particular conjunto de entrenamiento y la capacidad del modelo. La metodología de SVM tiene dos vertientes:

1. Si se trata de clasificar un conjunto de datos (representados en un plano n-dimensional) no separable linealmente, se toma el conjunto de datos y se mapea a un espacio de mayor dimensión donde se pueda separar linealmente (esto se realizará mediante las funciones Kernel). En el nuevo plano se busca el hiperplano que sea capaz de separar en 2 clases los datos de entrada, cumpliendo que el hiperplano de separación sea de máxima distancia posible a los puntos de ambas clases (los puntos más cercanos a este hiperplano de separación son los vectores de soporte).

2. Si se trata de hacer una regresión se toma el conjunto de datos y se transforma a un espacio de mayor dimensión (en el que sí se pueda hacer una regresión lineal). En la nueva dimensión se realiza la regresión lineal pero sin penalizar errores pequeños.

2.5.1. Separador lineal

Se considera para todos los casos que las variables independientes (o espacio de entrada) son de dimensión n, es decir, $\boldsymbol{X} \in \mathbb{R}^n$ la variable dependiente (o espacio de salida) es $\boldsymbol{Y} = [-1, +1]$, según el grupo de clasificación. Dentro de todos los hiperplanos de separación que pueden existir, tan sólo uno de ellos es óptimo. El hiperplano óptimo será aquel que cumpla que la distancia entre el hiperplano óptimo y el valor de entrada más cercano sea máxima (maximización del margen), véase el gráfico 2.1. Aquellos puntos sobre los cuales se apoya el margen máximo son los denominados vectores soporte.

Matemáticamente se expresa de la siguiente forma:

Dentro del conjunto de vectores de datos se elige un conjunto de vectores de entrenamiento $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, donde $\boldsymbol{x}_i \in \mathbb{R}^n$ e $y_i \in \{-1, +1\}$, para $i = 1, \ldots, n$ se dice

separable si existe algún hiperplano en $\mathbb{R}^n$ de dimensión $n-1$ que separa los vectores $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ con etiqueta $y_i = 1$ de aquellos con etiqueta $y_i = -1$. Dado un conjunto separable existe, al menos, un hiperplano

$$\pi_i : \boldsymbol{w}\boldsymbol{x} + b = 0$$

que separa los vectores $\boldsymbol{x}_i$, $i = 1, \ldots, n$. Siendo $\boldsymbol{w}$ el vector de pesos y b un umbral dado.

Dentro de los hiperplanos que cumplan la condición de separar los vectores por su clasificación, se buscará el hiperplano separador que maximice la distancia de separación entre los conjuntos $(\boldsymbol{x}_i, 1)$ y $(\boldsymbol{x}_i, -1)$.

Por tanto, el problema de optimización será:

$$\boldsymbol{x}_i\boldsymbol{w} + b \geq 1, \quad para \quad y_i = +1 \quad (regionA)$$

$$\boldsymbol{x}_i\boldsymbol{w} + b \leq -1, \quad para \quad y_i = -1 \quad (regionB)$$

Así pues, se obtienen dos hiperplanos:

$$\pi_1 : \boldsymbol{x}_i\boldsymbol{w} + b = 1$$

$$\pi_2 : \boldsymbol{x}_i\boldsymbol{w} + b = -1,$$

y el hiperplano de margen máximo:

$$H : \boldsymbol{x}_i\boldsymbol{w} + b = 0$$

Los vectores soporte son aquellos que definen los hiperplanos de separación π_1 y π_2.

De esta forma la mínima separación entre los vectores y el hiperplano separador es la unidad, por tanto las dos desigualdades se pueden expresar como:

$$y_i(\boldsymbol{x}_i\boldsymbol{w} + b) - 1 \geq 0,$$

siendo $i = 1, \ldots, n$

Esta será la restricción que deberá cumplir el hiperplano objetivo H. En el gráfico 2.1 se puede observar.

La distancia entre el hiperplano objetivo H y cada hiperplano separador π_1 y π_2 es $\frac{1}{\|w\|}$, por tanto el margen es $\frac{2}{\|w\|}$.

El objetivo será buscar los valores $\boldsymbol{w}$ y b que permitan maximizar el margen, la distancia entre π_1 y π_2, siendo $\frac{2}{\|w\|}$ o lo que es lo mismo matemáticamente, minimizar $\frac{1}{2}(\|w\|)^2$.

Por tanto se tendrá que resolver:

$$\left.\begin{array}{cc} \min_{\boldsymbol{w}\in\mathbb{R}^n} & \frac{1}{2}(\|w\|)^2 \\ sujeto \quad a \quad & y_i(\boldsymbol{x}_i\boldsymbol{w} + b) - 1 \geq 0, \quad i = 1, \ldots, n \end{array}\right\}$$

Para resolver este problema de optimización con restricciones se utilizan los multiplicadores de Lagrange.

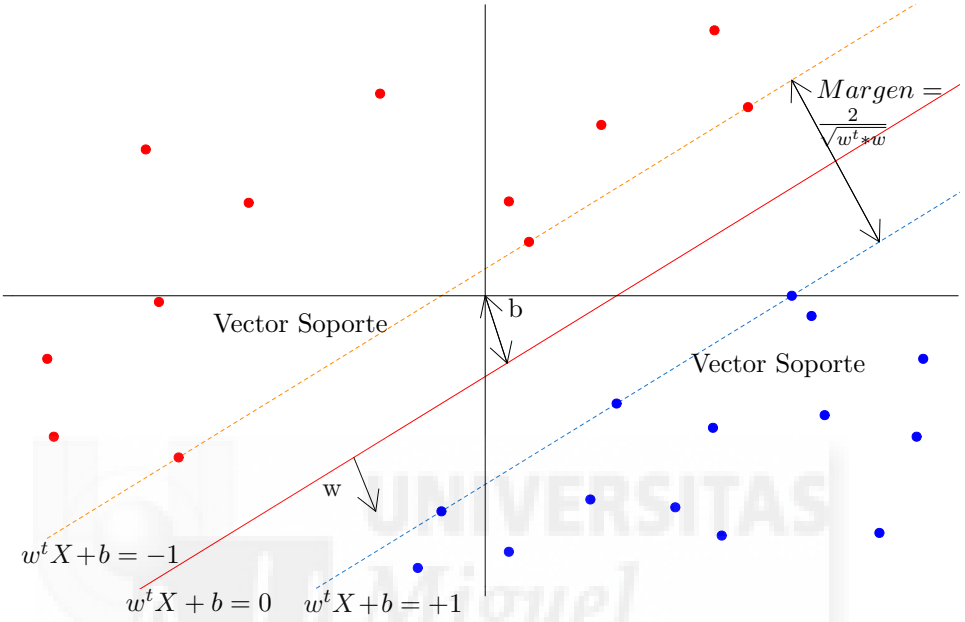

Gráfico 2.1: SVM lineal

2.5.2. Separador lineal para datos no separables

En la práctica lo habitual es que existan vectores de una clase dentro de la región correspondiente a los vectores de otra clase, por tanto no es posible separar por medio de hiperplanos. La solución para convertir en separable el conjunto de vectores no separable es la introducción de una variable de holgura ξ_i en las restricciones:

$$\boldsymbol{x}_i\boldsymbol{w} + b \geq 1 + \xi_i, \quad para \quad y_i = +1$$

$$\boldsymbol{x}_i\boldsymbol{w} + b \geq -1 + \xi_i, \quad para \quad y_i = -1$$

$$\xi_i \geq 0 \quad \forall_i = 1, \ldots, n$$

En este caso, se admite que necesariamente se van a cometer errores de clasificación, por tanto se reflejará en la función objetivo un parámetro C que denotará una cota máxima de errores. Por tanto el problema de optimización se planteará con la siguiente función:

$$\left.\begin{array}{rl} \min\limits_{\boldsymbol{w}\in\mathbb{R}^n} & \frac{1}{2}(\|w\|)^2 + C\sum\limits_{i=1}^{n}\xi_i \\ \text{sujeto a:} \quad y_i(\boldsymbol{x}_i\boldsymbol{w}+b) - 1 + \xi_i \geq 0, & i = 1,\ldots,n \\ & \xi_i \geq 0 \quad \forall i \end{array}\right\}$$

Para resolver este problema de optimización con restricciones se utilizará igualmente multiplicadores de Lagrange obteniéndose las condiciones Karush-Kuhn-Tucker (KKT).

2.5.3. Separador no lineal

Los datos no pueden ser separados linealmente a través de un hiperplano óptimo en el espacio de entrada. Se realiza una transformación no lineal del espacio de entradas mediante una función llamada Kernel, en la que sí que será posible la separación lineal, véase gráficamente en 2.2.

La función Kernel o núcleo (K) es un producto interno en el espacio característico de dimensión superior que tiene su equivalente en el espacio de entrada (Gunn, 1997). La función Kernel tiene que ser una función simétrica y definida positiva. Kernel será una función de transformación no lineal provista de un producto escalar en un espacio de mayor dimensionalidad. Siendo $\varphi(x)$ la transformación no lineal, la función Kernel será:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)^T\varphi(\boldsymbol{x}_j)$$

Las distintas funciones Kernel más empleadas son las siguientes:

1 Lineal $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T\boldsymbol{x}_j$

2 Gausiano $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right)$

3 Polinómica $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T\boldsymbol{x}_j + 1)^n$

$$k(\boldsymbol{x}, \boldsymbol{x}^{'}) = \theta(\boldsymbol{x})\theta(\boldsymbol{x}^{'}) = \{\theta(\boldsymbol{x}), \theta(\boldsymbol{x}^{'})\}_H$$

Para la clasificación del riesgo de crédito mediante SVM se utilizará el paquete e1071 (Meyer et al., 2014) de R.

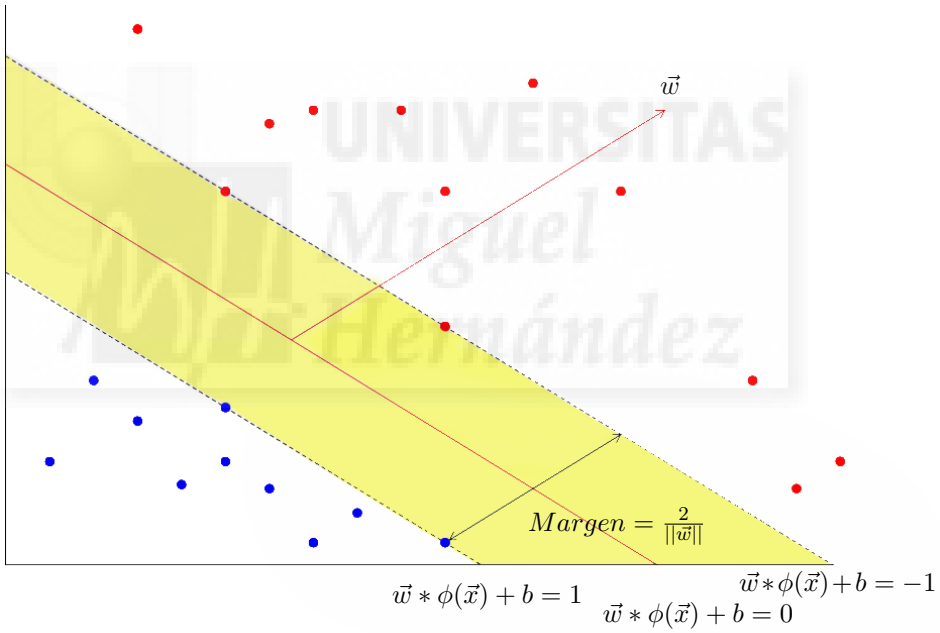

Gráfico 2.2: SVM Kernel

UNIVERSITAS
Miguel
Hernández

3

Evaluación de la robustez en modelos lineales

3.1. Introducción

Para evaluar el riesgo de crédito, existen una variedad de metodologías disponibles, desde el estudio personalizado de un analista de riesgo con experiencia, a diferentes métodos estadísticos y econométricos, tal y como se ha visto en el capítulo 1. Al mismo tiempo, la literatura científica aún no ha resuelto el problema de asignar un método estadístico o econométrico para resolver de manera eficiente la estimación del riesgo de crédito bancario en entidades financieras.

En este capítulo se proponen 3 experimentos para analizar la robustez de las estimaciones obtenidas mediante modelos de regresión lineal. Se proponen dos modelos, un modelo lineal simple con efectos fijos (3.4) y un modelo lineal mixto con un factor aleatorio (3.2). El objetivo es poner a prueba la robustez de las estimaciones de la variable objetivo y de los parámetros del modelo mediante un amplio abanico de modificaciones de las condiciones básicas que tienen que cumplir los datos para poder ser tratados mediante estos modelos. Al mismo tiempo también se realiza un estudio de dos métodos para el ajuste de estos modelos, la máxima verosimilitud (ml) y la máxima verosimilitud restringida (reml).

Para evaluar la adecuación de estos experimentos, se realizan simulaciones Monte Carlo que generan muestras atendiendo a una amplia diversidad en la casuística en la generación de la variable objetivo, manteniendo las variables explicativas que intervienen, así se pueden comparar ciertas propiedades entre ellos (Morales Gonzalez et al., 2013).

En la siguiente tabla 3.1 se resumen los experimentos llevados a cabo con la casuística de generación de la variable objetivo y de los errores:

Los objetivos que se pretenden alcanzar en este capítulo son:

	Generación variable objetivo		Modelo y método empleado en el ajuste
	Variable objetivo	Distribución del error	
Experimento 1	Como si se tratase de un modelo lineal simple (Y0) y un modelo lineal mixto (Y1)	$N_n(\mathbf{0}, \sigma_0^2 \boldsymbol{I}_n)$	ml (3.4)
Experimento 2	Como si se tratase de un modelo lineal simple (Y0) y un modelo lineal mixto (Y1)	$N_n(\mathbf{0}, \sigma_0^2 \boldsymbol{I}_n)$	ml y reml (3.2)
Experimento 3	Como si se tratase de un modelo lineal simple (Y0) y un modelo lineal mixto (Y1)	$Ga(1,1)$ y $We(1,1)$	ml y reml (3.2)

Tabla 3.1: Experimentos

a Establecer una prioridad del modelo y método a elegir según qué casos, en base a ciertos parámetros de eficacia o si al menos la elección de uno incorrecto afecta demasiado a las estimaciones.

b Comprobar si la falta de alguna de las condiciones básicas de los modelos, o cierto alejamiento de estas, influye en la capacidad de ajuste.

Al final del capítulo se presentan las conclusiones obtenidas en base a la casuística simulada en los distintos experimentos.

3.2. Modelos

Se considera el siguiente modelo lineal mixto con un efecto aleatorio como el modelo principal. Los efectos aleatorios poseen I niveles ($i = 1, \ldots, I$), y para cada nivel i se tienen n_i individuos o registros. La expresión matemática del modelo es la siguiente:

$$y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + u_i + w_{ij}^{-1/2} e_{ij}, \quad i = 1, \ldots, I, j = 1, \ldots, n_i, \tag{3.1}$$

donde y_{ij} es la variable objetivo para cada una de las observaciones j en el nivel i, $\boldsymbol{x}_{ij}$ es un vector fila de efectos fijos que contienen los valores de las p variables auxiliares o explicativas, w_{ij} son los pesos conocidos que aportan heterocedasticidad al modelo y $\boldsymbol{\beta}$ es un vector columna con los parámetros de regresión. Además, los efectos aleatorios u_i y los errores de e_{ij} se asume que son mutuamente independientes con distribución $u_i \sim N(0, \sigma_1^2)$ y $e_{ij} \sim N(0, \sigma_0^2)$ respectivamente. El modelo (3.1) se puede escribir de forma matricial como:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{W}^{-1/2}\boldsymbol{e}, \tag{3.2}$$

donde, $\boldsymbol{u} = \boldsymbol{u}_{1,I\times 1} \sim N_I(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I}_I)$, $\boldsymbol{e} = \boldsymbol{e}_{n\times 1} \sim N_n(\boldsymbol{0}, \sigma_0^2 \boldsymbol{I}_n)$ son independientes, $\boldsymbol{y} = \boldsymbol{y}_{n\times 1}$, $\boldsymbol{X} = \underset{1\leq i\leq I}{\text{col}}(\boldsymbol{X}_i)$ con $\text{rg}(\boldsymbol{X}) = p$, $\boldsymbol{X}_i = \underset{1\leq j\leq n_i}{\text{col}}(\boldsymbol{x}_{ij})$, $\boldsymbol{\beta} = \boldsymbol{\beta}_{p\times 1}$, $\boldsymbol{Z} = \underset{1\leq i\leq I}{\text{diag}}(\boldsymbol{1}_{n_i})$, $n = \sum_{i=1}^{I} n_i$, $\boldsymbol{I}_a$ es la matriz de identidad de orden a, $\boldsymbol{1}_a$ es el vector columna de dimensión a cuyos elementos son todos iguales a 1, $\boldsymbol{W} = \underset{1\leq i\leq I}{\text{diag}}(\boldsymbol{W}_i)$, $\boldsymbol{W}_i = \underset{1\leq j\leq n_i}{\text{diag}}(w_{ij})$ con $w_{ij} > 0$ conocido.

Bajo las condiciones del modelo (3.2), se cumple que:

$$\boldsymbol{V} = var(\boldsymbol{y}) = \boldsymbol{Z} var(\boldsymbol{u}) \boldsymbol{Z}^t + \sigma_0^2 \boldsymbol{W}^{-1} = \underset{1\leq i\leq I}{\text{diag}}(\boldsymbol{V}_i),$$

donde $\boldsymbol{V}_i = \sigma_1^2 \boldsymbol{1}_{n_i} \boldsymbol{1}_{n_i}^t + \sigma_0^2 \boldsymbol{W}_i^{-1}$.

Por tanto se tiene $y \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$. Si σ_0^2 y σ_1^2 son conocidas el *Predictor lineal insesgado óptimo* (BLUE) de $\boldsymbol{\beta}$ es el siguiente:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{V}^{-1} \boldsymbol{y},$$

pero habitualmente no se cumple en la realidad.

De forma general, se considerará el modelo lineal simple no ponderado por pesos,

$$y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + e_{ij}, \quad i = 1, \ldots, I, j = 1, \ldots, n_i, \tag{3.3}$$

o bien en notación matricial,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{3.4}$$

donde todas las definiciones de las variables son las mismas que en el modelo anterior. Nótese que son casos particulares de los modelos (3.1) o (3.2) respectivamente, cuando los efectos aleatorios no existen y los errores de la varianza son iguales (homocedasticidad).

Con el objetivo de abreviar esta sección y visto que es muy sencillo encontrar los estimadores máximo verosímiles de los modelos (3.3) y (3.4), se descarta el desarrollo de estos usando el modelo matricial (3.2) para la obtención de las estimaciones de los parámetros.

3.2.1. Estimación máximo verosímil (ml) en un modelo lineal mixto

La estimación máximo verosímil (ml) de $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \boldsymbol{\sigma}^t)^t = (\boldsymbol{\beta}^t, \sigma_0^2, \sigma_1^2)^t$ puede obtenerse mediante la maximización de la función log-verosimilitud

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

con el algoritmo Fisher-scoring

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \boldsymbol{F}(\boldsymbol{\theta}^{(k)})^{-1} \boldsymbol{S}(\boldsymbol{\theta}^{(k)}),$$

donde $\boldsymbol{S}(\boldsymbol{\theta})$ y $\boldsymbol{F}(\boldsymbol{\theta})$ son el $(p+2)\times 1$ vector de puntuación y la $(p+2)\times(p+2)$ matriz de información de Fisher respectivamente. Los elementos de $\boldsymbol{S}(\boldsymbol{\theta})$ y $\boldsymbol{F}(\boldsymbol{\theta})$ son

$$S_{\boldsymbol{\beta}} = \boldsymbol{X}^t\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) = \sum_{i=1}^{I} \boldsymbol{X}_i{}^t\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}),$$

$$S_{\sigma_1^2} = -\frac{1}{2}\sum_{i=1}^{I}\operatorname{tr}\left\{\boldsymbol{V}_i^{-1}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t\right\} + \frac{1}{2}\sum_{i=1}^{I}\left(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}\right)^t\boldsymbol{V}_i^{-1}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t\boldsymbol{V}_i^{-1}\left(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}\right),$$

$$S_{\sigma_0^2} = -\frac{1}{2}\sum_{i=1}^{I}\operatorname{tr}\left\{\boldsymbol{V}_i^{-1}\boldsymbol{W}_i^{-1}\right\} + \frac{1}{2}\sum_{i=1}^{I}\left(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}\right)^t\boldsymbol{V}_i^{-1}\boldsymbol{W}_i^{-1}\boldsymbol{V}_i^{-1}\left(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}\right),$$

$$F_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sum_{i=1}^{I}\boldsymbol{X}_i^t\boldsymbol{V}_i^{-1}\boldsymbol{X}_i, \qquad F_{\sigma_1^2\sigma_1^2} = \frac{1}{2}\sum_{i=1}^{I}\operatorname{tr}\left\{\left(\boldsymbol{V}_i^{-1}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t\right)^2\right\},$$

$$F_{\sigma_1^2\sigma_0^2} = \frac{1}{2}\sum_{i=1}^{I}\operatorname{tr}\left\{\boldsymbol{V}_i^{-1}\boldsymbol{W}_i^{-1}\boldsymbol{V}_i^{-1}\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t\right\}, \qquad F_{\sigma_0^2\sigma_0^2} = \frac{1}{2}\sum_{i=1}^{I}\operatorname{tr}\left\{\left(\boldsymbol{V}_i^{-1}\boldsymbol{W}_i^{-1}\right)^2\right\}.$$

3.2.2. Estimación máximo verosímil restringida (reml) en un modelo lineal mixto

El método de estimación máximo verosímil residual (reml) reduce el sesgo de los componentes de la varianza que se tienen con el método ml. Este método consiste en la estimación, por una parte los componentes de la varianza y por otro lado los efectos fijos (Bartlett, 1937; Searle et al., 1982 y Njuho and Milliken, 2005). El estimador reml de $\boldsymbol{\sigma} = (\sigma_0^2, \sigma_1^2)^t$ puede obtenerse mediante la maximización de la función log-verosimilitud

$$\ell(\boldsymbol{\sigma}) = -\frac{1}{2}(n-p)\log 2\pi - \frac{1}{2}\log|\boldsymbol{K}^t\boldsymbol{V}\boldsymbol{K}| - \frac{1}{2}\boldsymbol{y}^t\boldsymbol{P}\boldsymbol{y},$$

donde $\boldsymbol{K} = \boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{W}$ y $\boldsymbol{P} = \boldsymbol{K}(\boldsymbol{K}^t\boldsymbol{V}\boldsymbol{K})^{-1}\boldsymbol{K}^t$. La ecuación de actualización del algoritmo Fisher-scoring es:

$$\boldsymbol{\sigma}^{(k+1)} = \boldsymbol{\sigma}^{(k)} + \boldsymbol{F}(\boldsymbol{\sigma}^{(k)})^{-1}\boldsymbol{S}(\boldsymbol{\sigma}^{(k)}).$$

En este caso los elementos de $\boldsymbol{S}(\boldsymbol{\sigma})$ y $\boldsymbol{F}(\boldsymbol{\sigma})$ son:

$$\begin{aligned}
S_{\sigma_1^2} &= -\frac{1}{2}\operatorname{tr}\{\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t)\} + \frac{1}{2}\boldsymbol{y}^t\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t)\boldsymbol{P}\boldsymbol{y}\,,\\
S_{\sigma_0^2} &= -\frac{1}{2}\operatorname{tr}\{\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\boldsymbol{W}_i^{-1})\} + \frac{1}{2}\boldsymbol{y}^t\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\boldsymbol{W}_i^{-1})\boldsymbol{P}\boldsymbol{y}\,,\\
F_{\sigma_0^2\sigma_0^2} &= \frac{1}{2}\operatorname{tr}\{\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\boldsymbol{W}_i^{-1})\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\boldsymbol{W}_i^{-1})\},\,,\\
F_{\sigma_0^2\sigma_1^2} &= \frac{1}{2}\operatorname{tr}\{\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\boldsymbol{W}_i^{-1})\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t)\}\,,\\
F_{\sigma_1^2\sigma_1^2} &= \frac{1}{2}\operatorname{tr}\{\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t)\boldsymbol{P}\operatorname*{diag}_{1\le i\le I}(\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t)\}\,.
\end{aligned}$$

Posteriormente, la estimación del vector de efectos fijos $\widehat{\boldsymbol{\beta}}$ se realiza como si σ_0^2 y σ_1^2 fuesen conocidas.

3.3. Experimentos de simulación

En esta sección se realizan los 3 experimentos de simulación para comprobar la robustez de los modelos (3.2) y (3.4).

3.3.1. Simulación de muestras y cálculo de medidas de eficiencia

Las muestras se simulan de la siguiente forma. Para $i = 1, \ldots, I$, $j = 1, \ldots, n_i$:

1. Simulación de la variable explicativa,

$$x_{ij} = (b_i - a_i)U_{ij} + a_i \text{ con } U_{ij} = \frac{j}{n_i + 1}$$

$$a_i = 1, b_i = 1 + \frac{1}{I}\left(I + i\right),$$

2. Simulación de los pesos: $w_{ij} = 1/x_{ij}^{\ell}$, $\ell = 0, 1/2, 1, 2$, (4 posibilidades, 1 caso de homocedasticidad y 3 casos de heterocedasticidad).

3. Simulación de los errores y los efectos aleatorios:

$$e_{ij} \sim N(0, \sigma_0^2 = 1).$$

$$u_i \sim N(0, \sigma_1^2 = 1).$$

4. Generación de la variable objetivo. Para $i = 1, \ldots, I$ y $j = 1, \ldots, n_i$ se calcula:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + w_{ij}^{-1/2} e_{ij}, \quad \text{con } \beta_0 = \beta_1 = 1. \tag{Y0}$$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + w_{ij}^{-1/2} e_{ij}, \quad \text{con } \beta_0 = \beta_1 = 1. \tag{Y1}$$

Los pasos del experimento de simulación son los siguientes:

1. Simulación de la variable explicativa y los pesos.

2. Se repite $K = 10^6$ veces ($k = 1, \ldots, K$).

   2.1. Se genera una muestra aleatoria de tamaño $n = \sum_{i=1}^{I} \sum_{j=1}^{n_i} n_{ij}$ formada a partir de la variable objetivo que corresponda (Y0 o Y1).

2.2. Se calculan los valores estimados de los parámetros

$$\widehat{\tau}_{(k)} \in \{\widehat{\boldsymbol{\beta}}_{(k)}, \widehat{\sigma}^2_{0(k)}, \widehat{\sigma}^2_{1(k)}\}$$

y el valor estimado de la variable objetivo $\widehat{\mu}$, usando el método de ajuste que corresponda según el experimento a desarrollar.

3. Se calcula el error cuadrático medio empírico (EMSE) y el sesgo empírico (BIAS) para los parámetros estimados del modelo $\widehat{\tau}$ y la estimación de la variable objetivo $\widehat{\mu}$.

$$EMSE(\widehat{\tau}) = \frac{10^6}{K}\sum_{k=1}^{K}(\widehat{\tau}_{(k)} - \tau)^2, \qquad BIAS(\widehat{\tau}) = \frac{10^6}{K}\sum_{k=1}^{K}(\widehat{\tau}_{(k)} - \tau),$$

$$EMSE(\widehat{\mu}) = \frac{10^6}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\widehat{y}_{ij(k)} - y_{ij})^2,$$

$$BIAS(\widehat{\mu}) = \frac{10^6}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\widehat{y}_{ij(k)} - y_{ij}).$$

Las simulaciones se realizan para las 7 combinaciones de tamaño que aparecen en la tabla 3.2.

g	1	2	3	4	5	6	7
$I^{(g)}$	5	7	10	20	30	50	75
n_i	100	100	100	100	100	100	100
n	500	700	1 000	2 000	3 000	5 000	7 500

Tabla 3.2: Tamaño de las diferentes datasets

Para la obtención de los resultados numéricos del estudio se ha utilizado un servidor dedicado, Intel Xeon E52420 server en Linux Debian jessie operating system de 64 bits, 12 CPUs a 1.90 GHz y 32 GB Ddr3 RAM con el software libre de R Core Team (2015).

3.3.2. Experimento de simulación 1

El alcance de este experimento de simulación es investigar el impacto de la presencia o ausencia de pesos, es decir, de heterocedasticidad, cuando los datos se modelizan mediante el modelo lineal (3.4) usando el método ml bajo los dos tipos de variables objetivo Y0 y Y1.

En este experimento de simulación se pone el foco de atención en dos medidas, el error cuadrático medio empírico (EMSE) y el sesgo empírico (BIAS).

Los gráficos 3.1 y 3.2 contienen los resultados del EMSE y BIAS, respectivamente para este experimento de simulación. Estos gráficos están divididas en cuatro partes, una para cada

parámetro $\widehat{\boldsymbol{\beta}}_0$ (arriba izquierda), $\widehat{\boldsymbol{\beta}}_1$ (arriba derecha), $\widehat{\sigma}_0^2$ (abajo izquierda) y $\widehat{\mu}$ (abajo derecha). Cada parte a su vez, está dividida en cuatro secciones, la más a la izquierda para la homocedasticidad ($l = 0$) y las otras tres para cada una de las heterocedasticidades ($l = 1/2, 1, 2$). Para una mejor interpretación de los gráficos, el tamaño de los datasets está incorporado en el eje horizontal de acuerdo con la tabla 3.2.

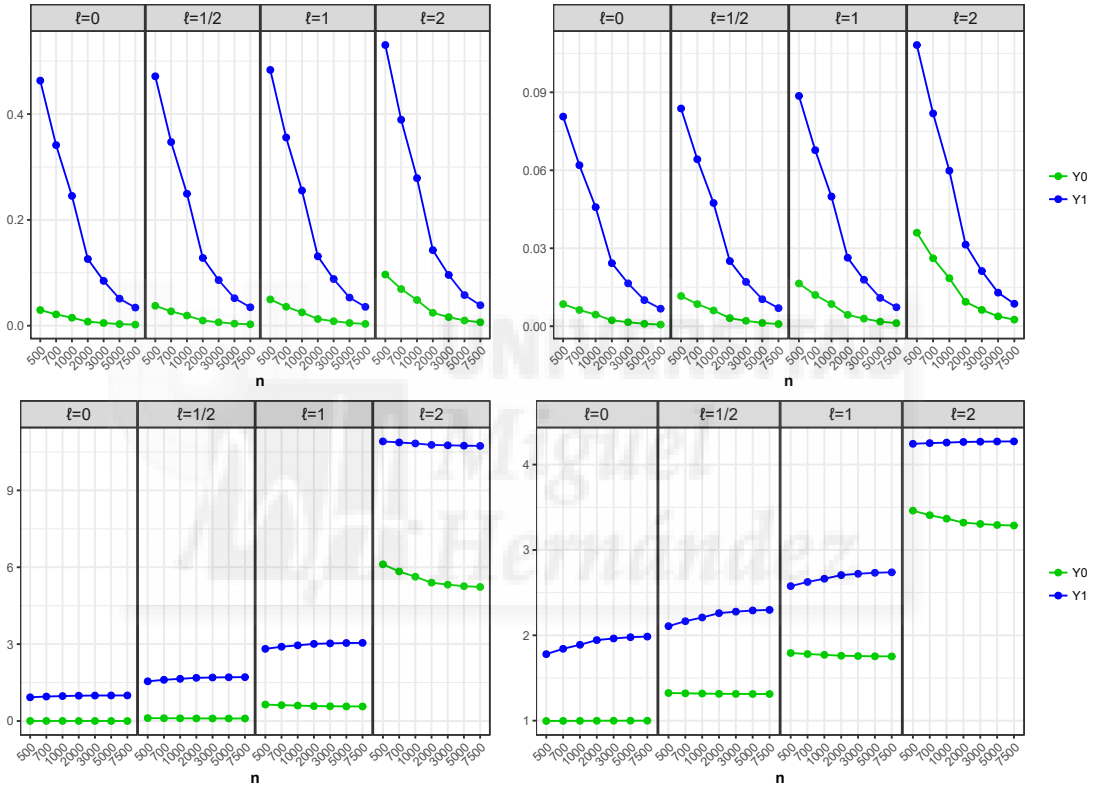

Gráfico 3.1: EMSE para $\widehat{\boldsymbol{\beta}_0}$ (arriba izquierda), $\widehat{\boldsymbol{\beta}_1}$ (arriba derecha), $\widehat{\sigma}_0^2$ (abajo izquierda) y $\widehat{\mu}$ (abajo derecha) para $\ell = 0, 1/2, 1, 2$, en experimento 1.

Respecto al EMSE, el gráfico 3.1 muestra dos conclusiones diferentes. La primera de todas es que el modelo ajustado es mejor para la variable objetivo Y0, es decir, es mejor cuando la variable objetivo simuladada solo tiene efectos fijos. Esto es bastante lógico, si el modelo propuesto es el modelo (3.4), cuando se ajustan unos datos que puedan provenir de una relación con un efecto aleatorio, el EMSE crecerá.

Además puede verse una cosa extremadamente curiosa, y es que el EMSE se reduce en los parámetros del modelo a medida que aumenta el tamaño del fichero de datos. Esto indica que cuando se está bajo la influencia de un problema BigData, la elección del modelo lineal más adecuado pasa a un segundo lugar en cuanto al EMSE.

La segunda conclusión es sobre la heterocedasticidad. La presencia de heterocedasticidad

($l = 1/2, 1, 2$) perjudica la estimación de todos los parámetros en ambas variables objetivo Y0 e Y1, pero sobre todo a la estimación de la componente de la varianza y el valor estimado de las variables objetivo $\widehat{\mu}$.

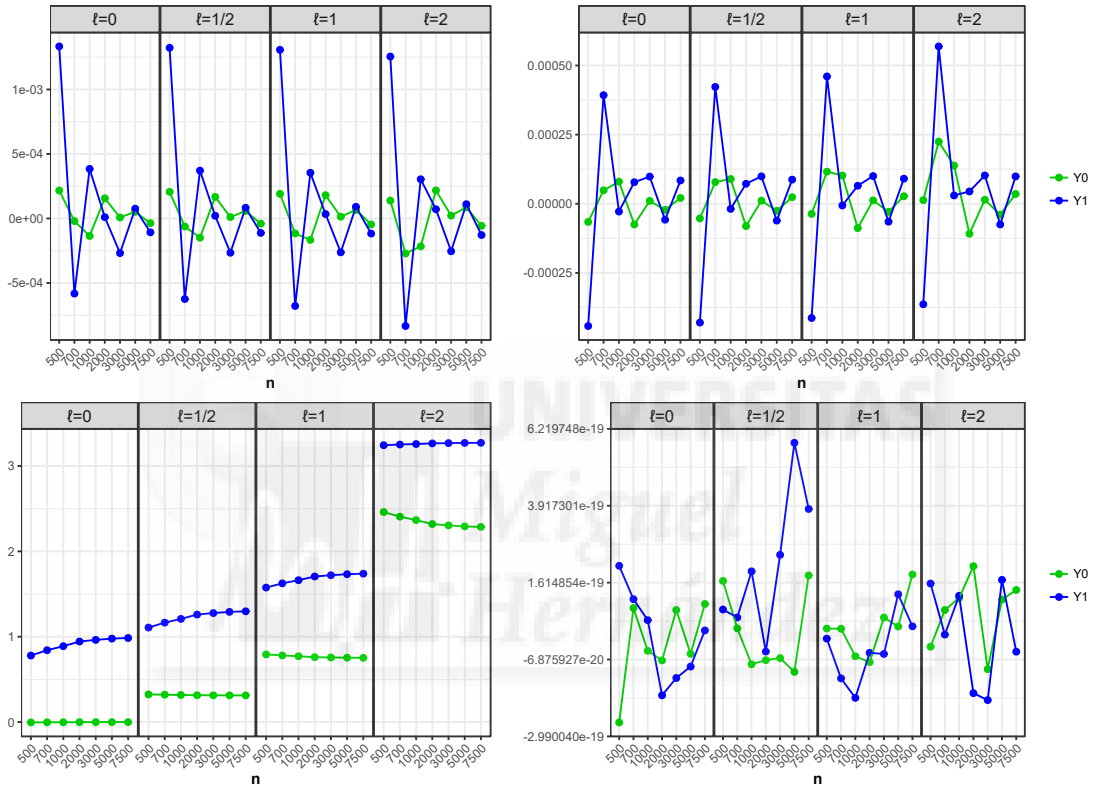

Gráfico 3.2: BIAS para $\widehat{\boldsymbol{\beta}_0}$ (arriba izquierda), $\widehat{\boldsymbol{\beta}_1}$ (arriba derecha), $\widehat{\sigma}_0^2$ (abajo izquierda) y $\widehat{\mu}$ (abajo derecha) para $\ell = 0, 1/2, 1, 2$, en experimento 1.

Respecto al sesgo, el gráfico 3.2 muestra cosas curiosas. Si el objetivo es estimar $\boldsymbol{\beta}$ con el modelo (3.4), la variable Y0 (línea verde) presenta menor sesgo que la variable Y1 (línea azul) y además para todo tamaño de fichero y presencia o ausencia de heterocedasticidad, está centrado en el cero. La presencia de heterocedasticidad ($l = 1/2, 1, 2$) causa un gran sesgo en la estimación de $\widehat{\sigma}_0^2$ y aún más en el caso de la variable objetivo Y1. Este es uno de los motivos por los que bajo un modelo lineal mixto (LMM) se utiliza como método de ajuste la máxima verosimilitud restringida, ya que reduce el sesgo de las componentes de la varianza (Searle et al., 1982; Jiang, 2007 y Pérez-Martín, 2008).

Sin embargo, si se mira la bondad de ajuste del modelo, ambas variables objetivo estan muy cerca de la insesgadez (gráfico 3.2 bajo derecha), aunque por supuesto, es mejor para Y0.

Las tablas con los valores numéricos se encuentran en el apéndice A.1.

3.3.3. Experimento de simulación 2

Este experimento de simulación ha sido diseñado para estudiar el comportamiento del modelo lineal mixto (3.2) bajo los métodos de ajuste ml y reml.

Las variables objetivo (Y0 e Y1), las variables explicativas, los pesos y los errores han sido generados del mismo modo que en el primer experimento de simulación. Los pasos del experimento de simulación también son los mismos pero teniendo en cuenta que en el apartado 2.2 se calculan los parámetros del modelo, $\widehat{\tau}_{(k)} \in \{\widehat{\boldsymbol{\beta}}_{(k)}, \widehat{\sigma}^2_{0(k)}, \widehat{\sigma}^2_{1(k)}\}$, y los valores ajustados, $\widehat{\mu}$, usando los métodos ml y reml con el modelo lineal mixto (3.2).

Las simulaciones se han llevado a cabo para las 7 combinaciones de tamaño que aparecen en la tabla 3.2.

En este experimento de simulación se pone el foco de atención en ambos métodos de estimación (ml y reml). Los gráficos 3.3 y 3.4 contienen los resultados del EMSE y BIAS respectivamente para el método reml. Los gráficos tienen la misma estructura que en el experimento previo. Se han omitido unos gráficos similares para el método ml pues la forma de las líneas, su tendencia y los valores que toman, son casi idénticas. Son tan similares, que para poder distinguir qué método da mejores resultados en EMSE y BIAS, se ha procedido a la creación del gráfico 3.5, con las diferencias (resta) entre ml y reml. De este modo valores positivos de estas diferencias se interpretan como mayor sesgo o error cuadrático medio del método ml, mientras que valores negativos se interpretan como mayor valor del reml. Los superíndices ml y reml son simplemente usados para denotar que el EMSE o BIAS han sido calculados usando el método ml o el reml.

Respecto al método reml y sus gráficos 3.3 y 3.4, se pueden establecer tres conclusiones. La primera de todas es que el modelo ajustado es ligeramente mejor para la variable objetivo Y1 (véase las partes derechas). El EMSE es sólo mejor para la varianza del efecto aleatorio. La segunda es que en el EMSE vuelve a verse una reducción cuando aumentan los tamaños de los datasets. Se vuelve a comprobar que la elección del modelo no es prioritaria bajo escenarios BigData. Además la estimación de μ empeora cuando se está en presencia de heterocedasticidad para ambas variables objetivo (gráfico 3.3 abajo a la derecha). La tercera, es sobre el sesgo (gráfico 3.4). Sólo para la variable objetivo Y1 se está en presencia de insesgadez. Es decir, cuando se tengan evidencias de relación entre una variable objetivo, y un factor con muchos niveles, susceptible de ser considerado como aleatorio, dejar de introducirlo como tal o emplear un método que no lo soporte, lleva a una situación de sesgo negativo. La presencia o ausencia de heterocedasticidad no afecta lo más mínimo al sesgo.

Respecto a el gráfico 3.5 en la que se comparan los métodos ml y reml, se puede comprobar que prácticamente todas las diferencias son positivas, lo cual lleva a la conclusión de que el método de estimación máximo verosímil restringido (reml) es mejor que el máximo verosímil (ml). Esto es especialmente relevante bajo la variable objetivo Y1 como puede verse en el gráfico 3.5 esquina superior izquierda. Aquí además se nota la reducción del sesgo de la componente de la varianza con el método reml (Searle et al., 1982; Jiang, 2007 y Pérez-Martín, 2008).

Las tablas con los valores numéricos se encuentran en el apéndice A.1.

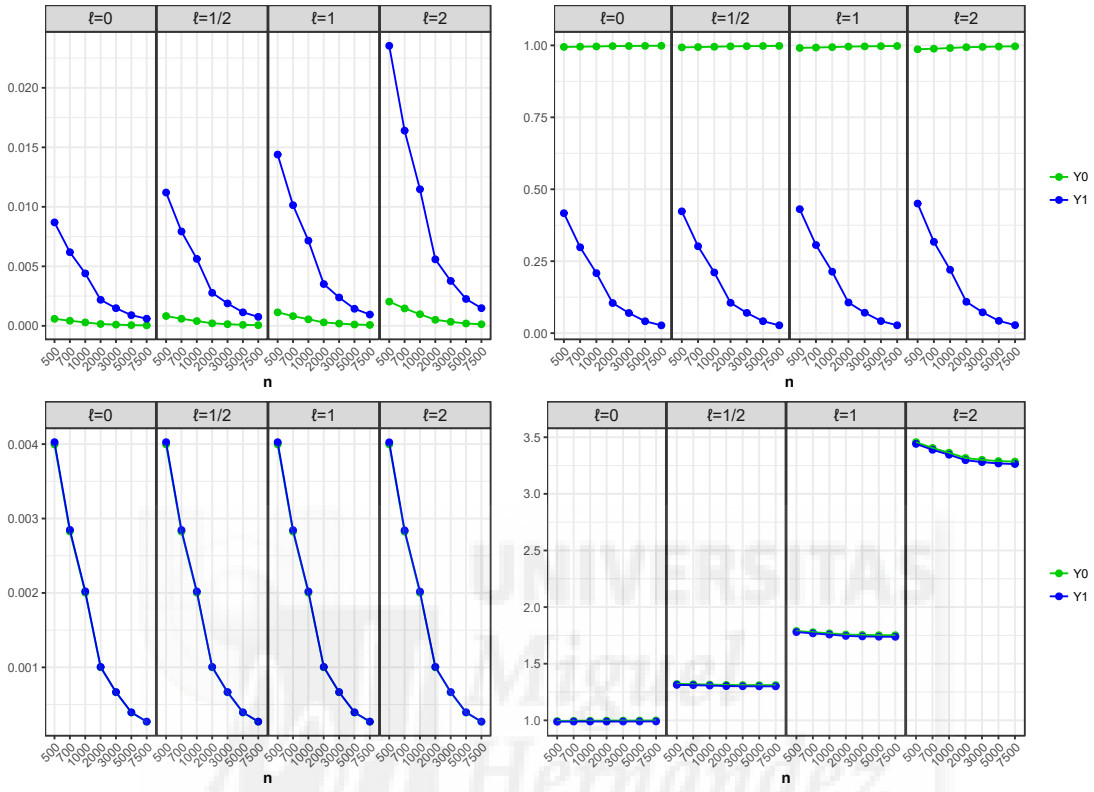

Gráfico 3.3: EMSE $\widehat{\beta_1}$ (arriba izquierda), $\widehat{\sigma}_1^2$ (arriba derecha), $\widehat{\sigma}_0^2$ (abajo izquierda) y $\widehat{\mu}$ (abajo derecha), para $\ell = 0, 1/2, 1, 2$, en experimento 2.

3.3.4. Experimento de simulación 3 de robustez

Este experimento de simulación ha sido diseñado para estudiar la robustez del modelo lineal mixto (3.2) en los casos de ajuste mediante el método ml y el método reml.

Se repite el experimento de simulación 2 en todos sus pasos pero cambiando los errores distribuidos según una normal de media cero y varianza uno, $N(\mathbf{0}, \sigma_0^2 = 1)$, por distribuciones Gamma y Weibull. Esto da información sobre la adecuación del modelo (3.2) y los métodos de ajuste ml y reml respecto a desviaciones de la asunción de normalidad de los errores. Las distribuciones Gamma y Weibull han sido convenientemente parametrizadas para tener la misma media y varianza que en el caso normal, $Ga(1,1)$ y $We(1,1)$. Nótese que las distribuciones Gamma y Weibull tienen soporte en la recta real positiva, dato de relevancia en caso de que en algún momento se desee estimar el riesgo de crédito como una cuantía monetaria en lugar de como la probabilidad de impago (default).

En el gráfico 3.6 se presentan los resultados obtenidos para este experimento. El primer hecho que se observa es que los resultados son prácticamente idénticos tanto para ml (izquierda) como para reml (derecha) en EMSE y sesgo. Por tanto los comentarios siguientes son extensibles a

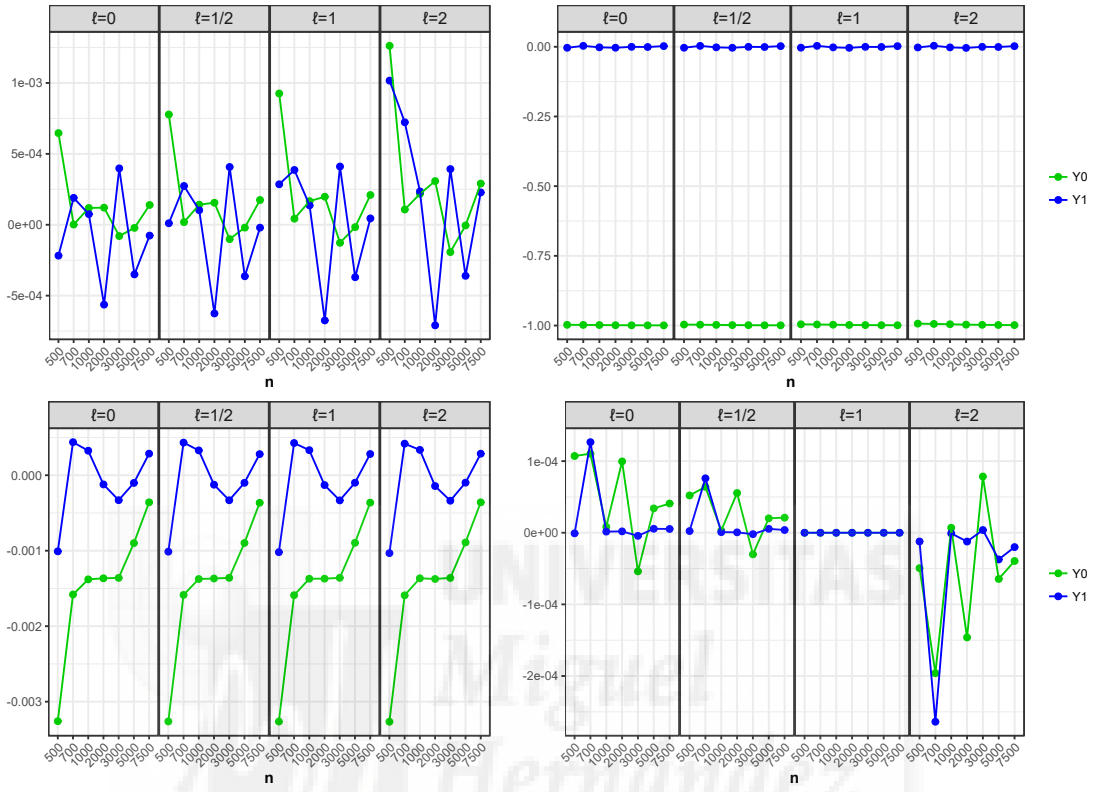

Gráfico 3.4: BIAS $\widehat{\beta_1}$ (arriba izquierda), $\widehat{\sigma}_1^2$ (arriba derecha), $\widehat{\sigma}_0^2$ (abajo izquierda) y $\widehat{\mu}$ (abajo derecha), para $\ell = 0, 1/2, 1, 2$, en experimento 2.

ambos métodos.

En presencia de heterocedasticidad EMSE crece mientras que al sesgo le sucede lo contrario (se acerca a cero). Sólo bajo homocedasticidad ($l = 0$) o cercano a ésta ($l = 1/2$), el caso de errores normales es insesgado, mientras que bajo heterocedasticidad el sesgo no se ve afectado ante desviaciones de la normalidad. Otro dato curioso es que tampoco se ve afectado el sesgo ni el EMSE ante el aumento del tamaño de los dataset, salvo el caso más heterocedástico ($l = 2$) para el EMSE que se reduce ligeramente. Por último, y como respuesta al principal objetivo de este experimento, se puede observar que alteraciones de la hipótesis de normalidad en los errores no afecta en ningún momento al EMSE (las tres líneas se sobreponen constantemte) y tan sólo afecta al sesgo, que aumenta, bajo homocedasticidad o cercano a ésta.

Las tablas con los valores numéricos se encuentran en el apéndice A.1.

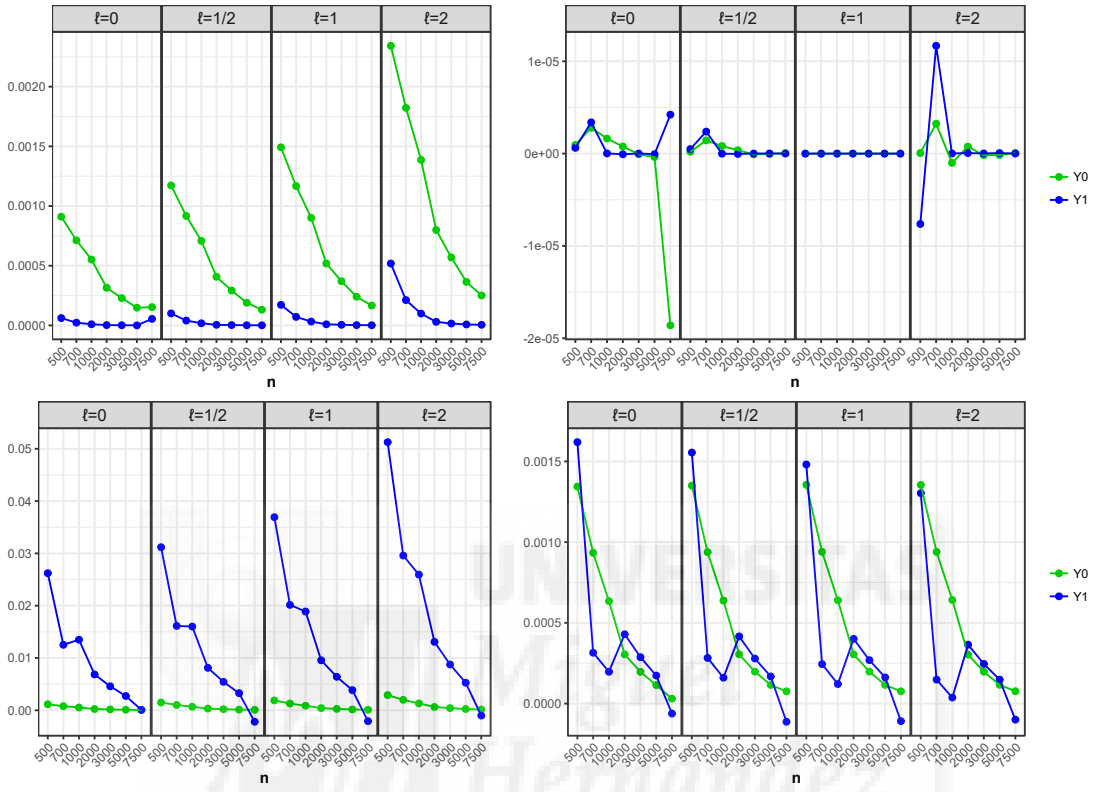

Gráfico 3.5: $EMSE^{ml} - EMSE^{reml}$ para μ (arriba izquierda), $BIAS^{ml} - BIAS^{reml}$ para μ (arriba derecha), $BIAS^{ml} - BIAS^{reml}$ para σ_1^2 (abajo izquierda) y $BIAS^{ml} - BIAS^{reml}$ para σ_0^2 (abajo derecha), for $\ell = 0, 1/2, 1, 2$, en experimento 2.

3.4. Conclusiones

El objetivo de los tres experimentos desarrollados en este capítulo ha sido estudiar la robustez y método de ajuste más adecuado para la estimación del riesgo de crédito, en los modelos lineales, mediante dos medidas, el error cuadrático medio empírico y el sesgo empírico. Se han planteado distintos escenarios de comportamiento de generación de la variable objetivo y el error del modelo, y se han comparado para observar en qué escenario se produce un mejor comportamiento de estas. Se ha conseguido con ello dar respuesta al objetivo número 2 de la sección "Objetivos" del "Prólogo".

A partir de los resultados obtenidos en los 3 experimentos anteriores se puede recomendar prioritariamente el uso de estimación máximo verosimil residual bajo modelos lineales mixtos.

Bajo situaciones de ficheros muy voluminosos (problemas BigData) la elección de un modelo lineal mixto (LMM) garantiza sesgos reducidos y valores pequeños del EMSE.

Se recomienda también el uso del método de máxima verosimilitud restringido (reml) ya que

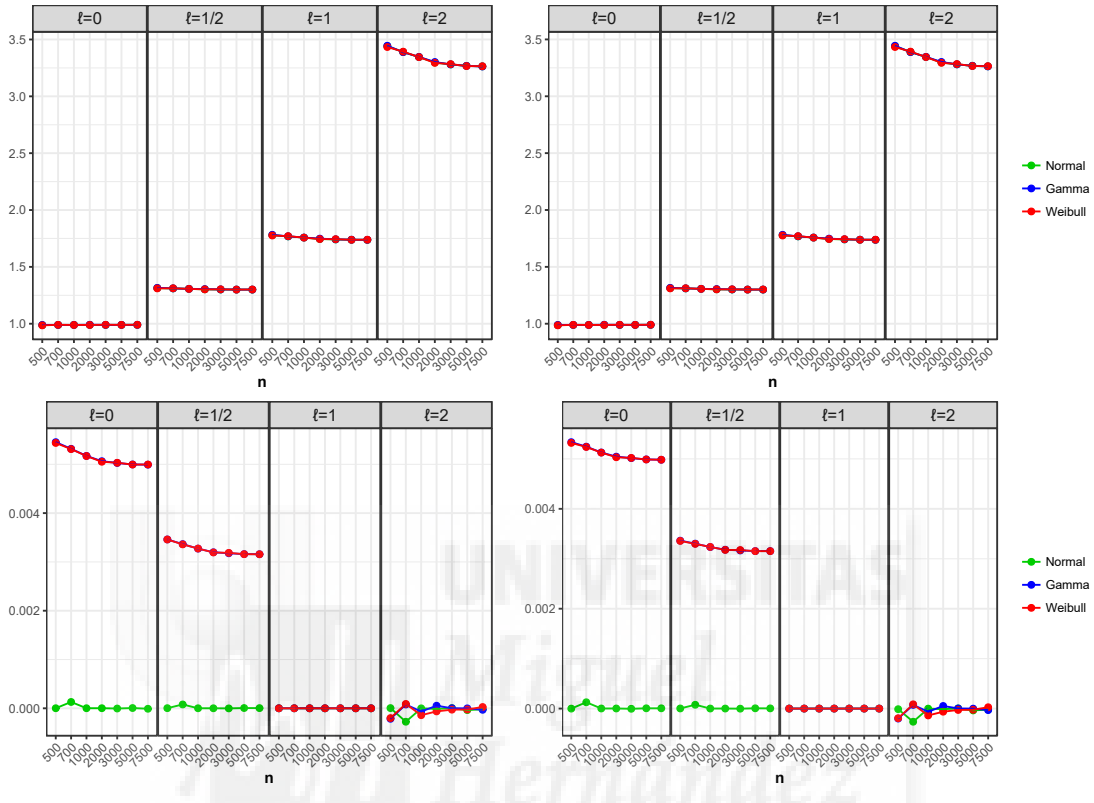

Gráfico 3.6: $EMSE(\widehat{\mu})^{ml}$ (arriba izquierda), $EMSE(\widehat{\mu})^{reml}$ (arriba derecha), $BIAS(\widehat{\mu})^{ml}$ (abajo izquierda), $BIAS(\widehat{\mu})^{reml}$ (abajo derecha), $\ell = 0, 1/2, 1, 2$ para casos de Normal, Gamma, y Weibull, en experimento 3.

mejora al de máxima verosimilitud de manera constante bajo diferente casuística (véase gráfico 3.5).

Por último no debe de preocupar excesivamente la asunción de errores distribuidos según una Normal, ya que alteraciones de ésta poco o nada afectan a las estimaciones de la variable objetivo. Esto justifica el uso de este modelo (LMM) con el método reml en capítulos posteriores.

UNIVERSITAS
Miguel
Hernández

4

Evaluación de métodos estadísticos para la estimación del riesgo de crédito

4.1. Introducción

Como se ha visto en el capítulo 1, existe una amplía variedad de técnicas y métodos empleados para la estimación del riesgo de crédito. Pero ninguna de las investigaciones realizadas aporta luz sobre cuál es el método que se debería de usar en todo momento o que sea el óptimo para algún tipo de métrica dada.

El objetivo principal de este capítulo es la evaluación del modelo lineal mixto (LMM) visto en (3.2) su ajuste mediante máxima verosimilitud restringida (reml) y su comparación con otros métodos tradicionales descritos en la literatura del tema.

Los modelos lineales mixtos son muy adecuados cuando se dispone de una variable explicativa categórica con muchos niveles. Si se ajustase mediante un modelo lineal simple habría que estimar, para sólo este factor, tantos parámetros como número de niveles tenga el factor menos uno, mientras que con un LMM tan sólo se estima un único parámetro, el de la componente de varianza (σ_1^2) del factor aleatorio.

Para evaluar este tipo de modelos y además compararlo con otros procedimientos tradicionales, se plantea el estudio de su eficacia y su eficiencia computacional. La eficacia se mide mediante la tasa de acierto del método y el error cuadrático medio de las estimaciones de este, mientras que la eficiencia se mide mediante el tiempo de ejecución del procedimiento. Estas tres métricas servirán para comparar los métodos entre sí (Pérez-Martín and Vaca, 2017).

De los distintos métodos para la estimación del riesgo de crédito que existen en la literatura, se han seleccionado los ya descritos en el capítulo 2 al ser los más comunmente utilizados y estar presentes en investigaciones que buscan la comparativa de ellos. Así se tendrá:

a. Análisis lineal discriminante (LDA)

b. Árboles de clasificación (CART)

c. Modelo lineal mixto, efectos fijos y aleatorios (LMM)

d. Modelo lineal generalizado con nexo logit (GLMlogit)

e. Máquinas de vectores soporte con kernel lineal (LSVM)

Para llevar a cabo la comparativa se ha diseñado un exhaustivo experimento de simulación Monte Carlo bajo diversas condiciones tanto en los datos como en su estructura.

En la mayoría de investigaciones sobre elección de métodos para medir la probabilidad de incumplimiento en un préstamo bancario, se emplea la tasa de acierto, calculada a partir de la matriz de confusión. Se puede citar entre otros trabajos los siguientes: Abdou and El Masry (2008), Crook et al. (2007), Galindo and Tamayo (2000), Huang and Wang (2007), Laitinen and Kankaanpää (1999), Wang et al. (2011), West (2000). Los autores Marqués et al. (2013) realizan un estudio teórico de métodos utilizados para predecir la probabilidad de incumplimiento. Observan que en más del 60 % de los trabajos estudiados se considera la tasa de acierto como criterio significativo para evaluar la clasificación entre cumplimiento o no del pago del préstamo. Sin embargo, señalan que existen evidencias empíricas y teóricas que han demostrado que la tasa de acierto está fuertemente sesgada con respecto al desequilibrio en la distribución de clases. En la misma línea Fawcett and Provost (1997) apuntan que, dentro del aprendizaje automático y la minería de datos, la tasa de acierto es una métrica estándar y no es suficiente, siendo los costes de una clasificación errónea una medida óptima. Los autores West (2000) y Baesens et al. (2003b) consideran tanto la tasa de acierto como los errores tipo I y tipo II, ya que en aplicaciones reales se cree que la clase de clientes que no atienden el pago no está suficientemente representada en comparación con la clase de clientes que atienden el pago. Además, los costos asociados con los errores de tipo II (clientes clasificados como éxito y son fracaso) son mucho más altos que los costos de clasificación errónea asociados a errores de tipo I (clientes clasificados como fracaso y son éxito).

En base a las observaciones realizadas por Loterman et al. (2012), en cuanto a las métricas de rendimiento se refiere, se decide utilizar el error cuadrático medio por ser un criterio de calibración. Al mismo tiempo se observa que aplicando la medición con la tasa de aciertos a través del cálculo de la matriz de confusión, existe una pérdida de calidad en los aciertos. La matriz de confusión incluye un criterio con cierto grado de subjetividad que es el punto de corte. Los resultados que se obtienen al hacer la predicción son probabilidades, no clasificaciones, excepto con LDA y CART. SVM es un método que la predicción la realiza con probabilidad o con clasificación, dependiendo de cómo se introduzca la variable respuesta (numérica o categórica). En los experimentos de esta tesis se ha realizado considerando la variable respuesta numérica, es decir, se ha obtenido una predicción en términos de probabilidad, por obtener mejores resultados. Por tanto, los métodos cuya predicción obtiene un resultado en términos de probabilidad son GLMlogit, LMM y SVM. En estos métodos para clasificar la predicción del modelo y compararla con los valores observados de la muestra, se elige un punto de corte. Si la probabilidad predicha por el modelo es mayor que el punto de corte se clasifica como éxito y si es menor como fracaso. En consecuencia, de la elección de un punto de corte u otro dependerá la clasificación de los resultados en la matriz de confusión, por tanto se tendrá diferentes tasas de aciertos. En este capítulo se ha considerado como punto de corte la mediana.

4.2. Algoritmo de simulación

En esta sección se detallan los pasos del algoritmo de simulación para el experimento de simulación de la siguiente sección. El algoritmo que se describe a continuación se realiza dos veces, una para generar un conjunto de datos aleatorios al que se denominará dataset de aprendizaje, entrenamiento o training (m_A) y otro de las mismas características al que se denominará dataset de test o prueba (m_T).

El conjunto de datos m_A se utilizará para ajustar el modelo o método de estimación y así obtener sus parámetros, que se usarán en el conjunto de datos m_T para predecir la variable objetivo, que será posteriormente comparada con su valor real en m_T, en el cálculo del error cuadrático medio y la tasa de acierto.

Cada conjunto de datos se genera de la siguiente forma. Para $i = 1, \ldots, I$, $j = 1, \ldots, n_i$:

- Simulación de la primera variable explicativa:

$$x_{ij1} = (b_i - a_i)U_{ij} + a_i \text{ con } U_{ij} = \frac{j}{n_i + 1}.$$

$$a_i = 1, b_i = 1 + \frac{1}{I}\left(I + i\right).$$

- Simulación del resto de variables explicativas, x_{ij_2} hasta x_{ij_p}. Se generan siguiendo una distribución uniforme

$$x_{ij} \sim Unif(0, 1).$$

- Simulación de los pesos:

$$w_{ij} = \frac{1}{x_{ij}}, \quad (heterocedasticidad)$$

- Simulación de los efectos aleatorios y errores:

$$u_i \sim N(0, \sigma_1^2 = 1).$$
$$e_{ij} \sim N(0, \sigma_1^2 = 1).$$

- Generación de la variable objetivo.

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \ldots + \beta_p x_{ijp} + u_i + w_{ij}^{-1/2} e_{ij}, \quad \text{con } \beta_0 = \ldots = \beta_p = (-0{,}95, 1)$$

- Recategorización de la variable objetivo por éxito o fracaso:

$$p_{ij} = (e^{y_{ij}})/(1 + e^{y_{ij}})$$

Al ser una probabilidad el resultado de p_{ij} está entre 0 y 1. La variable respuesta objeto de este estudio es una variable discreta dicotómica, es decir, es 0 en caso de éxito y 1 en caso de fracaso, default o caída en mora. Se necesita una recategorización de la variable dependiente. Para la transformación de p_{ij} en variable dicotómica, 0 y 1, se realiza el siguiente proceso:

- Se transforma la probabilidad en número de ocurrencias de un suceso, para ello se multiplica por 100 la probabilidad anterior, obteniendo el número de veces que ocurre un suceso. Se trunca el número de ocurrencias del suceso.

$$P_{ij} = floor(p_{ij} * 100)$$

- Se asigna P_{ij} al suceso fracaso y $100 - P_{ij}$ al suceso éxito, obteniendo la base de datos completa con la variable respuesta dicotómica. Este P_{ij} será la nueva variable objetivo a analizar.

4.3. Experimento de simulación

Los pasos a seguir para realizar los experimentos propuestos son los siguientes:

1. Simulación de la variable o variables explicativas según el algoritmo de simulación de tamaños $n = \sum_{i=1}^{I} n_i$.
2. Se repite cada uno de ellos $K = 10^4$ veces $(k = 1, \ldots, K)$

2.1. Se genera una muestra de entrenamiento, m_A y una muestra de test, m_T.

2.2. Se calculan los valores estimados de los parámetros de cada uno de los métodos, LDA, CART, LMM, GLMlogit y LSVM, con sus respectivas funciones de R, para la muestra de entrenamiento o aprendizaje m_A.

2.3. Se calcula para estos métodos la predicción de la variable objetivo $\widehat{P}_{ij}^T$ con las variables explicativas de la muestra de test m_T.

2.4. Se obtiene la matriz de confusión y de esta se obtiene la tasa de aciertos, como el número de casos bien clasificados sobre el total de casos con la muestra de test m_T.

2.5. Se calcula la raíz cuadrada del error cuadrático medio (RMSE) con el conjunto de datos de test m_T del siguiente modo:

$$RMSE_l = \sqrt{\frac{1}{n_T}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\widehat{P}_{ijl}^T - P_{ijl}^T)^2}$$

con $l \in \{GLMlogit, LMM, LDA, CART, LSVM\}$, donde n_T es el tamaño de la muestra m_T y $\widehat{P}_{ijl}^T$ es el valor estimado de la variable objetivo de la muestra de testeo para el procedimiento de estimación l.

2.6. Se registra el tiempo de ejecución de cada uno de los procedimientos de estimación.

3. Se calcula la media del tiempo empleado para cada método.

Las simulaciones se realizan para las 5 combinaciones de tamaño que aparecen en la tabla 4.1.

g	1	2	3	4	5
$I^{(g)}$	2	5	10	25	50
n_i	100	100	100	100	100
n	2 000	5 000	10 000	25 000	50 000

Tabla 4.1: Tamaño del conjunto de base de datos.

Para cada combinación de la tabla 4.1 se han incluido 5 grupos de variables explicativas x_{ij}. El número de variables explicativas es $p = 1, 2, 10, 50, 100$. Con estos valores finalmente se generan y analizan 50 conjuntos de datos, 10^4 veces, a lo largo de los cinco métodos.

La matriz, tabla de confusión o de contingencia (Stehman, 1997b y Stehman, 1997a) incluye tanto los aciertos como los errores para cada una de las clasificaciones, en nuestro caso las clases predichas comparadas con los valores reales de la muestra de test. Esta tabla es la fuente de información para evaluar la bondad de la predicción (Foody, 2002 y Liu and Kumar, 2007). La matriz de confusión es una tabla de contingencia 2x2 en la que se realiza un conteo de elementos

predichos clasificados según respondan al crédito (pagadores) y los que no lo hagan (default) frente a los verdaderos valores que toman obtenidos de la variable objetivo medida en la muestra test.

	Éxito	Fracaso
Éxito	VP	FN
Fracaso	FP	VN

Tabla 4.2: Matriz de Confusión

Siendo VP los valores verdaderos positivos, FP falsos positivos, FN falsos negativos y VN verdaderos negativos. Cuanto mayor sea VP y VN mejor será el desempeño de la técnica clasificatoria. Esta tabla puede ser extensible a k clases. Se suele identificar y separar los errores de clasificación en función de su distancia de la diagonal (Koh, 1992 y Bessis, 2002).

A partir de la matriz de confusión se pueden definir distintos índices o tasas:

1. La tasa de acierto vendrá dado por el número de individuos correctamente clasificados entre el número total de individuos. Es decir, todos aquellos clasificados como verdaderos positivos (VP) y verdaderos negativos (VN) en 4.2.

$$aciertos = \frac{\sum(diag(MC))}{\sum(MC)}$$

2. La tasa de error será el complementario del acierto. O bien, todos aquellos individuos clasificados erróneamente.

3. Especificidad es la proporción entre los valores clasificados como default correctos y el total de valores default o la probabilidad que una clase no relevante sea identificada correctamente por el modelo. (VN/FP+VN).

4. Sensibilidad es la razón entre los valores clasificados como éxito y el total de valores postivos. (VP/VP+FN).

5. Tasa de falsos errores. El error tipo I se identifica con el error de clasificar mal a los individuos, un individuo que siendo pagador se predice que va a ser default(FN) falso negativo.

6. Tasa de falsos aciertos. El error tipo II se identifica con falso positivo, es decir, aquel individuo que siendo no pagador se predice como éxito (FP). Se considera importante dentro de los resultados obtener las tasas de falsos positivos y falsos negativos, porque la entidad financiera debe de valorar el coste que le supone clasificar mal al individuo.

La medición de estos dos errores es importante ya que cada uno de ellos supondrá un coste o pérdida para la institución financiera. El error tipo I indicará la clasificación de un cliente bueno como malo. Generará un coste de oportunidad, el crédito no se cursará por tanto la institución financiera no percibirá la utilidad de ese potencial crédito bueno.

El error tipo II indicará la clasificación de un cliente malo como bueno. Generará una pérdida para la institución financiera al tener que aumentar el riesgo total de la cartera aumentando la posibilidad de no recuperar el monto prestado.

Para la obtención de los resultados numéricos de los procedimientos realizados en esta tesis se ha utilizado un servidor dedicado Intel Xeon E52420 server en Linux Debian jessie operating system 64 bits, 12 CPUs at 1.90 GHz and 32 GB Ddr3 RAM con R.

4.4. Resultados

Los resultados con los valores numéricos de esta sección se encuentran en el apéndice A, sección A.2.

En estos experimentos de simulación en primer lugar se analiza el error cuadrático medio (RMSE) para todos los métodos.

Se observa en el gráfico 4.1 cómo el RMSE decrece conforme aumenta el número de variables para todos los métodos analizados. Al mismo tiempo RMSE crece conforme aumenta el tamaño del conjunto de datos (n_T).

Todos los métodos obtienen peor RMSE para 1 variable explicativa, excepto CART para el tamaño del conjunto de datos de $I = 2\,000$. En 3 de los 5 métodos estudiados, el valor más alto de RMSE se da con 1 variable y 50 000 registros, siendo el peor método LDA con un valor de 0,68386519. El método LDA mejora considerablemente a medida que aumenta el número de variables acercándose al resto de métodos. Pasa de ser el método con el valor más alto de RMSE para 1 variable a obtener un valor de 0,3831344 con 100 variables y 10 000 registros, aunque en todos los casos estudiados es el método con peor RMSE, obtiene peores resultados respecto al RMSE, en 24 de los 25 experimentos estudiados.

Con los métodos LMM, GLMlogit y CART se obtienen valores inferiores de RMSE en 11, 5 y 9 casos respectivamente de los 25 estudiados. El método CART obtiene un RMSE menor que el resto de métodos cuando se tienen pocas variables explicativas (1, 2 y 10 variables) y número de registros menores o iguales a 10 000.

Los métodos LMM y GLMlogit tienen un comportamiento muy similar. Con el método LMM se obtienen mejores resultados hasta 50 variables. A partir de 100 variables, el RMSE de GLMlogit se iguala con LMM hasta 25 000 registros, después vuelve a ser menor el RMSE de LMM.

El método CART a partir de un conjunto de datos superior a 10 000 aumenta su RMSE muy rápido. El método LSVM es el segundo con resultados más altos de RMSE, en 23 de los 25 experimentos resulta el peor.

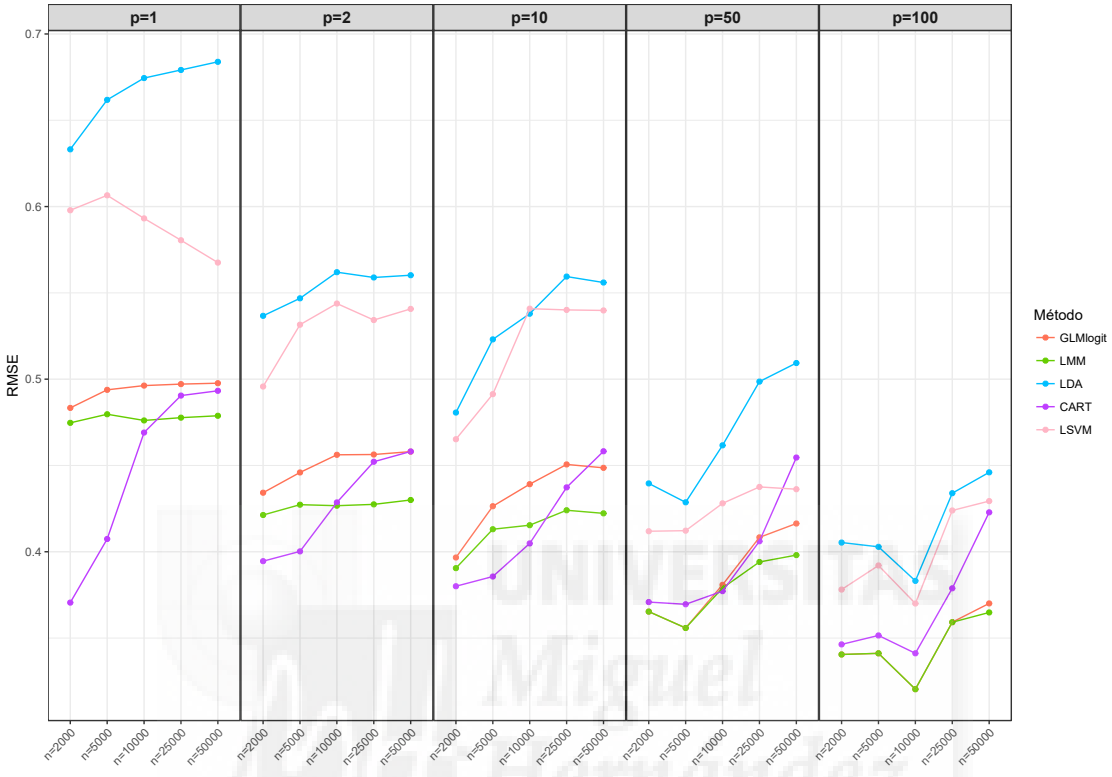

Gráfico 4.1: RMSE para los métodos LDA, CART, LMM, GLMlogit, LSVM

En cuanto al porcentaje de aciertos para todos los métodos, como se puede comprobar en el gráfico 4.2, se comportan exactamente igual, conforme aumenta el número de variables, el porcentaje de acierto aumenta. En el caso de CART es similar el comportamiento, excepto para 1 variable, que el porcentaje de acierto es el más alto con 1 variable y tamaño del conjunto de datos mínimo, es decir para $n_T = 2\,000$, empeorando el porcentaje de aciertos conforme aumenta n_T.

Observando los resultados en la tabla A.9, LDA es el mejor método, obteniendo una tasa de acierto máxima en 12 de los 25 experimentos. El método con mejor porcentaje de aciertos después de LDA es LSVM con 9 casos.

El porcentaje máximo de acierto se obtiene con el método LDA con 100 variables y 50 000 registros, siendo 76.19 % seguido de GLMlogit con un 76.05 %. Se observa que para los métodos lineales conforme aumenta el número de variables y el tamaño del conjunto de datos mejora el acierto.

Un aspecto importante para las instituciones financieras es el tiempo de respuesta para las posibles operaciones. Por tanto, un buen método es aquel que minimiza tanto el tiempo de

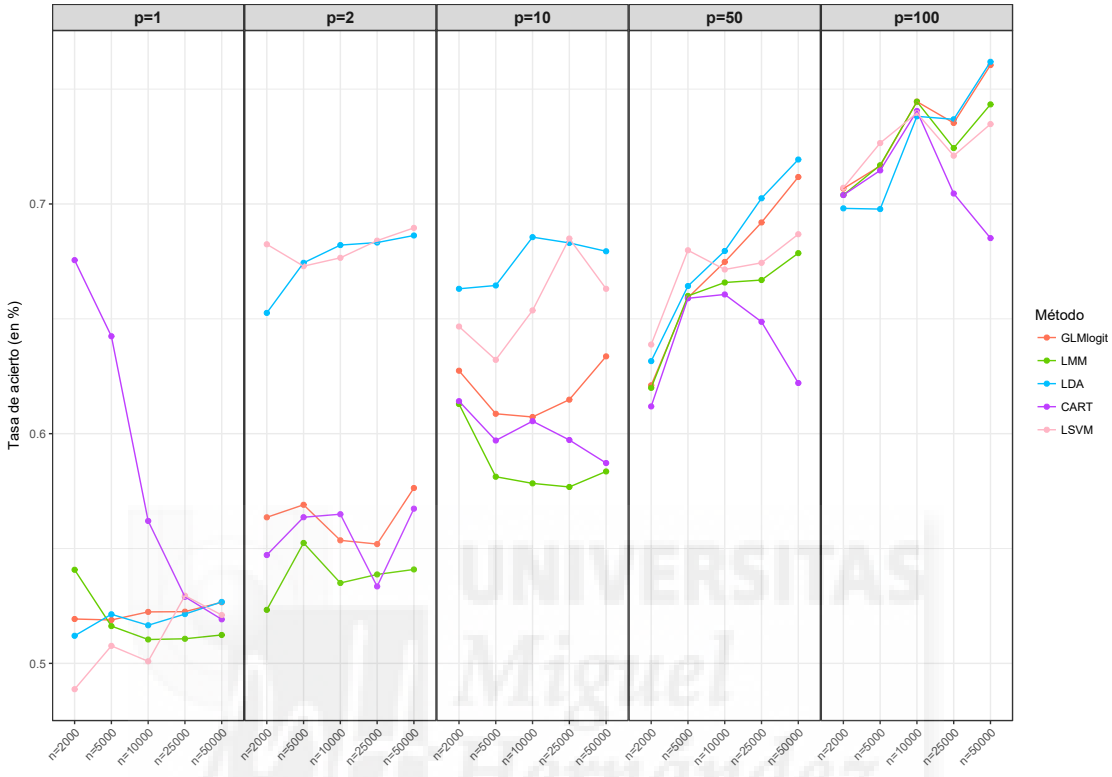

Gráfico 4.2: Tasa de acierto

respuesta al cliente como el riesgo del cliente en la entidad financiera.

En la tabla A.11 se puede observar cómo aumenta el tiempo de ejecución conforme aumenta p, la relación entre incrementos en p e incremento en tiempo de ejecución puede ser:

- *Constante*. Se cumpliría si la relación del tiempo fuese siempre la misma. Se puede comprobar en la tabla A.11 que el tiempo aumenta con el número de variables, por lo que no es constante.
- *Lineal*. La relación entre los tiempos y las variables explicativas debería de aumentar en la misma proporción.
- *Exponencial*. La relación entre el número de variables y el tiempo crece de forma multiplicativa.

El método computacionalmente más eficiente en el 36.00 % de los casos es CART, seguido de LDA y GLMlogit. GLMlogit en gran parte de los casos estudiados es más eficaz que LMM, pero se observa que a partir de 2 variables explicativas y 50 000 registros LMM se aproxima en tiempos a GLMlogit. Y para 100 variables y tamaños de 2 000 y 5 000 registros LMM es más eficiente.

El método más lento computacionalmente en todos los casos es LSVM, incluso como se puede observar en los resultados, ver apéndice A.11, el tiempo aumenta exponencialmente. Se vuelve totalmente ineficiente conforme aumenta el número de variables y también el número de registros.

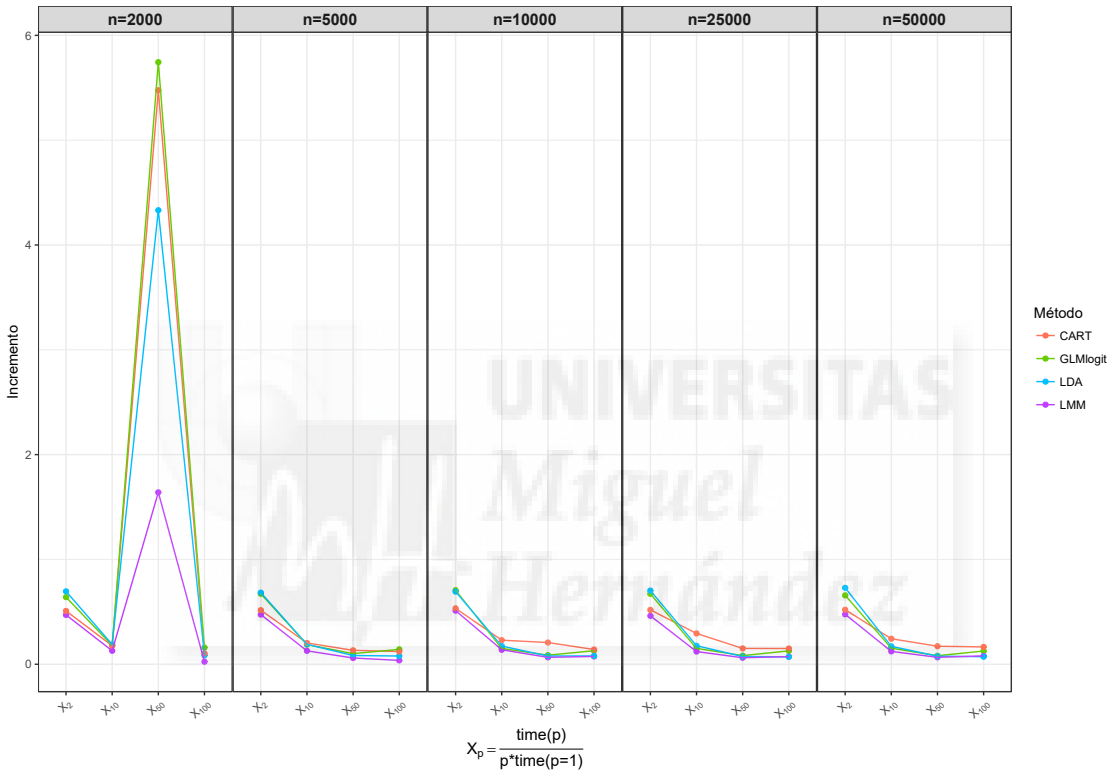

Gráfico 4.3: Incremento relativo del tiempo sobre el caso $p = 1$

En el gráfico 4.3 se estudia el tiempo relativo transcurrido (OX axis). Se calcula la relación entre el tiempo total para cada p en función del tiempo total en el caso $p = 1$, se pondera la relación del tiempo para el caso de $p = 1$ como sigue:

$$\text{Incremento relativo para } p = \frac{tiempo(p)}{p \times tiempo(p=1)} = X_p$$

Por ejemplo, la relación del aumento del tiempo para $p = 100$ es $\frac{tiempo(p=100)}{100 \times tiempo(p=1)}$.

Para observar mejor los resultados se ha eliminado del gráfico los resultados del método LSVM.

Para todos los métodos y exceptuando el caso de 2 000 registros se tiene aproximadamente el mismo comportamiento con respecto a los incrementos relativos X_p. El tiempo total aumenta

conforme aumenta el tamaño del conjunto de datos hasta $n_T = 25\,000$ para después decrecer. En todos ellos, el máximo se alcanza para X_2, descendiendo hasta X_{50}, para después aumentar muy ligeramente en X_{100} excepto CART que en todos los casos sigue descendiendo o se mantiene. El tiempo mínimo total se da en el método LMM con 100 variables y tamaño del conjunto de datos de 2 000. La pendiente de X_2 a X_{10} es mayor que la pendiente en el resto de casos, para todos los tamaños del conjunto de datos y métodos, por tanto, se observa que a partir de X_{10} el incremento de tiempo es proporcionalmente menor al incremento de variables. El tiempo relativo mínimo en el 65 % de los casos se produce con el método LMM, lo que significa que es un método más eficiente cuando se incrementa el número de variables.

El tiempo relativo cuando el número de registros es de 2 000 y X_{50} tiene un comportamiento distinto al resto, alcanza el máximo en ese punto para después volver a descender para X_{100} a niveles mínimos para todos los métodos.

4.5. Conclusiones

En este capítulo se ha realizado un estudio comparativo de los métodos LDA, CART, LMM, GLMlogit y LSVM para medir su eficacia y eficiencia computacional. La eficacia de cada uno de los métodos se ha medido a través del error cuadrático medio y la tasa de acierto, la eficiencia computacional se ha medido calculando el tiempo medio de ejecución. Se realiza el estudio con datos generados sintéticamente mediante experimentos Monte Carlo, en diferentes volúmenes tanto de registros, como de variables. Con esto, se ha cumplido la parte sintética del objetivo número 3 de la sección “Objetivos” del “Prólogo”.

En el experimento 4.3 se simulan desde $N = 2\,000$ a $N = 50\,000$ registros y desde $p = 1$ a $p = 100$ variables explicativas para intentar reproducir el máximo número de posibles solicitudes de préstamos que puedan exitir en cualquier entidad financiera.

Se proponen cinco métodos LDA, CART, LMM, LSVM y GLMlogit. Se calculan medidas de eficiencia y eficacia.

Los métodos más eficaces son LMM y GLMlogit. Mientras que LDA, LMM y GLMlogit son los métodos más eficientes computacionalmente, seguidos de CART. Cuando aumenta el número de datos LMM es más eficiente computacionalmente, y en caso de un número de datos menor el método mejor es GLMlogit.

De los métodos estudiados en el experimento de este capítulo los que obtienen mejores resultados en términos de RMSE y el tiempo total transcurrido son LMM, GLMlogit y LDA. Se descarta LDA en relación con LMM porque este último es más eficaz que LDA. Aunque LDA sea más eficiente computacionalmente, es menos eficaz y es el que peores resultados obtiene en términos de RMSE. LMM es similar a GLMlogit, aunque para grandes conjuntos de datos, LMM es mejor y obtiene mejores resultados en términos de RMSE.

Por tanto se propone LMM y GLMlogit como mejores métodos para evaluar el riesgo de crédito bajo las hipótesis de la simulación de datos realizada en el experimento de este capítulo.

UNIVERSITAS
Miguel
Hernández

5 Aplicación caso semi-real y real

5.1. Introducción

El objetivo de este capítulo es comparar los mismos métodos empleados con la muestra sintética del capítulo 4, para evaluar el riesgo de crédito aplicado a una base de datos semi-real y otra real. Las técnicas que se emplean son técnicas estadísticas tradicionales que se basan en métodos paramétricos y no paramétricos (LDA, CART), métodos estadísticos (LMM, GLMlogit) y técnicas de inteligencia artificial (LSVM). La comparación de las diferentes técnicas se realiza a través del cálculo de los mismos indicadores o índices empleados en el capítulo 4:

1. El error cuadrático medio (RMSE).
2. Tasa de acierto.
3. El error I. Este error indica la clasificación de un cliente bueno como malo.
4. El error tipo II. Este error indica la clasificación de un cliente malo como bueno
5. El tiempo medio de ejecución empleado por cada método.

La base de datos semi-real que se ha utilizado proviene de una base de datos elaborada por el Instituto de Estudios Fiscales (I.E.F) de una muestra de microdatos fiscales correspondiente a las declaraciones de IRPF del territorio español de régimen fiscal común desde los años 2008 a 2013, sin incluir Comunidad Foral de Navarra y País Vasco. Los documentos de trabajo de los distintos años son los recogidos en: Picos Sánchez et al. (2011), Pérez López et al. (2012),

Pérez López et al. (2013), Pérez López et al. (2014), Pérez López et al. (2015) y Pérez López et al. (2016).

Las bases de datos reales utilizadas para el estudio son *German Credit* y *Australian Credit* del repositorio de la UCI (Lichman, 2013). Las bases de datos están disponibles en la página web `http://www.archive.ics.uci.edu/ml/datasets.html`.

Teniendo en cuenta el capítulo 3, el modelo que se va a utilizar para la obtención de la variable objetivo o dependiente es el modelo mixto, efectos fijos y aleatorios (3.2).

Los cálculos realizados para la obtención de la matriz de confusión y el error cuadrático medio tienen su justificación en la teoría y simulaciones computacionales desarrolladas en el capítulo 4.

5.2. Aplicación caso semi-real

5.2.1. Metodología

El primer obstáculo para poder evaluar el riesgo de crédito con diferentes técnicas o métodos de estimación ha sido encontrar una base de datos real para su aplicación. No se ha podido acceder a ninguna base de datos real sobre el comportamiento crediticio de los clientes de entidades financieras, con el fin de analizar si atienden los pagos de un préstamo hipotecario o incumplen, es decir, entran en default, que es el objeto de esta investigación. Dada la situación de partida, en primer lugar, se realiza una búsqueda de la normativa que establece el Banco de España sobre gestión del riesgo de crédito, política de concesión, modificación, evaluación, seguimiento y control de operaciones, estimación de las coberturas de las pérdidas por riesgo de crédito y obligaciones de las entidades financieras (BOE, 2016). Al mismo tiempo se realiza un estudio de las variables que tienen en cuenta las entidades financieras para la concesión o no de un préstamo hipotecario. En los distintos artículos y trabajos consultados Mylonakis (2010), Hand and Henley (1997), Malhotra and Malhotra (2003), Boj et al. (2009a), Juan Camilo Ochoa P. et al. (2010), Cabrera Cruz (2014), Moreno Valencia (2013), Salinas Flores (2005), Yu (2014), Gomes Goncalves (2009), Lee et al. (2002), Steenackers and Goovaerts (1989), Mures et al. (2005) y UCI Machine learning Repository (`http://archive.ics.uci.edu/ml/datasets.html`) se observa que existe unanimidad en la elección de determinadas variables explicativas. Las entidades financieras, en general, recogen información de los clientes para el estudio previo de la concesión de un préstamo de distintos ámbitos. Las variables que formarán parte del modelo deben de incluir los distintos factores que consideran las entidades financieras para medir el nivel de riesgo en las operaciones de crédito. Los factores a tener en cuenta para la selección de las variables explicativas son los siguientes:

1. Factores sociales y personales, se seleccionan variables relacionadas con el entorno social y personal del cliente tales como: sexo, edad, estado civil, número de personas a cargo, población de la vivienda habitual, vivienda habitual y otras viviendas.

2. Factores económicos, se seleccionan variables relacionadas con la disponibilidad de ingresos tales como: remuneraciones, sector económico donde ejerce su actividad e importe del préstamo solicitado.

Una vez identificadas las variables utilizadas por las entidades financieras se ha buscado una base de datos que cumpla con los siguientes requisitos:

1. Fuese accesible para la investigación.

2. Contuviese gran parte de las variables que son utilizadas por las entidades financieras.

3. Cumpliese requisitos estadísticos de muestreo para que sea representativa.

La base de datos de la que se parte es las declaraciones de IRPF del territorio español de régimen fiscal común (sin incluir Comunidad Foral de Navarra y País Vasco), facilitada por el Instituto de Estudios Fiscales dependiente del Ministerio de Hacienda y Función Pública.

Siguiendo los documentos de trabajo del Instituto de estudios fiscales, la población objetivo de la muestra utilizada son las declaraciones del Impuesto sobre la Renta de las Personas Físicas desde el año 2008 al 2013. El ámbito geográfico constituye el Territorio español de régimen fiscal común. El muestreo utilizado por el I.E.F. ha sido un muestreo estratificado aleatorio. El muestreo estratificado es aquel en que los elementos de la muestra son proporcionales a su presencia en la población. La presencia de un elemento en un estrato excluye su presencia en otro. Para este tipo de muestreo, se divide a la población en varios grupos o estratos con el fin de dar representatividad a los distintos factores que integran el universo de estudio. Para la selección de los elementos o unidades representantes, se utiliza el método de muestreo aleatorio. Es decir, se separa la población en grupos, denominados estratos, y se elige una muestra aleatoria simple en cada estrato. Las muestras aleatorias simples constituyen la muestra. Para los estratos han considerado, en primer lugar, las provincias españolas del Territorio Fiscal Común, total 48 ya que Ceuta y Melilla se han considerado de forma conjunta. En un segundo nivel de estratificación se han considerado 12 tramos de renta y en un tercer nivel de estratificación el tipo de declaración, individual o conjunta. Por tanto, el numero de estratos de ultimo nivel es $48 * 2 * 12 = 1\,152$, no existiendo estratos vacíos.

Sobre los tamaños de diseño se ha impuesto una restricción de secreto estadístico. El reparto de la muestra en los estratos se ha realizado mediante afijación de mínima varianza.

Tal y como se destaca en los documentos de trabajo de las muestras empleadas, las ventajas de la utilización de los datos fiscales son las siguientes:

1. Gran representatividad, debida al muestreo estratificado.

2. Ausencia de problemas de infrarepresentación y falta de respuesta.

3. Alta precisión debida al origen fiscal de los datos.

Los inconvenientes que se indican desde el I.E.F. son los siguientes (cabe señalar que para el objeto de esta investigación los dos primeros no son tales inconvenientes):

1. Imposibilidad de separar las rentas individuales en las declaraciones conjuntas debido a la unidad de análisis (declaración IRPF, modelo 100). En este estudio debido a la variable económica que utilizan las entidades financieras respecto a la renta no supone ningún inconveniente. Las entidades financieras tienen interés en la renta familiar, teniendo en cuenta el número de miembros en la unidad familiar.

2. Imposibilidad de construir declaraciones conjuntas a partir de individuales, ni unir a los declarantes en hogares, debido a la inexistencia de información que relacione las declaraciones. Igual que en el apartado anterior, la información necesaria para las entidades financieras es la renta del cliente y su unidad familiar.

3. Falta de cualquier información extrafiscal no necesaria para la liquidación del impuesto correspondiente. En este caso, sí que ha sido un inconveniente en la investigación desarrollada en esta tesis, ya que algunas variables tanto sociales como económicas que utilizan las entidades financieras para el cálculo del riesgo de crédito no están disponibles en la base de datos utilizada. En la sección 5.2.2 se expone la creación de algunas de las variables de forma semi-sintética para completar la base de datos.

El número de declaraciones y número total de variables de las que se parte se muestra en la siguiente tabla, para los distintos años:

año	**num. clientes**	**num. variables**
2008	1 867 594	388
2009	1 928 494	410
2010	1 904 554	388
2011	2 036 186	447
2012	2 074 225	475
2013	2 161 647	462

Tabla 5.1: Tamaño original de la base de datos.

5.2.2. Construcción base de datos

Los datos de origen son los especificados en el tabla 5.1. El número de variables empleado en cada uno de los años objeto de estudio no coincide porque la Ley del Impuesto sobre la Renta de las Personas Físicas en los distintos años sufre modificaciones y por tanto el modelo 100 de elaboración del IRPF también, introduciendo variaciones en las variables del modelo.

Los pasos a seguir para construir la submuestra que se utiliza para la aplicación de los distintos procedimientos son los siguientes:

1. Se realiza la homogeneización en las variables para todos los años de estudio. Siendo las variables objeto de la muestra las especificadas en el apéndice A.3.

2. Se seleccionan las variables de interés para nuestra investigación, transformando algunas variables previas o bien descartando otras, de acuerdo con las variables que tienen en cuenta las entidades financieras para la concesión de un préstamo. Dentro de este proceso, se observó en la muestra de origen que existían clientes que su base imponible general provenía principalmente de rentas obtenidas por inmuebles de su propiedad (rentas capital inmobiliario). Se considera interesante separar de la base imponible general del cliente la parte de la renta obtenida por rentas de inmuebles. Se crea una variable que tan sólo refleja rentas de capital inmobiliario y la base imponible general la denominamos renta familiar (incluye el resto de rentas que obtiene el cliente, en general serán rentas que provienen de remuneración salarial o actividad empresarial y/o capital mobiliario). El objetivo principal es detectar los clientes que sus rentas vienen determinadas por capital inmobiliario principalmente y si además obtienen rentas por conceptos salariales.

3. Se imputan tres variables a las seleccionadas en el paso anterior.

    3.1. Corrector base imponible especial (BIE). La base imponible especial viene referida a las ganancias y pérdidas de patrimonio que se generan, por ejemplo, en el caso de venta de un inmueble del que se ha sido propietario durante un período de tiempo superior a un año. Por tanto al ser algo extraordinario, que no suele ocurrir anualmente, se considera una variable categórica. Es una variable dicotómica que asigna a todos aquellos clientes que tengan una base imponible especial menor a 120 000 € el valor 0 y los que su base imponible especial sea mayor a 120 000 € se le asigna el valor 1. Se introduce esta variable con el objetivo de discriminar aquellas rentas que puntualmente aparecen (en un año concreto), pero no es habitual en la renta del cliente. Siguiendo los criterios que establece el Banco de España respecto a la valoración de las entidades financieras de la capacidad de pago del prestatario. Tal y como se indica en el BOE (2016), anejo IX, artículos 12 y 13, la capacidad de pago se valora por los flujos de efectivo procedentes de los negocios del prestatario o fuentes de renta habituales u ordinarias. Se considera fuente principal o habitual de renta los fondos procedentes de su trabajo habitual y otras fuentes recurrentes de generación de flujos de efectivo.

    3.2. Corrector base imponible general (BIG). Se añade esta variable siguiendo los criterios que establece el Banco de España respecto a la valoración de las entidades financieras sobre la base del análisis de la capacidad de pago del prestatario y condiciones de concesión de operaciones BOE (2016), anexo IX, artículo 15. En dicho artículo se establece expresamente que dentro de los planes de amortización sobre préstamos que oferten las entidades financieras a los prestatarios, se tendrán en cuenta todos los pagos recurrentes para atender todas las obligaciones financieras, tanto las que posea

con dicha entidad como con otras. La renta disponible del prestatario disminuirá, pero sin que suponga una limitación notoria para cubrir sus gastos familiares. El Gobierno español considera que la renta mínima disponible para atender los gastos familiares es el salario mínimo interprofesional. El corrector de la base imponible general es una variable dicotómica que recategoriza el salario mínimo interprofesional. Se asigna a todos aquellos clientes que tengan una base imponible general menor que el salario mínimo interprofesional el valor de 1 y los que su base imponible general sea mayor al salario mínimo interprofesional se le asigna el valor 0. El salario mínimo interprofesional utilizado es el publicado en el BOE para cada año (BOE, 2007; BOE, 2008; BOE, 2009; BOE, 2010; BOE, 2011; BOE, 2012).

3.3. Importe del préstamo medio solicitado. Esta variable se ha generado en base a las estadísticas trimestrales publicadas por el Banco de España sobre índice de precios de la vivienda en base al año 2005 (`http://www.bde.es/webbde/es/estadis/infoest/indeco.html`.)

Una vez realizados los cambios indicados en las variables, el número total de variables seleccionadas es de 24, tabla 5.2. Las variables seleccionadas son las variables explicativas en el modelo, algunas son numéricas y otras son categóricas. Las variables categóricas se transforman en variables dummy, es decir para cada una de las variables categóricas se crean tantos niveles como categorías tienen las variables menos una, por tanto, el número de variables totales en el modelo es de 103 (la matriz de variables dummy del modelo se representa por $\boldsymbol{X}$).

Descripción	Tipo de Variable	Clase Variable	Opción
Provincia	Categórica	Social	02 Albacete 03 Alacant 04 Almería 33 Asturias 05 Ávila 06 Badajoz 07 Illes Balears 08 Barcelona 09 Burgos 10 Cáceres 11 Cádiz 39 Cantabria 12 Castelló 13 Ciudad Real 14 Córdoba 15 A Coruña 16 Cuenca 17 Girona 18 Granada 19 Guadalajara 21 Huelva 22 Huesca 23 Jaén 24 León 25 Lleida 27 Lugo 28 Madrid 29 Málaga 30 Murcia 32 Ourense 34 Palencia 35 Las Palmas 36 Pontevedra 26 La Rioja

Continúa en la página siguiente...

Descripción	Tipo de Variable	Clase Variable	Opción
			37 Salamanca 38 Santa Cruz de Tenerife 40 Segovia 41 Sevilla 42 Soria 43 Tarragona 44 Teruel 45 Toledo 46 València 47 Valladolid 49 Zamora 50 Zaragoza 51 Ceuta 99 No residentes
Estado civil	Categórica	Social	1= Soltero 2=Casado 3=Viudo 4=Divorciado o separado legalmente
Actividad 1	Categórica	Económica	0=Trabajador por cuenta ajena 1=Industrial 2=Profesional 4=Agricultura y Ganadería
Actividad 2	Categórica	Económica	0=Trabajador por cuenta ajena 1=Industrial 2=Profesional 4=Agricultura y Ganadería
Situación de la vivienda habitual del declarante	Categórica	Social	1=Propiedad 9=Sin información 2=Usufructo 3=Arrendada 4=Otra situación
Porcentaje de propiedad de la vivienda habitual (declarante)	Numérica (en %)	Económica	-
Porcentaje de propiedad de la vivienda habitual (conyuge)	Numérica (en %)	Económica	-
Deduciones relacionadas con la vivienda	Numérica (en €)	Económica	-
Identificación de titularidad del inmueble 1	Categórica	Económica	0=Sin información 1=Conjunta 2=Titular 1 3=Titular 2 4=Hijos

Continúa en la página siguiente...

Descripción	Tipo de Variable	Clase Variable	Opción
Identificación de titularidad del inmueble 2	Categórica	Económica	0=Sin información 1=Conjunta 2=Titular 1 3=Titular 2 4=Hijos
Identificación de titularidad del inmueble 3	Categórica	Económica	0=Sin información 1=Conjunta 2=Titular 1 3=Titular 2 4=Hijos
Identificación de titularidad del inmueble 4	Categórica	Económica	0=Sin información 1=Conjunta 2=Titular 1 3=Titular 2 4=Hijos
Identificación de titularidad del inmueble 5	Categórica	Económica	0=Sin información 1=Conjunta 2=Titular 1 3=Titular 2 4=Hijos
Identificación de titularidad del inmueble 6	Categórica	Económica	0=Sin información 1=Conjunta 2=Titular 1 3=Titular 2 4=Hijos
Renta del Capital Inmobiliario	Numérica	Económica	-
Renta Familiar	Numérica	Económica	-
Número de Inmuebles	Numérica	Económica	-
Número de Miembros de la Unidad Familiar	Numérica	Social	-
Edad del individuo	Categórica	Social	Edad <20 Edad 20-30 Edad 31-40 Edad 41-50 Edad 51-60 Edad 61-70 Edad >70
Miembros de la muestra con una Renta Especial mayor de 120 000 €	Categórica	Estadística	0. BIE < 120 000 € 1. BIE ≥120 000 €
Titularidad Media de los inmuebles	Numérica (en %)	Económica	-
El importe propuesto por el cliente	Numérica (en €)	Económica	-

Continúa en la página siguiente...

Descripción	Tipo de Variable	Clase Variable	Opción
Miembros de la muestra con un salario menor del SMI	Categórica	Estadística	0. BIG < SMI 1. BIG > SMI
Año	Categórica	Social	-

Tabla 5.2: Variables Seleccionadas de la muestra

4. Se depura la base de datos para subsanar errores encontrados (por ejemplo actividad del cliente distinta a las establecidas, porcentaje de titularidad de la vivienda mayor 100 %, fechas de nacimiento incoherentes), falta de información en variables relevantes seleccionadas para el estudio (celdas en blanco por ejemplo respecto a la base imponible general), datos atípicos u outliers.

   El número de clientes objeto de estudio para los distintos años serán los que aparecen en el tabla 5.3.

año	num. clientes
2008	1 418 299
2009	1 437 274
2010	1 391 890
2011	1 458 386
2012	1 480 959
2013	1 507 784

Tabla 5.3: Tamaño de la submuestra de la base de datos.

5. Una vez depurada la base de datos y dada la capacidad computacional disponible, se realiza un muestreo estratificado del 10 % de los datos por variables categóricas para que todas ellas tengan representación en la muestra final, es decir, que no existan estratos vacíos. Por tanto, el número final de observaciones por cada año se recoge en la siguiente tabla, 5.4

   Finalmente se tiene una base de datos de 869 469 clientes.

6. Se fijan los valores de los parámetros que denotan la importancia relativa de cada una de las variables a estudiar. Esta importancia relativa se asigna teniendo en cuenta la relación entre las variables explicativas y la variable respuesta. La variable respuesta se construye sintéticamente. Se considera éxito cuando el cliente cumple con el pago del préstamo y fracaso o default cuando el cliente incumple con el pago del préstamo. Es decir, si denotamos como:

$$p_i = \begin{cases} 0 & \text{éxito} \\ 1 & \text{fracaso} \end{cases}$$

año	num. clientes
2008	141 829
2009	143 727
2010	139 189
2011	145 839
2012	148 097
2013	150 788

Tabla 5.4: Tamaño de la muestra final de base de datos.

el signo que tenga el parámetro $\boldsymbol{\beta}_i$ denotará la influencia de la variable explicativa en el éxito o incumplimiento del cliente en el préstamo. Si el signo es positivo, significa que la variable explicativa es directamente proporcional a la probabilidad de fracaso o default en el préstamo.

Para fijar el valor de cada uno de los parámetros se realiza una búsqueda de información sobre las distintas variables seleccionadas en el modelo objeto de esta investigación y su influencia en el riesgo de crédito. En la normativa del Banco de España se obtienen datos respecto al riesgo de crédito de las entidades financieras (BOE, 2016), en cuanto a la morosidad de clientes de entidades financieras sobre préstamos e hipotecas se utiliza la información obtenida a través del Banco de España (`http://www.//www.bde.es/bde/es/areas/estadis/`). La morosidad por provincias de España se encuentra en los datos publicados por el Instituto Nacional de Estadística, `http://www.ine.es/inebmenu/mnu_financie.htm`. Al mismo tiempo se consultan diversos trabajos sobre el riesgo de crédito bancario empleando variables utilizadas en esta tesis como Mylonakis (2010), Hand and Henley (1997), Malhotra and Malhotra (2003), Boj et al. (2009a), Juan Camilo Ochoa P. et al. (2010), Cabrera Cruz (2014), Moreno Valencia (2013), Salinas Flores (2005), Yu (2014), Gomes Goncalves (2009), Lee et al. (2002), Steenackers and Goovaerts (1989) y Mures et al. (2005). Con los datos recopilados en los trabajos detallados se calculan los valores de los parámetros, siendo los que se detallan en la tabla 5.5.

6.1. Provincia. Se consideran las provincias españolas, sin incluir Comunidad Foral de Navarra y País Vasco, y una provincia como no residentes (son los casos de españoles que tienen su residencia habitual en cualquier país extranjero pero puede que soliciten un préstamo hipotecario en España). Para cada una de las provincias se asigna un valor del parámetro $\boldsymbol{\beta}_i$, siendo todos ellos positivos. Según el informe estadístico sobre ejecuciones hipotecarias por provincias de los años 2014 y 2015, elaborado por el Instituto Nacional de Estadística, junto con la tasa de morosidad de préstamo inmobiliario de los mismos años publicado por el Banco de España, se puede comprobar que la tasa de morosidad por provincias no varía a lo largo de los años. Para su cálculo se ha utilizado la Estadística del INE sobre ejecuciones hipotecarias. Se realiza un

media ponderada por la tasa de morosidad del año 2005. En el gráfico 5.1 se puede apreciar la distribución de la morosidad en las provincias objeto de estudio.

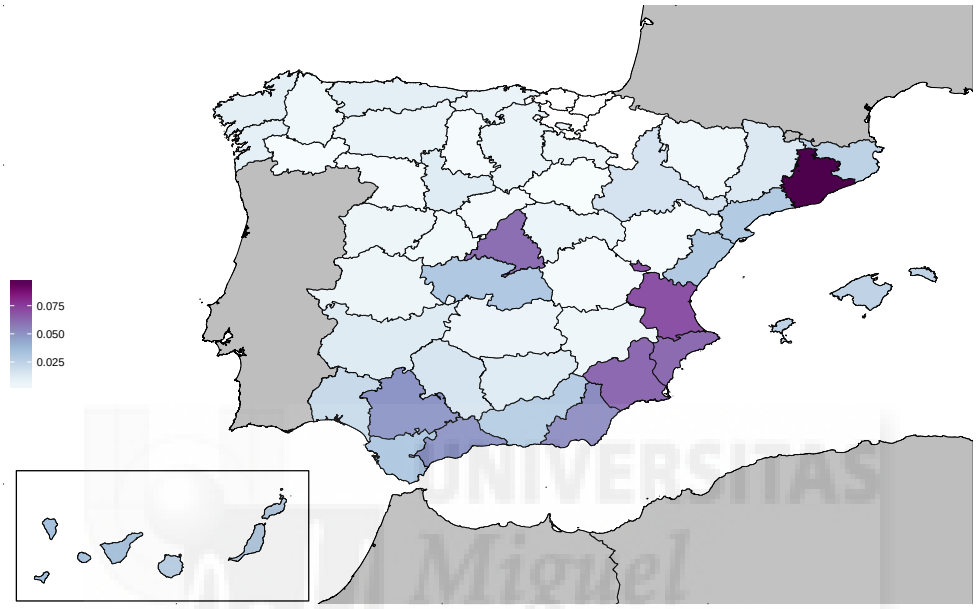

Gráfico 5.1: Distribución de la morosidad por provincias

6.2. Estado civil. Las categorías incluídas en el estudio son: soltero, casado, viudo, divorciado o separado legalmente. Se considera que el único estado civil que influye en el fracaso del préstamo hipotecario es el de divorciado, debido a que probablemente tiene más cargas económicas y financieras. El resto de estado civil no tiene ninguna influencia en el préstamo hipotecario.

6.3. Actividad desarrollada. Las actividades categorizadas se reducen a: trabajador por cuenta ajena, industrial, profesional independiente, y agricultura-ganadería (sector primario). Según los autores Abell (1980), Porter (1982) y López and Martín (1996) en la industria existen altos niveles de incertidumbre debido a las fuertes inversiones en capital fijo que conlleva los constantes cambios tecnológicos, por tanto la actividad industrial posee un mayor riesgo en cuanto a estabilidad laboral se refiere. Se considera que el parámetro asociado a la categoría de personal industrial tiene más riesgo de incumplimiento en un préstamo debido a la inestabilidad laboral señalada. La siguiente actividad valorada con riesgo, aunque menor que el caso de industrial, es el sector primario debido a que su actividad, por tanto sus rentas, están sujetas a riesgos diversos como por ejemplo inclemencias meteorológicas, plagas, etc. La categoría de trabajador profesional independiente se ha valorado positivamente (aunque menos que las anteriores) porque las rentas a percibir no suelen ser fijas anualmente, dependerán de la coyuntura económica del momento, de la profesión, etc.. es decir,

siempre tendrá un pequeño riesgo. La categoría de trabajador por cuenta ajena no se le ha dado ningún valor, porque supone más estabilidad laboral, en esta categoría se encuentran también el personal que trabaja para la administración pública que es el que consideran las entidades financieras de nivel de riesgo mínimo.

6.4. Titular de la vivienda habitual. Las categorías establecidas para esta variable son las siguientes: propietario de la vivienda habitual, usufructuario, arrendatario, sin información y cualquier otra situación. Los clientes que no poseen vivienda habitual o no se tenga información se les asigna un parámetro máximo de 1, el cliente que disfruta de usufructo en la vivienda habitual se le asigna la mitad y el cliente que es propietario se le asigna 0, es decir, no influye en el éxito o fracaso del cumplimiento del préstamo.

6.5. Porcentaje de la titularidad de la vivienda habitual. Esta variable es numérica y el parámetro asignado es en el mismo sentido que la titularidad de la vivienda habitual. Cuanto más porcentaje de la vivienda habitual se dispone, menos riesgo de fracaso conlleva.

6.6. Porcentaje del cónyuge en la titularidad de la vivienda habitual. Esta variable es numérica y el parámetro asignado es en el mismo sentido que el porcentaje de titularidad de la vivienda habitual. Cuanto más porcentaje de la vivienda habitual se dispone, menos riesgo de fracaso conlleva, aunque en este caso se reduce respecto al anterior.

6.7. Deducción en vivienda habitual. Esta variable es numérica. Se considera que al disponer de un préstamo en vivienda habitual y cumplir con el mismo, aportando una deducción que se traduce en más renta disponible para el cliente, se ve reducido el riesgo de fracaso en una pequeña proporción.

6.8. Titularidad del resto de inmuebles. El resto de inmuebles, excluída la vivienda habitual, son 6 posibles. Las variables categóricas referentes a inmuebles se clasifican en: sin información, titularidad conjunta, titularidad del cliente, titularidad del cónyuge o titularidad de los hijos. Los parámetros asignados a cada una de las categorías se establecen considerando el máximo de riesgo de fracaso en el caso de la categoría sin información. Le sigue el caso que la titularidad pertenezca a los hijos o descendientes, ya que el cliente no dispone de su pleno dominio. La titularidad del cónyuge es el siguiente parámetro de mayor valor, y por último titularidad conjunta. El mínimo riesgo se da si el titular es el declarante ya que no afectaría al fracaso.

6.9. Renta capital inmobiliario. Es una variable numérica. El parámetro se considera negativo, ya que el cliente que percibe rentas por los inmuebles que forman parte de su unidad familiar, tiene un riesgo negativo, es decir, que la probabilidad de éxito es mayor.

6.10. Renta familiar. Es una variable numérica. El parámetro asignado es negativo, ya que

la renta es inversamente proporcional al riesgo de fracaso. Cuanto mayor renta posee el cliente, menor riesgo de incumplir o mayor éxito, por tanto es negativa.

6.11. Número de inmuebles. Es una variable numérica. El parámetro se considera negativo. Cuantos más inmuebles posee el cliente menor riesgo de fracaso, por tanto es negativa.

6.12. Número de familiares a cargo. Es una variable numérica. El parámetro se considera positivo. Influye en la misma proporción que la variable número de inmuebles pero en sentido contrario. Cuanto más familiares a cargo se tiene, más probabilidad de fracaso.

6.13. Edad. Es una variable categórica. Se establecen siete categorías empezando con menor de 20 años. Las franjas aumentan de 10 en 10 años, hasta los mayores de 70 años. Antes y después de las edades establecidas como mínima y máxima se consideran casos aislados los posibles clientes que se pueda dar el caso de solicitud de préstamo hipotecario. Siguiendo los estudios antes citados, el parámetro que se considera que no influye en el cumplimiento o incumplimiento de préstamos es la categoría de 20 a 30 años. A partir de esta edad va aumentando exponencialmente hasta el máximo que se sitúa en los mayores de 70 años. Los menores de 20 años se consideran con un riesgo de fracaso mayor, ya que en el caso de tener rentas sería algo puntual.

6.14. Corrector base imponible especial. Es una variable categórica. Se asigna 1 si la renta especial supera los 120 000 €, en otro caso no tendría efecto en el modelo. En relación con el concepto de la propia variable creada para la muestra objeto de estudio, el valor que se le asigna al parámetro será positivo, se entiende que existe un riesgo pequeño de fracaso.

6.15. Porcentaje medio de titularidad. Es una variable numérica y el parámetro asignado es en el mismo sentido que el porcentaje de titularidad de la vivienda habitual. Cuanto más porcentaje de titularidad, respecto a los inmuebles distintos de la vivienda habitual, tenga el declarante, más probabilidad de éxito tendrá o menos probabilidad de incumplir con el préstamo.

6.16. Importe del préstamo hipotecario. Es una variable numérica y el parámetro asignado es positivo. Cuanto mayor sea el importe del préstamo mayor probabilidad de incumplimiento.

6.17. Corrector base imponible general. Es una variable categórica creada. El parámetro asignado tendrá un efecto nulo, en el caso que la renta del cliente sea mayor al salario mínimo interprofesional, en el caso que la renta del cliente sea menor al salario mínimo interprofesional será positivo, teniendo un riesgo alto de fracaso.

En la tabla 5.5 se muestran los valores de los parámetros $\boldsymbol{\beta}$:

Variable		**Beta**
Provincia	02 Albacete	$2{,}063 \cdot 10^{-05}$
	03 Alacant	$2{,}388 \cdot 10^{-04}$
	04 Almería	$2{,}029 \cdot 10^{-04}$
	33 Asturias	$4{,}169 \cdot 10^{-05}$
	05 Ávila	$1{,}537 \cdot 10^{-05}$
	06 Badajoz	$5{,}116 \cdot 10^{-05}$
	07 Illes Balears	$9{,}591 \cdot 10^{-05}$
	08 Barcelona	$3{,}805 \cdot 10^{-04}$
	09 Burgos	$2{,}956 \cdot 10^{-05}$
	10 Cáceres	$2{,}235 \cdot 10^{-05}$
	11 Cádiz	$1{,}220 \cdot 10^{-04}$
	39 Cantabria	$2{,}968 \cdot 10^{-05}$
	12 Castelló	$1{,}160 \cdot 10^{-04}$
	13 Ciudad Real	$2{,}572 \cdot 10^{-05}$
	14 Córdoba	$6{,}226 \cdot 10^{-05}$
	15 A Coruña	$4{,}275 \cdot 10^{-05}$
	16 Cuenca	$7{,}731 \cdot 10^{-06}$
	17 Girona	$9{,}966 \cdot 10^{-05}$
	18 Granada	$1{,}091 \cdot 10^{-04}$
	19 Guadalajara	$2{,}710 \cdot 10^{-05}$
	21 Huelva	$8{,}392 \cdot 10^{-05}$
	22 Huesca	$1{,}520 \cdot 10^{-05}$
	23 Jaén	$4{,}972 \cdot 10^{-05}$
	24 León	$2{,}960 \cdot 10^{-05}$
	25 Lleida	$5{,}527 \cdot 10^{-05}$
	27 Lugo	$1{,}702 \cdot 10^{-05}$
	28 Madrid	$2{,}338 \cdot 10^{-04}$
	29 Málaga	$1{,}940 \cdot 10^{-04}$
	30 Murcia	$2{,}414 \cdot 10^{-04}$
	32 Ourense	$3{,}570 \cdot 10^{-06}$
	34 Palencia	$7{,}046 \cdot 10^{-06}$
	35 Las Palmas	$1{,}066 \cdot 10^{-04}$
	36 Pontevedra	$3{,}931 \cdot 10^{-05}$
	26 La Rioja	$2{,}264 \cdot 10^{-05}$
	37 Salamanca	$2{,}204 \cdot 10^{-05}$

Continúa en la página siguiente...

Variable		Beta
	38 Santa Cruz de Tenerife	$1{,}328 \cdot 10^{-04}$
	40 Segovia	$7{,}553 \cdot 10^{-06}$
	41 Sevilla	$1{,}927 \cdot 10^{-04}$
	42 Soria	$3{,}080 \cdot 10^{-06}$
	43 Tarragona	$1{,}135 \cdot 10^{-04}$
	44 Teruel	$5{,}914 \cdot 10^{-06}$
	45 Toledo	$1{,}183 \cdot 10^{-04}$
	46 València	$2{,}689 \cdot 10^{-04}$
	47 Valladolid	$4{,}980 \cdot 10^{-05}$
	49 Zamora	$6{,}618 \cdot 10^{-06}$
	50 Zaragoza	$6{,}619 \cdot 10^{-05}$
	51 Ceuta	$2{,}143 \cdot 10^{-06}$
	99 No residentes	$3{,}805 \cdot 10^{-04}$
Estado civil	1= Soltero	0,00
	2=Casado	0,00
	3=Viudo	0,00
	4=Divorciado o separado legalmente	0,20
Actividad1	0=Trabajador por cuenta ajena	0,00
	1=Industrial	0,70
	2=Profesional	0,30
	4=Agricultura y Ganadería	0,60
Actividad2	0=Trabajador por cuenta ajena	0,00
	1=Industrial	0,70
	2=Profesional	0,30
	4=Agricultura y Ganadería	0,60
Situación de la vivienda habitual del declarante	1=Propiedad	0,00
	9=Sin información	1,00
	2=Usufructo	0,50
	3=Arrendada	1,00
	4=Otra situación	1,00
Porcentaje de propiedad de la vivienda habitual del declarante		$-0{,}01$
Porcentaje de propiedad de la vivienda habitual del cónyuge		$-5 \cdot 10^{-03}$
Deducciones relacionadas con la vivienda		$-1 \cdot 10^{-03}$
Identificación de titularidad del inmueble 1	0=Sin información	1,00
	1=Conjunta	0,10
	2=Titular 1	0,00
	3=Titular 2	0,60
	4=Hijos	0,75

Continúa en la página siguiente...

Variable		**Beta**
Identificación de titularidad del inmueble 2	0=Sin información	1,00
	1=Conjunta	0,10
	2=Titular 1	0,00
	3=Titular 2	0,60
	4=Hijos	0,75
Identificación de titularidad del inmueble 3	0=Sin información	1,00
	1=Conjunta	0,10
	2=Titular 1	0,00
	3=Titular 2	0,60
	4=Hijos	0,75
Identificación de titularidad del inmueble 4	0=Sin información	1,00
	1=Conjunta	0,10
	2=Titular 1	0,00
	3=Titular 2	0,60
	4=Hijos	0,75
Identificación de titularidad del inmueble 5	0=Sin información	1,00
	1=Conjunta	0,10
	2=Titular 1	0,00
	3=Titular 2	0,60
	4=Hijos	0,75
Identificación de titularidad del inmueble 6	0=Sin información	1,00
	1=Conjunta	0,10
	2=Titular 1	0,00
	3=Titular 2	0,60
	4=Hijos	0,75
Renta del Capital Inmobiliario		$-1 \cdot 10^{-03}$
Renta Familiar		$-1 \cdot 10^{-04}$
Número de Inmuebles		−0,10
Número de Miembros de la Unidad Familiar		0,10
Edad del individuo	Edad <20	0,90
	Edad 20-30	0,00
	Edad 31-40	0,062 5
	Edad 41-50	0,125
	Edad 51-60	0,25
	Edad 61-70	2,00
	Edad >70	10,00
Rentas Extraordinarias	0. BIE < 120 000 €	0,00
	1. BIE ≥ 120 000 €	0,10
Titularidad Media de los inmuebles		−0,01
Importe del préstamo		$1 \cdot 10^{-06}$

Continúa en la página siguiente...

Variable		Beta
Miembros de la muestra con un salario menor del SMI	0. BIG $\geq$ SMI	0
	1. BIG $<$ SMI	5
Año		0

Tabla 5.5: Valores parámetros Betas

7. Se calcula la variable respuesta para cada uno de los clientes. La variable respuesta se define como la probabilidad de fracaso o default, p_i, indicando si el cliente alcanza el estado de morosidad o no. Es una variable dicotómica, donde los clientes clasificados como fracaso o default se codifican con el valor 1 y los clientes clasificados como éxito, es decir que cumplen los pagos del préstamo, se codifican con el valor 0.

   Se calcula la probabilidad de fracaso p_i de la siguiente forma:

   7.1. Las variables independientes son las definidas en la tabla 5.2, y la matriz de las variables independientes es $\boldsymbol{X}$.

   7.2. Los valores de los parámetros que se fijan son los que aparecen en la tabla 5.5, el vector de los parámetros es $\boldsymbol{\beta}$.

   7.3. La variable objetivo se obtiene como regresión lineal, siendo $\boldsymbol{y}$ el vector respuesta que se obtiene como: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$, teniendo en cuenta que el vector de perturbaciones $\boldsymbol{e}$ es un vector que explica las diferencias entre los valores observados y los valores previstos. Se considera que se distribuye de la siguiente forma: $\boldsymbol{e} \sim \mathcal{N}_n(\mathbf{0}; 0{,}15)$.

   7.4. Se obtiene la probabilidad de ocurrencia del fracaso o incumplimiento como:

$$p_i = (e^{\boldsymbol{y}_i})/(1 + e^{\boldsymbol{y}_i})$$

   Al ser probabilidad el resultado de p_i está entre 0 y 1. La variable respuesta objeto de este estudio es una variable discreta dicotómica, es decir, es 0 en caso de éxito y 1 en caso de fracaso o default. Se necesita una transformación de la variable dependiente. Para la transformación de p_i en variable dicotómica se realiza el siguiente proceso:

   1. Se trunca la probabilidad a un sólo decimal el resultado $p_i = floor(p_i, 1)$
   2. Se transforma la probabilidad en número de ocurrencias de un suceso, para ello se multiplica por 10 la probabilidad anterior $P_i = p_i * 10$, obteniendo el número de veces que ocurre un suceso.
   3. Se asigna P_i al suceso fracaso y $10 - P_i$ al suceso éxito, obteniendo la base de datos completa con la variable respuesta dicotómica. La base de datos completa está compuesta por un total de 8 694 690 clientes, de los cuales 3 720 985 se espera dejen de pagar el préstamo, es decir, el 42,80 % serán default.

	Éxito			Default o incumplimiento		
	$\bar{X}$	M_e	σ	$\bar{X}$	M_e	σ
Porcentaje de propiedad de la vivienda habitual del declarante	51,41	50,00	32,51	39,13	50,00	35,86
Porcentaje de propiedad de la vivienda habitual del cónyuge	27,50	50,00	27,17	20,33	0,00	26,65
deducción vivienda habitual	162,67	0,00	394,84	25,37	0,00	165,85
Renta del capital inmobiliario	351,63	0,00	921,00	138,97	0,00	463,63
Renta familiar	45 889,61	54 555,75	31 378,59	9 309,08	3 525,07	19 561,42
Número de inmuebles	1,36	1,00	1,88	0,73	0,00	1,36
Número de miembros de la unidad familiar	2,11	2,00	1,15	1,79	1,00	1,05
Titularidad media de los inmuebles	34,91	14,28	39,16	23,90	0,00	36,94
Importe del préstamo	88 366,40	89 842,86	12 782,64	86 924,90	86 768,13	12 729,36

Tabla 5.6: Análisis descriptivo variables cuantitativas

8. Se separa la muestra definitiva en dos, los años 2008 a 2012 m_A se emplea como muestra de aprendizaje y la muestra del año 2013 m_T como muestra de prueba o test.

5.2.3. Comportamiento de la base de datos

En esta sección se realiza un análisis descriptivo de la muestra definitiva.

En la tabla 5.6 se muestran la media, la mediana y la desviación de las variables cuantitativas seleccionadas en el estudio tanto éxito como default o incumplimiento. En general, para todas las variables, los importes de la media y de la mediana de los clientes que incumplen se corresponden con valores menores que los clientes que cumplen con el préstamo. Los resultados coinciden con lo esperado.

La media del porcentaje de propiedad de la vivienda habitual, en el caso de los clientes que se clasifican como éxito es un poco mayor, sin embargo la mediana es igual. En cuanto al valor de la desviación típica en los clientes que se clasifican como fracaso tienen más variabilidad.

En cuanto a la media del porcentaje de propiedad del cónyuge en la vivienda habitual existe menos diferencia, siendo mayor en los clientes clasificados como éxito, coincide con lo previsto ya que tiene más peso en el modelo propuesto el declarante o cliente que el cónyuge. La mediana, sin embargo, en el caso de los clientes clasificados como éxito es de 50 frente a 0 en el caso de los clientes clasificados como fracaso. La desviación típica es similar, siendo un poco mayor en los clientes clasificados como éxito.

Las variables deducción de la vivienda habitual y renta del capital inmobiliario, tienen el mismo comportamiento respecto a la media, mediana y desviación típica. En el caso de los clientes clasificados como éxito es mayor que en el caso de los clasificados como fracaso, siendo la desviación típica muy alta, es decir, existe mucha variabilidad en los datos.

La media y la mediana para la variable renta familiar de los clientes clasificados como fracaso

es mucho menor que para los clientes clasificados como éxito. Tiene sentido este resultado, ya que en principio se espera que cuanta más renta posea el cliente, más probabilidad de éxito en el cumplimiento del préstamo existe. La desviación típica en el caso de los clientes clasificados como fracaso es muy alta, mayor que la media y la mediana, lo cual significa que no se agrupan en torno a un punto medio, es decir que existe mucha variabilidad en los valores.

La media de la variable número de inmuebles es mayor en los clientes clasificados como éxito, la mediana en los clientes clasificados como éxito se sitúa en un único inmueble y para el caso de fracaso en cero inmuebles. En cuanto a la desviación típica también es mayor en los clientes clasificados como éxito, siendo más alta que la media para las dos opciones, lo cual denota mucha variabilidad.

En cuanto al número de miembros de la unidad familiar, llama la atención que en los clientes clasificados como éxito es mayor su media y su mediana. En principio se puede interpretar que esto es contrario a lo esperado, ya que cuantos más miembros en la unidad familiar, menos renta disponible, por tanto, más probabilidad de incumplimiento en un préstamo. En la construcción del modelo, se ha penalizado el número de miembros de la unidad familiar. Pero viendo el resultado y analizando el origen de los datos que se emplean para el estudio, el número de miembros de la unidad familiar proviene de la suma, de número de descendientes, número de ascendientes, declarante y en el caso de declaración conjunta se le sumaba el cónyuge. El caso que el cónyuge, descendientes y/o ascendientes obtengan rentas todas ellas aparecen sumadas, por tanto, se puede entender que en el caso de la muestra extraída aunque se haya penalizado el número de miembros de la unidad familiar, han obtenido más rentas y ha resultado positivo la existencia de más miembros. Se observa que en los clientes clasificados como incumplimiento, la media del número de miembros de la unidad familiar es 1,79; frente al 2,11 de los clientes clasificados como éxito. La desviación típica es mayor para los clientes clasificados como éxito.

Respecto al porcentaje medio de titularidad de los inmuebles (excluída la vivienda habitual), la media se comporta como el resto de variables, es mayor para los clientes clasificados como éxito. La mediana en los clientes clasificados como fracaso es 0. La desviación típica también es mayor que la media en los dos casos, siendo incluso mayor en los clientes clasificados como fracaso. Es decir, existe mucha variabilidad en los datos.

En cuanto a la variable importe del préstamo, la media, mediana y desviación típica tienen unos valores similares tanto en los clientes clasificados como éxito como fracaso. Este resultado indica que el importe del préstamo no influye para que sea atendido o no el mismo.

En cuanto al análisis descriptivo de las variables independientes cualitativas, se puede observar en la tabla 5.7 los resultados del mismo. En general, se obtienen resultados que se corresponden con lo esperado, aunque en algunos casos, dentro de las categorías, no coincide con el peso que se le daba a las mismas.

Variable		Éxito	Default	Total	% Éxito	% Default
Provincia	Melilla	27 610	15 700	43 310	63,75 %	36,25 %
	Albacete	59 610	70 690	130 300	45,75 %	54,25 %
	Alacant	132 742	110 948	243 690	54,47 %	45,53 %
	Almería	71 428	76 962	148 390	48,14 %	51,86 %
	Asturias	38 777	48 393	87 170	44,48 %	55,52 %
	Ávila	70 834	91 546	162 380	43,62 %	56,38 %
	Badajoz	123 232	80 978	204 210	60,35 %	39,65 %
	Illes Balears	545 081	169 819	714 900	76,25 %	23,75 %
	Barcelona	61 671	59 719	121 390	50,80 %	49,20 %
	Burgos	61 139	76 191	137 330	44,52 %	55,48 %
	Cáceres	103 512	102 258	205 770	50,30 %	49,70 %
	Cádiz	82 239	76 301	158 540	51,87 %	48,13 %
	Cantabria	65 498	77 412	142 910	45,83 %	54,17 %
	Castelló	79 305	89 075	168 380	47,10 %	52,90 %
	Ciudad Real	125 704	96 426	222 130	56,59 %	43,41 %
	Córdoba	40 320	53 370	93 690	43,04 %	56,96 %
	A Coruña	94 016	70 024	164 040	57,31 %	42,69 %
	Cuenca	88 972	96 348	185 320	48,01 %	51,99 %
	Girona	50 617	46 043	96 660	52,37 %	47,63 %
	Granada	64 409	76 181	140 590	45,81 %	54,19 %
	Guadalajara	45 479	50 401	95 880	47,43 %	52,57 %
	Huelva	67 025	90 185	157 210	42,63 %	57,37 %
	Huesca	64 997	66 423	131 420	49,46 %	50,54 %
	Jaén	65 225	62 115	127 340	51,22 %	48,78 %
	León	61 699	59 811	121 510	50,78 %	49,22 %
	Lleida	50 292	66 848	117 140	42,93 %	57,07 %
	Lugo	798 281	197 509	995 790	80,17 %	19,83 %
	Madrid	121 721	103 219	224 940	54,11 %	45,89 %
	Málaga	120 148	99 612	219 760	54,67 %	45,33 %
	Murcia	49 067	63 573	112 640	43,56 %	56,44 %
	Ourense	121 692	100 578	222 270	54,75 %	45,25 %

Continúa en la página siguiente...

Variable		Éxito	Default	Total	% Éxito	% Default
	Palencia	38 685	41 645	80 330	48,16 %	51,84 %
	Las Palmas	92 025	78 465	170 490	53,98 %	46,02 %
	Pontevedra	91 194	85 736	176 930	51,54 %	48,46 %
	La Rioja	59 413	63 307	122 720	48,41 %	51,59 %
	Salamanca	91 015	86 105	177 120	51,39 %	48,61 %
	Santa Cruz de Tenerife	85 620	72 400	158 020	54,18 %	45,82 %
	Segovia	37 840	40 790	78 630	48,12 %	51,88 %
	Sevilla	160 222	110 598	270 820	59,16 %	40,84 %
	Soria	30 249	31 901	62 150	48,67 %	51,33 %
	Tarragona	92 625	78 225	170 850	54,21 %	45,79 %
	Teruel	37 351	42 869	80 220	46,56 %	53,44 %
	Toledo	74 126	81 534	155 660	47,62 %	52,38 %
	Valéncia	246 766	128 584	375 350	65,74 %	34,26 %
	Valladolid	85 194	73 436	158 630	53,71 %	46,29 %
	Zamora	39 573	51 327	90 900	43,53 %	56,47 %
	Zaragoza	125 355	91 625	216 980	57,77 %	42,23 %
	Ceuta	27 212	13 838	41 050	66,29 %	33,71 %
	No residentes	6 898	3 942	10 840	63,63 %	36,37 %
Estado civil	Soltero	871 639	1 047 911	1 919 550	45,41 %	54,59 %
	Casado	3 601 500	2 211 280	5 812 780	61,96 %	38,04 %
	Viudo	201 647	280 293	481 940	41,84 %	58,16 %
	Divorciado o separado legalmente	298 919	181 501	480 420	62,22 %	37,78 %
Actividad 1	Trabajador por cuenta ajena	4 194 006	3 243 164	7 437 170	56,39 %	43,61 %
	Industrial	337 324	316 916	654 240	51,56 %	48,44 %
	Profesional	416 694	144 356	561 050	74,27 %	25,73 %
	Agricultura y Ganadería	25 681	16 549	42 230	60,81 %	39,19 %
Actividad 2	Trabajador por cuenta ajena	4 354 248	3 345 822	7 700 070	56,55 %	43,45 %
	Industrial	537 379	338 591	875 970	61,35 %	38,65 %
	Profesional	67 185	29 295	96 480	69,64 %	30,36 %
	Agricultura y Ganadería	14 893	7 277	22 170	67,18 %	32,82 %
Titular de la Vivienda Habitual	Propiedad	4 106 333	2 339 557	6 445 890	63,70 %	36,30 %
	Usufructo	29 836	34 444	64 280	46,42 %	53,58 %
	Arrendada	230 666	231 194	461 860	49,94 %	50,06 %
	Otra situación	541 426	930 544	1 471 970	36,78 %	63,22 %
	Sin información	65 444	185 246	250 690	26,11 %	73,89 %

Continúa en la página siguiente...

Variable		Éxito	Default	Total	% Éxito	% Default
Inmueble Titular 1	Sin información	2 444 986	2 452 874	4 897 860	49,92 %	50,08 %
	Conjunta	263 117	145 423	408 540	64,40 %	35,60 %
	Titular 1	2 157 484	1 046 586	3 204 070	67,34 %	32,66 %
	Titular 2	107 756	75 964	183 720	58,65 %	41,35 %
	Hijos	362	138	500	72,40 %	27,60 %
Inmueble Titular 2	Sin información	3 387 484	3 070 376	6 457 860	52,46 %	47,54 %
	Conjunta	147 472	70 128	217 600	67,77 %	32,23 %
	Titular 1	1 291 267	487 943	1 779 210	72,58 %	27,42 %
	Titular 2	147 223	92 427	239 650	61,43 %	38,57 %
	Hijos	259	111	370	70,00 %	30,00 %
Inmueble Titular 3	Sin información	3 965 629	3 381 411	7 347 040	53,98 %	46,02 %
	Conjunta	86 354	35 916	122 270	70,63 %	29,37 %
	Titular 1	844 900	260 510	1 105 410	76,43 %	23,57 %
	Titular 2	76 658	43 082	119 740	64,02 %	35,98 %
	Hijos	164	66	230	71,30 %	28,70 %
Inmueble Titular 4	Sin información	4 282 866	3 515 404	7 798 270	54,92 %	45,08 %
	Conjunta	52 177	19 033	71 210	73,27 %	26,73 %
	Titular 1	572 960	151 650	724 610	79,07 %	20,93 %
	Titular 2	65 556	34 864	100 420	65,28 %	34,72 %
	Hijos	146	34	180	81,11 %	18,89 %
Inmueble Titular 5	Sin información	4 491 138	3 596 772	8 087 910	55,53 %	44,47 %
	Conjunta	32 335	10 625	42 960	75,27 %	24,73 %
	Titular 1	415 288	96 722	512 010	81,11 %	18,89 %
	Titular 2	34 826	16 844	51 670	67,40 %	32,60 %
	Hijos	118	22	140	84,29 %	15,71 %
Inmueble Titular 6	Sin información	4 615 891	3 637 659	8 253 550	55,93 %	44,07 %
	Conjunta	19 622	5 958	25 580	76,71 %	23,29 %
	Titular 1	307 732	63 578	371 310	82,88 %	17,12 %
	Titular 2	30 343	13 777	44 120	68,77 %	31,23 %
	Hijos	117	13	130	90,00 %	10,00 %
Edad	Edad $<$ 20	218 744	438 846	657 590	33,26 %	66,74 %
	Edad 20-30	926 819	648 141	1 574 960	58,85 %	41,15 %
	Edad 31-40	1 349 209	678 131	2 027 340	66,55 %	33,45 %
	Edad 41-50	1 339 455	635 485	1 974 940	67,82 %	32,18 %
	Edad 51-60	749 512	540 898	1 290 410	58,08 %	41,92 %
	Edad 61-70	385 063	759 097	1 144 160	33,65 %	66,35 %
	Edad $>$ 70	4 903	20 387	25 290	19,39 %	80,61 %
Corrector BIE	BIE $<$ 120 000 €	4 919 026	3 704 814	8 623 840	57,04 %	42,96 %
	BIE $\geq$ 120 000 €	54 679	16 171	70 850	77,18 %	22,82 %

Continúa en la página siguiente...

Variable		Éxito	Default	Total	% Éxito	% Default
Corrector Renta Familiar	BIG $\geq$ SMI	4 030 884	457 886	4 488 770	89,80 %	10,20 %
	BIG $<$ SMI	942 821	3 263 099	4 205 920	22,42 %	77,58 %
Año	2008	873 361	544 929	1 418 290	61,58 %	38,42 %
	2009	858 696	578 574	1 437 270	59,74 %	40,26 %
	2010	800 784	591 106	1 391 890	57,53 %	42,47 %
	2011	832 875	625 515	1 458 390	57,11 %	42,89 %
	2012	796 614	684 356	1 437 270	53,79 %	46,21 %
	2013	811 375	696 505	1 507 880	53,81 %	46,19 %
TOTAL		**4 973 705**	**3 720 985**	**8 694 690**	**57,20 %**	**42,80 %**

Tabla 5.7: Análisis descriptivo variables cualitativas

Respecto a la categoría provincia se observa que en general no existe ninguna que destaque por incumplimiento del préstamo hipotecario, es decir, que no hay ninguna provincia de residencia del cliente en la que el suceso de incumplimiento ocurra más. Sin embargo, sí que destaca el caso contrario, existen provincias en las que el cumplimiento en el pago del préstamo hipotecario sucede más que el incumplimiento, son los casos siguientes: las ciudades autónomas de Ceuta y Melilla (63,75 %), Islas Baleares (76,25 %), Lugo (80,17 %), Valencia (65,74 %) y los no residentes (63,63 %).

Se realiza un gráfico 5.2 agrupando las provincias por comunidades autónomas y se observa el comportamiento en cuanto a cumplimiento o incumplimiento del préstamo hipotecario en las distintas Comunidades Autónomas. La Comunidad Autónoma de Murcia es en la que más incumplimiento existe en la muestra objeto de estudio, seguida de Asturias y Cantabria. En cuanto a comunidades con mayor porcentaje de éxito o cumplimiento se sitúa Baleares seguido de Galicia y las ciudades autónomas de Ceuta y Melilla.

En cuanto a la variable estado civil, se observa que tanto el soltero como el viudo son las categorías que más se produce el incumplimiento en el préstamo, siendo en las categorías de casado y divorciado o separado legalmente en las que se producen más éxitos en cuanto al cumplimiento con el préstamo. En el caso del divorciado se produce lo contrario a lo esperado, se ha considerado que el parámetro del estado civil divorciado influye en el fracaso del préstamo hipotecario.

Las variables actividades profesionales, sea como primera o segunda actividad, son similares. La categoría de profesional es la que más éxito alcanza en el cumplimiento. En todos los casos es mayor los casos de éxito que de fracaso, aunque se aprecia que en la categoría de industrial la diferencia es muy pequeña. El trabajador por cuenta ajena es el siguiente mayor en cuanto a incumplimiento del préstamo, seguido de la actividad agricultura y ganadería.

La variable titularidad de la vivienda habitual refleja el comportamiento esperado, la categoría de propietario es la que más éxito refleja, siendo el resto de categorías las que más fracaso reflejan, sin información con un 73,89 %, seguida de otra situación con 63,22 %, usufructuario

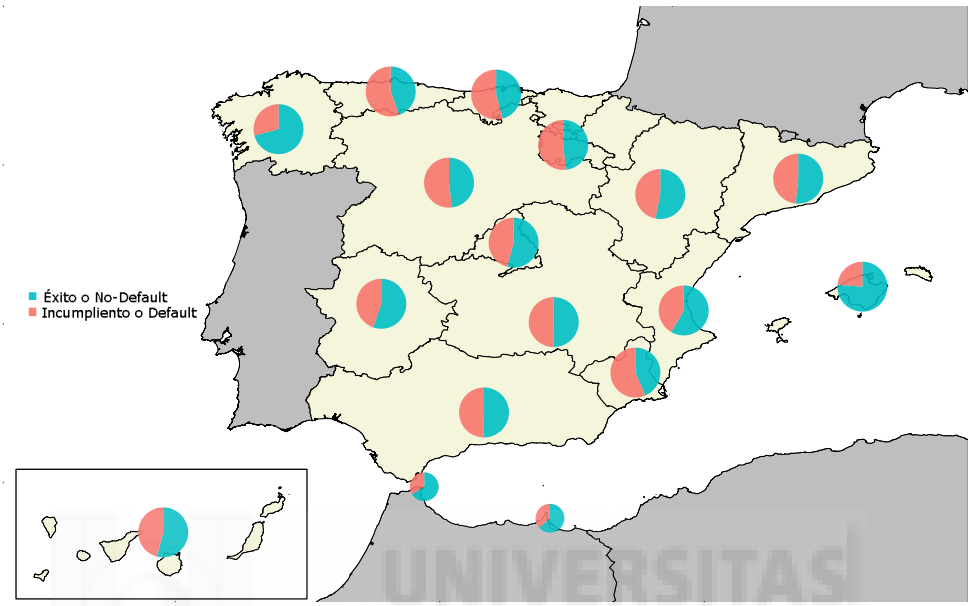

Gráfico 5.2: Tasa de cumplimiento/incumplimiento por comunidades autónomas

con un 53,58 % y arrendado con un 50,06 %.

Las variables titularidad del resto de inmuebles tienen un comportamiento similar, aunque se observa que conforme se tienen más inmuebles aumenta el éxito o cumplimiento con el préstamo hipotecario en todas las categorías. La categoría con mayor fracaso es la asignada sin información, seguida de la categoría titular 2, referido al cónyuge como titular. Después se sitúa la categoría de titularidad conjunta, titularidad del cliente y como mayor éxito cuando la titularidad la tengan los hijos. Estos resultados difieren de lo esperado ya que se había penalizado el parámetro cuando la titularidad del resto de inmuebles la tenían los hijos, respecto a los otros casos de titularidad. La explicación se encuentra en la muestra, los casos que la titularidad de los inmuebles la posean los hijos es una proporción muy pequeña de clientes, por tanto el resultado no es muy representativo, porque esos pocos casos se compensan con otras variables que tienen más peso en la variable dependiente (representa aproximadamente el 0,01 % del total de categorías).

En cuanto a la variable edad, se observa que la categoría que obtiene un alto porcentaje de fracaso es la de mayores de 70 años, tal y como se esperaba, seguido de los menores de 20. El comportamiento de la variable edad en cuanto a éxito, va creciendo desde más de 20 años hasta alcanzar el máximo en 41 y 50 años para volver a decrecer y situarse en el mínimo que se corresponde con los mayores de 70 años.

La variable corrector de la base imponible especial obtiene un porcentaje mayor de éxito en el caso que la categoría sea mayor a 120 000 €, tal y como se esperaba. Cuando es menor sigue siendo mayor el porcentaje de éxito que el fracaso aunque el fracaso se sitúa en 42,96 %.

La variable corrector de la base imponible general, cuando la base imponible general es

menor que el salario mínimo interprofesional, obtiene un porcentaje alto de fracaso, tal y como se esperaba (77,58 %). En el caso que la base imponible general sea mayor al salario mínimo interprofesional se comporta en el otro sentido, se obtiene un porcentaje de éxito alto, del 89,80 %.

La variable año se comporta de una forma similar durante todo el período de estudio de la base de datos. El porcentaje de éxito se sitúa entorno al 55 %. A destacar que los años 2008 y 2009 tienen un porcentaje de éxito mayor, descendiendo progresivamente en los años sucesivos.

5.2.4. Procedimiento

El objetivo de esta tesis es buscar el método estadístico más eficaz y eficiente para evaluar la predicción de default o incumplimiento en una entidad financiera para préstamos hipotecarios. Se construye el modelo con la muestra de los años 2008 a 2012 m_A como muestra de aprendizaje, se calcula posteriormente la predicción del modelo con la muestra del año 2013 m_T como muestra de prueba o test.

Los pasos seguidos para realizar el estudio han sido los siguientes:

1. Se aplican los siguientes métodos estadísticos a los datos semi-reales obtenidos:
   a. Análisis lineal discriminante (LDA). Para la obtención de los resultados se aplica el paquete de R software MASS, de los autores Venables and Ripley (2002), de R, con dos variables discriminantes.
   b. Árboles de clasificación (CART). Se utiliza el paquete rpart, de los autores Therneau et al. (2014), de R.
   c. Modelo lineal mixto, efectos fijos y aleatorios (LMM). Para su aplicación se considera variable aleatoria la provincia. Para su aplicación se utilizan los paquetes lme4, de los autores de Bates et al. (2015b) de R.
   d. Modelo lineal generalizado con nexo logit (GLMlogit). Se utiliza el paquete básico de R, que tiene implementada la función GLM, dentro del paquete stats.
   e. Máquinas de vectores soporte (LSVM). Se utiliza LSVM con un kernel lineal (LSVM). Se emplea el paquete e1071, de Meyer et al. (2014), de R.
2. Se calcula el valor estimado de la variable objetivo $\widehat{\boldsymbol{y}}$ para cada uno de los métodos con la muestra de aprendizaje m_A.
3. Se calcula el error cuadrático medio(RMSE) con la muestra de aprendizaje m_A.
4. Se obtiene la predicción del modelo con la muestra de prueba o test m_T y el modelo ajustado en cada caso.
5. Una vez realizada la predicción se procede a evaluar su desempeño y comparar sus resultados, es decir, se cuantifica la bondad de la predicción. En el procedimiento empleado se

utiliza la matriz de confusión junto a la medición del éxito logrado, fracaso y los errores de haber identificado mal al cliente (error tipo I y error tipo II). En la matriz, tabla de confusión o de contingencia se materializa tanto los aciertos como los errores para cada una de las clasificaciones, en el caso objeto de estudio, las clases predichas comparadas con los valores reales de la muestra de test.

6. Se mide el tiempo medio total empleado para cada una de las técnicas. El objetivo es encontrar un método que minimice el error cuadrático medio, obtenga el máximo acierto y se ejecute en el menor tiempo posible.

5.2.5. Resultados numéricos

Para cada uno de los métodos estadísticos se calculan las medidas de eficiencia planteadas. Dentro de los resultados numéricos se ha eliminado el método LSVM. Después de 72 días procesándose no terminó de ejecutarse y teniendo en cuenta que las entidades financieras cada 3 meses rinden cuentas, el método LSVM no resulta operativo por ser ineficiente computacionalmente. En todo caso, se ha realizado una prueba con el 1 % de la muestra estratificada, para comprobar si los resultados pudiesen ser óptimos. Se obtuvo una tasa de acierto de 78,03 %, un error cuadrático medio de 46,86819 % y el tiempo de ejecución fue de 7 257 600 segundos (2016 horas o su equivalente en días 84).

Una vez analizados los resultados obtenidos con la muestra semi-real, se comparan los resultados con los obtenidos en el experimiento 4 con la muestra sintética.

Respecto al error cuadrático medio, como se observa en la tabla 5.8, el mejor resultado es el que se obtiene con el método CART, seguido de GLMlogit, LMM y el peor resultado se obtiene con el método LDA. La diferencia es muy pequeña, excepto en el caso del análisis lineal discriminante. Se observa que los métodos estadísticos (LMM, GLMlogit), obtienen resultados muy similares.

Respecto a la tasa de acierto se observa que el método que obtiene una mayor tasa de acierto es LMM, GLMlogit y LDA, y el peor resultado se obtiene con CART. Al igual que ocurre con el error cuadrático medio se observa que los métodos LMM y GLMLogit obtienen el mismo porcentaje de acierto, siendo bastante alto, 83,35 %.

Método	**Acierto**	**RMSE**	**Error Tipo I**	**Error Tipo II**
LMM	83,35 %	36,34 %	179 177	71 872
GLMlogit	83,35 %	36,16 %	179 177	71 872
LDA	83,35 %	40,80 %	179 177	71 872
CART	50,88 %	35,76 %	57 423	91 419

Tabla 5.8: Resultados numéricos

Por último, vamos a analizar el tiempo medio total que se ha empleado por cada uno de los métodos. Las entidades financieras necesitan métodos que sean tanto eficientes como eficaces.

La eficacia se mide a través del error cuadrático medio y la tasa de acierto y la eficiencia con el tiempo que tarda en obtener los resultados.

Según se observa en el gráfico 5.3 el método más eficiente es CART seguido de LMM, GLM-logit y el menos eficaz LDA.

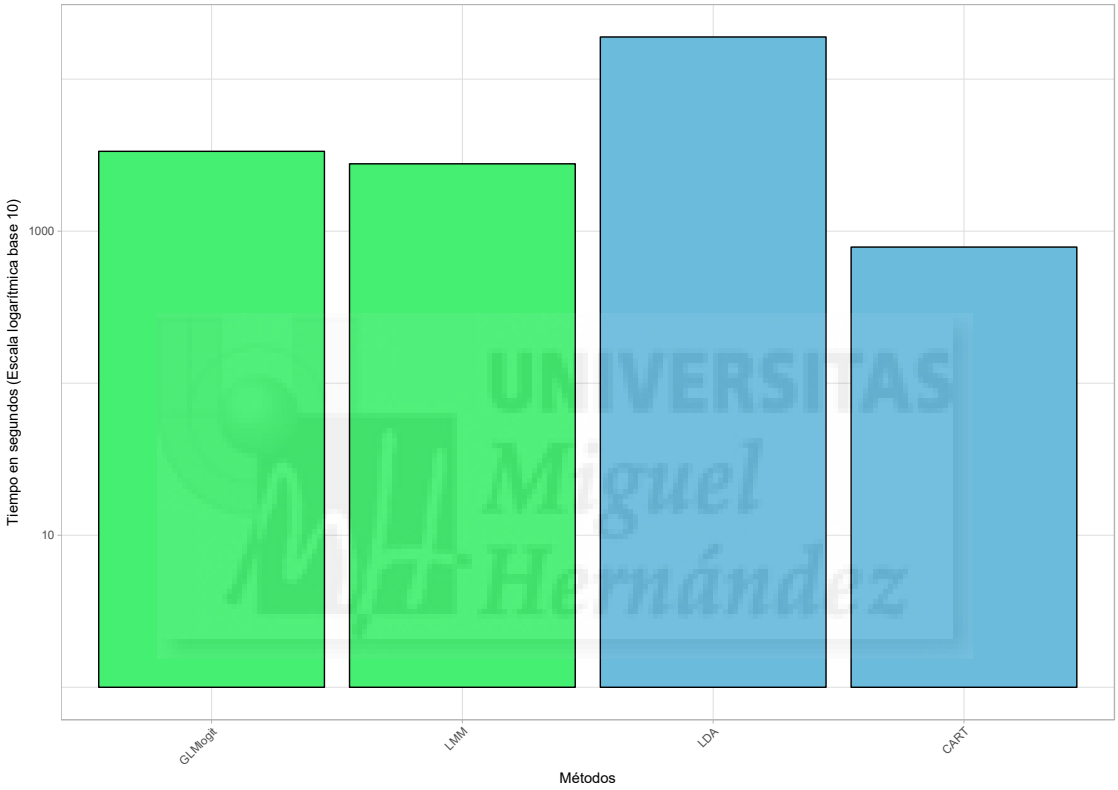

Gráfico 5.3: Resultados eficiencia

Analizando todos los resultados, el método LDA queda descartado tanto por su ineficacia como su ineficiencia. Con el método CART se obtienen buenos resultados tanto para el error cuadrático medio, como para el tiempo de ejecución, pero un mal resultado en cuanto al acierto, a pesar de ser un método que clasifica, es decir no influye el punto de corte en la obtención de la matriz de confusión.

Dentro de la matriz de confusión se obtiene el error tipo I y el error tipo II. Es interesante analizar los resultados obtenidos en los mismos para todos los métodos, ya que van a medir el número de clientes que se clasifican mal, siendo el error tipo I el número de clientes que atenderían el pago pero se clasifican como incumplimiento y el error tipo II, es el número de clientes clasificados como éxito (atienden el pago), pero realmente son clientes que incumplen con el pago.

Se puede ver en la tabla 5.8, que el método que da un error menor de tipo I es el CART. Los

resultados más altos se dan en los métodos LDA, GLMlogit y LMM. En cuanto al error tipo II, los que presentan un menor error son GLMlogit, LMM y LDA. CART presenta un error mayor a los anteriores. Tanto para el erro tipo I como tipo II, los resultados obtenidos para LMM, GLMlogit y LDA son iguales. Dependerá de la política de riesgos que quiera asumir la entidad financiera, le interesará un método u otro. Si la entidad financiera es más conservadora en cuanto al riesgo elegirá los métodos que menor error tipo II tengan (GLMlogit, LMM y LDA), aunque rechacen más clientes que se clasifican como default y sí que atenderían el pago. Si la entidad financiera asume mayor riesgo, sería al contrario y elegiría el método CART.

El siguiente paso es comparar los resultados obtenidos en la muestra semi-real (o semi-sintética) con los resultados obtenidos con la muestra sintética del experimento 4.

Para poder comparar los resultados obtenidos, se realiza el mismo experimento de las muestras sintéticas pero con 103 variables y 5 niveles, ya que son las variables y niveles de los que se dispone en la muestra semi-real.

En cuanto a la tasa de acierto, según se observa en el gráfico 5.4, se obtienen mejores resultados con la muestra semi-real excepto con el método CART. Cuanto más datos disponibles existan, mejor ajuste se puede realizar, en el caso de la muestra semi-real se dispone de muchos más datos de posibles clientes 8 694 690 frente a 50 000.

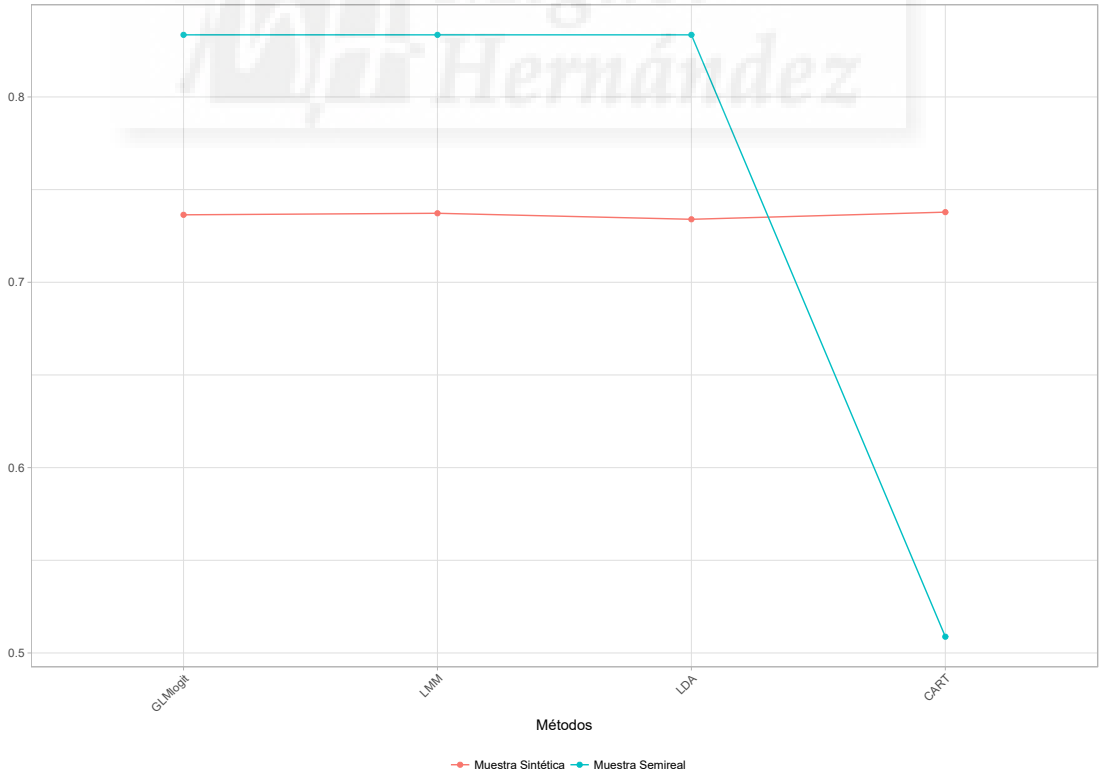

Gráfico 5.4: Tasa de acierto

En cuanto al error cuadrático medio, según se observa en el gráfico 5.5, se comprueba que es un poco mejor para todos los métodos en el caso de la muestra sintética que en la muestra semi-real, excepto LDA, que coincide con el método que peores resultados obtiene respecto al error cuadrático medio de la muestra semi-real. Se puede comprobar en el gráfico 4.1 de la muestra sintética que conforme aumentan los niveles (I) se aproximan más los resultados obtenidos con la muestra semi-real, siendo GLMlogit y LMM los mejores métodos.

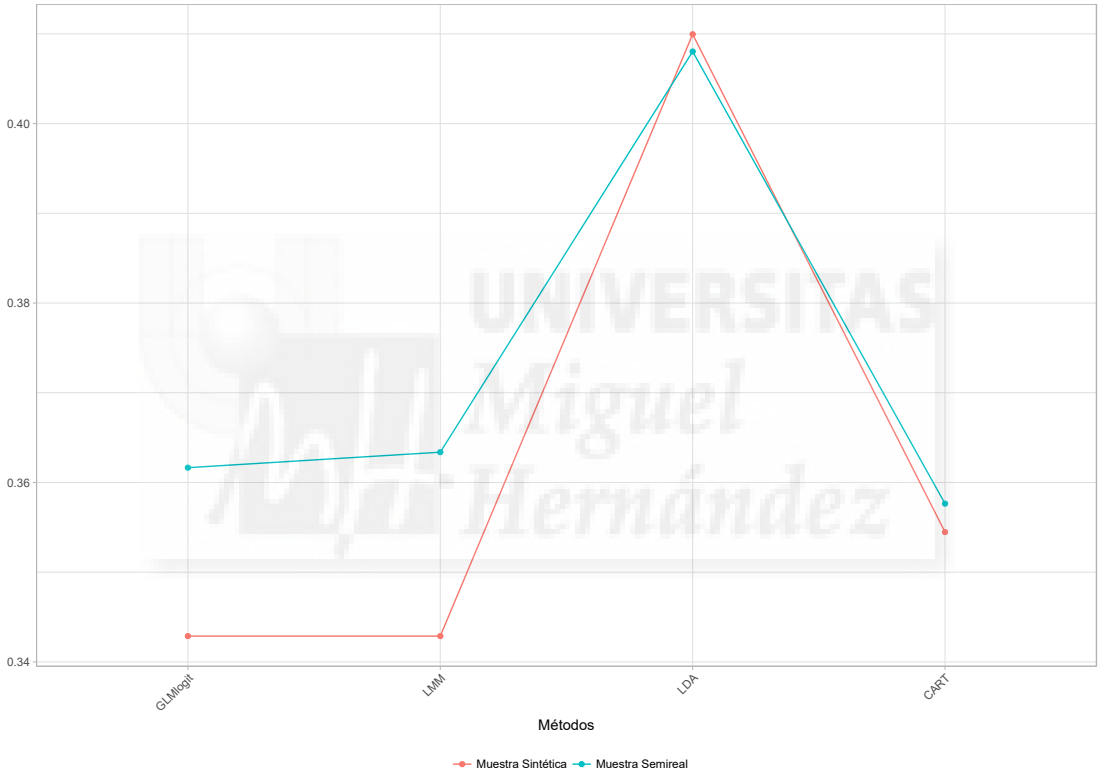

Gráfico 5.5: Error cuadrático medio

5.2.6. Análisis del sesgo y el error cuadrático medio

Una vez analizada la eficacia y eficiencia de cada uno de los métodos aplicados en esta investigación, se estudia el sesgo y error cuadrático medio de la variable objetivo, el incumplimiento del préstamo hipotecario, y de los parámetros de las variables explicativas. Se realiza el análisis para los dos métodos que resultan óptimos en cuanto a eficiencia y eficacia, es decir con LMM y GLMlogit para poder tener más pruebas estadísticas que indiquen qué método es el más adecuado para predecir el incumplimiento de un préstamo.

El procedimiento utilizado para obtener los sesgos y error cuadrático medio son los siguientes:

1. Iteraciones. Se repite $K = 10^3$ veces $(k = 1, \ldots, K)$

2. Error. Se genera un error aleatorio para cada vuelta.

$$\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma_0^2 = 0{,}15).$$

3. Variable objetivo. Se calcula la variable objetivo partiendo de la $\boldsymbol{y}$ obtenida en la muestra semi-real completa (desde el 2008 al 2013),

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

, para cada una de las vueltas.

4. Probabilidad de ocurrencia. Se obtiene la probabilidad de ocurrencia del fracaso o incumplimiento como:

$$p_i = (e^{\boldsymbol{y}_i})/(1 + e^{\boldsymbol{y}_i})$$

5. Número de ocurrencias. Se transforma la probabilidad en número de ocurrencias de un suceso, para ello se multiplica por 10 la probabilidad anterior, obteniendo el número de veces que ocurre un suceso. El resultado se trunca $P_i = floor(p_i * 10)$.

6. Variable dicotómica. Se asigna P_i al suceso fracaso y $10 - P_i$ al suceso éxito, obteniendo la variable respuesta dicotómica.

7. Procedimientos. Se calculan las regresiones con R, para los dos procedimientos LMM (con el paquete de Bates et al. (2015b)) y GLMlogit (con la función GLM, paquete stats, dentro del paquete básico de R).

8. Valores estimados. Se calculan los valores estimados de los parámetros y el valor estimado de la variable objetivo $\widehat{\mu}$ para cada uno de los procedimientos.

9. BIAS y EMSE. Se calcula, para cada parámetro estimado $\tau \in \{\boldsymbol{\beta}\}$ y para el valor estimado de la variable objetivo $\widehat{\mu}$ el error cuadrático medio (EMSE) y el sesgo (BIAS).

$$EMSE(\widehat{\tau}) = \frac{10^3}{K}\sum_{k=1}^{K}(\widehat{\tau}_{(k)} - \tau)^2, \qquad BIAS(\widehat{\tau}) = \frac{10^3}{K}\sum_{k=1}^{K}(\widehat{\tau}_{(k)} - \tau).$$

$$EMSE(\widehat{\mu}) = \frac{10^3}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\widehat{y}_{ij(k)} - y_{ij})^2,$$

$$BIAS(\widehat{\mu}) = \frac{10^3}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\widehat{y}_{ij(k)} - y_{ij}).$$

	SESGO	EMSE
GLMlogit	$4{,}240\,6 \cdot 10^{-07}$	0,13211954
LMM	$1{,}087\,1 \cdot 10^{-07}$	0,13381353

Tabla 5.9: BIAS y EMSE $\widehat{\mu}$

Los resultados numéricos obtenidos para cada una de las variables dummy de la muestra semi-real son los que aparecen en el apéndice A.4.

Los resultados numéricos del sesgo y error cuadrático medio para los dos métodos de la variable objetivo se muestran en la tabla 5.9:

Cuanto más próximo a cero sea el sesgo mejor se ha ajustado el modelo, se puede comprobar que con los dos métodos se obtienen resultados óptimos, siendo LMM el mejor. En cuanto a EMSE también se obtienen resultados buenos, el error es pequeño en todos los métodos, siendo GLMlogit el mejor.

Por tanto, se comprueba que con los dos métodos se ha ajustado bien la muestra semi-real.

Se grafican los parámetros estimados de cada una de las variables y métodos junto con los parámetros muestrales obtenidos en la tabla (5.5) que se han aplicado en la investigación. Las variables que corresponden a las provincias no se han tenido en cuenta porque distorsionan el gráfico debido a que el efecto aleatorio influye en la variable provincia. Se puede observar en el gráfico 5.6 que los parámetros muestrales no difieren mucho de los parámetros estimados en los dos métodos LMM y GLMlogit. Entre el 87,27 % y 90,91 % de las variables presentan una diferencia menor a 1, y tan sólo dos variables presentan una diferencia mayor a 2, que se corresponde con la variable edad 6 (franja de edad de 61 a 70 años) y renta correctora.

Se puede concluir que con los dos métodos LMM y GLMlogit el modelo planteado se ajusta de una manera óptima.

5.2.7. Conclusiones

En este capítulo, al igual que en el anterior, se ha realizado un estudio comparativo de los métodos LDA, CART, LMM, GLMlogit y LSVM para medir su eficacia y eficiencia computacional. La eficacia se ha medido a través del error cuadrático medio y la tasa de acierto, la eficiencia computacional se ha medido calculando el tiempo medio de ejecución. Para comprobar la aplicabilidad de los resultados obtenidos en el capítulo anterior se realiza el estudio con una base de datos semi-real y en el punto siguiente con dos bases de datos reales. Con esto, se ha cumplido la parte real del objetivo número 3 de la sección “Objetivos” del “Prólogo”.

Una vez analizados los resultados obtenidos, los métodos recomendados tanto por su eficacia

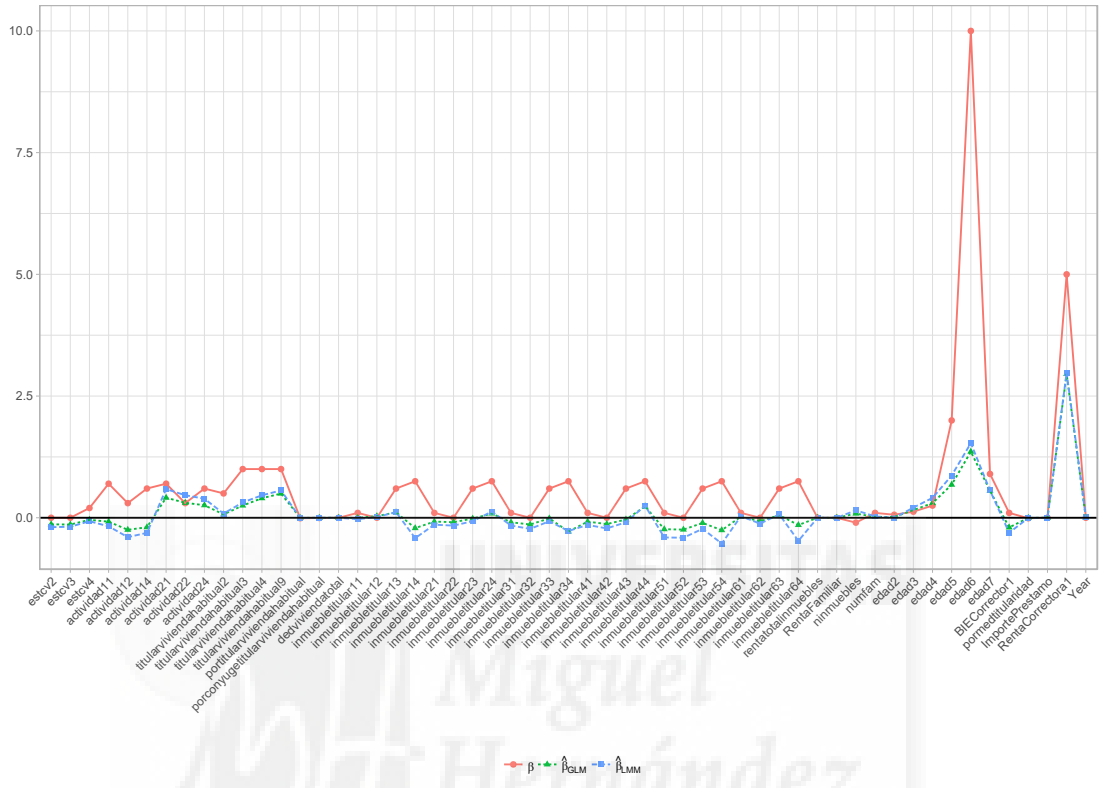

Gráfico 5.6: Comparación parámetros estimados y muestrales

como por su eficiencia son los estadísticos (LMM, GLMlogit) junto con el método no paramétrico CART. Se obtienen mejores resultados en cuanto a error cuadrático medio, porcentaje de acierto y tiempos de ejecución.

En cuanto a los errores tipo I y tipo II, es interesante para la entidad financiera la valoración de los mismos, ya que si se elige un método en el que su error tipo I sea más alto, en los casos analizados, LDA, GLMlogit y LMM, la entidad financiera estaría rechazando clientes potenciales que le reportarían ingresos. Sin embargo con el error tipo II, el método con el que se obtiene mayor error es CART, la entidad financiera estaría aceptando un cliente que va a entrar en default, lo que va a suponer unas pérdidas, y un coste de oportunidad en cuanto a nivel máximo de crédito que puede otorgar, ya que parte de ese crédito se lo está concediendo a clientes que incumplen y deja de concedérselo a clientes que podrían no incumplir. Los métodos que mejor clasifican los clientes en cuanto al error tipo II, son GLMlogit, LMM y LDA.

Tal y como se ha comprobado con los resultados obtenidos, es comparable las conclusiones obtenidas con la muestra semi-real y la sintética. Por tanto, LMM y GLMlogit son los métodos más recomendables, ya que CART conforme aumentan los niveles (en el caso objeto de estudio los niveles vienen determinados por el número de años) empeora sus resultados en cuanto a

acierto y error cuadrático medio.

Para una entidad financiera resulta de vital importancia disponer de una herramienta estadística que le permita predecir de una forma rápida el comportamiento de los clientes antes de otorgarles un préstamo hipotecario. Con un método potente predictor, la entidad financiera cumple tres funciones importantes:

1. Se garantiza una rentabilidad con las obligaciones crediticias, por lo tanto aumento del beneficio.

2. Oportunidad de nuevos clientes, por tanto de crecimiento de mercado.

3. Cumplen una labor social respecto a la vivienda, en el sentido de facilitar préstamos para su adquisición a todos aquellos clientes que reúnan características mínimas que garanticen la devolución de los mismos, aunque tengan que aumentar el plazo de devolución. Este tipo de clientes podrían quedar fuera del sistema si las entidades financieras no aplican métodos de predicción adecuados.

5.3. Aplicación a casos reales

5.3.1. Procedimiento

El objetivo es poner a prueba los métodos estudiados y comprobar que el método más eficaz y eficiente para determinar la probabilidad de default en estas dos bases de datos reales, coincide con lo expuesto en capítulos anteriores.

La base de datos *Australian Credit* hace referencia a datos sobre incumplimiento en tarjetas de crédito. Está compuesta por 690 observaciones, una vez eliminadas las observaciones con valores perdidos. Por confidencialidad de los datos no aparecen en la muestra ni la descripción, ni los nombres de las variables que la forman. La base de datos es de 14 variables, de las cuales 6 son contínuas, 8 categóricas y una variable objetivo dicotómica, donde esta úlima se ditribuye en un 55,5 % en las que el cliente responde al crédito y un 44,5 % no lo hace (default).

La base de datos *German Credit* hace referencia a datos sobre incumplimiento en préstamos bancarios. Está compuesta por 1.000 observaciones, donde el número de variables es de 20, de las cuales 7 son contínuas, 13 categóricas y una variable objetivo dicotómica, donde esta última se distribuye en un 70,00 % en las que el cliente atiende a la devolución del préstamo y un 30,00 % no lo hace (default).

Los pasos efectuados con ambas bases de datos han sido los siguientes:

1. Se realiza una partición aleatoria de la muestra, el 70 % como datos de aprendizaje y el 30 % restante como datos de prueba o test.

$$n_A = 0,7 * n$$
$$n_T = 0,3 * n$$

Tal que $n = n_A + n_T$, donde n_A es el tamaño de la datos de aprendizaje, n_T es el tamaño de la datos de test y n el tamaño de la muestra total.

2. Se aplican los métodos empleados en el capítulo 4y 5 a ambas bases de datos de aprendizaje (*Australian Credit* y *German Credit*):

   - Análisis lineal discriminante (LDA).
   - Árboles de clasificación (CART).
   - Modelo lineal mixto, efectos fijos y aleatorios (LMM). Para su aplicación se considera como variable aleatoria en *German Credit* la variable categórica personalstatussex (es la variable que asigna sexo a cada cliente) y en *Australian Credit* la variable categórica V5.
   - Modelo lineal generalizado con nexo logit (GLMlogit).
   - Máquinas de vectores soporte (LSVM).

3. Se calcula el valor estimado de la variable objetivo $\widehat{\boldsymbol{y}}$ para cada uno de los métodos con los datos de aprendizaje.

4. Se calcula el error cuadrático medio (RMSE) con los datos de aprendizaje.

5. Con los parámetros obtenidos mediante los datos de aprendizaje, se predice la variable objetivo mediante las variables explicativas del conjunto de datos test. Se obtiene la predicción del modelo con los datos de prueba o test, m_T, y el modelo ajustado en cada caso.

6. Una vez realizada la predicción se procede a evaluar su desempeño y comparar sus resultados, es decir, se cuantifica la bondad de la predicción. Esto se hace de dos modos:

   a) Calculando la matriz de confusión para los datos de test, como se puede ver en la tabla 4.2

   b) Calculando la raíz cuadrada del error cuadrático medio de la predicción:

   $$RMSE_j = \sqrt{\frac{1}{n_T}\sum_{i=1}^{n_T}(\widehat{y}_i - y_i)^2}$$

   con $j \in \{GLMlogit, LMM, LDA, CART, LSVM\}$ donde $\widehat{y}_i$ es la predicción de la variable objetivo para el elemento i-ésimo de la muestra de test e y_i es el valor que toma la variable objetivo en la misma.

7. Se calcula:

   a) La tasa de acierto como el cociente entre la suma de los conteos de la diagonal principal de la matriz de confusión (elementos clasificados, o aciertos) y el total de elementos en la muestra de test (n_T).

Método	Tiempo	RMSE	Acierto	errorTI	errorTII
GLMlogit	0,056	0,4371	65,00 %	89	16
LMM	0,126	0,42945	62,34 %	94	19
LDA	0,052	0,52915	72,00 %	31	53
CART	0,068	0,53541	71,34 %	30	56
LSVM	0,391	0,47272	60,67 %	103	15

Tabla 5.10: Resultados numéricos *German Credit*

Método	Tiempo	RMSE	Acierto	errorTI	errorTII
GLMlogit	0,057	0,33463	85,92 %	20	9
LMM	0,077	0,33236	83,98 %	24	9
LDA	0,028	0,38792	84,95 %	24	7
CART	0,335	0,36868	86,41 %	14	14
LSVM	0,127	0,35673	81,07 %	6	33

Tabla 5.11: Resultados numéricos *Australian Credit*

b) El error tipo I como cociente entre falsos negativos y el total de elementos en la muestra de test (n_T).

c) El error tipo II como cociente entre falsos positivos y el total de elementos en la muestra de test (n_T).

8. Se mide el tiempo medio total empleado para cada una de las técnicas.

El objetivo es encontrar un método que minimice el error cuadrático medio, obtenga el máximo acierto y se ejecute en el menor tiempo posible.

5.3.2. Resultados numéricos

Para la obtención de los resultados numéricos se ha utilizado el mismo hardware, software y paquetes estadísticos empleados en el capítulo 4.

En las tablas 5.10 y 5.11 se muestran los resultados obtenidos para los indicadores calculados en los datos de *German Credit* y *Australian Credit* respectivamente. Los métodos que obtienen una mejor eficacia en cuanto a error cuadrático medio se refiere, son LMM y GLMlogit, coincidiendo claramente con lo obtenido en los capítulos anteriores.

Los métodos con una eficiencia computacional mejor son LDA, GLMlogit y LMM, siendo el más lento el LSVM.

En cuanto a la eficacia observando la tasa de acierto, se comprueba que los mejores métodos son CART, LDA y GLMlogit, para ambas bases de datos. Para el cálculo de la matriz de confusión en los métodos que en lugar de clasificar devuelven probabilidades de default, se ha

utilizado como punto de corte para otogar una clasificación en base a las probabilidad la mediana de estas. Pero tal y como se refleja en el capítulo 4, el punto de corte elegido en la matriz de confusión incluye un cierto grado de subjetividad, dependiendo del punto de corte elegido se obtiene una tasa de acierto u otra.

5.3.3. Problemática en la elección del punto de corte

Se aplican distintos puntos de corte en la matriz de confusión para obtener la tasa de acierto con los 5 métodos para las bases de datos *German Credit* y *Australian Credit*. Para realizar esta prueba se decide fijar distintas semillas aleatorias para comprobar si los resultados dependen de las muestras de entrenamiento aleatoriamente seleccionadas y como puntos de corte valores desde el 1 % hasta 99 % de probabilidad, en saltos de 0,1 % en 0,1 %. Se clasifica como éxito aquellos valores mayores o iguales a ese punto de corte y como fracaso los menores. De este modo se dispondrá de una matriz de confusión para cada uno de estos pasos. La elección del punto de corte óptimo se realiza en función de la mejor matriz de confusión, es decir, aquella que tiene una tasa de acierto mayor. Este problema de la elección del punto de corte ha sido muy controvertido dentro de los departamentos de riesgo y de negocio de las entidades de crédito. En muchas ocasiones se han elegido dos puntos de corte dividiendo así el panorama de decisión en tres zonas, la de denegación del crédito (zona de impagos), la aceptación del riesgo (zonas de pago) y una zona intermedia en la que la decisión pasaba por una segunda fase de algún departamento de análisis. Con el procedimiento que se propone, ya se está dando el mejor de los puntos de corte, el punto de corte óptimo, ya que maximiza la tasa de acierto para una clasificación.
En los gráficos 5.7 y 5.8 (*German* y *Australian Credit* respectivamente) se pueden ver las tasas de acierto para diferentes puntos de corte en los métodos propuestos. Se observa lo siguiente:

1. Los métodos que clasifican no se ven afectados por el punto de corte, tal y como se señala en el capítulo 4. Para la base de datos *German Credit*, la mayor tasa de acierto es de 72 % y se corresponde con el método LDA. Para el caso de la base de datos *Australian Credit*, la tasa de acierto mayor es 86,4 % y se corresponde con el método CART.

2. En cuanto a los métodos que no clasifican, es decir, que obtienen probabilidades se observa lo siguiente:

   - Para la base de datos *German Credit*, la mayor tasa de acierto es de 73 % y se corresponde con los métodos GLMlogit y LMM y puntos de corte desde 51 % a 60 %.
   - Para la base de datos *Australian Credit* la mayor tasa de acierto es de 85,9 % y se corresponde con los métodos GLMlogit y LMM pero con distintos puntos de corte, en el caso de GLMlogit sucede desde 31 % a 40 % y en el caso de LMM en el punto de corte 61 % a 69 %.

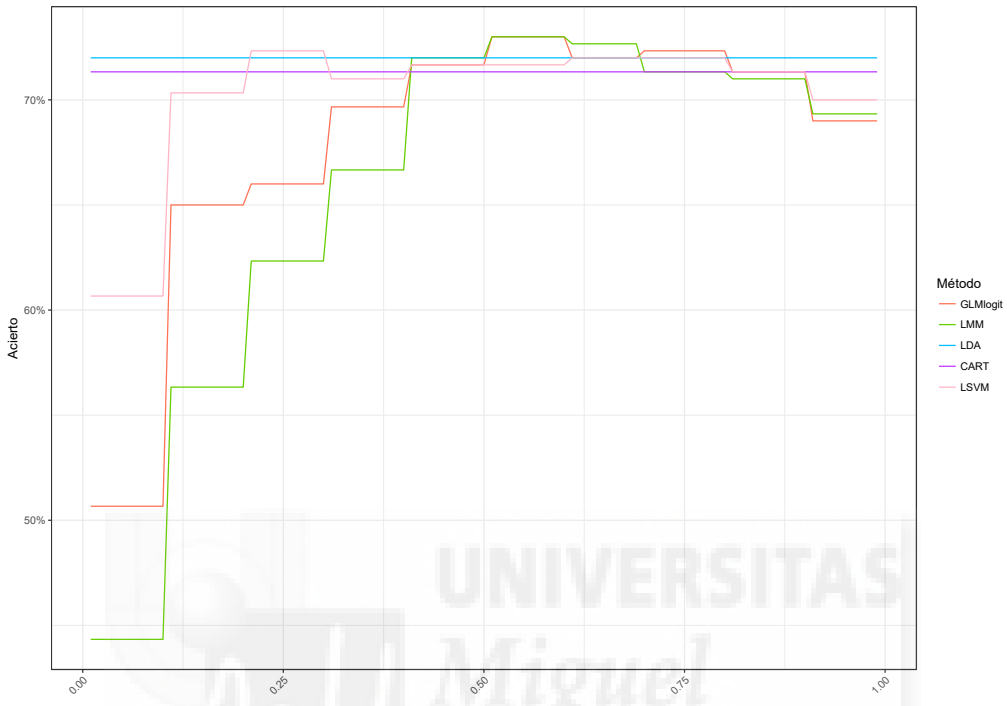

Gráfico 5.7: *German Credit* puntos de corte

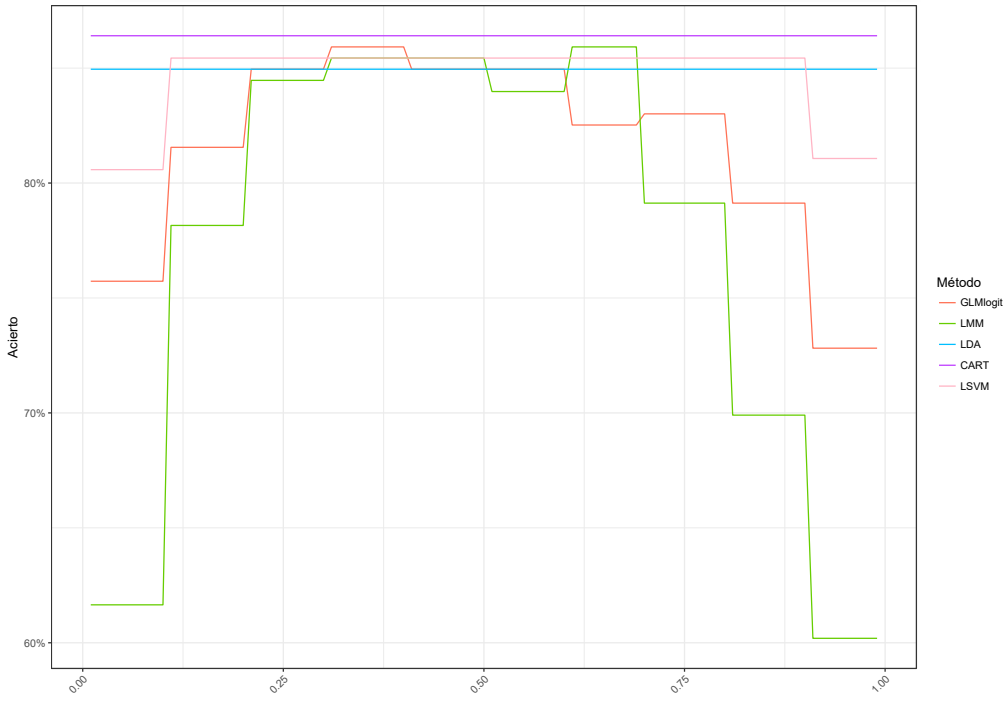

Gráfico 5.8: *Australian Credit* puntos de corte

Como se puede observar, la tasa de acierto no es la mejor manera de medir la eficacia de un método que no clasifica, ya que dependiendo del punto de corte que se establezca se obtiene una tasa de acierto u otra, al igual que cambiará el punto de corte óptimo en función del método que se utilice. Al contrario que el error cuadrático medio, que siempre es el mismo, independientemente del punto de corte elegido.

5.3.4. Conclusiones

Se observa que las bases de datos utilizadas *Australian Credit* y *German Credit*, aunque procedan de entidades financieras, no se corresponden con el número de observaciones reales que existen en las bases de datos de cualquier entidad financiera (son bases de datos muy pequeñas). En consecuencia las conclusiones que se pueden derivar son aproximadas y tienen esa limitación. En todo caso, sí que se comprueba que las conclusiones obtenidas en el capítulo 4 son coherentes con los resultados obtenidos con las dos bases de datos reales que además se corresponden con la mejor eficiencia computacional. Si nos fijamos en el número de aciertos CART, GLMlogit, LDA son los mejores métodos de predicción, pero si nos fijamos en el error cuadrático medio los mejores métodos de predicción son LMM y GLMlogit.

UNIVERSITAS
Miguel
Hernández

6 Selección de variables

6.1. Introducción

Las entidades financieras poseen una cantidad extraordinaria de datos y resulta muy difícil seleccionar qué datos son necesarios o aportan información útil a los modelos de predicción y qué datos son irrelevantes. El manejo y procesamiento de gran cantidad de datos crea un problema de capacidad de procesamiento en los equipos y ralentiza su funcionamiento. Las entidades financieras necesitan dar una respuesta rápida a sus posibles prestatarios, por tanto, los modelos de predicción deben cumplir dos requisitos, eficiencia computacional y eficacia predictiva. En este capítulo se estudia el problema del volumen de datos masivos (BigData) y un método para seleccionar las variables óptimas a formar parte del modelo con la finalidad de reducir la dimensión del dataset, manteniendo la eficacia del modelo con el objetivo de aumentar su eficiencia computacional.

Una vez obtenidos los resultados de la muestra semi-real, se plantea un nuevo objetivo de análisis. Se va a determinar la importancia de las variables seleccionadas y si la aportación de las mismas es significativa para el modelo. En primer lugar se clasifican las variables por orden de importancia y se aplica el método LMM para predecir (porque es el más eficaz y eficiente junto con GLMlogit) calculando el tiempo y el error cuadrático medio, introduciendo las variables por orden de su contribución al modelo. El problema que surge cuando se dispone de mucha información es la necesidad de extraer sólo la información relevante de forma automática y reducir el tiempo de computación manteniendo la eficacia de los métodos. La solución viene dada por métodos analíticos de Minería de Datos (o Data Mining). Por tanto se realiza una selección de variables y se comprueba si mejora la predicción y si esa mejora se corresponde con una mayor eficacia computacional. Con ello se espera conseguir pronósticos más rápidos

y eliminar la información que no aporte ninguna ganancia al modelo. Se trata de aplicar un algoritmo de Minería de Datos que genere y reduzca sistemas de reglas de clasificación y consiga predicciones precisas, pero con un coste temporal y de consumo de memoria sustancialmente menor.

La selección automática de características es un conjunto de técnicas analíticas que generan las combinaciones de variables que tienen una mayor incidencia en el comportamiento de una determinada variable objetivo. La selección de variables relevantes se puede realizar aplicando diferentes metodologías como árboles de decisión, sistemas de reglas y técnicas de análisis de componentes principales con un enfoque incremental donde los algoritmos empleados buscan ir mejorando una determinada métrica de significancia (Martinez-Murcia et al., 2014).

En esta investigación se va a emplear:

1. El método Gain Ratio (relación de ganancia), implementado en Weka y en R. Es un método para selección automática de características (Automatic Feature Selection) que evalúa cada atributo del dataset, midiendo su razón de ganancia con respecto a la variable de clase. Los atributos a evaluar son de tipo discreto (no numérico). La variable de clase por supuesto debe ser también de tipo discreto (concretamente binaria, en este caso). El resultado es un ranking de los atributos, ordenados de mayor a menor influencia sobre la variable de clase.

   El método de selección de variables es parte del proceso de elaboración del sistema de árboles de decisión C4.5 (Quinlan, 1993), en este caso se ha usado el algoritmo implementado en WEKA, J48, (Hall et al., 2009). Este algoritmo busca aquellas ramas que se encuentran más puras, es decir, existe un amplio número de individuos en el conjunto training que responde a esa categoría. El método viene definido por comprobar la relación entre:

   $$GainRatio = \frac{H(Class) + H(Attribute) - H(Class, Attribute)}{H(Attribute)},$$

   donde las H representan la entropía, la incertidumbre de los datos. El autor Shannon (1951) definió la entropía H de una variable discreta X con una función de probabilidad P(X)

   $$H(X) = E[I(X)] = E[-ln(P(X))]$$

   donde E es la esperanza matemática del operador e I es la función de información.

2. Método de Componentes principales (Dunteman, 1989), desarrollado por Pearson (1901) en el siglo XIX. Para estudiar las relaciones que se presentan entre p variables correlacionadas (que miden información común) se puede transformar el conjunto original de variables en otro conjunto de nuevas variables incorreladas entre sí (que no existe redundancia en la información) llamado conjunto de componentes principales.

   Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra,

por tanto, las primeras explican por sí solas la mayor parte de la variabilidad del conjunto inicial de datos. A una mayor variabilidad de los datos se considera que existe mayor información, lo cual está relacionado con el concepto de entropía, que usa el método Gain Ratio.

Una vez implementado el método Gain Ratio, se aplica un conjunto de reglas de clasificación a través del algoritmo CREA-RBS (Rabasa Dolado, 2009). Se realiza para aquellos atributos sobre los que existe una mayor influencia sobre la variable de clase.

El algoritmo CREA ("Classification Rules Extraction Algorithm") se utiliza para generación, reducción y ordenación de reglas de clasificación, que combina en un único proceso del método ID3 de generación de reglas de clasificación y el método RBS de reducción y ordenación de reglas de clasificación (Rodríguez Sala, 2014; Almiñana et al., 2014).

El algoritmo CREA-RBS calcula los soportes y confianzas de los hechos (tuplas) que refleja la base de datos a partir de un conteo de sus frecuencias de repetición generando reglas de clasificación, cada una de las cuales debe estar caracterizada por su soporte y su confianza. El soporte de una regla se entiende como la probabilidad que se dé el antecedente de la misma regla en la base de datos original. Confianza es la probabilidad que se dé el consecuente dentro de los ítems de la base de datos original con el mismo antecedente. Se parte de un conjunto de datos con "N" ítems, caracterizados por una serie de "C" atributos o variables, donde cada ítem pertenece a una determinada "Clase" o respuesta. A partir de dicho conjunto de datos se genera un conjunto de reglas de clasificación, cada una de las cuales debe estar caracterizada por su soporte y su confianza. Las reglas de clasificación son: $V_1, V_2, \ldots, V_n \rightarrow C$, siendo A_i el valor del atributo o variable i, y C un determinado valor de la variable de clase. El significado de la regla es que si ocurre que el atributo 1 toma el valor V_1 junto con el atributo 2 el valor V_2, y así sucesivamente, entonces la variable de clase toma el valor C.

Posteriormente, el algoritmo segmenta las reglas generadas en función de la importancia de las mismas, que es medida a partir de los umbrales aceptables de soporte y confianza en cada caso. Una vez obtenidas las reglas, junto con los valores numéricos de soporte y confianza asociados, cada regla se puede representar en un plano utilizando dichos valores como coordenadas. Por tanto, se puede obtener distintas regiones en el plano delimitadas por los ejes de confianza y de soporte, respectivamente. El autor de las reglas de clasificación Rabasa Dolado (2009) denomina a estas regiones "regiones de significancia" y la interpretación de las regiones es la siguiente:

1. Región 1. Reglas de descarte. Estas reglas tienen un alto soporte y baja confianza, son por tanto reglas que pueden ser empleadas para descartar ciertas situaciones debido a la baja probabilidad de ocurrencia del consecuente. Es decir, en los datos de entrenamiento, el consecuente inferido por estas reglas se da muy pocas veces (para el antecedente considerado) por lo que se puede asegurar con alta probabilidad de acierto que dicho consecuente no va a producirse.

2. Región 2. Reglas directas. En la región 2 las reglas tienen tanto un alto soporte como una alta confianza, es decir, son confiables y de aplicación directa. Por tanto son reglas que

proporcionan información fiable acerca del comportamiento de los datos de entrenamiento con los que se ha generado el modelo.

3. Región 3. Reglas a observar. Las reglas de la región 3 son aquellas que tienen un soporte muy bajo, se debe entender que cualquier conclusión que pudiera extraerse de ellas no se puede considerar segura, ya que no hay suficientes ejemplos en el conjunto de datos inicial que cumplan el antecedente, por lo que no se pueden extraer conclusiones estadísticamente fiables.

4. Región 0. Reglas descartables. Por último, las reglas de la región 0 tienen niveles de soporte y confianza intermedios. Son éstas las que el algoritmo RBS elimina para reducir el conjunto de reglas proporcionado como entrada del algoritmo, dando lugar al conjunto reducido de reglas.

El método por último define y calcula "ACI", que es el Índice de correlación de atributos, que permite ponderar de manera global cómo están de correlacionadas las variables del antecedente del sistema de reglas con la variable consecuente, por tanto, "ACI" es una métrica aplicable al conjunto total de reglas reducidas.

6.2. Selección de variables Gain Ratio y algoritmo CREA-RBS

Para realizar el procedimiento y poder interpretar mejor las reglas, en primer lugar resulta necesario realizar un preprocesado de la muestra. Se agregará a la misma, una serie de prefijos a las variables categóricas para el correcto funcionamiento del algoritmo de selección, así como para facilitar la interpretación de las reglas resultantes. Al mismo tiempo se procede a discretizar todas las variables, al aplicarse un proceso de clasificación, es conveniente que las variables sean discretas y no numéricas, al igual que la variable objetivo es necesario que sea discreta. En la tabla 6.1, se pueden observar los cambios que se han realizado en las variables numéricas de la muestra semi-real.

Una vez procesada la muestra se aplica el algoritmo Gain Ratio. Para ello, se utiliza el software R, y el software Weka, y se obtiene la contribución de las variables en el modelo. Los resultados se muestran en la tabla 6.2.

Variables Originales	**Clasificadas en**	**Pasa a tener valor**
portitularviviendahabitual	Iguales a 0 %	ptvh-0
	Iguales al 50 %	ptvh-50
	Iguales al 100 %	ptvh-100
	Entre 0 % y 50 %	ptvh-<50
	Entre 50 % y 100 %	ptvh->50
porconyugetitularviviendahabitual	Iguales a 0 %	pctvh-0
	Iguales al 50 %	pctvh-50
	Iguales al 100 %	pctvh-100
	Entre 0 % y 50 %	pctvh-<50
	Entre 50 % y 100 %	pctvh->50
pormedtitularidad	Iguales a 0 %	pormedtit-0
	Iguales al 50 %	pormedtit-50
	Iguales al 100 %	pormedtit-100
	Entre 0 % y 50 %	pormedtit-<50
	Entre 50 % y 100 %	pormedtit->50
dedviviendatotal	No dispone	dvt-0
	Dispone de Dedución	dvt-Dist0
rentatotalinmuebles	No dispone	rti-0
	Dispone de Renta	rti-Dist0
rentafamiliar	Negativa	renfam-<0
	Igual a 0 €	renfam-0
	Entre 0 y 3 701 € (1 Cuartil)	renfam-1stQuartile
	Entre 3 701 € y 10 730 €	renfam-<Media
	Mayor> 10 730 €	renfam-+Media
ImportePrestamo	Primer Cuartil	IP-1stQ
	Segundo Cuartil	IP-2stQ
	Tercer Cuartil	IP-3stQ
	Cuarto Cuartil	IP-4stQ

Tabla 6.1: Procesamiento

Variables	Importancia
RentaCorrectora	0,3385742
rentafamiliar	0,1834061
dedviviendatotal	0,0866876
titularviviendahabitual	0,0318049
inmuebletitular6	0,0276966
inmuebletitular5	0,0261169
inmuebletitular4	0,0232536
inmuebletitular3	0,0215097
edad	0,0209065
portitularviviendahabitual	0,0194543
inmuebletitular2	0,0170477
rentatotalinmuebles	0,0140341
inmuebletitular1	0,0135106
estcv	0,0133969
BIECorrector	0,0127670
ninmuebles	0,0124063
porconyugetitularviviendahabitual	0,0120175
pormedtitularidad	0,0108874
numfam	0,0096911
actividad1	0,0083553
prov	0,0075127
actividad2	0,0022774
ImportePrestamo	0,0008614
Year	0,0008512

Tabla 6.2: Selección de variables

Gráficamente se observa cómo se reduce el Gain Ratio a partir de la introducción de la tercera variable que corresponde a la deducción de la vivienda habitual, gráfico 6.1.

Los resultados obtenidos están dentro de la previsión realizada con la definición del modelo. La variable que más influye es la renta correctora. La renta correctora es una variable que se introdujo en el modelo para poder discriminar aquellas rentas menores al salario mínimo interprofesional, ya que con esa mínima renta resulta imposible poder atender una cuota hipotecaria. Por tanto es la variable que más influencia tiene en el modelo. La segunda variable, por orden de importancia en el modelo, es la renta familiar, la renta del individuo influye en el modelo ya que cuanta más renta tenga el individuo, mejor puede atender sus pagos y viceversa. La tercera variable en importancia es la deducción de la vivienda total que es una variable económica que refleja un ingreso al individuo a través de su renta; por tanto, influye en el modelo al ser una parte más de la renta. El resto de variables influyen comparativamente muy poco en el modelo. Cabe destacar que las variables que menos influyen son el número de miembros de la unidad familiar, la profesión principal, provincia, la profesión secundaria, el importe del préstamo soli-

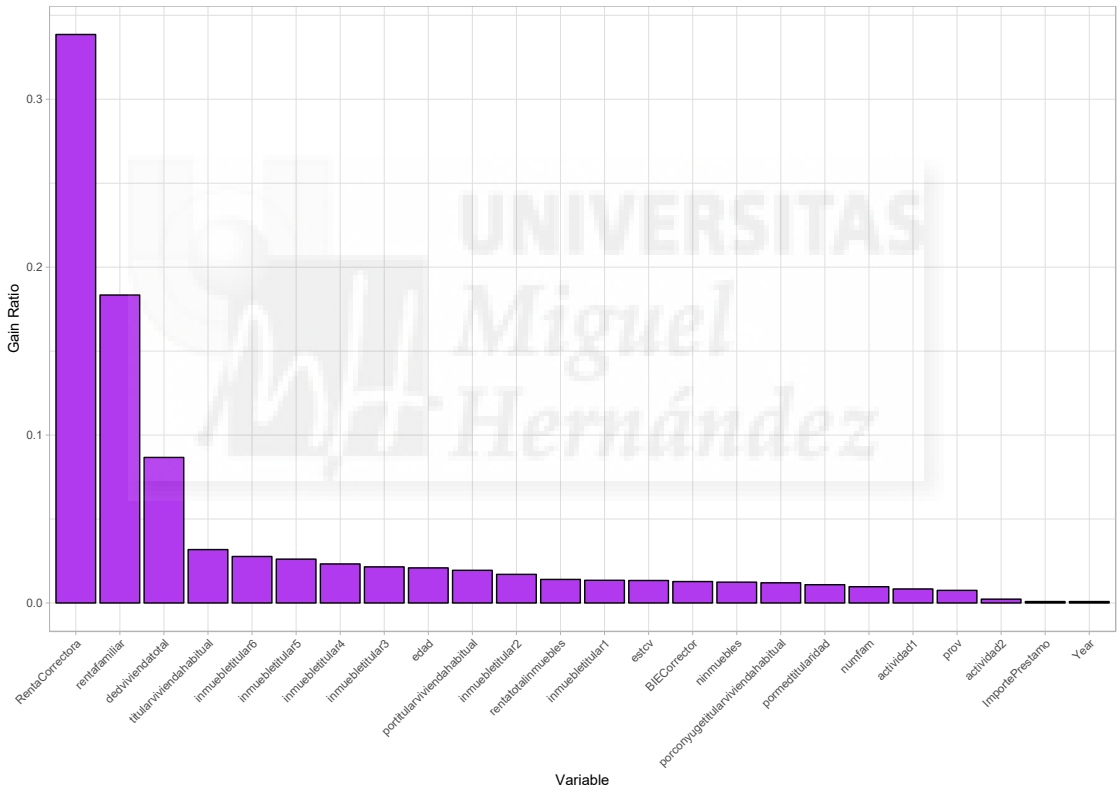

Gráfico 6.1: Selección de variables proporcionadas por GR

citado y el año. En cuanto a estos resultados tan sólo resaltar que el importe del préstamo no es una variable que influye en el modelo, pero teniendo en cuenta que esa variable se ha generado sintéticamente aplicando una media y una varianza de los importes medios de los precios de los inmuebles en los distintos años del análisis, sí que tiene sentido el resultado, porque tiene un peso igual en todos los casos. En cuanto a la profesión principal o secundaria del individuo y el número de miembros de la unidad familiar, de forma indirecta están reflejadas en la renta familiar. Tienen menos importancia las variables sociales que las económicas. El año de la muestra es la variable menos importante del modelo, indica que no influye el año en dejar de atender un préstamo o que dicho comportamiento es relativamente constante en el período de estudio.

Una vez seleccionadas las variables por importancia se aplica el algoritmo CREA-RBS. Los resultados obtenidos se muestran en la siguiente tabla 6.3:

Soporte	Confianza	Importancia	RentaCorrectora	rentafamiliar	dedviviendatotal	Y
29,86	100	0,299	RC-0	renfam-+Media	dvt-0	**Y-0**
22.431	99,15	0,222	RC-1	renfam-1stQuartile	dvt-0	**Y-1**
12.816	100	0,128	RC-0	cualquier renfam	dvt-Dist0	**Y-0**
9,98	99.632	0,099	RC-0	renfam-<Media	dvt-0	**Y-0**
22.431	0,85	-0,998	RC-1	renfam-1stQuartile	dvt-0	**Y-0**
9,98	0,368	-1,000	RC-0	renfam-<Media	dvt-0	**Y-1**
0,1	99.025	-1,999	RC-1	renfam- < 0	dvt-0	**Y-1**
0,032	82.895	-2	RC-1	renfam-1stQuartile	dvt-Dist0	**Y-1**
0,02	95.833	-2	RC-1	renfam-0	dvt-Dist0	**Y-1**
0,032	17.105	-2	RC-1	renfam-1stQuartile	dvt-Dist0	**Y-0**
0,003	100	-2	RC-0	renfam-1stQuartile	dvt-0	**Y-0**
0,002	100	-2	RC-0	renfam-0	dvt-0	**Y-0**
0,1	0,975	-2	RC-1	renfam- < 0	dvt-0	**Y-0**
0,02	4.167	-2	RC-1	renfam-0	dvt-Dist0	**Y-0**
0,001	100	-2	RC-1	renfam- < 0	dvt-Dist0	**Y-1**
0,001	100	-2	RC-0	renfam- < 0	dvt-0	**Y-0**

Tabla 6.3: Algoritmo CREA-RBS: Reglas resultantes

Observando los resultados se concluye lo siguiente:

1. Los cuatro primeros resultados se encuentran en la región 2 (aquellos que tienen el color azul en la tabla 6.3). Tienen tanto un alto soporte como una alta confianza. Se trata de reglas que proporcionan información fiable acerca del comportamiento de los datos de entrenamiento con los que se ha generado el modelo, por tanto son las que mejor sirven para modelar el comportamiento de la variable objetivo. La primera regla, por ejemplo, se puede interpretar así: "si la renta correctora es 0, la renta familiar se sitúa por encima de la media y la deducción de la vivienda habitual es 0, entonces la variable respuesta será 0, es decir, se atenderá el préstamo, con un soporte del 29,86 % y una confianza del 100 %", es decir que el antecedente (renta correctora es 0, la renta familiar se sitúa en la media y la deducción de la vivienda habitual es 0) se cumple el 29,86 % de las veces y que el 100 % de las veces que se cumple el antecedente, la variable de clase adquiere el valor 0.

2. Los dos siguientes resultados se encuentran en la región 1 (aquellos que tienen el color gris en la tabla 6.3), tienen soportes suficientes pero confianzas muy bajas, por tanto son reglas que pueden ser empleadas para descartar ciertas situaciones debido a la baja probabilidad de ocurrencia del consecuente dado el antecedente considerado. Se puede asegurar que con

alta probabilidad de acierto el consecuente no va a producirse. Esta regla, por ejemplo, se puede interpretar como: si la renta correctora es 1, la renta familiar se sitúa en el primer cuartil y la deducción de la vivienda habitual es 0, entonces la variable respuesta será 0, es decir, se atenderá el préstamo, con un soporte del 22,431 % y una confianza del 0,85 %, es decir que el antecedente (renta correctora es 1, la renta familiar se sitúa en el primer cuartil y la deducción de la vivienda habitual es 0) se cumple el 22,431 % de las veces y que el 0,85 % de las veces que se cumple el antecedente, la variable de clase adquiere el valor 0. Este tipo de reglas sirven para descartar ciertos patrones pues se conoce que son muy poco probables en el dataset.

3. El resto de reglas se sitúan en la región 3 (aquellos que tienen el color lila en la tabla 6.3). Son aquellas reglas que tienen un soporte muy bajo, aparecen muy pocas veces en la muestra. Se debe entender que cualquier conclusión que pudiera extraerse de ellas no debe considerarse segura ya que no hay suficientes ejemplos en el conjunto de datos inicial que cumplan el antecedente, y por lo tanto deben considerarse como casos anómalos a partir de los que no se puede inferir ningún valor de la variable de clase. Esta regla, por ejemplo, se puede interpretar como: si la renta correctora es 1, la renta familiar menor que 0 y la deducción de la vivienda habitual es 0, entonces la variable respuesta será 1, es decir, no se atenderá el préstamo, con un soporte del 0,1 % y una confianza del 99,025 %, es decir que el antecedente (renta correctora es 1, la renta familiar menor que 0 y la deducción de la vivienda habitual es 0) se cumple el 0,1 % de las veces. Sin embargo, el 99,025 % de las veces que se cumple el antecedente, la variable de clase adquiere el valor 1.

El sistema de reglas RBS es un sistema de apoyo a la decisión, nunca un sistema experto. Con las reglas obtenidas, la entidad financiera puede clasificar a través de 3 variables: renta correctora, renta familiar y deducción de la vivienda habitual con un grado de confianza y soporte determinados si el prestatario paga o no el préstamo. Si el prestatario se encuentra en las regiones 1, 2 y 3 la entidad financiera puede adoptar una decisión y el error cuadrático medio sería 0, ya que conoce el consecuente (variable objetivo) y con qué probabilidad sucede, dados los antecedentes (variables independientes). Si el prestatario se encuentra en la región 0, es decir, en la zona de reglas descartables, tienen niveles de soporte y confianza intermedios, la entidad financiera no tiene una regla de decisión por tanto puede optar por:

- No tomar ninguna decisión sobre la concesión o no del préstamo realizando otro estudio.
- Rechazar al prestatario. En este caso, si rechazamos al prestatario y se cumple que el consecuente es 1 (no paga), el error cuadrático medio es $0,3032674$.
- Rechazar al prestatario. En este caso, si rechazamos al prestatario y se cumple que el consecuente es 0 (paga), el error cuadrático medio es $0,3944287$.
- Rechaza al prestatario. En este caso, si rechazamos al prestatario y aleatoriamente unas

	Completa		Discretizada	
Variables	**ECM**	**Telapsed**	**ECM**	**Telapsed**
1	0,371335	44,028	0,478317	65,276
2	0,371197	56,134	0,358973	115,702
3	0,371162	59,35	0,358872	117,494
4	0,370285	65,679	0,358101	136,996
5	0,370273	79,863	0,35809	176,688
6	0,37027	94,59	0,358088	218,548
7	0,370257	108,018	0,358008	236,951
8	0,370246	122,146	0,358079	266,101
9	0,370241	121,695	0,358052	330,871
10	0,366408	150,58	0,352597	331,666
11	0,366397	181,9	0,35255	396,274
12	0,366376	170,772	0,352471	405,76
13	0,366359	213,256	0,352407	454,256
14	0,366351	200,592	0,352397	463,681
15	0,366295	187,149	0,352369	517,45
16	0,366295	194,708	0,352369	578,115
17	0,366288	211,396	0,352365	672,949
18	0,366248	216,241	0,352316	733,61
19	0,366238	232,929	0,352277	979,271
20	0,36612	247,128	0,35215	1 012,181
21	0,366022	269,853	0,352117	1 132,341
22	0,366022	249,862	0,352117	1 155,923
23	0,366022	368,844	0,352112	1 184,734

Tabla 6.4: Resultados LMM Gain ratio

veces resulta el consecuente 1 (no paga) y otras es 0 (paga), el error cuadrático medio es $0,3203057$.

Posteriormente se procede a ajustar la muestra semi-real con el método LMM. Se obtiene el tiempo de ejecución y el error cuadrático medio introduciendo las variables por orden de su contribución al modelo. Se calcula tanto con la muestra sin el procesamiento o discretización (muestra sin discretizar, con variables discretas y numéricas) como con la muestra procesada o discretizada (muestra discretizada o procesada, con todas las variables discretas). Los resultados se muestran en la tabla 6.4.

En primer lugar se puede observar en el gráfico 6.2 que el error cuadrático medio es similar en las dos muestras, siendo un poco menor para la muestra discretizada (1,2 %), excepto para la variable 1. Dado que tienen un comportamiento similar las dos muestras y el error cuadrático menor se produce con la muestra discretizada, se van a analizar los datos con la muestra discretizada.

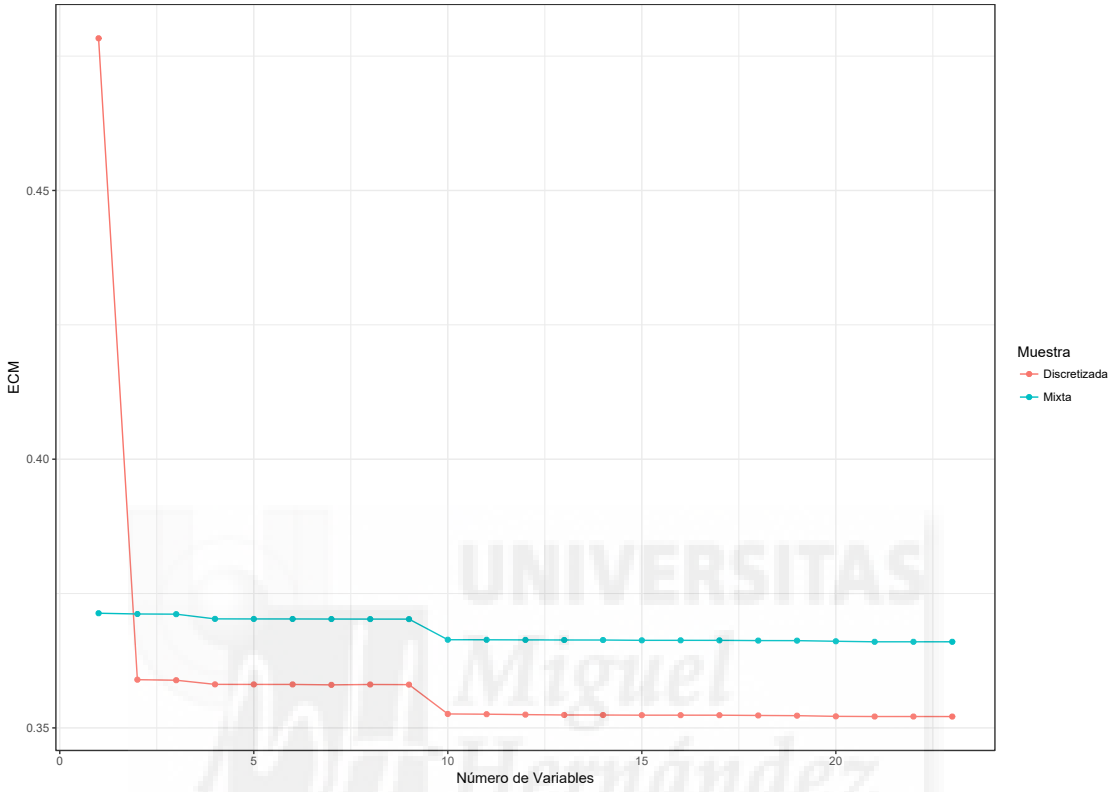

Gráfico 6.2: LMM ECM Gain ratio

En segundo lugar se observa que al añadir la segunda variable, renta familiar, se produce una disminución mayor del error cuadrático medio, se reduce el 33,25 %, pasa de $0,4783167$ a $0,358973$.

A partir de la segunda variable hasta la incorporación de la variable 10, que se corresponde con el porcentaje titularidad de la vivienda habitual, el error cuadrático medio no experimenta apenas ningún cambio, es decir, añadir las variables al modelo no supone una disminución significativa del error. Una vez incorporada la variable porcentaje titularidad de la vivienda habitual el error cuadrático medio se mantiene entre $0,352112$ y $0,352597$, es decir pasar de 10 variables a 24 supone una disminución porcentual del 0,14 %. Por tanto, a partir de esta variable apenas mejora el error introduciendo más variables, es decir, se debe de valorar si el disminuir un porcentaje muy pequeño el error cuadrático medio compensa el aumento del tiempo de ejecución.

En cuanto al tiempo de ejecución de los procesos, se puede observar en el gráfico 6.3, que para las dos muestras sigue la misma tendencia, siendo menor los tiempos de la muestra sin discretizar. Tiene sentido el resultado ya que en la muestra discretizada todas las variables son discretas (dummy), y tarda más la ejecución de cualquier proceso con estas variables. Sin embargo en la muestra mixta, existen variables discretas y numéricas.

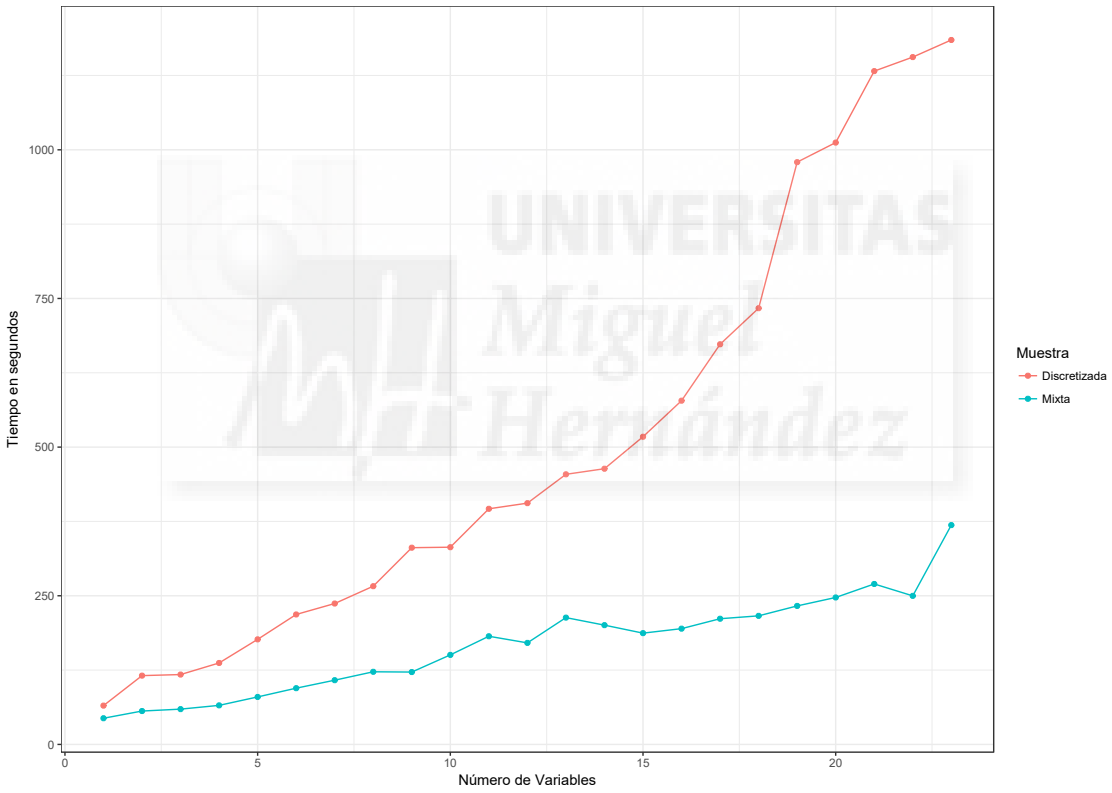

Gráfico 6.3: LMM tiempos Gain ratio

El primer incremento de tiempo importante se produce al introducir la variable 11 (tanto para la muestra discretizada como para la no discretizada), pasando en la muestra discretizada de 331, 67 segundos a 396, 27, el aumento del tiempo es de 16,30 % frente a una disminución del error cuadrático medio del 0,013 %.

6.3. Análisis de componentes de principales

Para aplicar el análisis de componentes principales se utiliza el software de R, con el paquete FactoMineR de Lê et al. (2008).

Se procede a aplicar el análisis de componentes principales a la muestra semi-real con el método LMM. Se obtiene el tiempo de ejecución, el error cuadrático medio y el porcentaje acumulado de la varianza explicada para cada una de las componentes principales. Se calcula tanto para la muestra sin el procesamiento o discretización (muestra sin discretizar, con variables discretas y numéricas) como con la muestra procesada o discretizada (muestra discretizada o procesada, con todas las variables discretas). Los resultados se muestran en la tabla 6.5.

	Mixta				Discretizada		
	ECM	Tiempo	Var. Exp.		ECM	Tiempo	Var. Exp.
1	0,4644195	83,785	10,914	1	0,4609405	72,973	7,0503
2	0,4615357	83,924	17,288	2	0,4572005	81,889	11,4247
3	0,4303272	103,065	22,359	3	0,4466995	89,743	15,0462
4	0,4302277	95,297	27,090	4	0,4466314	91,162	18,0347
5	0,4191166	102,728	31,210	5	0,440626	95,544	20,9958
6	0,4178691	122,12	35,033	6	0,4127611	105,922	23,9230
7	0,4169157	124,812	38,621	7	0,4074352	118,539	26,4710
8	0,4083538	128,207	41,871	8	0,404877	122,396	28,8869
9	0,3965872	134,469	44,863	9	0,394889	134,488	31,0143
10	0,396318	146,428	47,356	10	0,3874884	131,685	32,9923
11	0,3958834	154,065	49,697	11	0,387199	141,693	34,9424
12	0,3955392	165,027	51,964	12	0,3826237	152,65	36,7829
13	0,3952439	162,518	54,209	13	0,380284	157,568	38,5853
14	0,3940331	165,946	56,412	14	0,3802684	170,497	40,3384
15	0,3936744	171,979	58,531	15	0,379884	180,03	41,9934
16	0,3936381	192,874	60,536	16	0,3781402	191,393	43,5949
17	0,393599	189,371	62,530	17	0,3781158	201,737	45,1480
18	0,3929238	195,241	64,510	18	0,3700562	213,088	46,6526
19	0,3927772	211,239	66,422	19	0,3697161	227,889	48,0824
20	0,3887169	206,92	68,274	20	0,3694512	223,858	49,4844
21	0,3884838	230,603	70,064	21	0,369429	247,872	50,8825
22	0,388404	226,006	71,826	22	0,3694227	254,384	52,2636

Continúa en la página siguiente...

	Mixta				Discretizada		
	ECM	Tiempo	Var. Exp.		ECM	Tiempo	Var. Exp.
23	0,3883726	240,319	73,568	23	0,3693353	267,598	53,6252
24	0,3879011	251,491	75,284	24	0,3679057	270,045	54,9509
25	0,3852205	256,155	76,941	25	0,3678491	284,078	56,2535
26	0,3852138	269,187	78,546	26	0,3676172	295,875	57,5371
27	0,3852061	273,805	80,074	27	0,3673457	313,085	58,7940
28	0,3851746	289,252	81,476	28	0,3672062	314,227	60,0211
29	0,3849441	290,359	82,840	29	0,3668745	332,041	61,2375
30	0,384928	302,017	84,159	30	0,3665283	339,078	62,4292
31	0,3843119	321,379	85,378	31	0,3659819	359,485	63,5971
32	0,3838835	328,307	86,559	32	0,3657284	363,137	64,7540
33	0,3815298	335,199	87,732	33	0,3656851	372,299	65,8927
34	0,380095	347,603	88,877	34	0,3655914	382,842	67,0309
35	0,380092	381,107	89,991	35	0,3655665	391,829	68,1688
36	0,3776317	372,739	91,040	36	0,365565	426,17	69,3060
37	0,3776095	395,318	92,050	37	0,365565	433,902	70,4427
38	0,3776062	394,029	92,977	38	0,365561	436,717	71,5791
39	0,377606	408,693	93,891	39	0,3655606	462,597	72,7154
40	0,3772108	446,383	94,790	40	0,3655523	460,964	73,8513
41	0,3770424	461,242	95,667	41	0,3654093	471,377	74,9841
42	0,3770246	476,687	96,424	42	0,3653934	497,915	76,1132
43	0,3770232	503,702	97,061	43	0,3653149	498,617	77,2308
44	0,3770231	498,247	97,621	44	0,3653024	523,22	78,3407
45	0,3770226	532,53	98,079	45	0,3652876	531,15	79,4393
46	0,3770179	532,649	98,531	46	0,3648456	541,979	80,5274
47	0,3663901	588,42	98,889	47	0,3648366	559,756	81,5914
48	0,3663843	566,541	99,181	48	0,3648205	586,238	82,6293
49	0,366381	601,711	99,400	49	0,364366	602,303	83,6579
50	0,366381	608,598	99,600	50	0,3640443	611,339	84,6403
51	0,3663689	624,528	99,734	51	0,3639953	629,544	85,5948
52	0,3660588	642,819	99,852	52	0,3638785	640,448	86,5299
53	0,366022	685,482	99,939	53	0,3638709	644,176	87,4098
54	0,3660211	709,945	100,000	54	0,363853	684,411	88,2812
55	0,3660211	727,872	100,000	55	0,3638285	697,846	89,1403
				56	0,3638072	715,382	89,9791
				57	0,3638069	735,007	90,8029
				58	0,3638058	741,217	91,5405
				59	0,3637967	765,572	92,2624
				60	0,3637226	771,486	92,9689
				61	0,3625888	809,539	93,6038
				62	0,3625034	826,625	94,2349

Continúa en la página siguiente...

	Discretizada		
	ECM	Tiempo	Var. Exp.
63	0,3622611	839,239	94,8503
64	0,3622554	864,327	95,4218
65	0,3622534	889,054	95,9671
66	0,3617001	924,303	96,4866
67	0,3616497	950,389	96,9738
68	0,3616422	937,37	97,4435
69	0,3614439	954,831	97,9052
70	0,3614438	972,267	98,2987
71	0,3614428	981,218	98,6462
72	0,3537258	1 020,61	98,8969
73	0,3536997	1 047,392	99,1246
74	0,3536091	1 141,262	99,3423
75	0,3536043	1 120,246	99,4973
76	0,3536029	1 134,057	99,6347
77	0,3536028	1 129,348	99,7462
78	0,3535842	1 206,386	99,8572
79	0,3532854	1 206,871	99,9307
80	0,3521117	1 241,379	99,9678
81	0,3521117	1 263,225	99,9949
82	0,3521116	1 285,347	99,9998
83	0,3521116	1 293,627	99,9999
84	0,3521116	1 277,015	99,9999
85	0,3521116	1 319,148	100,0000
86	0,3521116	1 348,282	100,0000
87	0,3521113	1 399,301	100,0000
88	0,3521113	1 356,532	100,0000

Tabla 6.5: Resultados LMM PCA

Se observa en el gráfico 6.4 que el error cuadrático medio de la muestra sin discretizar es mayor excepto de la tercera a la quinta componente principal. Es decir la muestra discretizada tiene un comportamiento mejor respecto al error cuadrático medio. La pendiente de la gráfica para las dos muestras es mayor hasta llegar a la componente 10, después se suaviza, sobre todo en la muestra sin discretizar que coincide con el 47,36 % de varianza acumulada. En la muestra discretizada se suaviza la pendiente, aunque sigue descendiendo hasta la componente 18, que coincide con una varianza explicada de 46,65 %. A partir de los puntos comentados, el descenso del error cuadrático medio es menor y la varianza explicada aumenta en menor proporción.

Para la muestra sin discretizar, a partir de la componente 47, se observa lo siguiente:

1. La disminución del error cuadrático medio es muy pequeña 0,10 %.
2. El aumento del tiempo de ejecución es muy grande 19,16 %.

3. La varianza explicada es muy alta, del 98,89 %.

Por tanto, pasar de 47 a 55 componentes supone una ganancia mínima en cuanto a disminución del error cuadrático medio y aumento de la varianza explicada, frente a un incremnto mayor del tiempo de ejecución. El objetivo es minimizar la dimensión de la base de datos perdiendo una mínima información y conseguir reducir los tiempos de ejecución, por tanto, se puede indicar que en la componente 47 se ha conseguido el óptimo buscado.

Para la muestra discretizada, a partir de la componente 72, se observa que:

1. La disminución del error cuadrático medio es muy pequeña 0,46 %.

2. El aumento del tiempo de ejecución es muy grande 24,76 %.

3. La varianza explicada muy alta, del 98,90 %.

Por tanto, pasar de 72 a 88 componentes supone una ganancia mínima en cuanto a disminución del error cuadrático medio y aumento de la varianza explicada, frente a un incremento mayor del tiempo de ejecución. El objetivo es mínimizar la dimensión de la base de datos perdiendo una mínima información y conseguir reducir los tiempos de ejecución, por tanto, se puede indicar que en la componente 47 se ha conseguido el óptimo buscado.

Los resultados de la matriz de correlaciones de las componentes principales se pueden comprobar en el apéndice A.5 . Para la muestra no discretizada, se observa que hasta la componente 28 la aportación máxima de alguna de las variables a la componente principal están por encima del 50 %, concretamente el 57,93 %. A partir de esta componente pasa a ser menor al 40,42 %. En la componente 48 el máximo de aportación que se produce es del 24,66 % por la variable titularidad del inmueble 3 del prestatario, pasando a aportar menos del 18,51 % las variables al resto de componentes principales. En cuanto a la muestra discretizada, se observa que hasta la componente 54 la aportación máxima de alguna de las variables a la componente principal están por encima del 50 %, concretamente el 57,05 %. A partir de esta componente pasa a ser menor al 47,24 %. En la componente 72 el máximo de aportación que se produce es del 36,90 % por la variable renta correctora, pasando a aportar cada vez menos las variables al resto de componentes principales.

En cuanto a los tiempos, tal y como se observa en la gráfica 6.5 son similares para ambas muestras. Se observa que el tiempo aumenta de forma lineal al número de componentes principales, conforme aumentan las componentes aumenta el tiempo proporcionalmente.

6.4. Conclusiones

En este capítulo se aplican dos métodos para reducir la dimensión de la base de datos, selección de variables Gain Ratio y análisis de componentes principales. Para comprobar que el método LMM, al reducir la dimensión del dataset, sigue siendo eficaz, se ha procedido a comprobar el error cuadrático medio y la varianza explicada cuando se reducen las variables

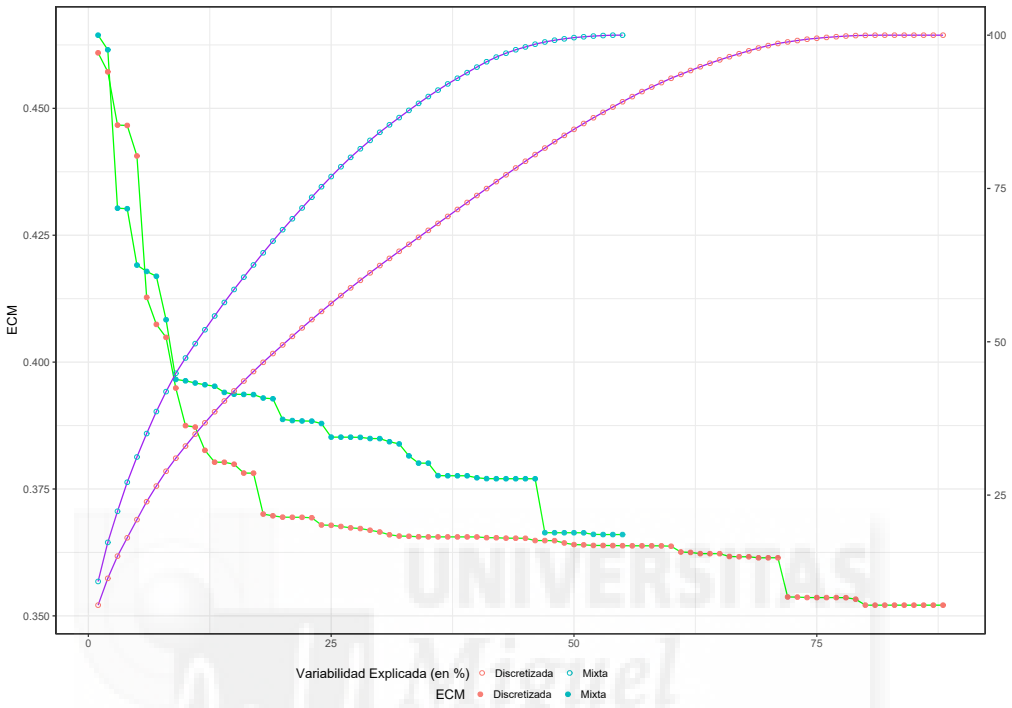

Gráfico 6.4: RMSE y varianza explicada para las muestras resultantes de las dimensiones del PCA ajustadas con LMM

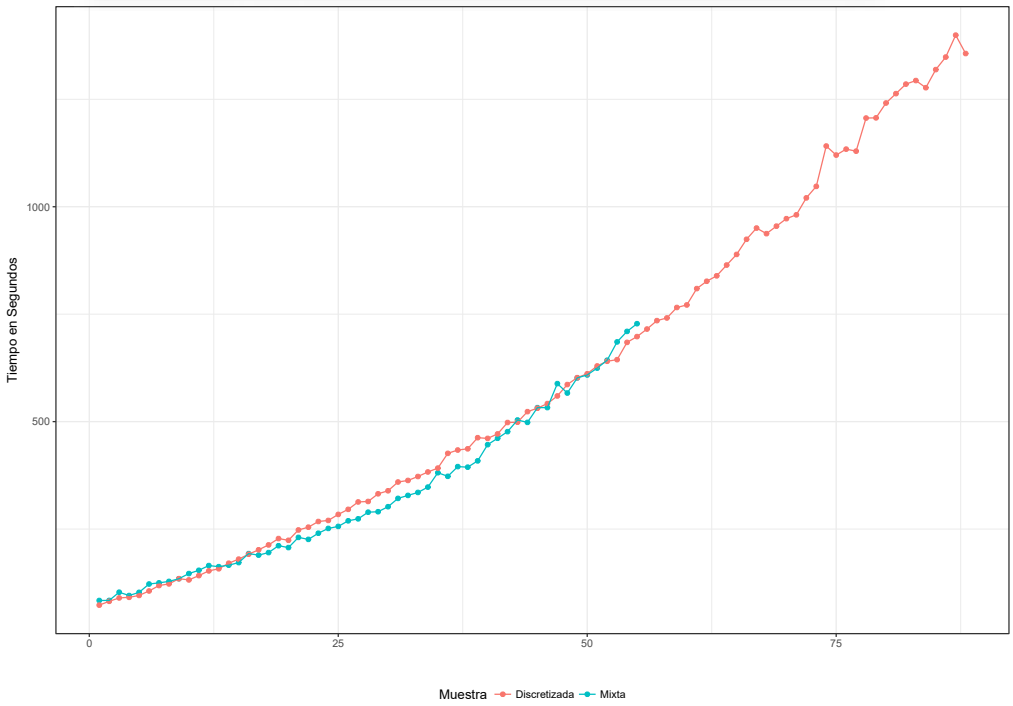

Gráfico 6.5: Tiempo para las muestras resultantes de las dimensiones del PCA ajustadas con LMM

explicativas, midiendo a la vez el tiempo medio para comprobar su reducción, y por tanto se convierte en eficiente computacionalmente. Se ha conseguido con ello dar respuesta al objetivo número 4 de la sección "Objetivos" del "Prólogo".

Una vez analizados los datos con los dos métodos empleados para la selección de variables se puede concluir que:

1. Las variables que tienen más importancia en la base de datos empleada en esta tesis, de cara a predecir la probabilidad de fracaso en un préstamo, son la renta correctora y la renta familiar. Estas variables alcanzan aproximadamente el 51 % del peso, respecto al total de variables. Aplicando el método LMM tanto para Gain Ratio como para PCA se observa que el error cuadrático medio desciende más hasta la introducción de la variable 10 (Gain Ratio) o componente principal 10 (PCA).

2. Introducida la variable 10, representa un 47,35 % de variabilidad explicada. Una vez introducidas las 10 primeras variables o componentes, la variación del error cuadrático medio es mucho menor, es decir por cada variable o componente principal que se introduce la disminución del error cuadrático medio es muy pequeño. Sin embargo, mediante el sistema CREA-RBS se han obtenido mejores resultados que con las 3 primeras componentes del PCA al usar el LMM y utilizando el orden proporcionado por el GR para las 3 primeras variables se mejora el error cuadrático medio.

3. Se observa que las variables económicas son las que explican mejor el modelo. Al mismo tiempo, a partir de la componente principal 48 en la base de datos completa o la 72 en la base de datos discretizada la ganancia, en cuanto a minimización del error cuadrático medio y aumento de varianza explicada se refiere, es muy pequeña comparativamente con el tiempo de ejecución.

7

Conclusiones generales y posibles lineas futuras de investigación

7.1. Conclusiones generales

En el estudio presentado en esta tesis, se puede concluir en varios aspectos:

1. Ajuste de los modelos. Se plantea la utilización de dos modelos: el modelo lineal (efectos fijos) (3.4) y el modelo lineal mixto (efectos fijos y aleatorios)(3.2) en el caso que la variable objetivo sea dicotómica, es decir tome valores 0 y 1. Teniendo en cuenta que los conjuntos de datos a utilizar en esta investigación se refieren a préstamos, es muy común encontrar factores con un mayor número de niveles, y por tanto, se justifica la utilización del modelo lineal mixto. Se recomienda el uso del modelo de estimación reml, para modelos lineales mixtos con efectos aleatorios (3.2).

2. Aplicando el modelo lineal mixto (3.2), con efectos fijos y aleatorios, y heterocedástico ($l = 1$) en una base de datos sintética, para 5 métodos estadísticos (Análisis lineal discriminante (LDA), Árboles de clasificación (CART), Máquinas de vectores soporte (LSVM), Modelo lineal generalizado con nexo logit (GLMlogit), Modelo lineal mixto, efectos fijos y aleatorios (LMM)), con el objetivo de elegir qué método es más eficaz y eficiente para la predicción de la probabilidad de fracaso en un préstamo, se encuentra que los métodos GLMlogit y LMM son mejores métodos para evaluar el riesgo de crédito bajo las hipótesis de las simulaciones de datos realizadas. Son los métodos más eficientes y eficaces al obtener un EMSE menor, tasa de acierto más alta y un tiempo de ejecución más reducido.

3. Se aplica el modelo lineal mixto (3.2) con una base de datos semi-real con el objetivo de

elegir el método más eficaz y eficiente para la predicción de la probabilidad de fracaso en un préstamo. Se concluye que los mejores métodos de evaluación de la predicción de riesgo de crédito hipotecario son GLMlogit y LMM.

4. Debido al problema que surge cuando se trabaja con gran cantidad de datos (Big data), se plantea la selección de variables con el objetivo de reducir el tiempo de computación manteniendo la eficacia. Se comprueba que una óptima selección de variables permite una mejora de la eficiencia sin un menoscabo importante de la eficacia.

Por tanto, los métodos LMM y GLMlogit además de comprobarse eficaces y eficientes, tienen una ventaja añadida, y es que en lugar de clasificar como lo hacen CART, LDA y LSVM, devuelven la probabilidad de que un crédito no sea atendido; es decir, devuelven la probabilidad de default.

Se posicionan como uno de los mejores métodos para el cálculo del riesgo de crédito el modelo lineal mixto (LMM) que hasta la fecha no se había utilizado en esta área de las finanzas.

7.2. Futuras líneas de investigación

Para futuras investigaciones se plantea profundizar en el estudio de los métodos GLMlogit y LMM para mejorar sus resultados.

Se ha comprobado a lo largo de esta investigación que los métodos vectoriales obtienen muy buenos resultados, pero son totalmente ineficientes. Se plantea investigar en dichos métodos y su aplicación en el credit scoring.

Al mismo tiempo se ha considerado la posiblidad de ampliar variables explicativas. Se pretende realizar una investigación sobre las variables macroeconómicas que influyen en la economía familiar y, por tanto, que indirectamente tienen efecto en la posibilidad de cumplir o no con la devolución de un préstamo.

Por último aplicar esta investigación a conjuntos de datos reales correspondientes a series más largas introduciendo las variables macroeconómicas.

A

Apéndices

A.1. Resúmenes numéricos del capitulo 3

		$\hat{\beta_0}$		$\hat{\beta_1}$		$\hat{\sigma_0}$		$\hat{\mu}$	
		Y_0	Y_1	Y_0	Y_1	Y_0	Y_1	Y_0	Y_1
$\ell = 0$	500	0,029 6	0,463 3	0,008 5	0,080 7	0,004 0	0,925 0	0,996 0	1,779 4
	700	0,021 4	0,341 4	0,006 3	0,062 0	0,002 9	0,954 5	0,997 2	1,841 9
	1000	0,015 1	0,245 4	0,004 5	0,045 8	0,002 0	0,972 9	0,998 0	1,889 1
	2000	0,007 6	0,126 1	0,002 3	0,024 3	0,001 0	0,987 7	0,999 0	1,943 9
	3000	0,005 1	0,084 8	0,001 5	0,016 5	0,000 7	0,993 1	0,999 3	1,962 9
	5000	0,003 1	0,051 1	0,000 9	0,010 1	0,000 4	0,996 4	0,999 6	1,977 8
	7500	0,002 1	0,034 2	0,000 6	0,006 7	0,000 3	0,997 5	0,999 7	1,985 1
$\ell = \frac{1}{2}$	500	0,037 6	0,471 4	0,011 6	0,083 8	0,112 2	1,548 9	1,324 0	2,107 3
	700	0,027 2	0,347 2	0,008 5	0,064 2	0,107 8	1,607 1	1,320 5	2,165 2
	1000	0,019 1	0,249 4	0,006 1	0,047 4	0,104 4	1,646 3	1,317 7	2,208 8
	2000	0,009 7	0,128 1	0,003 1	0,025 1	0,100 7	1,684 4	1,314 6	2,259 4
	3000	0,006 5	0,086 1	0,002 1	0,017 0	0,099 4	1,697 7	1,313 4	2,277 1
	5000	0,003 9	0,052 0	0,001 3	0,010 4	0,098 5	1,707 3	1,312 7	2,290 9
	7500	0,002 6	0,034 8	0,000 8	0,007 0	0,098 0	1,711 3	1,312 3	2,297 7
$\ell = 1$	500	0,049 7	0,483 5	0,016 4	0,088 6	0,642 1	2,815 8	1,792 6	2,575 9
	700	0,035 8	0,355 8	0,012 0	0,067 7	0,619 0	2,897 7	1,780 5	2,625 2
	1000	0,025 2	0,255 5	0,008 5	0,049 9	0,601 5	2,953 4	1,771 2	2,662 4
	2000	0,012 7	0,131 1	0,004 4	0,026 3	0,581 9	3,009 4	1,760 7	2,705 5
	3000	0,008 5	0,088 1	0,002 9	0,017 9	0,575 3	3,029 1	1,757 0	2,720 6
	5000	0,005 1	0,053 2	0,001 8	0,010 9	0,570 2	3,043 3	1,754 2	2,732 5
	7500	0,003 4	0,035 6	0,001 2	0,007 3	0,567 7	3,049 5	1,752 8	2,738 2
$\ell = 2$	500	0,096 7	0,530 6	0,036 0	0,108 2	6,114 9	10,910 9	3,460 4	4,243 7
	700	0,069 3	0,389 3	0,026 1	0,081 9	5,835 9	10,870 2	3,407 0	4,251 7
	1000	0,048 6	0,278 9	0,018 5	0,059 8	5,629 7	10,830 8	3,366 6	4,257 8
	2000	0,024 4	0,142 8	0,009 4	0,031 4	5,397 6	10,774 9	3,320 3	4,265 1
	3000	0,016 3	0,095 9	0,006 3	0,021 2	5,320 7	10,759 2	3,304 7	4,268 3
	5000	0,009 8	0,057 9	0,003 8	0,012 9	5,260 3	10,744 1	3,292 3	4,270 6
	7500	0,006 5	0,038 7	0,002 5	0,008 7	5,230 7	10,735 4	3,286 3	4,271 7

Tabla A.1: EMSE de $\hat{\beta_0},\ \hat{\beta_1},\ \hat{\sigma_0}$ *y* $\hat{\mu}$ con $\ell = 0,\ \frac{1}{2},\ 1$ *y* 2 en experimento 1.

		$\hat{\beta_0}$		$\hat{\beta_1}$		$\hat{\sigma_0}$		$\hat{\mu}$	
		Y_0	Y_1	Y_0	Y_1	Y_0	Y_1	Y_0	Y_1
$\ell = 0$	500	$2{,}17 \cdot 10^{-04}$	$1{,}3 \cdot 10^{-03}$	$-6{,}7 \cdot 10^{-05}$	$-4{,}2 \cdot 10^{-04}$	$-3{,}5 \cdot 10^{-03}$	$7{,}9 \cdot 10^{-01}$	$-2{,}7 \cdot 10^{-19}$	$2{,}2 \cdot 10^{-19}$
	700	$-2{,}1 \cdot 10^{-05}$	$-5{,}2 \cdot 10^{-04}$	$4{,}5 \cdot 10^{-05}$	$3{,}3 \cdot 10^{-04}$	$-2{,}1 \cdot 10^{-03}$	$8{,}2 \cdot 10^{-01}$	$8{,}4 \cdot 10^{-20}$	$1{,}2 \cdot 10^{-19}$
	1000	$-1{,}6 \cdot 10^{-04}$	$3{,}4 \cdot 10^{-04}$	$7{,}9 \cdot 10^{-05}$	$-2{,}3 \cdot 10^{-05}$	$-2{,}4 \cdot 10^{-03}$	$8{,}9 \cdot 10^{-01}$	$-4{,}7 \cdot 10^{-20}$	$4{,}0 \cdot 10^{-20}$
	2000	$1{,}7 \cdot 10^{-04}$	$1{,}2 \cdot 10^{-05}$	$-7{,}1 \cdot 10^{-05}$	$7{,}9 \cdot 10^{-05}$	$-9{,}8 \cdot 10^{-04}$	$9{,}4 \cdot 10^{-01}$	$-7{,}6 \cdot 10^{-20}$	$-1{,}6 \cdot 10^{-19}$
	3000	$7{,}9 \cdot 10^{-06}$	$-2{,}8 \cdot 10^{-04}$	$1{,}2 \cdot 10^{-05}$	$9{,}1 \cdot 10^{-05}$	$-7{,}0 \cdot 10^{-04}$	$9{,}3 \cdot 10^{-01}$	$7{,}4 \cdot 10^{-20}$	$-1{,}4 \cdot 10^{-19}$
	5000	$5{,}2 \cdot 10^{-05}$	$7{,}1 \cdot 10^{-05}$	$-2{,}2 \cdot 10^{-05}$	$-5{,}8 \cdot 10^{-05}$	$-4{,}0 \cdot 10^{-04}$	$9{,}8 \cdot 10^{-01}$	$-5{,}8 \cdot 10^{-20}$	$-8{,}9 \cdot 10^{-20}$
	7500	$-3{,}1 \cdot 10^{-05}$	$-1{,}7 \cdot 10^{-04}$	$2{,}7 \cdot 10^{-05}$	$8{,}2 \cdot 10^{-05}$	$-2{,}9 \cdot 10^{-04}$	$9{,}5 \cdot 10^{-01}$	$9{,}5 \cdot 10^{-20}$	$1{,}3 \cdot 10^{-20}$
$\ell = \frac{1}{2}$	500	$2{,}6 \cdot 10^{-04}$	$1{,}2 \cdot 10^{-03}$	$-5{,}2 \cdot 10^{-05}$	$-4{,}0 \cdot 10^{-04}$	$3{,}4 \cdot 10^{-01}$	1,11	$1{,}7 \cdot 10^{-19}$	$8{,}9 \cdot 10^{-20}$
	700	$-6{,}3 \cdot 10^{-05}$	$-6{,}4 \cdot 10^{-04}$	$7{,}3 \cdot 10^{-05}$	$4{,}2 \cdot 10^{-04}$	$3{,}0 \cdot 10^{-01}$	1,17	$2{,}6 \cdot 10^{-20}$	$5{,}2 \cdot 10^{-20}$
	1000	$-1{,}9 \cdot 10^{-04}$	$3{,}1 \cdot 10^{-04}$	$8{,}6 \cdot 10^{-05}$	$-1{,}6 \cdot 10^{-05}$	$3{,}8 \cdot 10^{-01}$	1,21	$-8{,}9 \cdot 10^{-20}$	$1{,}5 \cdot 10^{-19}$
	2000	$1{,}7 \cdot 10^{-04}$	$2{,}9 \cdot 10^{-05}$	$-8{,}8 \cdot 10^{-05}$	$7{,}2 \cdot 10^{-05}$	$3{,}5 \cdot 10^{-01}$	1,26	$-7{,}7 \cdot 10^{-20}$	$-4{,}0 \cdot 10^{-20}$
	3000	$1{,}7 \cdot 10^{-05}$	$-2{,}6 \cdot 10^{-04}$	$1{,}1 \cdot 10^{-05}$	$9{,}0 \cdot 10^{-05}$	$3{,}3 \cdot 10^{-01}$	1,28	$-6{,}4 \cdot 10^{-20}$	$2{,}5 \cdot 10^{-19}$
	5000	$5{,}0 \cdot 10^{-05}$	$8{,}9 \cdot 10^{-05}$	$-2{,}5 \cdot 10^{-05}$	$-6{,}0 \cdot 10^{-05}$	$3{,}3 \cdot 10^{-01}$	1,29	$-1{,}6 \cdot 10^{-19}$	$5{,}0 \cdot 10^{-19}$
	7500	$-4{,}7 \cdot 10^{-05}$	$-1{,}2 \cdot 10^{-04}$	$2{,}8 \cdot 10^{-05}$	$8{,}3 \cdot 10^{-05}$	$3{,}2 \cdot 10^{-01}$	1,30	$1{,}3 \cdot 10^{-19}$	$3{,}2 \cdot 10^{-19}$
$\ell = 1$	500	$1{,}1 \cdot 10^{-04}$	$1{,}1 \cdot 10^{-03}$	$-3{,}7 \cdot 10^{-05}$	$-4{,}3 \cdot 10^{-04}$	$7{,}3 \cdot 10^{-01}$	1,58	$2{,}8 \cdot 10^{-20}$	$-6{,}3 \cdot 10^{-21}$
	700	$-1{,}6 \cdot 10^{-04}$	$-6{,}8 \cdot 10^{-04}$	$1{,}6 \cdot 10^{-04}$	$4{,}0 \cdot 10^{-04}$	$7{,}1 \cdot 10^{-01}$	1,63	$2{,}4 \cdot 10^{-20}$	$-1{,}5 \cdot 10^{-19}$
	1000	$-1{,}6 \cdot 10^{-04}$	$3{,}4 \cdot 10^{-04}$	$1{,}2 \cdot 10^{-04}$	$-6{,}2 \cdot 10^{-06}$	$7{,}1 \cdot 10^{-01}$	1,66	$-5{,}1 \cdot 10^{-20}$	$-1{,}4 \cdot 10^{-19}$
	2000	$1{,}1 \cdot 10^{-04}$	$3{,}1 \cdot 10^{-05}$	$-8{,}0 \cdot 10^{-05}$	$6{,}1 \cdot 10^{-05}$	$7{,}1 \cdot 10^{-01}$	1,71	$-7{,}6 \cdot 10^{-20}$	$-4{,}7 \cdot 10^{-20}$
	3000	$1{,}8 \cdot 10^{-05}$	$-2{,}3 \cdot 10^{-04}$	$1{,}2 \cdot 10^{-05}$	$1{,}0 \cdot 10^{-04}$	$7{,}7 \cdot 10^{-01}$	1,72	$5{,}9 \cdot 10^{-20}$	$-5{,}2 \cdot 10^{-20}$
	5000	$6{,}0 \cdot 10^{-05}$	$9{,}8 \cdot 10^{-05}$	$-2{,}4 \cdot 10^{-05}$	$-6{,}9 \cdot 10^{-05}$	$7{,}4 \cdot 10^{-01}$	1,73	$3{,}1 \cdot 10^{-20}$	$1{,}7 \cdot 10^{-19}$
	7500	$-4{,}8 \cdot 10^{-05}$	$-1{,}7 \cdot 10^{-04}$	$2{,}3 \cdot 10^{-05}$	$9{,}7 \cdot 10^{-05}$	$7{,}3 \cdot 10^{-01}$	1,74	$1{,}6 \cdot 10^{-19}$	$3{,}6 \cdot 10^{-20}$
$\ell = 2$	500	$1{,}9 \cdot 10^{-04}$	$1{,}6 \cdot 10^{-03}$	$1{,}0 \cdot 10^{-05}$	$-3{,}4 \cdot 10^{-04}$	2,46	3,24	$-3{,}5 \cdot 10^{-20}$	$1{,}9 \cdot 10^{-19}$
	700	$-2{,}2 \cdot 10^{-04}$	$-8{,}4 \cdot 10^{-04}$	$2{,}5 \cdot 10^{-04}$	$5{,}9 \cdot 10^{-04}$	2,41	3,25	$7{,}4 \cdot 10^{-20}$	$5{,}3 \cdot 10^{-21}$
	1000	$-2{,}6 \cdot 10^{-04}$	$3{,}4 \cdot 10^{-04}$	$1{,}8 \cdot 10^{-04}$	$2{,}9 \cdot 10^{-05}$	2,37	3,26	$1{,}5 \cdot 10^{-19}$	$1{,}2 \cdot 10^{-19}$
	2000	$2{,}8 \cdot 10^{-04}$	$7{,}2 \cdot 10^{-05}$	$-1{,}9 \cdot 10^{-04}$	$4{,}5 \cdot 10^{-05}$	2,32	3,27	$2{,}1 \cdot 10^{-19}$	$-1{,}9 \cdot 10^{-19}$
	3000	$2{,}3 \cdot 10^{-05}$	$-2{,}4 \cdot 10^{-04}$	$1{,}5 \cdot 10^{-05}$	$1{,}2 \cdot 10^{-04}$	2,30	3,27	$-9{,}8 \cdot 10^{-20}$	$-1{,}0 \cdot 10^{-19}$
	5000	$8{,}0 \cdot 10^{-05}$	$1{,}1 \cdot 10^{-04}$	$-3{,}5 \cdot 10^{-05}$	$-7{,}0 \cdot 10^{-05}$	2,29	3,27	$1{,}1 \cdot 10^{-19}$	$1{,}0 \cdot 10^{-19}$
	7500	$-5{,}1 \cdot 10^{-05}$	$-1{,}8 \cdot 10^{-04}$	$3{,}4 \cdot 10^{-05}$	$9{,}8 \cdot 10^{-05}$	2,29	3,27	$1{,}9 \cdot 10^{-19}$	$-4{,}4 \cdot 10^{-20}$

Tabla A.2: BIAS de $\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\sigma_0}$ *y* $\hat{\mu}$ con $\ell = 0,\ \frac{1}{2},\ 1$ *y* 2 en experimento 1.

		$\hat{\beta_0}$		$\hat{\sigma_1}$		$\hat{\sigma_0}$		$\hat{\mu}$	
		Y_0	Y_1	Y_0	Y_1	Y_0	Y_1	Y_0	Y_1
$\ell = 0$	500	$5{,}9 \cdot 10^{-04}$	$8{,}9 \cdot 10^{-03}$	0,994 7	$4{,}7 \cdot 10^{-01}$	$4{,}0 \cdot 10^{-03}$	$4{,}3 \cdot 10^{-03}$	0,993 6	0,987 5
	700	$4{,}6 \cdot 10^{-04}$	$6{,}9 \cdot 10^{-03}$	0,995 5	$2{,}8 \cdot 10^{-01}$	$2{,}3 \cdot 10^{-03}$	$2{,}5 \cdot 10^{-03}$	0,995 9	0,989 4
	1000	$2{,}9 \cdot 10^{-04}$	$4{,}0 \cdot 10^{-03}$	0,996 3	$2{,}9 \cdot 10^{-01}$	$2{,}0 \cdot 10^{-03}$	$2{,}2 \cdot 10^{-03}$	0,996 6	0,989 6
	2000	$1{,}3 \cdot 10^{-04}$	$2{,}9 \cdot 10^{-03}$	0,997 5	$1{,}5 \cdot 10^{-01}$	$1{,}0 \cdot 10^{-03}$	$1{,}1 \cdot 10^{-03}$	0,997 3	0,989 6
	3000	$1{,}2 \cdot 10^{-04}$	$1{,}8 \cdot 10^{-03}$	0,997 9	$7{,}1 \cdot 10^{-02}$	$6{,}4 \cdot 10^{-04}$	$6{,}8 \cdot 10^{-04}$	0,997 6	0,989 5
	5000	$5{,}7 \cdot 10^{-05}$	$8{,}8 \cdot 10^{-04}$	0,998 4	$4{,}3 \cdot 10^{-02}$	$3{,}3 \cdot 10^{-04}$	$3{,}6 \cdot 10^{-04}$	0,998 3	0,989 8
	7500	$4{,}1 \cdot 10^{-05}$	$5{,}9 \cdot 10^{-04}$	0,998 7	$2{,}1 \cdot 10^{-02}$	$2{,}9 \cdot 10^{-04}$	$2{,}1 \cdot 10^{-04}$	0,999 0	0,990 3
$\ell = \frac{1}{2}$	500	$8{,}4 \cdot 10^{-04}$	$1{,}2 \cdot 10^{-02}$	0,993 1	$4{,}3 \cdot 10^{-01}$	$4{,}0 \cdot 10^{-03}$	$4{,}3 \cdot 10^{-03}$	1,320 9	1,313 1
	700	$5{,}5 \cdot 10^{-04}$	$7{,}3 \cdot 10^{-03}$	0,994 1	$3{,}2 \cdot 10^{-01}$	$2{,}2 \cdot 10^{-03}$	$2{,}5 \cdot 10^{-03}$	1,318 7	1,310 4
	1000	$4{,}2 \cdot 10^{-04}$	$5{,}2 \cdot 10^{-03}$	0,995 2	$2{,}1 \cdot 10^{-01}$	$2{,}0 \cdot 10^{-03}$	$2{,}2 \cdot 10^{-03}$	1,315 7	1,306 8
	2000	$2{,}2 \cdot 10^{-04}$	$2{,}7 \cdot 10^{-03}$	0,996 7	$1{,}6 \cdot 10^{-01}$	$1{,}0 \cdot 10^{-03}$	$1{,}1 \cdot 10^{-03}$	1,312 4	1,302 4
	3000	$1{,}1 \cdot 10^{-04}$	$1{,}8 \cdot 10^{-03}$	0,997 3	$7{,}5 \cdot 10^{-02}$	$6{,}4 \cdot 10^{-04}$	$6{,}8 \cdot 10^{-04}$	1,311 2	1,300 9
	5000	$8{,}5 \cdot 10^{-05}$	$1{,}4 \cdot 10^{-03}$	0,997 9	$4{,}6 \cdot 10^{-02}$	$3{,}3 \cdot 10^{-04}$	$3{,}6 \cdot 10^{-04}$	1,311 0	1,300 1
	7500	$5{,}5 \cdot 10^{-05}$	$7{,}5 \cdot 10^{-04}$	0,998 3	$2{,}2 \cdot 10^{-02}$	$2{,}9 \cdot 10^{-04}$	$2{,}1 \cdot 10^{-04}$	1,311 3	1,300 1
$\ell = 1$	500	$1{,}3 \cdot 10^{-03}$	$1{,}4 \cdot 10^{-02}$	0,991 2	$4{,}1 \cdot 10^{-01}$	$4{,}0 \cdot 10^{-03}$	$4{,}3 \cdot 10^{-03}$	1,788 9	1,779 1
	700	$8{,}7 \cdot 10^{-04}$	$1{,}1 \cdot 10^{-02}$	0,992 5	$3{,}6 \cdot 10^{-01}$	$2{,}2 \cdot 10^{-03}$	$2{,}4 \cdot 10^{-03}$	1,778 5	1,768 0
	1000	$5{,}0 \cdot 10^{-04}$	$7{,}6 \cdot 10^{-03}$	0,994 0	$2{,}4 \cdot 10^{-01}$	$2{,}0 \cdot 10^{-03}$	$2{,}2 \cdot 10^{-03}$	1,768 7	1,757 4
	2000	$2{,}9 \cdot 10^{-04}$	$3{,}1 \cdot 10^{-03}$	0,995 9	$1{,}7 \cdot 10^{-01}$	$1{,}0 \cdot 10^{-03}$	$1{,}1 \cdot 10^{-03}$	1,757 9	1,745 4
	3000	$1{,}2 \cdot 10^{-04}$	$2{,}8 \cdot 10^{-03}$	0,996 6	$7{,}0 \cdot 10^{-02}$	$6{,}4 \cdot 10^{-04}$	$6{,}8 \cdot 10^{-04}$	1,754 2	1,741 1
	5000	$1{,}2 \cdot 10^{-04}$	$1{,}3 \cdot 10^{-03}$	0,997 4	$4{,}9 \cdot 10^{-02}$	$3{,}3 \cdot 10^{-04}$	$3{,}6 \cdot 10^{-04}$	1,752 1	1,738 4
	7500	$7{,}4 \cdot 10^{-05}$	$9{,}0 \cdot 10^{-04}$	0,997 9	$2{,}4 \cdot 10^{-02}$	$2{,}9 \cdot 10^{-04}$	$2{,}1 \cdot 10^{-04}$	1,751 7	1,737 5
$\ell = 2$	500	$2{,}3 \cdot 10^{-03}$	$2{,}5 \cdot 10^{-02}$	0,986 4	$4{,}0 \cdot 10^{-01}$	$4{,}0 \cdot 10^{-03}$	$4{,}2 \cdot 10^{-03}$	3,456 6	3,441 3
	700	$1{,}6 \cdot 10^{-03}$	$1{,}4 \cdot 10^{-02}$	0,988 5	$3{,}7 \cdot 10^{-01}$	$2{,}2 \cdot 10^{-03}$	$2{,}4 \cdot 10^{-03}$	3,405 4	3,389 1
	1000	$9{,}7 \cdot 10^{-04}$	$1{,}5 \cdot 10^{-02}$	0,990 8	$2{,}0 \cdot 10^{-01}$	$2{,}0 \cdot 10^{-03}$	$2{,}2 \cdot 10^{-03}$	3,363 2	3,345 6
	2000	$5{,}9 \cdot 10^{-04}$	$5{,}9 \cdot 10^{-03}$	0,993 7	$1{,}9 \cdot 10^{-01}$	$1{,}0 \cdot 10^{-03}$	$1{,}1 \cdot 10^{-03}$	3,316 7	3,297 6
	3000	$3{,}6 \cdot 10^{-04}$	$3{,}8 \cdot 10^{-03}$	0,994 8	$7{,}4 \cdot 10^{-02}$	$6{,}4 \cdot 10^{-04}$	$6{,}8 \cdot 10^{-04}$	3,300 4	3,280 4
	5000	$1{,}7 \cdot 10^{-04}$	$2{,}6 \cdot 10^{-03}$	0,996 0	$4{,}8 \cdot 10^{-02}$	$3{,}3 \cdot 10^{-04}$	$3{,}6 \cdot 10^{-04}$	3,289 3	3,268 4
	7500	$1{,}2 \cdot 10^{-04}$	$1{,}9 \cdot 10^{-03}$	0,996 8	$2{,}0 \cdot 10^{-02}$	$2{,}9 \cdot 10^{-04}$	$2{,}1 \cdot 10^{-04}$	3,285 1	3,263 5

Tabla A.3: EMSE de $\hat{\beta_0},\ \hat{\sigma_1},\ \hat{\sigma_0}$ *y* $\hat{\mu}$ con $\ell = 0,\ \frac{1}{2},\ 1$ *y* 2 en experimento 2.

		$\hat{\beta}_0$		$\hat{\sigma_1}$		$\hat{\sigma_0}$		$\hat{\mu}$	
		Y_0	Y_1	Y_0	Y_1	Y_0	Y_1	Y_0	Y_1
$\ell = 0$	500	$6{,}6 \cdot 10^{-04}$	$-2{,}8 \cdot 10^{-04}$	$-9{,}7 \cdot 10^{-01}$	$-3{,}9 \cdot 10^{-03}$	$-3{,}6 \cdot 10^{-03}$	$-1{,}1 \cdot 10^{-03}$	$1{,}7 \cdot 10^{-04}$	$-8{,}2 \cdot 10^{-07}$
	700	$1{,}0 \cdot 10^{-06}$	$1{,}0 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$3{,}7 \cdot 10^{-03}$	$-1{,}8 \cdot 10^{-03}$	$4{,}8 \cdot 10^{-04}$	$1{,}0 \cdot 10^{-04}$	$1{,}6 \cdot 10^{-04}$
	1000	$1{,}9 \cdot 10^{-04}$	$7{,}0 \cdot 10^{-05}$	$-9{,}8 \cdot 10^{-01}$	$-1{,}0 \cdot 10^{-03}$	$-1{,}8 \cdot 10^{-03}$	$3{,}5 \cdot 10^{-04}$	$8{,}4 \cdot 10^{-06}$	$1{,}0 \cdot 10^{-06}$
	2000	$1{,}1 \cdot 10^{-04}$	$-5{,}4 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-3{,}9 \cdot 10^{-03}$	$-1{,}7 \cdot 10^{-03}$	$-1{,}2 \cdot 10^{-04}$	$9{,}6 \cdot 10^{-05}$	$1{,}3 \cdot 10^{-06}$
	3000	$-7{,}7 \cdot 10^{-05}$	$3{,}8 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-2{,}9 \cdot 10^{-04}$	$-1{,}6 \cdot 10^{-03}$	$-3{,}1 \cdot 10^{-04}$	$-5{,}9 \cdot 10^{-05}$	$-4{,}2 \cdot 10^{-06}$
	5000	$-2{,}2 \cdot 10^{-05}$	$-3{,}0 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-7{,}9 \cdot 10^{-04}$	$-8{,}9 \cdot 10^{-04}$	$-1{,}2 \cdot 10^{-04}$	$3{,}0 \cdot 10^{-05}$	$5{,}4 \cdot 10^{-06}$
	7500	$1{,}0 \cdot 10^{-04}$	$-7{,}4 \cdot 10^{-05}$	$-9{,}9 \cdot 10^{-01}$	$2{,}0 \cdot 10^{-03}$	$-3{,}9 \cdot 10^{-04}$	$2{,}6 \cdot 10^{-04}$	$4{,}8 \cdot 10^{-05}$	$5{,}0 \cdot 10^{-06}$
$\ell = \frac{1}{2}$	500	$7{,}8 \cdot 10^{-04}$	$1{,}7 \cdot 10^{-05}$	$-9{,}7 \cdot 10^{-01}$	$-3{,}6 \cdot 10^{-03}$	$-3{,}6 \cdot 10^{-03}$	$-1{,}1 \cdot 10^{-03}$	$5{,}1 \cdot 10^{-05}$	$2{,}7 \cdot 10^{-06}$
	700	$1{,}8 \cdot 10^{-05}$	$2{,}4 \cdot 10^{-04}$	$-9{,}7 \cdot 10^{-01}$	$3{,}6 \cdot 10^{-03}$	$-1{,}9 \cdot 10^{-03}$	$4{,}2 \cdot 10^{-04}$	$6{,}1 \cdot 10^{-05}$	$7{,}0 \cdot 10^{-05}$
	1000	$1{,}2 \cdot 10^{-04}$	$1{,}3 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-1{,}6 \cdot 10^{-03}$	$-1{,}8 \cdot 10^{-03}$	$3{,}9 \cdot 10^{-04}$	$3{,}3 \cdot 10^{-06}$	$8{,}2 \cdot 10^{-07}$
	2000	$1{,}5 \cdot 10^{-04}$	$-6{,}6 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-3{,}4 \cdot 10^{-03}$	$-1{,}7 \cdot 10^{-03}$	$-1{,}6 \cdot 10^{-04}$	$5{,}5 \cdot 10^{-05}$	$5{,}0 \cdot 10^{-07}$
	3000	$-1{,}1 \cdot 10^{-04}$	$4{,}8 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-2{,}7 \cdot 10^{-04}$	$-1{,}6 \cdot 10^{-03}$	$-3{,}2 \cdot 10^{-04}$	$-3{,}1 \cdot 10^{-05}$	$-1{,}4 \cdot 10^{-06}$
	5000	$-2{,}1 \cdot 10^{-05}$	$-3{,}3 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-7{,}2 \cdot 10^{-04}$	$-8{,}7 \cdot 10^{-04}$	$-1{,}1 \cdot 10^{-04}$	$2{,}2 \cdot 10^{-05}$	$5{,}7 \cdot 10^{-06}$
	7500	$1{,}5 \cdot 10^{-04}$	$-2{,}4 \cdot 10^{-05}$	$-9{,}9 \cdot 10^{-01}$	$2{,}6 \cdot 10^{-03}$	$-3{,}6 \cdot 10^{-04}$	$2{,}0 \cdot 10^{-04}$	$2{,}1 \cdot 10^{-05}$	$3{,}4 \cdot 10^{-06}$
$\ell = 1$	500	$9{,}6 \cdot 10^{-04}$	$2{,}6 \cdot 10^{-04}$	$-9{,}6 \cdot 10^{-01}$	$-3{,}0 \cdot 10^{-03}$	$-3{,}6 \cdot 10^{-03}$	$-1{,}2 \cdot 10^{-03}$	$8{,}4 \cdot 10^{-18}$	$-1{,}2 \cdot 10^{-17}$
	700	$4{,}7 \cdot 10^{-05}$	$3{,}6 \cdot 10^{-04}$	$-9{,}6 \cdot 10^{-01}$	$3{,}1 \cdot 10^{-03}$	$-1{,}9 \cdot 10^{-03}$	$4{,}7 \cdot 10^{-04}$	$-1{,}4 \cdot 10^{-17}$	$9{,}3 \cdot 10^{-18}$
	1000	$1{,}7 \cdot 10^{-04}$	$1{,}7 \cdot 10^{-04}$	$-9{,}7 \cdot 10^{-01}$	$-2{,}2 \cdot 10^{-03}$	$-1{,}7 \cdot 10^{-03}$	$3{,}2 \cdot 10^{-04}$	$-3{,}0 \cdot 10^{-17}$	$-8{,}0 \cdot 10^{-18}$
	2000	$1{,}7 \cdot 10^{-04}$	$-6{,}5 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-3{,}9 \cdot 10^{-03}$	$-1{,}7 \cdot 10^{-03}$	$-1{,}0 \cdot 10^{-04}$	$-1{,}9 \cdot 10^{-17}$	$4{,}5 \cdot 10^{-18}$
	3000	$-1{,}7 \cdot 10^{-04}$	$4{,}1 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-2{,}8 \cdot 10^{-04}$	$-1{,}6 \cdot 10^{-03}$	$-3{,}3 \cdot 10^{-04}$	$-4{,}8 \cdot 10^{-17}$	$-9{,}2 \cdot 10^{-18}$
	5000	$-1{,}6 \cdot 10^{-05}$	$-3{,}0 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-7{,}1 \cdot 10^{-04}$	$-8{,}6 \cdot 10^{-04}$	$-9{,}7 \cdot 10^{-05}$	$-6{,}6 \cdot 10^{-17}$	$7{,}5 \cdot 10^{-18}$
	7500	$2{,}0 \cdot 10^{-04}$	$4{,}8 \cdot 10^{-05}$	$-9{,}9 \cdot 10^{-01}$	$2{,}3 \cdot 10^{-03}$	$-3{,}4 \cdot 10^{-04}$	$2{,}1 \cdot 10^{-04}$	$-3{,}8 \cdot 10^{-17}$	$5{,}6 \cdot 10^{-18}$
$\ell = 2$	500	$1{,}6 \cdot 10^{-03}$	$1{,}2 \cdot 10^{-03}$	$-9{,}3 \cdot 10^{-01}$	$-2{,}6 \cdot 10^{-03}$	$-3{,}7 \cdot 10^{-03}$	$-1{,}3 \cdot 10^{-03}$	$-4{,}4 \cdot 10^{-05}$	$-1{,}2 \cdot 10^{-05}$
	700	$1{,}7 \cdot 10^{-04}$	$7{,}3 \cdot 10^{-04}$	$-9{,}4 \cdot 10^{-01}$	$3{,}5 \cdot 10^{-03}$	$-1{,}9 \cdot 10^{-03}$	$4{,}8 \cdot 10^{-04}$	$-1{,}6 \cdot 10^{-04}$	$-2{,}4 \cdot 10^{-04}$
	1000	$2{,}8 \cdot 10^{-04}$	$2{,}5 \cdot 10^{-04}$	$-9{,}5 \cdot 10^{-01}$	$-2{,}2 \cdot 10^{-03}$	$-1{,}7 \cdot 10^{-03}$	$3{,}6 \cdot 10^{-04}$	$6{,}9 \cdot 10^{-06}$	$-8{,}7 \cdot 10^{-07}$
	2000	$3{,}8 \cdot 10^{-04}$	$-7{,}0 \cdot 10^{-04}$	$-9{,}7 \cdot 10^{-01}$	$-3{,}9 \cdot 10^{-03}$	$-1{,}7 \cdot 10^{-03}$	$-1{,}3 \cdot 10^{-04}$	$-1{,}6 \cdot 10^{-04}$	$-1{,}3 \cdot 10^{-05}$
	3000	$-1{,}3 \cdot 10^{-04}$	$3{,}3 \cdot 10^{-04}$	$-9{,}7 \cdot 10^{-01}$	$-2{,}5 \cdot 10^{-04}$	$-1{,}6 \cdot 10^{-03}$	$-3{,}8 \cdot 10^{-04}$	$7{,}5 \cdot 10^{-05}$	$3{,}9 \cdot 10^{-06}$
	5000	$-4{,}9 \cdot 10^{-06}$	$-3{,}0 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-6{,}6 \cdot 10^{-04}$	$-8{,}1 \cdot 10^{-04}$	$-9{,}3 \cdot 10^{-05}$	$-6{,}5 \cdot 10^{-05}$	$-3{,}3 \cdot 10^{-05}$
	7500	$2{,}1 \cdot 10^{-04}$	$2{,}8 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$2{,}7 \cdot 10^{-03}$	$-3{,}9 \cdot 10^{-04}$	$2{,}5 \cdot 10^{-04}$	$-3{,}4 \cdot 10^{-05}$	$-2{,}0 \cdot 10^{-05}$

Tabla A.4: BIAS de $\hat{\beta}_0,\ \hat{\sigma_1},\ \hat{\sigma_0}$ *y* $\hat{\mu}$ con $\ell = 0,\ \frac{1}{2},\ 1$ *y* 2 en experimento 2.

		$\hat{\beta}_0$		$\hat{\sigma_1}$		$\hat{\sigma_0}$		$\hat{\mu}$	
		Y_0	Y_1	Y_0	Y_1	Y_0	Y_1	Y_0	Y_1
$\ell = 0$	500	$5{,}7 \cdot 10^{-04}$	$8{,}8 \cdot 10^{-03}$	0,997 0	0,415 0	$3{,}9 \cdot 10^{-03}$	$4{,}2 \cdot 10^{-03}$	0,994 6	0,987 6
	700	$4{,}3 \cdot 10^{-04}$	$6{,}1 \cdot 10^{-03}$	0,997 0	0,297 2	$2{,}2 \cdot 10^{-03}$	$2{,}4 \cdot 10^{-03}$	0,996 6	0,989 4
	1000	$2{,}8 \cdot 10^{-04}$	$4{,}1 \cdot 10^{-03}$	0,997 4	0,208 1	$2{,}0 \cdot 10^{-03}$	$2{,}2 \cdot 10^{-03}$	0,997 2	0,989 6
	2000	$1{,}3 \cdot 10^{-04}$	$2{,}9 \cdot 10^{-03}$	0,998 0	0,104 6	$1{,}0 \cdot 10^{-03}$	$1{,}1 \cdot 10^{-03}$	0,997 6	0,989 6
	3000	$1{,}2 \cdot 10^{-04}$	$1{,}8 \cdot 10^{-03}$	0,998 2	0,070 0	$6{,}4 \cdot 10^{-04}$	$6{,}8 \cdot 10^{-04}$	0,997 8	0,989 5
	5000	$5{,}7 \cdot 10^{-05}$	$8{,}8 \cdot 10^{-04}$	0,998 6	0,041 3	$3{,}3 \cdot 10^{-04}$	$3{,}5 \cdot 10^{-04}$	0,998 4	0,989 8
	7500	$4{,}3 \cdot 10^{-05}$	$6{,}6 \cdot 10^{-04}$	0,998 8	0,026 9	$2{,}9 \cdot 10^{-04}$	$2{,}1 \cdot 10^{-04}$	0,999 1	0,990 3
$\ell = \frac{1}{2}$	500	$8{,}9 \cdot 10^{-04}$	$1{,}3 \cdot 10^{-02}$	0,996 0	0,420 5	$3{,}9 \cdot 10^{-03}$	$4{,}2 \cdot 10^{-03}$	1,322 1	1,313 2
	700	$5{,}1 \cdot 10^{-04}$	$7{,}7 \cdot 10^{-03}$	0,996 1	0,300 3	$2{,}2 \cdot 10^{-03}$	$2{,}4 \cdot 10^{-03}$	1,319 6	1,310 5
	1000	$4{,}1 \cdot 10^{-04}$	$5{,}4 \cdot 10^{-03}$	0,996 6	0,210 2	$2{,}0 \cdot 10^{-03}$	$2{,}2 \cdot 10^{-03}$	1,316 5	1,306 8
	2000	$2{,}1 \cdot 10^{-04}$	$2{,}8 \cdot 10^{-03}$	0,997 4	0,105 4	$1{,}0 \cdot 10^{-03}$	$1{,}1 \cdot 10^{-03}$	1,312 8	1,302 5
	3000	$1{,}1 \cdot 10^{-04}$	$1{,}8 \cdot 10^{-03}$	0,997 7	0,070 4	$6{,}4 \cdot 10^{-04}$	$6{,}8 \cdot 10^{-04}$	1,311 5	1,300 9
	5000	$8{,}5 \cdot 10^{-05}$	$1{,}4 \cdot 10^{-03}$	0,998 2	0,041 6	$3{,}3 \cdot 10^{-04}$	$3{,}6 \cdot 10^{-04}$	1,311 2	1,300 1
	7500	$5{,}4 \cdot 10^{-05}$	$7{,}5 \cdot 10^{-04}$	0,998 5	0,027 2	$2{,}9 \cdot 10^{-04}$	$2{,}1 \cdot 10^{-04}$	1,311 4	1,300 1
$\ell = 1$	500	$1{,}1 \cdot 10^{-03}$	$1{,}6 \cdot 10^{-02}$	0,994 9	0,426 9	$3{,}9 \cdot 10^{-03}$	$4{,}2 \cdot 10^{-03}$	1,790 4	1,779 3
	700	$8{,}2 \cdot 10^{-04}$	$1{,}2 \cdot 10^{-02}$	0,995 1	0,303 9	$2{,}2 \cdot 10^{-03}$	$2{,}4 \cdot 10^{-03}$	1,779 6	1,768 0
	1000	$5{,}8 \cdot 10^{-04}$	$7{,}8 \cdot 10^{-03}$	0,995 7	0,212 6	$2{,}0 \cdot 10^{-03}$	$2{,}2 \cdot 10^{-03}$	1,769 6	1,757 4
	2000	$2{,}8 \cdot 10^{-04}$	$3{,}2 \cdot 10^{-03}$	0,996 7	0,106 3	$1{,}0 \cdot 10^{-03}$	$1{,}1 \cdot 10^{-03}$	1,758 4	1,745 4
	3000	$1{,}1 \cdot 10^{-04}$	$2{,}8 \cdot 10^{-03}$	0,997 1	0,070 9	$6{,}4 \cdot 10^{-04}$	$6{,}8 \cdot 10^{-04}$	1,754 6	1,741 1
	5000	$1{,}2 \cdot 10^{-04}$	$1{,}3 \cdot 10^{-03}$	0,997 7	0,041 9	$3{,}3 \cdot 10^{-04}$	$3{,}6 \cdot 10^{-04}$	1,752 4	1,738 4
	7500	$7{,}4 \cdot 10^{-05}$	$9{,}0 \cdot 10^{-04}$	0,998 1	0,027 4	$2{,}9 \cdot 10^{-04}$	$2{,}1 \cdot 10^{-04}$	1,751 9	1,737 5
$\ell = 2$	500	$2{,}0 \cdot 10^{-03}$	$2{,}0 \cdot 10^{-02}$	0,992 1	0,442 5	$3{,}9 \cdot 10^{-03}$	$4{,}2 \cdot 10^{-03}$	3,458 9	3,441 9
	700	$1{,}5 \cdot 10^{-03}$	$1{,}5 \cdot 10^{-02}$	0,992 4	0,312 8	$2{,}2 \cdot 10^{-03}$	$2{,}4 \cdot 10^{-03}$	3,407 2	3,389 3
	1000	$9{,}4 \cdot 10^{-04}$	$1{,}5 \cdot 10^{-02}$	0,993 4	0,218 2	$2{,}0 \cdot 10^{-03}$	$2{,}2 \cdot 10^{-03}$	3,364 6	3,345 7
	2000	$5{,}8 \cdot 10^{-04}$	$5{,}9 \cdot 10^{-03}$	0,994 9	0,108 6	$1{,}0 \cdot 10^{-03}$	$1{,}1 \cdot 10^{-03}$	3,317 5	3,297 6
	3000	$3{,}6 \cdot 10^{-04}$	$3{,}8 \cdot 10^{-03}$	0,995 6	0,072 1	$6{,}4 \cdot 10^{-04}$	$6{,}8 \cdot 10^{-04}$	3,301 0	3,280 4
	5000	$1{,}7 \cdot 10^{-04}$	$2{,}6 \cdot 10^{-03}$	0,996 5	0,042 7	$3{,}3 \cdot 10^{-04}$	$3{,}6 \cdot 10^{-04}$	3,289 7	3,268 4
	7500	$1{,}2 \cdot 10^{-04}$	$1{,}9 \cdot 10^{-03}$	0,997 1	0,027 9	$2{,}9 \cdot 10^{-04}$	$2{,}1 \cdot 10^{-04}$	3,285 3	3,263 5

Tabla A.5: EMSE de $\hat{\beta}_0,\ \hat{\sigma_1},\ \hat{\sigma_0}$ *y* $\hat{\mu}$ con $\ell = 0,\ \frac{1}{2},\ 1$ *y* 2 en experimento 2.

		$\hat{\beta_0}$		$\hat{\sigma_1}$		$\hat{\sigma_0}$		$\hat{\mu}$	
		Y_0	Y_1	Y_0	Y_1	Y_0	Y_1	Y_0	Y_1
$\ell = 0$	500	$6{,}1 \cdot 10^{-04}$	$-2{,}0 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-2{,}8 \cdot 10^{-02}$	$-4{,}0 \cdot 10^{-03}$	$-2{,}3 \cdot 10^{-03}$	$1{,}8 \cdot 10^{-04}$	$1{,}4 \cdot 10^{-06}$
	700	$-3{,}9 \cdot 10^{-07}$	$2{,}9 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-1{,}8 \cdot 10^{-02}$	$-2{,}1 \cdot 10^{-03}$	$-7{,}3 \cdot 10^{-04}$	$1{,}3 \cdot 10^{-04}$	$1{,}0 \cdot 10^{-04}$
	1000	$1{,}9 \cdot 10^{-04}$	$7{,}5 \cdot 10^{-05}$	$-9{,}9 \cdot 10^{-01}$	$-1{,}4 \cdot 10^{-02}$	$-2{,}1 \cdot 10^{-03}$	$-5{,}2 \cdot 10^{-04}$	$9{,}7 \cdot 10^{-06}$	$1{,}3 \cdot 10^{-06}$
	2000	$1{,}3 \cdot 10^{-04}$	$-5{,}4 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-1{,}2 \cdot 10^{-02}$	$-1{,}7 \cdot 10^{-03}$	$-5{,}1 \cdot 10^{-04}$	$1{,}0 \cdot 10^{-04}$	$1{,}6 \cdot 10^{-06}$
	3000	$-8{,}0 \cdot 10^{-05}$	$3{,}8 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-4{,}0 \cdot 10^{-03}$	$-1{,}6 \cdot 10^{-03}$	$-6{,}8 \cdot 10^{-04}$	$-5{,}8 \cdot 10^{-05}$	$-4{,}3 \cdot 10^{-06}$
	5000	$-2{,}5 \cdot 10^{-05}$	$-3{,}9 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-3{,}9 \cdot 10^{-03}$	$-1{,}1 \cdot 10^{-03}$	$-2{,}5 \cdot 10^{-04}$	$3{,}6 \cdot 10^{-05}$	$5{,}8 \cdot 10^{-06}$
	7500	$1{,}5 \cdot 10^{-04}$	$3{,}2 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-2{,}7 \cdot 10^{-03}$	$-3{,}9 \cdot 10^{-04}$	$2{,}4 \cdot 10^{-04}$	$-2{,}2 \cdot 10^{-05}$	$-9{,}2 \cdot 10^{-06}$
$\ell = \frac{1}{2}$	500	$7{,}3 \cdot 10^{-04}$	$1{,}9 \cdot 10^{-05}$	$-9{,}8 \cdot 10^{-01}$	$-3{,}6 \cdot 10^{-02}$	$-4{,}1 \cdot 10^{-03}$	$-2{,}7 \cdot 10^{-03}$	$5{,}3 \cdot 10^{-05}$	$2{,}6 \cdot 10^{-06}$
	700	$1{,}3 \cdot 10^{-05}$	$3{,}4 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-1{,}4 \cdot 10^{-02}$	$-2{,}2 \cdot 10^{-03}$	$-7{,}5 \cdot 10^{-04}$	$6{,}5 \cdot 10^{-05}$	$7{,}4 \cdot 10^{-05}$
	1000	$1{,}2 \cdot 10^{-04}$	$1{,}2 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-1{,}0 \cdot 10^{-02}$	$-2{,}1 \cdot 10^{-03}$	$-4{,}9 \cdot 10^{-04}$	$3{,}5 \cdot 10^{-06}$	$8{,}6 \cdot 10^{-07}$
	2000	$1{,}8 \cdot 10^{-04}$	$-6{,}6 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-1{,}7 \cdot 10^{-02}$	$-1{,}7 \cdot 10^{-03}$	$-5{,}2 \cdot 10^{-04}$	$5{,}9 \cdot 10^{-05}$	$5{,}5 \cdot 10^{-07}$
	3000	$-1{,}1 \cdot 10^{-04}$	$4{,}7 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-5{,}5 \cdot 10^{-03}$	$-1{,}6 \cdot 10^{-03}$	$-6{,}1 \cdot 10^{-04}$	$-3{,}0 \cdot 10^{-05}$	$-1{,}5 \cdot 10^{-06}$
	5000	$-1{,}9 \cdot 10^{-05}$	$-3{,}3 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-4{,}1 \cdot 10^{-03}$	$-1{,}1 \cdot 10^{-03}$	$-2{,}9 \cdot 10^{-04}$	$2{,}1 \cdot 10^{-05}$	$5{,}8 \cdot 10^{-06}$
	7500	$1{,}5 \cdot 10^{-04}$	$-2{,}4 \cdot 10^{-05}$	$-9{,}9 \cdot 10^{-01}$	$1{,}1 \cdot 10^{-04}$	$-4{,}1 \cdot 10^{-04}$	$1{,}7 \cdot 10^{-04}$	$2{,}1 \cdot 10^{-05}$	$3{,}4 \cdot 10^{-06}$
$\ell = 1$	500	$9{,}1 \cdot 10^{-04}$	$2{,}0 \cdot 10^{-04}$	$-9{,}7 \cdot 10^{-01}$	$-4{,}1 \cdot 10^{-02}$	$-4{,}2 \cdot 10^{-03}$	$-2{,}0 \cdot 10^{-03}$	$7{,}7 \cdot 10^{-18}$	$-2{,}7 \cdot 10^{-18}$
	700	$3{,}6 \cdot 10^{-05}$	$4{,}7 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-2{,}5 \cdot 10^{-02}$	$-2{,}3 \cdot 10^{-03}$	$-6{,}1 \cdot 10^{-04}$	$4{,}1 \cdot 10^{-18}$	$-6{,}9 \cdot 10^{-18}$
	1000	$1{,}7 \cdot 10^{-04}$	$1{,}5 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-2{,}9 \cdot 10^{-02}$	$-2{,}1 \cdot 10^{-03}$	$-4{,}3 \cdot 10^{-04}$	$2{,}2 \cdot 10^{-18}$	$-2{,}0 \cdot 10^{-17}$
	2000	$2{,}1 \cdot 10^{-04}$	$-6{,}5 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-1{,}3 \cdot 10^{-02}$	$-1{,}8 \cdot 10^{-03}$	$-5{,}1 \cdot 10^{-04}$	$4{,}9 \cdot 10^{-17}$	$8{,}7 \cdot 10^{-18}$
	3000	$-1{,}7 \cdot 10^{-04}$	$4{,}0 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-6{,}3 \cdot 10^{-03}$	$-1{,}6 \cdot 10^{-03}$	$-6{,}2 \cdot 10^{-04}$	$-3{,}1 \cdot 10^{-17}$	$-1{,}8 \cdot 10^{-17}$
	5000	$-1{,}4 \cdot 10^{-05}$	$-3{,}0 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-4{,}7 \cdot 10^{-03}$	$-1{,}1 \cdot 10^{-03}$	$-2{,}2 \cdot 10^{-04}$	$-2{,}6 \cdot 10^{-17}$	$5{,}7 \cdot 10^{-18}$
	7500	$2{,}0 \cdot 10^{-04}$	$4{,}9 \cdot 10^{-05}$	$-9{,}9 \cdot 10^{-01}$	$-2{,}9 \cdot 10^{-04}$	$-4{,}0 \cdot 10^{-04}$	$1{,}3 \cdot 10^{-04}$	$-6{,}6 \cdot 10^{-17}$	$3{,}6 \cdot 10^{-18}$
$\ell = 2$	500	$1{,}6 \cdot 10^{-03}$	$8{,}0 \cdot 10^{-04}$	$-9{,}6 \cdot 10^{-01}$	$-5{,}6 \cdot 10^{-02}$	$-4{,}2 \cdot 10^{-03}$	$-2{,}4 \cdot 10^{-03}$	$-4{,}5 \cdot 10^{-05}$	$4{,}3 \cdot 10^{-06}$
	700	$1{,}6 \cdot 10^{-04}$	$8{,}8 \cdot 10^{-04}$	$-9{,}6 \cdot 10^{-01}$	$-3{,}2 \cdot 10^{-02}$	$-2{,}3 \cdot 10^{-03}$	$-5{,}6 \cdot 10^{-04}$	$-1{,}9 \cdot 10^{-04}$	$-2{,}6 \cdot 10^{-04}$
	1000	$2{,}9 \cdot 10^{-04}$	$2{,}3 \cdot 10^{-04}$	$-9{,}7 \cdot 10^{-01}$	$-2{,}1 \cdot 10^{-02}$	$-2{,}1 \cdot 10^{-03}$	$-3{,}3 \cdot 10^{-04}$	$5{,}8 \cdot 10^{-06}$	$-8{,}2 \cdot 10^{-07}$
	2000	$3{,}3 \cdot 10^{-04}$	$-7{,}9 \cdot 10^{-04}$	$-9{,}7 \cdot 10^{-01}$	$-1{,}1 \cdot 10^{-02}$	$-1{,}8 \cdot 10^{-03}$	$-5{,}8 \cdot 10^{-04}$	$-1{,}7 \cdot 10^{-04}$	$-1{,}3 \cdot 10^{-05}$
	3000	$-1{,}3 \cdot 10^{-04}$	$3{,}2 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-9{,}1 \cdot 10^{-03}$	$-1{,}6 \cdot 10^{-03}$	$-5{,}3 \cdot 10^{-04}$	$7{,}3 \cdot 10^{-05}$	$3{,}0 \cdot 10^{-06}$
	5000	$-4{,}5 \cdot 10^{-06}$	$-3{,}0 \cdot 10^{-04}$	$-9{,}8 \cdot 10^{-01}$	$-5{,}3 \cdot 10^{-03}$	$-1{,}1 \cdot 10^{-03}$	$-2{,}7 \cdot 10^{-04}$	$-6{,}4 \cdot 10^{-05}$	$-3{,}3 \cdot 10^{-05}$
	7500	$2{,}1 \cdot 10^{-04}$	$2{,}8 \cdot 10^{-04}$	$-9{,}9 \cdot 10^{-01}$	$-1{,}5 \cdot 10^{-03}$	$-4{,}6 \cdot 10^{-04}$	$1{,}6 \cdot 10^{-04}$	$-3{,}5 \cdot 10^{-05}$	$-2{,}0 \cdot 10^{-05}$

Tabla A.6: BIAS de $\hat{\beta}_0,\ \hat{\sigma_1},\ \hat{\sigma_0}$ *y* $\hat{\mu}$ con $\ell = 0,\ \frac{1}{2},\ 1$ *y* 2 en experimento 2.

		$EMSE(\hat{\mu})^{ML}$			$EMSE(\hat{\mu})^{REML}$		
		Normal	Gamma	Weibull	Normal	Gamma	Weibull
$\ell = 0$	500	0,987 6	0,989 2	0,985 6	0,987 5	0,989 1	0,985 6
	700	0,989 4	0,989 8	0,990 1	0,989 4	0,989 7	0,990 1
	1000	0,989 6	0,989 5	0,989 1	0,989 6	0,989 5	0,989 1
	2000	0,989 6	0,991 0	0,989 2	0,989 6	0,991 0	0,989 2
	3000	0,989 5	0,990 1	0,990 7	0,989 5	0,990 1	0,990 7
	5000	0,989 8	0,990 1	0,989 8	0,989 8	0,990 1	0,989 8
	7500	0,990 3	0,990 4	0,990 9	0,990 3	0,990 7	0,991 0
$\ell = \frac{1}{2}$	500	1,313 2	1,315 2	1,310 7	1,313 1	1,315 2	1,310 6
	700	1,310 5	1,310 9	1,311 8	1,310 4	1,310 9	1,311 7
	1000	1,306 8	1,306 9	1,306 4	1,306 8	1,306 9	1,306 4
	2000	1,302 5	1,304 5	1,302 0	1,302 4	1,304 5	1,302 0
	3000	1,300 9	1,301 7	1,302 7	1,300 9	1,301 7	1,302 7
	5000	1,300 1	1,300 6	1,300 1	1,300 1	1,300 6	1,300 1
	7500	1,300 1	1,300 4	1,301 1	1,300 1	1,300 4	1,301 1
$\ell = 1$	500	1,779 3	1,781 9	1,775 9	1,779 1	1,781 7	1,775 7
	700	1,768 0	1,768 7	1,770 2	1,768 0	1,768 6	1,770 1
	1000	1,757 4	1,758 0	1,757 3	1,757 4	1,757 9	1,757 2
	2000	1,745 4	1,748 3	1,744 8	1,745 4	1,748 3	1,744 7
	3000	1,741 1	1,742 2	1,743 6	1,741 1	1,742 2	1,743 6
	5000	1,738 4	1,739 0	1,738 3	1,738 4	1,739 0	1,738 3
	7500	1,737 5	1,738 0	1,739 0	1,737 5	1,738 0	1,739 0
$\ell = 2$	500	3,441 9	3,444 7	3,433 7	3,441 3	3,444 2	3,433 2
	700	3,389 3	3,388 7	3,392 9	3,389 1	3,388 5	3,392 7
	1000	3,345 7	3,346 7	3,345 2	3,345 6	3,346 6	3,345 1
	2000	3,297 6	3,301 7	3,294 2	3,297 6	3,301 7	3,294 1
	3000	3,280 4	3,280 7	3,283 5	3,280 4	3,280 7	3,283 5
	5000	3,268 4	3,268 0	3,266 1	3,268 4	3,268 0	3,266 1
	7500	3,263 5	3,262 5	3,264 9	3,263 5	3,262 5	3,264 9

Tabla A.7: EMSE de $\hat{\mu}$ con $\ell = 0,\ \frac{1}{2},\ 1\ y\ 2$ en experimento 3.

		$BIAS(\hat{\mu})^{ML}$			$BIAS(\hat{\mu})^{REML}$		
		Normal	Gamma	Weibull	Normal	Gamma	Weibull
$\ell = 0$	500	$1{,}4 \cdot 10^{-06}$	$5{,}5 \cdot 10^{-03}$	$5{,}4 \cdot 10^{-03}$	$-8{,}2 \cdot 10^{-07}$	$5{,}4 \cdot 10^{-03}$	$5{,}2 \cdot 10^{-03}$
	700	$1{,}0 \cdot 10^{-04}$	$5{,}2 \cdot 10^{-03}$	$5{,}1 \cdot 10^{-03}$	$1{,}6 \cdot 10^{-04}$	$5{,}5 \cdot 10^{-03}$	$5{,}4 \cdot 10^{-03}$
	1000	$1{,}3 \cdot 10^{-06}$	$5{,}7 \cdot 10^{-03}$	$5{,}7 \cdot 10^{-03}$	$1{,}0 \cdot 10^{-06}$	$5{,}3 \cdot 10^{-03}$	$5{,}3 \cdot 10^{-03}$
	2000	$1{,}6 \cdot 10^{-06}$	$5{,}6 \cdot 10^{-03}$	$5{,}5 \cdot 10^{-03}$	$1{,}3 \cdot 10^{-06}$	$5{,}4 \cdot 10^{-03}$	$5{,}3 \cdot 10^{-03}$
	3000	$-4{,}3 \cdot 10^{-06}$	$5{,}3 \cdot 10^{-03}$	$5{,}3 \cdot 10^{-03}$	$-4{,}2 \cdot 10^{-06}$	$5{,}2 \cdot 10^{-03}$	$5{,}2 \cdot 10^{-03}$
	5000	$5{,}8 \cdot 10^{-06}$	$5{,}0 \cdot 10^{-03}$	$5{,}0 \cdot 10^{-03}$	$5{,}4 \cdot 10^{-06}$	$4{,}9 \cdot 10^{-03}$	$4{,}9 \cdot 10^{-03}$
	7500	$-9{,}2 \cdot 10^{-06}$	$4{,}9 \cdot 10^{-03}$	$5{,}0 \cdot 10^{-03}$	$5{,}0 \cdot 10^{-06}$	$4{,}8 \cdot 10^{-03}$	$4{,}8 \cdot 10^{-03}$
$\ell = \frac{1}{2}$	500	$2{,}6 \cdot 10^{-06}$	$3{,}6 \cdot 10^{-03}$	$3{,}6 \cdot 10^{-03}$	$2{,}7 \cdot 10^{-06}$	$3{,}6 \cdot 10^{-03}$	$3{,}6 \cdot 10^{-03}$
	700	$7{,}4 \cdot 10^{-05}$	$3{,}6 \cdot 10^{-03}$	$3{,}6 \cdot 10^{-03}$	$7{,}0 \cdot 10^{-05}$	$3{,}0 \cdot 10^{-03}$	$3{,}0 \cdot 10^{-03}$
	1000	$8{,}6 \cdot 10^{-07}$	$3{,}7 \cdot 10^{-03}$	$3{,}7 \cdot 10^{-03}$	$8{,}2 \cdot 10^{-07}$	$3{,}4 \cdot 10^{-03}$	$3{,}4 \cdot 10^{-03}$
	2000	$5{,}5 \cdot 10^{-07}$	$3{,}0 \cdot 10^{-03}$	$3{,}9 \cdot 10^{-03}$	$5{,}0 \cdot 10^{-07}$	$3{,}8 \cdot 10^{-03}$	$3{,}8 \cdot 10^{-03}$
	3000	$-1{,}5 \cdot 10^{-06}$	$3{,}8 \cdot 10^{-03}$	$3{,}9 \cdot 10^{-03}$	$-1{,}4 \cdot 10^{-06}$	$3{,}7 \cdot 10^{-03}$	$3{,}7 \cdot 10^{-03}$
	5000	$5{,}8 \cdot 10^{-06}$	$3{,}6 \cdot 10^{-03}$	$3{,}6 \cdot 10^{-03}$	$5{,}7 \cdot 10^{-06}$	$3{,}5 \cdot 10^{-03}$	$3{,}5 \cdot 10^{-03}$
	7500	$3{,}4 \cdot 10^{-06}$	$3{,}6 \cdot 10^{-03}$	$3{,}6 \cdot 10^{-03}$	$3{,}4 \cdot 10^{-06}$	$3{,}5 \cdot 10^{-03}$	$3{,}5 \cdot 10^{-03}$
$\ell = 1$	500	$-2{,}7 \cdot 10^{-18}$	$4{,}4 \cdot 10^{-17}$	$2{,}0 \cdot 10^{-17}$	$-1{,}2 \cdot 10^{-17}$	$7{,}8 \cdot 10^{-17}$	$4{,}8 \cdot 10^{-17}$
	700	$-6{,}9 \cdot 10^{-18}$	$1{,}5 \cdot 10^{-17}$	$2{,}4 \cdot 10^{-17}$	$9{,}3 \cdot 10^{-18}$	$1{,}0 \cdot 10^{-17}$	$2{,}7 \cdot 10^{-17}$
	1000	$-2{,}0 \cdot 10^{-17}$	$-3{,}1 \cdot 10^{-17}$	$-3{,}8 \cdot 10^{-17}$	$-8{,}0 \cdot 10^{-18}$	$-1{,}3 \cdot 10^{-17}$	$-1{,}9 \cdot 10^{-17}$
	2000	$8{,}7 \cdot 10^{-18}$	$-9{,}7 \cdot 10^{-18}$	$-1{,}9 \cdot 10^{-17}$	$4{,}5 \cdot 10^{-18}$	$-1{,}8 \cdot 10^{-17}$	$-1{,}9 \cdot 10^{-18}$
	3000	$-1{,}8 \cdot 10^{-17}$	$-2{,}6 \cdot 10^{-17}$	$-3{,}3 \cdot 10^{-17}$	$-9{,}2 \cdot 10^{-18}$	$-3{,}5 \cdot 10^{-17}$	$-2{,}4 \cdot 10^{-17}$
	5000	$5{,}7 \cdot 10^{-18}$	$8{,}3 \cdot 10^{-19}$	$5{,}2 \cdot 10^{-18}$	$7{,}5 \cdot 10^{-18}$	$-9{,}4 \cdot 10^{-19}$	$-3{,}0 \cdot 10^{-18}$
	7500	$3{,}6 \cdot 10^{-18}$	$2{,}9 \cdot 10^{-17}$	$2{,}4 \cdot 10^{-17}$	$5{,}6 \cdot 10^{-18}$	$2{,}3 \cdot 10^{-17}$	$3{,}7 \cdot 10^{-17}$
$\ell = 2$	500	$4{,}3 \cdot 10^{-06}$	$-2{,}5 \cdot 10^{-04}$	$-2{,}0 \cdot 10^{-04}$	$-1{,}2 \cdot 10^{-05}$	$-2{,}0 \cdot 10^{-04}$	$-1{,}3 \cdot 10^{-04}$
	700	$-2{,}6 \cdot 10^{-04}$	$7{,}8 \cdot 10^{-05}$	$8{,}2 \cdot 10^{-05}$	$-2{,}4 \cdot 10^{-04}$	$7{,}2 \cdot 10^{-05}$	$8{,}3 \cdot 10^{-05}$
	1000	$-8{,}2 \cdot 10^{-07}$	$-6{,}4 \cdot 10^{-05}$	$-1{,}1 \cdot 10^{-04}$	$-8{,}7 \cdot 10^{-07}$	$-6{,}8 \cdot 10^{-05}$	$-1{,}8 \cdot 10^{-04}$
	2000	$-1{,}3 \cdot 10^{-05}$	$5{,}6 \cdot 10^{-05}$	$-6{,}2 \cdot 10^{-05}$	$-1{,}3 \cdot 10^{-05}$	$5{,}1 \cdot 10^{-05}$	$-6{,}9 \cdot 10^{-05}$
	3000	$3{,}0 \cdot 10^{-06}$	$1{,}6 \cdot 10^{-06}$	$-2{,}8 \cdot 10^{-05}$	$3{,}9 \cdot 10^{-06}$	$2{,}2 \cdot 10^{-06}$	$-2{,}9 \cdot 10^{-05}$
	5000	$-3{,}3 \cdot 10^{-05}$	$-1{,}5 \cdot 10^{-06}$	$-2{,}8 \cdot 10^{-05}$	$-3{,}3 \cdot 10^{-05}$	$-1{,}6 \cdot 10^{-06}$	$-2{,}8 \cdot 10^{-05}$
	7500	$-2{,}0 \cdot 10^{-05}$	$-3{,}7 \cdot 10^{-05}$	$2{,}7 \cdot 10^{-05}$	$-2{,}0 \cdot 10^{-05}$	$-3{,}6 \cdot 10^{-05}$	$2{,}7 \cdot 10^{-05}$

Tabla A.8: BIAS de $\hat{\mu}$ con $\ell = 0,\ \frac{1}{2},\ 1\ y\ 2$ en experimento 3.

A.2. Resúmenes numéricos del capitulo 4

		GLMlogit	LMM	CART	LDA	LSVM
1 variable	I=2	51,931 %	54,075 %	67,555 %	51,201 %	48,875 %
	I=5	51,889 %	51,622 %	64,241 %	52,136 %	50,762 %
	I=10	52,240 %	51,039 %	56,200 %	51,658 %	50,089 %
	I=25	52,251 %	51,070 %	52,897 %	52,146 %	52,947 %
	I=50	52,657 %	51,237 %	51,916 %	52,676 %	52,103 %
2 variables	I=2	56,358 %	52,330 %	54,716 %	65,258 %	68,245 %
	I=5	56,904 %	55,243 %	56,362 %	67,438 %	67,292 %
	I=10	55,357 %	53,500 %	56,493 %	68,210 %	67,656 %
	I=25	55,194 %	53,871 %	53,349 %	68,320 %	68,404 %
	I=50	57,635 %	54,088 %	56,734 %	68,628 %	68,959 %
10 variables	I=2	62,741 %	61,291 %	61,415 %	66,307 %	64,665 %
	I=5	60,866 %	58,126 %	59,703 %	66,451 %	63,214 %
	I=10	60,726 %	57,836 %	60,544 %	68,556 %	65,366 %
	I=25	61,481 %	57,679 %	59,725 %	68,305 %	68,495 %
	I=50	63,365 %	58,355 %	58,722 %	67,941 %	66,307 %
50 variables	I=2	62,105 %	61,995 %	61,186 %	63,159 %	63,880 %
	I=5	65,919 %	65,998 %	65,894 %	66,427 %	67,984 %
	I=10	67,477 %	66,581 %	66,060 %	67,953 %	67,147 %
	I=25	69,195 %	66,688 %	64,870 %	70,248 %	67,438 %
	I=50	71,174 %	67,859 %	62,206 %	71,935 %	68,683 %
100 variables	I=2	70,669 %	70,391 %	70,387 %	69,809 %	70,705 %
	I=5	71,655 %	71,686 %	71,460 %	69,778 %	72,652 %
	I=10	74,456 %	74,455 %	74,049 %	73,816 %	73,958 %
	I=25	73,528 %	72,435 %	70,455 %	73,686 %	72,101 %
	I=50	76,047 %	74,335 %	68,514 %	76,189 %	73,478 %

Tabla A.9: Tasa de acierto respecto al número de variables para los métodos GLMlogit, LMM, LDA, CART y LSVM

		GLMlogit	LMM	CART	LDA	LSVM
1 variable	I=2	48,335 %	47,470 %	63,315 %	37,055 %	59,785 %
	I=5	49,382 %	47,967 %	66,183 %	40,740 %	60,651 %
	I=10	49,628 %	47,608 %	67,446 %	46,907 %	59,315 %
	I=25	49,714 %	47,770 %	67,914 %	49,049 %	58,049 %
	I=50	49,766 %	47,879 %	68,387 %	49,326 %	56,754 %
2 variables	I=2	43,421 %	42,130 %	53,669 %	39,455 %	49,573 %
	I=5	44,595 %	42,727 %	54,684 %	40,022 %	53,156 %
	I=10	45,618 %	42,668 %	56,198 %	42,858 %	54,382 %
	I=25	45,634 %	42,745 %	55,891 %	45,216 %	53,425 %
	I=50	45,801 %	43,000 %	56,022 %	45,810 %	54,071 %
10 variables	I=2	39,667 %	39,051 %	48,066 %	38,002 %	46,520 %
	I=5	42,638 %	41,305 %	52,304 %	38,565 %	49,131 %
	I=10	43,915 %	41,538 %	53,792 %	40,483 %	54,084 %
	I=25	45,059 %	42,405 %	55,944 %	43,729 %	54,012 %
	I=50	44,863 %	42,225 %	55,599 %	45,822 %	53,981 %
50 variables	I=2	36,533 %	36,533 %	43,959 %	37,085 %	41,188 %
	I=5	35,584 %	35,584 %	42,864 %	36,957 %	41,222 %
	I=10	38,085 %	37,899 %	46,166 %	37,718 %	42,804 %
	I=25	40,834 %	39,408 %	49,858 %	40,609 %	43,757 %
	I=50	41,637 %	39,806 %	50,935 %	45,459 %	43,624 %
100 variables	I=2	34,049 %	34,049 %	40,534 %	34,633 %	37,810 %
	I=5	34,115 %	34,115 %	40,286 %	35,151 %	39,205 %
	I=10	32,037 %	32,037 %	38,313 %	34,116 %	37,001 %
	I=25	35,920 %	35,917 %	43,392 %	37,882 %	42,392 %
	I=50	37,006 %	36,483 %	44,602 %	42,283 %	42,936 %

Tabla A.10: Error cuadratico medio(RMSE) respecto al número de variables para los métodos GLMlogit, LMM, LDA, CART y LSVM

		GLMlogit	LMM	CART	LDA	LSVM
1 variable	I=2	0,0295	0,11014	0,0315	0,02746	0,2918
	I=5	0,06334	0,1715	0,06744	0,06	1,6657
	I=10	0,12046	0,27056	0,13082	0,11406	6,3878
	I=25	0,28472	0,60878	0,32372	0,29316	39,2177
	I=50	0,552	1,10248	0,65072	0,5721	177,6148
2 variables	I=2	0,03778	0,10368	0,04382	0,0286	0,2327
	I=5	0,08518	0,16254	0,09218	0,06292	1,4938
	I=10	0,17014	0,27684	0,1813	0,12406	5,5823
	I=25	0,3822	0,56252	0,45488	0,30414	31,7275
	I=50	0,72612	1,0511	0,94908	0,59702	131,3565
10 variables	I=2	0,05298	0,14178	0,05928	0,04968	1,3513
	I=5	0,11872	0,21936	0,12834	0,12398	5,8608
	I=10	0,19658	0,38734	0,24694	0,26006	16,1669
	I=25	0,43192	0,73794	0,57226	0,86124	83,9742
	I=50	0,84538	1,35344	1,1123	1,39646	494,2334
50 variables	I=2	0,16944	0,18062	0,13644	0,15388	1,1054
	I=5	0,32096	0,51194	0,2813	0,40492	11,7604
	I=10	0,52172	0,89174	0,51536	1,20508	431,8181
	I=25	1,17224	1,8861	1,19688	2,21452	1113,6922
	I=50	2,24336	3,63972	2,51214	4,933	3094,4282
100 variables	I=2	0,47136	0,2678	0,26724	0,29336	1,517
	I=5	0,89908	0,63	0,52936	0,74004	10,0025
	I=10	1,5527	2,01388	1,03634	1,63754	84,5579
	I=25	3,63716	4,43372	2,28854	4,40286	2331,9372
	I=50	6,97114	8,79636	4,69012	9,47974	6310,9878

Tabla A.11: Tiempo total respecto al número de variables para los métodos GLMlogit, LMM, LDA, CART y LSVM

A.3. Resúmenes numéricos del Capitulo 5

Descripción	**Tipo de Variable**
Datos adicionales de la vivienda en la que tiene su domicilio habitual: Titularidad (Clave)	Categórica
Datos adicionales de la vivienda en la que tiene su domicilio habitual: % participación primer declarante	Numérica (en %)
Datos adicionales de la vivienda en la que tiene su domicilio habitual: % participación cónyuge	Numérica (en %)
Código postal	Categórica
Tipo de declaración	Categórica
Ejercicio de nacimiento del cónyuge	Numérica
Ejercicio de nacimiento del declarante	Numérica
Estado civil del declarante	Categórica
Factor de elevación de la muestra	Categórica
Modelo de declaración	Categórica
Grado de minusvalía del cónyuge	Categórica
Grado de minusvalía del declarante	Categórica
Número de ascendientes	Numérica
Número de ascendientes sin minusvalía	Numérica
Número de ascendientes con minusvalía ≥ 33 y $< 65\,\%$ sin movilidad reducida	Numérica
Número de ascendientes con minusvalía ≥ 33 y $< 65\,\%$ con movilidad reducida	Numérica
Número de descendientes con minusvalía $\geq 65\,\%$	Numérica
Número total de descendientes	Numérica
Número de descendientes < 3 años	Numérica
Número de descendientes ≥ 16 y < 18 años	Numérica
Número de descendientes ≥ 18 y < 25 años	Numérica
Número de descendientes ≥ 3 y < 16 años	Numérica
Número de descendientes con edad desconocida	Numérica
Número de descendientes ≥ 65 años	Numérica
Número de descendientes sin minusvalía	Numérica
Número de descendientes con minusvalía ≥ 33 y $< 65\,\%$ sin movilidad reducida	Numérica
Número de descendientes con minusvalía ≥ 33 y $< 65\,\%$ con movilidad reducida	Numérica
Número de descendientes con minusvalía $\geq 65\,\%$	Numérica
Número de ascendientes con minusvalía	Numérica
Número de descendientes con minusvalía	Numérica
Número de ascendientes > 65 años	Numérica
Número de ascendientes > 75 años	Numérica
Rendimientos del Trabajo. Ingresos íntegros. Dinerarios	Numérica (Dineraria)
Rendimientos del Trabajo. Cuotas satisfechas a sindicatos.	Numérica (Dineraria)
Tipo de actividad/es realizada/s: Clave indicativa	Categórica

Continúa en la página siguiente...

Descripción	Tipo de Variable
Rendimientos de actividades económicas en régimen de E.D.: epígrafe IAE	Categórica
Modalidad aplicable para la determinación del rendimiento neto	Categórica
Rendimientos del Trabajo. Cuotas satisfechas a colegios profesionales.	Numérica (Dineraria)
Rendimientos del Trabajo. Gastos defensa jurídica	Numérica (Dineraria)
Rendimientos del Trabajo. Total gastos deducibles.	Numérica (Dineraria)
Rendimientos del Trabajo. Rendimiento neto.	Numérica (Dineraria)
Rendimientos de actividades económicas en estimación directa. Rendimiento neto reducido.	Numérica (Dineraria)
Rendimientos del Trabajo. Reducción rendimientos Copa America 2007.	Numérica (Dineraria)
Rendimientos de actividades económicas excepto agrícolas en régimen E.O.: epígrafe IAE	Numérica (Dineraria)
Rendimientos de actividades económicas (excepto agrícolas, ganaderas y forestales) en estimación objetiva. Rendimiento neto reducido.	Numérica (Dineraria)
Rendimientos de actividades agrícolas, ganaderas y forestales en estimación objetiva. Rendimiento neto reducido.	Numérica (Dineraria)
Rendimientos del Trabajo. Rendimiento neto reducido.	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Ingresos íntegros. Intereses de cuentas	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Ingresos íntegros. Intereses de activos	Numérica (Dineraria)
Régimen de atribución de rentas. Rendimientos capital mobiliario.	Numérica (Dineraria)
Régimen de atribución de rentas. Rendimientos capital inmobiliario.	Numérica (Dineraria)
Régimen de atribución de rentas. Rendimientos actividades económicas.	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Ingresos íntegros. Dividendos	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Ingresos íntegros. Letras	Numérica (Dineraria)
Imputaciones de agrupaciones de interés económico y uniones temporales de empresas. Imputación de bases imponibles y deducciones.	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Ingresos íntegros. Otros activos	Numérica (Dineraria)
Imputaciones de rentas positivas en el régimen de transparencia fiscal internacional.	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Ingresos íntegros. Contratos de seguro	Numérica (Dineraria)
Imputación de rentas por la cesión de derechos de imagen.	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Ingresos íntegros. Otros rendimientos	Numérica (Dineraria)
Imputación de rentas derivadas participación Instituciones Inversión Colectiva en paraísos fiscales	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Ingresos íntegros. Total	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Gastos deducibles. Gastos de administración	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Gastos deducibles. Otros gastos	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Gastos deducibles. Total	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Rendimiento neto	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Reducciones Art. 24.2 y 94 de la Ley del Impuesto.	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Reducciones Disp.Transitoria 5.	Numérica (Dineraria)
Rendimientos del Capital Mobiliario. Rendimiento Neto Reducido.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 1: Contribuyente titular	Categórica

Continúa en la página siguiente...

Descripción	**Tipo de Variable**
Identificación de inmuebles urbanos, inmueble 1: Titularidad (%)	Numérica (en %)
Identificación de inmuebles urbanos, inmueble 1: Uso	Categórica
Identificación de inmuebles urbanos, inmueble 1: Renta imputada.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 2: Contribuyente titular	Categórica
Ganancias y pérdidas patrimoniales. Suma de ganancias patrimoniales (parte general)	Numérica (Dineraria)
Ganancias y pérdidas patrimoniales. Suma de pérdidas patrimoniales (parte general)	Numérica (Dineraria)
Ganancias y pérdidas patrimoniales. Suma de ganancias patrimoniales (parte especial)	Numérica (Dineraria)
Ganancias y pérdidas patrimoniales. Suma de pérdidas patrimoniales (parte especial)	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 2: Titularidad (%)	Numérica (Dineraria)
Renta del período. Saldo neto positivo de ganancias y pérdidas patrimoniales imputables a 2005 a integrar en la Parte General de la Base Imponible.	Numérica (Dineraria)
Renta del período. Saldo neto negativo de ganancias y pérdidas patrimoniales de 2001 a 2004 a integrar en la Parte General de la Base Imponible.	Numérica (Dineraria)
Renta del período. Saldo neto de rendimientos e imputaciones de rentas.	Numérica (Dineraria)
Renta del período. Compensaciones. Resto saldo neto negativo de ganancias y perdidas patrimoniales 2001 a 2004 a integrar en la Parte General de la Base Imponible.	Numérica (Dineraria)
Renta del período. Compensaciones. Saldo neto negativo de ganancias y pérdidas patrimoniales imputables 2005 a integrar en la Parte General de la Base Imponible.	Numérica (Dineraria)
Renta del período. Parte general de la renta del período	Numérica (Dineraria)
Renta del período. Saldo neto positivo de ganancias y pérdidas patrimoniales imputables a 2005 integrar en la Parte Especial de la Base Imponible.	Numérica (Dineraria)
Renta del período. Saldo neto negativo de ganancias y pérdidas patrimoniales de 2001 a 2004 a integrar en la Parte Especial de la Base Imponible.	Numérica (Dineraria)
Renta del período. Parte especial de la renta del período.	Numérica (Dineraria)
Base imponible. Mínimo personal y familiar.	Numérica (Dineraria)
Base imponible. Mínimo personal y familiar aplicado a la parte general de la base imponible.	Numérica (Dineraria)
Base imponible. Parte general de la Base Imponible.	Numérica (Dineraria)
Base imponible. Resto del mínimo personal y familiar: Importe no aplicado en la minoración de la base imponible general.	Numérica (Dineraria)
Base imponible. Parte especial de la Base Imponible.	Numérica (Dineraria)
Reducciones de la base imponible. Reducción por rendimientos del trabajo.	Numérica (Dineraria)
Reducciones de la base imponible. Reducción por prolongación de la actividad laboral.	Numérica (Dineraria)
Reducciones de la base imponible. Reducción por movilidad geogrfica.	Numérica (Dineraria)
Reducciones de la base imponible. Reducción por cuidado de hijos.	Numérica (Dineraria)
Reducciones de la base imponible. Reducción por edad.	Numérica (Dineraria)
Reducciones de la base imponible. Reducción por asistencia.	Numérica (Dineraria)
Reducciones de la base imponible. Reducción por discapacidad del contribuyente.	Numérica (Dineraria)
Reducciones de la base imponible. Reducción por discapacidad de ascendientes o descendientes.	Numérica (Dineraria)

Continúa en la página siguiente...

Descripción	Tipo de Variable
Reducciones de la base imponible. Reducción por discapacidad de trabajadores activos.	Numérica (Dineraria)
Reducciones de la base imponible. Reducción por gastos de asistencia de los discapacitados.	Numérica (Dineraria)
Rendimientos del Trabajo. Ingresos íntegros. En especie.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 2: Uso	Categórica
Reducciones de la base imponible. Suma de reducciones por circunstancias laborales, personales y familiares.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 2: Renta imputada.	Numérica (Dineraria)
Reducciones de la base imponible. Reducciones por aportaciones a los patrimonios protegidos de personas con discapacidad.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 3: Contribuyente titular	Categórica
Identificación de inmuebles urbanos, inmueble 3: Titularidad (%)	Numérica (en %)
Reducciones de la base imponible. Reducciones por aportaciones a Planes de Pensiones y a Mutualidades de Previsión Social. Régimen general	Numérica (Dineraria)
Reducciones de la base imponible. Reducciones por aportaciones a Planes de Pensiones y a Mutualidades de Previsión Social del cónyuge.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 3: Uso	Categórica
Identificación de inmuebles urbanos, inmueble 3: Renta imputada.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 4: Contribuyente titular	Categórica
Reducciones de la base imponible. Reducciones por aportaciones a Planes de Pensiones y Mutualidades de Previsión Social a favor de minusvlidos.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 4: Titularidad (%)	Numérica (en %)
Rendimientos del Trabajo. Contribuciones empresariales a Planes de Pensiones y a Mutualidades de Previsión social.	Numérica (Dineraria)
Reducciones de la base imponible. Reducción por pensiones compensatorias al cónyuge y anualidades por alimentos.	Numérica (Dineraria)
Reducciones de la base imponible. Reducciones por aportaciones a Mutualidades de Previsión Social de deportistas profesionales o de alto nivel.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 4: Uso	Categórica
Base liquidable. Base liquidable general.	Numérica (Dineraria)
Base liquidable. Compensación de bases liquidables generales negativas de 2001 a 2004.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 4: Renta imputada.	Numérica (Dineraria)
Base liquidable. Base liquidable general sometida a gravamen.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 5: Contribuyente titular	Categórica
Base liquidable. Base liquidable especial.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 5: Titularidad (%)	Numérica (en %)
Rentas exentas del IRPF., excepto para determinar el tipo de gravamen aplicable a las dems rentas.	Numérica (Dineraria)
Anualidades por alimentos en favor de los hijos satisfechas por resolución judicial.	Numérica (Dineraria)
Cuota íntegra. Cuota estatal correspondiente a la base liquidable general	Numérica (Dineraria)
Cuota íntegra. Cuota autonómica o complementaria correspondiente a la base liquidable general	Numérica (Dineraria)
Cuota íntegra. Cuota estatal correspondiente a la base liquidable especial	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 5: Uso	Categórica
Continúa en la página siguiente...	

Descripción	**Tipo de Variable**
Cuota íntegra. Cuota autonómica o complementaria correspondiente a la base liquidable especial	Numérica (Dineraria)
Cuota íntegra. Cuota íntegra estatal.	Numérica (Dineraria)
Cuota íntegra. Cuota íntegra autonómica o complementaria.	Numérica (Dineraria)
Deducciones. Por inversiones o gastos en bienes de interés cultural parte estatal.	Numérica (Dineraria)
Deducciones. Por inversiones o gastos en bienes de interés cultural parte autonómica.	Numérica (Dineraria)
Deducciones. Por cantidades o bienes donados a determinadas entidades parte estatal.	Numérica (Dineraria)
Deducciones. Por cantidades o bienes donados a determinadas entidades parte autonómica.	Numérica (Dineraria)
Deducciones. Por adquisición o rehabilitación de la vivienda habitual, con financiación ajena, parte estatal.	Numérica (Dineraria)
Deducciones. Por adquisición o rehabilitación de la vivienda habitual, con financiación ajena, parte autonómica.	Numérica (Dineraria)
Deducciones. Por adquisición o rehabilitación de vivienda habitual, sin financiación ajena, parte estatal.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 5: Renta imputada.	Numérica (Dineraria)
Deducciones. Por adquisición o rehabilitación de vivienda habitual sin financiación ajena, parte autonómica.	Numérica (Dineraria)
Deducciones. Por construcción o ampliación de la vivienda habitual, parte estatal.	Numérica (Dineraria)
Deducciones. Por construcción o ampliación de la vivienda habitual, parte autonómica.	Numérica (Dineraria)
Deducciones. Por cantidades depositadas en cuenta vivienda, parte estatal.	Numérica (Dineraria)
Deducciones. Por cantidades depositadas en cuentas vivienda, parte autonómica.	Numérica (Dineraria)
Deducciones. Por adecuación de la vivienda habitual de minusvlido, con financiación ajena, parte estatal.	Numérica (Dineraria)
Deducciones. Por adecuación de la vivienda habitual de minusvlido, con financiación ajena, parte autonómica.	Numérica (Dineraria)
Deducciones. Por adecuación de la vivienda habitual de minusvlido, sin financiación ajena, parte estatal.	Numérica (Dineraria)
Deducciones. Por adecuación de la vivienda habitual de minusvlido, sin financiación ajena, parte autonómica.	Numérica (Dineraria)
Deducciones. Por incentivos y estímulos a la inversión empresarial, parte estatal.	Numérica (Dineraria)
Rendimientos del Trabajo. Aportaciones al patrimonio protegido de las personas con discapacidad.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 6: Contribuyente titular	Categórica
Deducciones. Por incentivos y estímulos a la inversión empresarial, parte autonómica.	Numérica (Dineraria)
Deducciones. Por dotaciones a la Reserva para Inversiones en Canarias, parte estatal.	Numérica (Dineraria)
Deducciones. Por dotaciones a la Reserva para Inversiones en Canarias, parte autonómica.	Numérica (Dineraria)
Deducciones. Por rendimientos derivados de la venta bienes corporales producidos en Canarias, parte estatal.	Numérica (Dineraria)
Deducciones. Por rendimientos derivados de la venta bienes corporales producidos en Canarias, parte autonómica.	Numérica (Dineraria)
Deducciones. Por rentas obtenidas en Ceuta y Melilla parte estatal.	Numérica (Dineraria)
Deducciones. Por rentas obtenidas en Ceuta y Melilla parte autonómica.	Numérica (Dineraria)
Continúa en la página siguiente...	

Descripción	Tipo de Variable
Deducciones. Por cantidades depositadas en cuentas ahorro-empresa parte estatal.	Numérica (Dineraria)
Deducciones. Por cantidades depositadas en cuentas ahorro-empresa parte autonómica.	Numérica (Dineraria)
Deducciones. Suma de deducciones autonómicas.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 6: Titularidad (%)	Numérica (en %)
Cuota líquida. Cuota líquida estatal.	Numérica (Dineraria)
Cuota líquida. Cuota líquida autonómica.	Numérica (Dineraria)
Cuota líquida. Importe de las deducciones de 1996 y ejercicios anteriores a las que se ha perdido el derecho.	Numérica (Dineraria)
Cuota líquida. Intereses demora de deducciones de 1996 y ejercicios anteriores a las que se ha perdido el derecho.	Numérica (Dineraria)
Cuota líquida. 85 % del importe de las deducciones generales de 1997 a 2004 a las que se ha perdido el derecho.	Numérica (Dineraria)
Cuota líquida. Intereses demora de deducciones generales de 1997 a 2004 a las que se ha perdido el derecho.	Numérica (Dineraria)
Cuota líquida. 15 % del importe de las deducciones generales 1997 a 2004 a las que se ha perdido el derecho.	Numérica (Dineraria)
Cuota líquida. Intereses de demora de deducciones generales de 1997 a 2004 a las que se ha perdido el derecho.	Numérica (Dineraria)
Cuota líquida. Importe de las deducciones autonómicas de 1998 a 2004 a las que se ha perdido el derecho.	Numérica (Dineraria)
Cuota líquida. Intereses demora de deducciones autonómicas de 1998 a 2004 a las que se ha perdido el derecho.	Numérica (Dineraria)
Cuota líquida. Cuota líquida estatal incrementada.	Numérica (Dineraria)
Cuota líquida. Cuota líquida autonómica incrementada.	Numérica (Dineraria)
Cuota líquida. Cuota líquida incrementada total.	Numérica (Dineraria)
Deducción doble imposición ejercicio 2005	Numérica (Dineraria)
Deducción por doble imposición de dividendos: importe que se aplica en esta declaración.	Numérica (Dineraria)
Deducción por doble imposición internacional, por las rentas obtenidas y gravadas en el extranjero.	Numérica (Dineraria)
Deducción por doble imposición internacional en los supuestos de aplicación del régimen de transparencia fiscal internacional.	Numérica (Dineraria)
Deducción por doble imposición en los supuestos de aplicaciones del régimen de imputación de rentas derivadas de la cesión de derechos de imagen.	Numérica (Dineraria)
Compensación fiscal a los contribuyentes arrendatarios de su vivienda habitual.	Numérica (Dineraria)
Compensación fiscal por deducción en la adquisición de la vivienda habitual.	Numérica (Dineraria)
Retenciones deducibles correspondientes a rendimientos bonificados.	Numérica (Dineraria)
Cuota resultante de la autoliquidación.	Numérica (Dineraria)
Retenciones e ingresos a cuenta. Por rendimientos del trabajo.	Numérica (Dineraria)
Retenciones e ingresos a cuenta. Por rendimientos del capital mobiliario.	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 6: Uso	Categórica
Retenciones e ingresos a cuenta. Por arrendamientos de inmuebles urbanos.	Numérica (Dineraria)
Retenciones e ingresos a cuenta. Por rendimientos de actividades económicas.	Numérica (Dineraria)
Pagos fraccionados realizados por actividades económicas.	Numérica (Dineraria)

Continúa en la página siguiente...

Descripción	**Tipo de Variable**
Retenciones e ingresos a cuenta atribuidos. Por aplicación del régimen especial de atribución de rentas.	Numérica (Dineraria)
Retenciones e ingresos a cuenta imputados. Por imputación de las agrupaciones de interés económico y uniones temporales de empresas.	Numérica (Dineraria)
Ingresos a cuenta. Por imputaciones de rentas derivadas de la cesión de derechos de imagen.	Numérica (Dineraria)
Suma de retenciones e ingresos a cuenta por ganancias patrimoniales.	Numérica (Dineraria)
Importe deducible por bonificaciones otorgadas conforme al programa PREVER.	Numérica (Dineraria)
Cuotas del IRNR de contribuyentes que han adquirido dicha condición por cambio de residencia.	Numérica (Dineraria)
Retenciones a cuenta efectivamente practicadas en virtud del artículo 11 de la Directiva 2003/48/CE del Consejo, de 3 de junio de 2003	Numérica (Dineraria)
Identificación de inmuebles urbanos, inmueble 6: Renta imputada.	Numérica (Dineraria)
Total pagos a cuenta.	Numérica (Dineraria)
Cuota diferencial.	Numérica (Dineraria)
Deducción por maternidad: importe de la deducción.	Numérica (Dineraria)
Deducción por maternidad: importe del abono anticipado de la deducción por maternidad.	Numérica (Dineraria)
Devoluciones acordadas por la Administración por tramitación de la solicitud de devolución (modelo 104) correspondientes a 2005.	Numérica (Dineraria)
Resultado de la declaración.	Numérica (Dineraria)
Identificación de inmuebles urbanos e imputación, si procede, de rentas inmobiliarias. Total rentas imputadas	Numérica (Dineraria)
Rendimientos del Trabajo. Reducciones (Art. 17, apartados 2 y 3 y 94 de la Ley del Impuesto).	Numérica (Dineraria)
Rendimientos del Capital Inmobiliario. Ingresos íntegros.	Numérica (Dineraria)
Partidas del modelo de IRPF de 2005 (DEDUCCIONES AUTONÓMICAS).	Numérica (Dineraria)
Rendimientos del Capital Inmobiliario. Gastos deducibles.	Numérica (Dineraria)
Rendimientos del Capital Inmobiliario. Rendimiento neto.	Numérica (Dineraria)
Rendimientos del Capital Inmobiliario. Reducciones Artículo 21.2.	Numérica (Dineraria)
Rendimientos del Capital Inmobiliario. Reducciones Artículo 21.3.	Numérica (Dineraria)
Rendimientos del Capital Inmobiliario. Rendimiento mínimo computable en caso de parentesco.	Numérica (Dineraria)
Rendimientos del Trabajo. Cotizaciones Seguridad Social o a Mutualidades Generales de Funcionarios, detracciones por derechos pasivos y cotizaciones a Colegios de Huérfanos o entidades similares.	Numérica (Dineraria)
Rendimientos del Capital Inmobiliario. Rendimiento Neto Reducido.	Numérica (Dineraria)
Provincia	Categórica
Variable de renta utilizada para el muestreo	Numérica (Dineraria)
Sexo del declarante	Categórica
Tramo de renta desagregado	Categórica
Tramo de renta agregado	Categórica

Tabla A.12: Variables Iniciales de la muestra

A.4. Resúmenes numéricos de la prueba de ajuste de las betas

Variables	GLMlogit		LMM	
	Sesgo	EMSE	Sesgo	EMSE
prov0	0.000000	0.000000	-24.590463	604.714815
prov2	0.127432	0.016245	-24.244646	587.826930
prov3	0.200004	0.040007	-24.147729	583.136841
prov4	0.093800	0.008805	-24.295241	590.282647
prov5	0.053952	0.002917	-24.334685	592.201016
prov6	0.119314	0.014242	-24.253886	588.274965
prov7	0.119314	0.014242	-24.251793	588.173361
prov8	0.196257	0.038521	-24.150389	583.265278
prov9	0.121754	0.014831	-24.248243	588.001340
prov10	0.121902	0.014865	-24.248457	588.011619
prov11	0.155499	0.024186	-24.204084	585.861579
prov12	0.136754	0.018707	-24.230664	587.149030
prov13	0.134587	0.018119	-24.234443	587.332198
prov14	0.129816	0.016858	-24.239827	587.593265
prov15	0.165099	0.027263	-24.192270	585.289983
prov16	0.063514	0.004040	-24.326232	591.789596
prov17	0.127829	0.016346	-24.239693	587.586730
prov18	0.165720	0.027469	-24.193563	585.352581
prov19	0.068391	0.004684	-24.321907	591.579113
prov21	0.095350	0.009099	-24.284243	589.748584
prov22	0.060478	0.003663	-24.331459	592.043830
prov23	0.144174	0.020791	-24.222495	586.753302
prov24	0.106103	0.011264	-24.265320	588.829833
prov25	0.124539	0.015517	-24.247237	587.952414
prov26	0.092335	0.008532	-24.286837	589.874488
prov27	0.112917	0.012756	-24.266076	588.866440
prov28	0.212599	0.045203	-24.126943	582.133387
prov29	0.147052	0.021630	-24.218388	586.554345
prov30	0.156140	0.024385	-24.205734	585.941508
prov32	0.107109	0.011478	-24.267105	588.916421
prov33	0.158024	0.024977	-24.200726	585.699170
prov34	0.061035	0.003733	-24.331883	592.064478
prov35	0.118816	0.014123	-24.250964	588.133119
prov36	0.119763	0.014349	-24.250590	588.115132
prov37	0.077422	0.006000	-24.307952	590.900564
prov38	0.109280	0.011948	-24.264870	588.807856
prov39	0.082229	0.006768	-24.303645	590.691310
prov40	0.024503	0.000608	-24.374748	594.152382
prov41	0.163370	0.026694	-24.193331	585.341230
prov42	0.033279	0.001115	-24.359121	593.390758
prov43	0.157404	0.024782	-24.201109	585.717597
prov44	0.034044	0.001165	-24.360325	593.449386
prov45	0.151155	0.022854	-24.212939	586.290358
prov46	0.192198	0.036945	-24.154563	583.466878
prov47	0.122297	0.014962	-24.245953	587.890300
prov49	0.056252	0.003171	-24.330750	592.009392
prov50	0.186911	0.034941	-24.162966	583.872948
prov51	-0.100356	0.010080	-24.569280	603.673367
prov99	-0.034226	0.001190	-24.415240	596.127966
estcv2	-0.135019	0.018231	-0.193430	0.037416
estcv3	-0.139069	0.019341	-0.184472	0.034031
estcv4	-0.242568	0.058840	-0.271457	0.073690
actividad11	-0.780320	0.608900	-0.875912	0.767224
actividad12	-0.547142	0.299365	-0.697653	0.486723
actividad14	-0.794609	0.631409	-0.910497	0.829021
actividad21	-0.288490	0.083228	-0.111275	0.012385
actividad22	0.007672	0.000061	0.163497	0.026737
actividad24	-0.342680	0.117438	-0.209269	0.043813
titularviviendahabitual2	-0.431430	0.186134	-0.412899	0.170488
titularviviendahabitual3	-0.748620	0.560433	-0.684713	0.468834
titularviviendahabitual4	-0.598454	0.358149	-0.538753	0.290256
titularviviendahabitual9	-0.503405	0.253418	-0.438249	0.192065
portitularviviendahabitual	0.007641	0.000058	0.007096	0.000050
porconyugetitularviviendahabitual	0.003473	0.000012	0.003284	0.000011
dedviviendatotal	0.001054	0.000001	0.001111	0.000001
inmuebletitular11	-0.102033	0.010412	-0.129540	0.016784
inmuebletitular12	0.032926	0.001085	0.015273	0.000234
inmuebletitular13	-0.488870	0.238995	-0.484634	0.234872
inmuebletitular14	-0.962717	0.927308	-1.167948	1.368567
inmuebletitular21	-0.178474	0.031854	-0.242483	0.058800
inmuebletitular22	-0.090973	0.008276	-0.160874	0.025881
inmuebletitular23	-0.617462	0.381260	-0.670723	0.449871
inmuebletitular24	-0.663164	0.439823	-0.622220	0.387401
inmuebletitular31	-0.192452	0.037041	-0.259932	0.067572
inmuebletitular32	-0.137940	0.019028	-0.228294	0.052119
inmuebletitular33	-0.613656	0.376575	-0.671624	0.451082

Continúa en la página siguiente...

Variables	GLMlogit		LMM	
	Sesgo	EMSE	Sesgo	EMSE
inmuebletitular34	-1.033005	1.067250	-1.021840	1.044694
inmuebletitular41	-0.179659	0.032281	-0.247273	0.061153
inmuebletitular42	-0.129150	0.016681	-0.218551	0.047767
inmuebletitular43	-0.635837	0.404290	-0.693387	0.480789
inmuebletitular44	-0.515569	0.265846	-0.502472	0.252498
inmuebletitular51	-0.335161	0.112340	-0.504510	0.254548
inmuebletitular52	-0.242905	0.059004	-0.410175	0.168249
inmuebletitular53	-0.706227	0.498760	-0.827007	0.683947
inmuebletitular54	-1.004788	1.009608	-1.279193	1.636850
inmuebletitular61	-0.077660	0.006040	-0.068341	0.004693
inmuebletitular62	-0.071952	0.005178	-0.119955	0.014392
inmuebletitular63	-0.547212	0.299445	-0.529851	0.280748
inmuebletitular64	-0.898685	0.807659	-1.224245	1.500001
rentatotalinmuebles	0.000962	0.000001	0.000957	0.000001
RentaFamiliar	0.001097	0.000001	0.001097	0.000001
ninmuebles	0.185095	0.034260	0.253318	0.064170
numfam	-0.081169	0.006588	-0.071801	0.005155
edad2	-0.047309	0.002239	-0.066099	0.004370
edad3	0.031890	0.001017	0.083525	0.006977
edad4	0.051270	0.002629	0.158406	0.025094
edad5	-1.320745	1.744367	-1.131937	1.281283
edad6	-8.641781	74.680376	-8.467226	71.693919
edad7	-0.344571	0.118739	-0.338688	0.114720
BIECorrector1	-0.298987	0.089395	-0.407100	0.165736
pormedtitularidad	0.007336	0.000054	0.006228	0.000039
ImportePrestamo	-0.000002	0.000000	0.000000	0.000000
RentaCorrectora1	-2.045114	4.182490	-2.017503	4.070320
Year	-0.001133	0.000001	0.010868	0.000118

Tabla A.13: EMSE y Sesgo para las pruebas realizadas de comprobación del ajuste

A.5. Resúmenes numéricos de las correlaciones del analisis de componentes principales

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10	Dim.11	Dim.12	Dim.13	Dim.14	Dim.15	Dim.16
estcv2	0,2363	0,4061	-0,4900	-0,0086	0,1980	-0,0568	0,0438	-0,5202	0,0834	-0,0085	-0,0857	-0,0040	0,0225	-0,0623	0,0801	-0,0210
estcv3	0,0538	-0,2496	0,3700	0,0039	-0,0032	-0,0505	-0,3708	0,1855	0,1835	0,3085	-0,0518	-0,0119	0,1061	0,1284	0,1388	0,0321
estcv4	-0,0150	-0,1315	0,0576	0,0152	-0,2173	0,0670	-0,0458	0,4245	-0,0759	-0,1678	0,2066	-0,0982	-0,0041	-0,0484	-0,4054	0,2029
act11	0,0160	0,0943	-0,0606	0,0139	-0,1408	0,1706	0,3820	0,1301	0,5918	0,1099	-0,0510	-0,1276	-0,1325	-0,0320	-0,1259	0,0171
act12	0,0913	0,0956	-0,1952	0,0016	-0,1422	-0,1517	0,2671	0,3237	0,2982	0,0400	-0,0451	0,1189	0,1083	-0,0366	0,1997	-0,0524
act14	0,0307	0,0160	0,0023	-0,0029	0,0029	-0,0381	0,0378	0,0128	0,0467	-0,0185	0,0255	0,5870	0,4309	-0,0789	-0,1377	0,0069
act21	0,0745	0,1249	-0,1677	0,0119	-0,2044	0,0466	0,4468	0,3052	0,6566	0,1113	-0,0803	-0,0790	-0,0839	-0,0523	0,0027	-0,0289
act22	0,0408	0,0500	-0,0630	0,0003	-0,0490	-0,0443	0,1210	0,1286	0,1247	0,0277	0,0005	0,1553	0,1376	-0,0083	0,1211	0,0153
act24	0,0100	0,0079	-0,0161	-0,0005	-0,0178	-0,0211	0,0485	0,0438	0,0390	-0,0118	0,0089	0,5964	0,4360	-0,0725	-0,0874	-0,0166
tvh2	0,0334	-0,0714	0,1144	0,0094	-0,0069	-0,0133	-0,1613	0,0531	0,1112	0,2262	-0,0910	-0,0006	0,0861	0,0944	0,0658	0,0676
tvh3	-0,1426	-0,0419	0,0413	-0,0006	-0,0284	0,0314	0,2177	0,0198	-0,1819	-0,0112	0,0565	-0,0226	-0,0181	0,1355	0,0982	0,6656
tvh4	-0,2842	-0,1669	0,3056	-0,0005	-0,0087	0,1134	0,4784	0,0925	-0,0907	-0,1777	0,3027	0,0320	0,0799	0,0630	0,2544	-0,2404
tvh9	-0,1649	-0,0914	0,1847	-0,0101	0,0719	-0,0478	0,0913	-0,0464	-0,1503	-0,0184	0,0023	-0,0672	0,0030	0,0008	-0,1390	-0,1358
ptvh	0,2802	0,0060	-0,1610	0,0094	-0,1406	-0,0780	-0,6077	0,2245	0,1836	0,2149	-0,2394	0,0172	-0,0803	-0,0844	-0,2648	-0,0155
pctvh	0,2628	0,3668	-0,4104	-0,0079	0,1983	-0,0515	-0,0699	-0,4599	0,1416	-0,0390	-0,1463	0,0005	0,0096	-0,0940	0,0070	-0,1137
dvt	0,0723	0,0890	-0,3446	0,0443	-0,3174	0,3141	-0,1906	0,1957	-0,2111	0,1016	-0,1064	0,1025	-0,0912	0,0727	0,0660	-0,1309
it11	0,1341	0,5913	0,1935	-0,0096	-0,0524	-0,0554	-0,0491	-0,0868	0,1340	-0,0844	0,1582	0,1251	-0,1305	0,4672	-0,1055	-0,0448
it12	0,6596	-0,2107	-0,0685	-0,0019	0,0141	0,0173	-0,2327	0,0640	0,1054	-0,2230	0,2631	-0,0528	-0,0105	-0,2117	0,2944	-0,0893
it13	0,0623	0,1540	-0,0722	-0,0005	0,2821	0,1000	0,0166	0,0322	0,0908	-0,0049	0,0030	0,0276	0,0725	-0,0283	0,0489	0,4382
it14	0,0023	0,0013	0,0133	0,5083	0,0135	-0,0183	-0,0051	-0,0062	-0,0005	0,0001	0,0133	0,0044	-0,0041	-0,0016	-0,0087	0,0082
it21	0,1601	0,6439	0,3000	-0,0114	-0,1107	-0,0412	-0,0229	-0,0388	0,0423	-0,0532	0,1092	0,1008	-0,1016	0,3849	-0,1007	-0,0387
it22	0,7433	-0,2612	-0,0023	-0,0044	-0,0518	0,0084	-0,0158	-0,0194	0,0218	-0,1088	0,1280	-0,0428	0,0001	-0,1252	0,1575	-0,0745
it23	0,1411	0,1613	-0,0915	-0,0060	0,4619	0,1581	-0,0202	0,1324	0,0175	-0,0217	0,0214	-0,0033	0,0336	-0,0665	0,0871	0,2168
it24	0,0060	0,0016	0,0217	0,7157	0,0192	-0,0285	0,0050	-0,0122	0,0032	-0,0076	0,0050	-0,0006	0,0011	0,0021	-0,0064	0,0057
it31	0,1698	0,6601	0,3640	-0,0108	-0,1460	-0,0234	0,0069	0,0110	-0,0534	-0,0011	0,0219	0,0288	-0,0298	0,1413	-0,0508	-0,0275
it32	0,7850	-0,2843	0,0443	-0,0063	-0,0351	0,0093	0,1510	-0,0716	-0,0562	0,0299	-0,0084	-0,0290	0,0091	-0,0250	-0,0016	-0,0607
it33	0,1548	0,1568	-0,0502	-0,0015	0,4905	0,1961	0,0104	0,2234	-0,0282	-0,0171	0,0166	0,0069	0,0110	-0,0263	0,0653	0,1489
it34	0,0053	-0,0002	0,0194	0,6674	0,0169	-0,0246	0,0051	-0,0063	0,0068	0,0014	-0,0038	0,0044	-0,0079	-0,0030	0,0056	0,0011
it41	0,1719	0,6422	0,3979	-0,0096	-0,1710	-0,0176	0,0389	0,0518	-0,1361	0,0571	-0,0719	-0,0600	0,0567	-0,1761	0,0275	-0,0016
it42	0,7673	-0,3053	0,0877	-0,0056	-0,0812	-0,0040	0,2366	-0,1435	-0,0884	0,1124	-0,0960	-0,0048	0,0184	0,0626	-0,1046	0,0158
it43	0,1859	0,1420	-0,0536	-0,0044	0,5961	0,2153	0,0504	0,3108	-0,0912	0,0403	-0,0235	-0,0083	-0,0175	0,0145	-0,0184	-0,0667
it44	0,0088	-0,0019	0,0199	0,5965	0,0194	-0,0242	0,0044	-0,0069	0,0034	0,0024	-0,0084	0,0002	-0,0001	0,0006	0,0070	-0,0091
it51	0,1536	0,5763	0,3809	-0,0122	-0,1744	-0,0140	0,0523	0,0698	-0,1698	0,0865	-0,1263	-0,1235	0,1044	-0,3851	0,0851	0,0187
it52	0,7279	-0,2930	0,1059	-0,0096	-0,0728	-0,0017	0,2745	-0,1554	-0,1112	0,1641	-0,1461	0,0039	0,0209	0,1060	-0,1591	0,0490
it53	0,1782	0,1248	-0,0160	-0,0001	0,5061	0,1807	0,0938	0,2891	-0,1243	0,0937	-0,0747	0,0031	-0,0254	0,0746	-0,1311	-0,1806
it54	0,0102	0,0113	0,0333	0,6727	0,0136	-0,0291	0,0068	-0,0117	-0,0053	-0,0045	-0,0061	-0,0028	0,0063	-0,0062	0,0041	0,0002
it61	0,1320	0,4915	0,3358	-0,0031	-0,1553	-0,0117	0,0534	0,0715	-0,1625	0,0898	-0,1360	-0,1306	0,1217	-0,4322	0,1009	0,0244
it62	0,6565	-0,2741	0,1123	-0,0058	-0,1036	-0,0118	0,2630	-0,1699	-0,1024	0,1673	-0,1490	0,0143	0,0208	0,1211	-0,1627	0,0876
it63	0,1786	0,1051	-0,0152	-0,0074	0,5007	0,1742	0,1040	0,2898	-0,1381	0,1136	-0,0904	-0,0097	-0,0307	0,0694	-0,1494	-0,2456
it64	0,0070	0,0018	0,0217	0,7522	0,0290	-0,0275	0,0074	-0,0056	0,0011	-0,0070	-0,0049	0,0041	-0,0075	-0,0045	-0,0035	-0,0056
rti	0,6372	-0,0022	0,1445	-0,0055	-0,0362	0,0232	0,0738	0,0046	-0,0196	0,0227	0,0274	0,0264	-0,0168	0,0779	0,0036	0,0763
rentafam	0,3871	0,1446	-0,5117	0,0259	-0,3141	-0,0883	-0,0939	0,3214	-0,2520	-0,0497	-0,0196	0,0349	-0,0148	0,0957	0,0890	-0,0075
ninmuebles	0,9620	0,0072	0,1366	-0,0009	0,0861	0,0593	0,0851	0,0033	-0,0345	-0,0171	0,0591	-0,0143	-0,0019	0,0055	0,0450	-0,0008
numfam	-0,0045	0,4067	-0,5031	0,0214	0,0348	0,0616	0,1554	-0,1945	-0,0523	0,2576	0,1315	-0,0707	0,1059	0,1098	0,0684	0,0789
edad2	-0,2382	-0,0461	-0,0361	0,0041	-0,1070	0,0381	0,2208	0,1011	-0,1991	-0,0130	-0,4272	0,2437	-0,3107	0,1317	0,4149	-0,0011
edad3	-0,0620	0,1153	-0,3509	0,0223	-0,1288	0,0473	0,0644	-0,0351	-0,1411	0,5381	0,6255	-0,0726	0,0395	-0,1404	-0,1486	-0,0280
edad4	0,1488	0,0957	-0,1865	0,0108	-0,0498	-0,0125	-0,0392	0,0649	0,0896	-0,5307	-0,2404	-0,4130	0,5334	0,1737	-0,1757	0,0079
edad5	0,2409	0,0355	0,1131	-0,0173	0,1138	-0,0414	-0,0731	-0,0866	0,1265	-0,2612	-0,0054	0,3672	-0,5131	-0,3418	-0,2650	0,1010
edad6	0,1044	-0,1397	0,3918	-0,0207	0,1997	-0,0530	-0,3729	-0,0855	0,2254	0,3429	-0,0526	-0,0443	0,1644	0,1685	0,2455	-0,0105
edad7	-0,0535	-0,0418	0,0877	-0,0040	0,0190	-0,0176	0,0876	0,0186	-0,0483	-0,0714	0,0927	-0,0344	0,0241	-0,0056	-0,0484	-0,2790
BIECorrector1	0,1473	-0,0559	0,0435	-0,0058	-0,0177	-0,0588	0,0155	0,0039	-0,0004	0,0172	0,0349	0,0333	0,0022	-0,0002	0,0007	0,0800
pormedtit	0,6230	0,1631	0,0807	-0,0029	0,0223	-0,0126	-0,2754	0,0711	0,2289	-0,2441	0,3613	0,0434	-0,0714	0,0849	0,2338	0,0474
IP	0,0177	0,0149	-0,0530	-0,0462	0,2288	-0,9147	0,1141	0,1605	-0,0486	0,0388	0,0246	-0,0430	-0,0445	0,0013	-0,0086	0,0021
RC	-0,3949	-0,1426	0,4909	-0,0277	0,2932	0,1007	0,0995	-0,2977	0,2817	0,0318	0,0369	-0,0537	0,0358	-0,1150	-0,0883	-0,0145
Year	-0,0173	-0,0145	0,0507	0,0470	-0,2305	0,9179	-0,1162	-0,1589	0,0461	-0,0390	-0,0245	0,0439	0,0439	-0,0012	0,0108	-0,0038

Tabla A.14: Correlaciones de las Dimensiones 1 a la 16 con las variables originales

	Dim.17	Dim.18	Dim.19	Dim.20	Dim.21	Dim.22	Dim.23	Dim.24	Dim.25	Dim.26	Dim.27	Dim.28	Dim.29	Dim.30	Dim.31	Dim.32
estcv2	-0,0259	-0,0746	-0,0435	0,0253	0,0345	-0,0074	0,0030	0,0214	0,0120	-0,0050	-0,0059	0,0195	0,0558	0,0007	0,0318	-0,0136
estcv3	0,0763	-0,0106	-0,0362	0,2281	0,0059	-0,0672	0,0005	-0,0634	-0,2347	0,0260	-0,0311	-0,0768	-0,2436	-0,0102	0,0471	-0,0903
estcv4	-0,0790	0,0132	-0,0677	-0,2383	-0,0581	0,0630	-0,0178	0,0665	0,2801	-0,0253	0,0697	0,1055	0,3632	0,0076	0,1229	-0,1010
act11	-0,2980	0,1067	0,1170	0,0592	0,0600	0,0948	0,2948	-0,1017	-0,2452	0,0120	0,0170	-0,1356	0,1322	0,0010	-0,0127	0,0111
act12	0,4709	-0,2371	-0,0538	-0,1022	-0,0502	-0,1304	-0,3654	0,1248	0,2508	-0,0247	-0,0030	0,0548	-0,0884	0,0000	0,0016	0,0007
act14	-0,2051	0,0142	0,0904	0,0292	-0,0066	0,0483	0,1350	-0,0450	-0,0935	-0,0024	0,0244	0,5793	-0,1276	-0,0058	0,0111	-0,0060
act21	-0,0672	-0,0188	0,1356	0,1205	-0,0133	-0,0924	-0,1741	0,0260	0,0051	-0,0069	0,0145	0,1317	-0,0016	-0,0008	-0,0081	0,0126
act22	0,4827	-0,1940	-0,1991	-0,4078	0,0835	0,2496	0,5207	-0,1098	-0,1813	0,0072	0,0101	-0,0934	0,0691	0,0021	0,0022	0,0023
act24	-0,1806	-0,0160	0,1253	0,0611	-0,0317	-0,0582	-0,1261	0,0495	0,0640	-0,0051	-0,0079	-0,5751	0,1482	0,0067	-0,0182	0,0053
tvh2	0,0827	-0,0223	-0,1209	0,5101	0,0504	0,1823	0,2563	-0,3201	0,5807	-0,0092	0,0353	0,0266	0,1180	0,0035	-0,0307	0,0320
tvh3	-0,0944	-0,5216	0,1195	0,1711	0,0488	-0,0455	0,0820	0,2253	-0,0931	0,0058	-0,0078	0,0063	-0,0010	0,0071	-0,0522	0,0484
tvh4	-0,0369	0,2925	-0,3720	0,0997	-0,0233	-0,0200	-0,0431	-0,0697	-0,0537	0,0107	0,0135	0,0030	0,0029	0,0037	0,0182	-0,0029
tvh9	0,2905	-0,0686	0,7287	-0,0439	0,0108	0,0391	-0,0882	-0,4333	-0,0332	0,0602	0,0135	0,0028	0,0324	-0,0084	0,0490	-0,0327
ptvh	-0,0849	0,0905	-0,0336	-0,2222	-0,0272	0,0344	0,0537	0,1514	0,0932	-0,0321	-0,0205	-0,0646	-0,2397	-0,0161	-0,0162	0,0108
pctvh	0,0446	-0,0401	-0,0502	0,0310	0,0277	-0,0178	-0,0292	0,0228	0,0350	-0,0207	0,0104	0,0757	0,3063	0,0143	0,0145	-0,0273
dvt	0,0300	0,1309	0,1137	-0,0217	0,0354	-0,0005	-0,0056	-0,0063	-0,0674	0,0184	0,0046	0,0634	0,1026	0,0391	-0,2026	0,1853
it11	0,0086	-0,0230	0,0097	-0,0291	-0,0173	0,0066	-0,0017	-0,0068	0,0197	-0,1100	0,0669	-0,0097	-0,0271	-0,0108	0,1331	-0,0343
it12	-0,1544	-0,1437	0,1136	-0,0348	-0,0152	0,0378	0,0524	-0,0212	0,0606	0,0507	0,0917	-0,0140	-0,0262	-0,0019	0,0025	0,0274
it13	0,2085	0,4991	0,0592	-0,0322	-0,0658	0,0204	-0,0452	0,0194	0,0231	0,3695	-0,4293	0,0255	0,0496	-0,0040	-0,1352	-0,0090
it14	0,0071	0,0142	0,0003	-0,0760	-0,0289	-0,6261	0,2548	-0,1083	0,0896	0,0084	-0,0128	0,0040	-0,0038	0,3305	0,0616	0,2384
it21	0,0188	-0,0131	0,0020	-0,0269	-0,0221	0,0054	-0,0192	-0,0008	0,0214	-0,0473	-0,0602	-0,0004	-0,0013	0,0040	-0,1138	0,0155
it22	-0,0921	-0,1145	0,0659	-0,0072	-0,0262	0,0270	0,0452	-0,0252	0,0430	-0,0699	-0,2360	0,0125	0,0121	0,0077	-0,1168	0,0014
it23	0,0922	0,2409	0,0577	-0,0155	-0,0390	0,0010	-0,0223	0,0247	-0,0004	0,1887	0,6458	-0,0303	-0,0479	0,0180	0,0663	0,0649
it24	0,0025	0,0092	-0,0013	-0,0427	-0,0152	-0,3071	0,1212	-0,0489	0,0349	0,0120	0,0072	-0,0002	0,0125	-0,2407	-0,0344	-0,0801
it31	0,0108	-0,0139	-0,0099	-0,0142	-0,0168	0,0014	-0,0182	0,0112	0,0160	0,0938	0,0186	0,0024	0,0021	0,0197	-0,2103	0,0645
it32	-0,0385	-0,0728	0,0005	-0,0114	-0,0370	0,0120	0,0199	0,0083	0,0285	0,0845	-0,0186	0,0089	-0,0039	0,0053	-0,1400	-0,0017
it33	0,0590	0,1731	0,0743	-0,0076	-0,0121	0,0006	0,0017	-0,0455	-0,0159	-0,6307	-0,1597	-0,0042	-0,0058	-0,0039	0,0534	0,0326
it34	-0,0032	-0,0002	0,0073	0,0187	0,0081	0,1623	-0,0617	0,0266	-0,0235	0,0051	-0,0073	-0,0006	-0,0040	0,3419	-0,0643	-0,5501
it41	-0,0171	-0,0069	-0,0091	0,0040	-0,0049	-0,0005	0,0024	0,0069	-0,0019	0,0549	-0,0073	0,0068	0,0074	0,0147	-0,0732	0,0447
it42	0,0335	0,0392	-0,0260	-0,0261	-0,0547	-0,0012	-0,0028	0,0206	0,0149	-0,0026	-0,0385	-0,0032	-0,0051	0,0022	0,0259	0,0192
it43	-0,0330	-0,0686	0,0030	0,0091	0,0178	-0,0085	0,0085	-0,0196	-0,0128	-0,2711	0,0960	0,0028	-0,0133	0,0070	-0,3192	-0,0643
it44	-0,0045	-0,0191	-0,0008	0,0671	0,0247	0,5067	-0,2002	0,0826	-0,0570	-0,0079	-0,0005	-0,0051	-0,0295	0,1891	0,0800	0,3902
it51	-0,0350	0,0035	-0,0031	0,0180	0,0030	-0,0041	0,0182	0,0010	-0,0104	-0,0236	0,0228	-0,0049	0,0009	0,0243	0,0025	-0,0280
it52	0,0705	0,0891	-0,0458	-0,0308	-0,0564	-0,0078	-0,0208	0,0311	0,0001	-0,0357	0,1213	-0,0068	-0,0185	0,0121	-0,0762	-0,0046
it53	-0,0745	-0,1888	-0,0706	0,0185	0,0417	0,0006	-0,0012	0,0055	-0,0009	0,1629	-0,3047	0,0016	0,0223	-0,0223	0,3803	0,0768
it54	0,0022	-0,0028	-0,0061	0,0048	0,0010	0,0483	-0,0251	0,0115	-0,0043	0,0002	0,0204	0,0028	0,0283	-0,6215	-0,0386	0,0146
it61	-0,0368	0,0015	-0,0020	0,0160	0,0112	-0,0017	0,0203	-0,0006	-0,0148	-0,0529	-0,0191	-0,0045	0,0034	-0,0329	0,1901	-0,0465
it62	0,0881	0,1348	-0,0427	-0,0267	-0,0443	-0,0079	-0,0274	0,0250	-0,0101	-0,0900	0,0700	-0,0102	-0,0133	-0,0058	0,0938	0,0235
it63	-0,0954	-0,2605	-0,0992	0,0188	0,0304	-0,0173	0,0103	0,0233	-0,0029	0,3674	0,0276	0,0070	-0,0001	0,0010	-0,0971	-0,0622
it64	-0,0022	-0,0085	0,0026	0,0046	0,0076	0,1231	-0,0570	0,0291	-0,0265	-0,0152	-0,0044	0,0017	-0,0031	0,1093	0,0149	0,0807
rti	0,0119	0,0852	0,0302	-0,0089	0,1067	-0,0119	-0,0280	-0,0258	-0,0115	-0,0006	0,0417	-0,0121	-0,0430	-0,0241	0,2833	-0,0726
rentafam	0,0828	0,0447	-0,0121	0,2310	0,0066	-0,0482	-0,0179	-0,0592	-0,1934	0,0092	0,0164	0,0122	0,1134	-0,0083	0,0649	-0,0544
ninmuebles	-0,0263	-0,0003	0,0400	-0,0346	-0,0528	0,0179	0,0167	0,0020	0,0400	-0,0017	-0,0062	-0,0035	-0,0137	0,0071	-0,0699	0,0221
numfam	-0,0973	0,0323	0,0297	-0,0462	-0,0042	0,0436	0,0708	-0,0158	0,1408	0,0180	0,0041	-0,0589	-0,2580	-0,0214	0,1364	-0,1141
edad2	-0,2619	0,0798	0,1447	-0,2414	-0,0405	0,0866	0,0993	0,0130	0,2457	0,0197	-0,0070	-0,0018	-0,0530	-0,0136	0,0570	-0,0598
edad3	0,0040	-0,0137	-0,0038	-0,0355	-0,0293	0,0191	-0,0154	-0,0339	0,0267	-0,0129	-0,0202	-0,0217	-0,0614	-0,0031	-0,0500	0,0278
edad4	-0,0165	-0,0149	-0,0532	0,0420	0,0239	0,0075	0,0219	-0,0619	-0,0083	0,0007	-0,0124	-0,0471	-0,1556	0,0056	-0,0296	0,0189
edad5	0,2147	-0,0569	-0,1643	0,2596	0,0235	-0,0356	-0,0103	-0,0674	-0,0911	0,0111	0,0041	-0,0308	-0,0743	-0,0061	0,0285	-0,0236
edad6	-0,0101	-0,0243	0,0137	-0,0819	0,0037	-0,0814	-0,1195	0,1475	-0,1635	-0,0133	0,0429	0,1208	0,4043	0,0118	0,0591	-0,0182
edad7	0,2449	0,1734	0,3224	0,2636	0,1991	-0,0002	0,3405	0,6658	0,1113	-0,0146	-0,0077	0,0031	-0,0096	-0,0060	0,0317	-0,0271
BIECorrector1	-0,0718	0,0670	-0,0491	-0,1357	0,9342	-0,0642	-0,1568	-0,0953	0,0634	0,0087	0,0019	0,0040	-0,0237	0,0028	-0,0563	0,0092
pormedtit	-0,0883	0,0268	0,1379	-0,0787	-0,0340	0,0401	0,0418	-0,0185	0,0741	0,1194	-0,0312	-0,0175	-0,0546	-0,0117	0,0837	-0,0055
IP	-0,0989	0,0771	0,0150	-0,0107	-0,0267	0,0220	0,0447	-0,0105	-0,0129	-0,0031	0,0013	-0,0053	0,0225	0,0087	-0,0464	0,0422
RC	-0,0464	-0,0748	0,0272	-0,2250	-0,0465	0,0446	0,0026	0,0611	0,2111	-0,0105	-0,0199	-0,0037	-0,1235	0,0083	-0,0758	0,0593
Year	0,1008	-0,0749	-0,0131	0,0099	0,0247	-0,0225	-0,0451	0,0101	0,0128	0,0028	-0,0001	0,0077	-0,0206	-0,0089	0,0407	-0,0348

Tabla A.15: Correlaciones de las Dimensiones 17 a la 32 con las variables originales

	Dim.33	Dim.34	Dim.35	Dim.36	Dim.37	Dim.38	Dim.39	Dim.40	Dim.41	Dim.42	Dim.43	Dim.44	Dim.45	Dim.46	Dim.47	Dim.48
estcv2	-0,0263	0,0011	-0,0198	0,0275	-0,0007	0,0118	0,0098	0,1821	0,1384	0,0304	0,0106	0,0000	-0,0478	0,3509	-0,0122	0,0033
estcv3	-0,1095	0,0528	0,3058	0,2789	0,0234	0,0175	0,0099	-0,1640	-0,0833	0,0070	0,0046	-0,0004	-0,0163	0,1467	-0,0092	-0,0007
estcv4	-0,1531	0,0420	0,2109	0,1925	0,0151	0,0043	0,0015	0,0267	0,0275	0,0145	0,0042	0,0002	-0,0115	0,0959	-0,0079	-0,0002
act11	0,0009	-0,0169	0,0006	-0,0042	0,0015	0,0021	0,0017	0,0043	0,0071	0,0048	-0,0011	-0,0002	0,0011	0,0001	-0,0247	0,0022
act12	0,0173	0,0150	-0,0069	0,0113	0,0001	0,0025	-0,0024	0,0027	0,0004	-0,0031	0,0015	0,0003	0,0004	0,0045	-0,0411	0,0021
act14	-0,0091	-0,0169	0,0036	-0,0114	-0,0008	0,0025	0,0000	-0,0129	-0,0038	0,0016	0,0021	0,0038	0,0014	-0,0002	-0,0042	-0,0008
act21	0,0121	-0,0150	-0,0059	-0,0122	0,0006	0,0060	0,0034	0,0037	0,0068	0,0038	0,0019	0,0001	-0,0006	0,0023	0,0269	-0,0042
act22	-0,0004	-0,0048	-0,0041	0,0081	0,0010	-0,0031	-0,0002	0,0030	0,0030	0,0014	0,0033	0,0007	0,0002	-0,0006	0,0115	-0,0017
act24	0,0039	0,0128	0,0005	0,0110	0,0012	-0,0033	-0,0007	0,0080	0,0002	-0,0010	-0,0007	-0,0011	0,0004	0,0011	0,0064	0,0004
tvh2	0,0604	-0,0002	-0,0743	-0,0502	-0,0191	-0,0004	0,0043	0,0293	0,0163	0,0015	-0,0010	-0,0010	-0,0011	-0,0019	-0,0004	0,0000
tvh3	0,0917	0,0426	-0,0387	0,0062	0,0013	0,0088	0,0022	0,0076	0,0081	-0,0010	-0,0005	0,0003	-0,0038	0,0104	0,0004	0,0012
tvh4	0,0036	0,0356	0,0143	0,0204	-0,0006	0,0043	0,0060	0,1016	0,0720	0,0075	0,0015	-0,0004	-0,0090	0,0485	-0,0012	0,0015
tvh9	-0,0621	-0,0058	0,0345	-0,0107	-0,0027	0,0104	0,0072	0,0704	0,0527	0,0046	0,0009	0,0002	-0,0062	0,0329	-0,0069	0,0012
ptvh	-0,0041	-0,0609	-0,0992	-0,1439	-0,0120	-0,0334	-0,0088	0,0979	0,0404	0,0040	-0,0003	-0,0002	0,0034	0,0212	0,0026	0,0007
pctvh	-0,0344	0,0041	0,1039	0,1590	0,0162	0,0233	-0,0020	-0,3120	-0,1889	-0,0224	-0,0036	0,0004	0,0037	-0,1115	0,0009	0,0050
dvt	0,3996	0,3152	-0,0052	0,2881	0,0288	0,0042	-0,0045	0,0743	0,0508	0,0068	0,0002	-0,0033	-0,0020	0,0048	-0,0092	0,0010
it11	0,0407	0,2937	0,0663	-0,1855	-0,0179	-0,0250	-0,0293	0,0281	-0,1510	-0,2177	-0,1179	0,0366	-0,0045	0,0351	0,0077	-0,0162
it12	0,0460	-0,1062	-0,0251	0,0621	0,0009	0,1643	0,0664	-0,0719	0,1170	0,0697	0,0562	-0,0142	-0,0377	-0,0165	-0,0026	-0,0139
it13	-0,0772	0,0650	-0,0349	-0,0474	-0,0074	0,0019	0,0058	-0,0431	0,0034	-0,0245	0,0001	-0,0033	0,0024	0,0140	0,0044	-0,0101
it14	-0,1107	0,0166	-0,0007	-0,0266	0,2443	-0,0382	0,1114	0,0013	-0,0077	-0,0047	0,0033	0,0002	-0,0012	0,0011	0,0001	-0,0002
it21	-0,0082	-0,0392	0,0137	-0,0323	0,0043	-0,0401	0,0062	-0,0948	0,0665	0,3182	0,2882	-0,1285	0,1200	0,0117	0,0016	-0,0381
it22	-0,0265	-0,0125	0,0942	0,0085	0,0085	-0,1666	-0,0685	0,1117	-0,1119	-0,0449	-0,1166	0,0624	0,2999	0,0296	0,0047	-0,1062
it23	0,1578	-0,0896	0,1164	-0,0727	-0,0024	-0,1375	-0,0476	-0,0013	-0,0788	0,0330	0,0114	0,0024	0,0892	0,0282	0,0036	-0,0249
it24	0,0447	-0,0092	0,0022	0,0206	-0,2752	0,1413	-0,4519	-0,0180	0,0438	0,0187	-0,0107	-0,0054	0,0024	-0,0003	0,0020	-0,0009
it31	0,0066	-0,3380	0,0137	0,1350	0,0165	-0,0762	-0,0070	-0,0361	0,1194	0,0621	-0,2489	0,2526	-0,0494	-0,0069	-0,0041	0,0767
it32	-0,0364	0,0781	0,1404	-0,0832	-0,0055	-0,2775	-0,0982	0,0589	-0,0828	-0,0079	0,0926	-0,1033	-0,0860	-0,0101	-0,0110	0,2466
it33	0,0929	-0,1097	0,0378	0,0104	0,0110	0,1451	0,0181	0,1143	-0,1894	0,1420	-0,0310	0,0212	-0,0321	0,0012	-0,0033	0,0800
it34	0,2268	-0,0449	-0,0337	-0,0311	0,2038	0,0015	-0,0056	0,0017	0,0133	-0,0055	0,0005	-0,0034	0,0011	-0,0009	0,0001	0,0033
it41	0,0291	-0,2414	0,0241	0,1107	0,0066	0,0981	0,0067	0,1470	-0,1260	-0,2229	-0,0322	-0,3368	-0,0473	-0,0035	0,0009	-0,0362
it42	0,0964	0,0000	0,1322	-0,0786	-0,0037	-0,0403	-0,0120	0,0295	-0,0561	0,1120	0,0224	0,0856	-0,2556	-0,0338	0,0041	-0,1158
it43	-0,2437	0,1188	-0,0949	0,0228	-0,0004	-0,1344	-0,0164	-0,1533	0,2479	-0,1817	0,0155	-0,0386	-0,0600	-0,0018	0,0010	-0,0507
it44	-0,1960	0,0389	0,0198	-0,0117	0,1670	0,0804	-0,2433	-0,0133	0,0169	0,0105	-0,0046	-0,0007	-0,0052	0,0001	0,0012	-0,0019
it51	-0,0396	0,1187	0,0027	-0,0317	-0,0259	0,0766	-0,0067	0,0528	-0,0677	-0,1357	0,3090	0,2740	0,0044	-0,0009	-0,0002	-0,0221
it52	-0,0373	0,0476	-0,0091	-0,0194	-0,0051	0,1960	0,0730	-0,0371	0,0477	0,0297	-0,0924	-0,0512	-0,0111	0,0006	0,0088	-0,1406
it53	0,2767	-0,1496	0,0929	-0,0202	-0,0049	-0,0979	-0,0216	-0,0902	0,1481	-0,1030	0,0552	0,0012	0,0030	0,0049	0,0020	-0,0348
it54	0,0155	-0,0031	0,0012	-0,0133	0,3136	-0,0641	0,2179	0,0185	-0,0283	-0,0208	0,0188	0,0127	-0,0009	0,0009	-0,0010	-0,0037
it61	0,0185	0,2949	0,0071	-0,1447	0,0021	-0,0697	0,0161	-0,1520	0,1396	0,2247	-0,2241	-0,0970	0,0314	0,0047	-0,0003	0,0302
it62	0,0583	-0,0426	-0,0450	0,0312	0,0026	0,2204	0,0808	-0,1052	0,1721	-0,1549	0,0359	0,0182	0,1990	0,0211	-0,0037	0,1600
it63	-0,1414	0,1350	-0,0987	-0,0024	0,0046	0,2561	0,0634	0,1251	-0,2150	0,1528	-0,0592	0,0324	0,0484	0,0067	-0,0016	0,0498
it64	-0,0302	0,0056	0,0057	0,0430	-0,4972	-0,1156	0,3572	0,0113	-0,0388	0,0004	-0,0035	-0,0010	0,0039	0,0006	-0,0003	0,0028
rti	-0,1000	0,0161	-0,5094	0,3239	0,0182	-0,2057	-0,0789	0,0139	-0,1181	0,0354	0,0146	0,0012	-0,0208	0,0003	0,0046	-0,0146
rentafam	-0,1250	-0,0878	0,0145	-0,1123	-0,0104	0,0098	0,0076	0,0082	0,0115	-0,0009	0,0011	0,0013	-0,0068	0,0042	0,3140	0,0122
ninmuebles	0,0327	-0,0144	0,0866	-0,0393	-0,0014	-0,0203	-0,0085	0,0026	-0,0131	0,0078	0,0068	0,0035	0,0066	0,0041	0,0029	-0,0467
numfam	-0,1620	0,0654	0,2251	0,1976	0,0117	-0,0033	0,0144	0,2289	0,1594	0,0254	0,0018	-0,0018	0,0269	-0,2135	0,0061	-0,0016
edad2	-0,1122	-0,0546	0,0728	0,0140	-0,0008	0,0106	0,0028	-0,1079	-0,0616	-0,0041	0,0008	0,0012	-0,0112	0,0986	0,0074	0,0002
edad3	0,0399	-0,0491	-0,0647	-0,0577	-0,0050	0,0034	-0,0027	-0,1531	-0,1000	-0,0110	-0,0007	0,0014	-0,0131	0,1116	-0,0023	0,0003
edad4	0,0519	0,0351	-0,0021	0,0470	0,0055	0,0061	0,0052	-0,0141	-0,0027	-0,0002	0,0018	-0,0010	-0,0072	0,0418	-0,0051	0,0011
edad5	-0,0221	0,0570	0,0882	0,0940	0,0084	0,0098	0,0072	0,1149	0,0882	0,0121	0,0018	-0,0019	-0,0019	-0,0191	0,0032	0,0004
edad6	-0,0528	0,0031	-0,0082	-0,0450	-0,0055	-0,0096	-0,0018	0,2068	0,1270	0,0149	0,0017	0,0012	0,0032	-0,0409	0,0092	-0,0001
edad7	-0,0464	-0,0103	0,0309	0,0152	0,0007	0,0062	0,0030	0,0012	0,0064	0,0019	0,0009	-0,0004	-0,0034	0,0205	-0,0023	0,0002
BIECorrector1	0,0086	-0,0099	0,0842	-0,0550	-0,0021	0,0255	0,0087	-0,0109	-0,0005	-0,0011	-0,0025	0,0009	0,0032	0,0005	0,0133	-0,0001
pormedtit	0,0224	0,1161	-0,0529	-0,0428	-0,0082	0,1903	0,0708	-0,0652	0,0616	-0,0800	-0,0202	-0,0024	-0,0935	-0,0228	-0,0099	0,0531
IP	0,0889	0,0602	-0,0033	0,0760	0,0083	-0,0032	-0,0021	0,0132	0,0063	0,0052	-0,0004	0,0000	0,0021	-0,0028	0,0011	-0,0003
RC	0,1391	0,0971	-0,0074	0,1399	0,0144	-0,0046	-0,0007	-0,0120	-0,0061	0,0041	0,0040	-0,0010	-0,0032	0,0206	0,3058	0,0125
Year	-0,0740	-0,0507	0,0009	-0,0633	-0,0062	0,0026	0,0016	-0,0095	-0,0051	-0,0041	0,0006	0,0002	-0,0016	0,0040	-0,0025	0,0000

Tabla A.16: Correlaciones de las Dimensiones 33 a la 48 con las variables originales

	Dim.49	Dim.50	Dim.51	Dim.52	Dim.53	Dim.54	Dim.55
estcv2	-0,0021	0,0028	0,0082	-0,0437	-0,0102	0,0001	0,0000
estcv3	-0,0019	0,0005	-0,0005	-0,0243	0,0019	0,0002	0,0000
estcv4	-0,0026	0,0004	-0,0004	-0,0198	-0,0006	0,0001	0,0000
act11	0,1851	-0,0029	-0,0024	-0,0008	0,0000	0,0004	0,0000
act12	0,1613	-0,0015	-0,0013	-0,0007	-0,0002	-0,0001	0,0000
act14	0,0314	-0,0008	-0,0007	-0,0007	0,0001	0,0000	0,0000
act21	-0,2241	0,0031	0,0011	0,0008	0,0001	-0,0002	0,0000
act22	-0,0787	0,0006	0,0000	0,0001	0,0000	-0,0001	0,0000
act24	-0,0332	0,0003	0,0000	0,0001	-0,0003	0,0001	0,0000
tvh2	0,0000	0,0000	0,0013	0,0001	0,0027	-0,0001	0,0000
tvh3	0,0010	0,0016	0,0225	-0,0038	0,0730	-0,0002	0,0000
tvh4	-0,0013	0,0029	0,0368	-0,0049	0,1203	-0,0003	0,0000
tvh9	0,0031	0,0014	0,0188	0,0002	0,0609	0,0000	0,0000
ptvh	0,0004	0,0028	0,0329	-0,0208	0,1167	0,0006	0,0000
pctvh	0,0003	0,0029	0,0363	-0,0051	0,0845	0,0002	0,0000
dvt	0,0002	0,0000	0,0004	0,0035	0,0007	-0,0037	0,0000
it11	-0,0028	-0,0142	-0,0965	-0,0032	0,0187	-0,0001	0,0004
it12	-0,0009	-0,0239	-0,1578	-0,0067	0,0332	-0,0006	0,0010
it13	-0,0005	-0,0059	-0,0543	-0,0021	0,0110	-0,0001	0,0003
it14	0,0004	-0,0003	-0,0025	-0,0001	0,0003	-0,0001	0,0000
it21	0,0012	0,0182	0,0131	0,0003	-0,0019	0,0000	0,0003
it22	0,0023	0,0481	0,0389	0,0011	-0,0062	-0,0001	0,0008
it23	0,0016	0,0170	0,0202	0,0001	-0,0043	-0,0002	0,0003
it24	-0,0008	0,0009	0,0009	0,0001	-0,0002	0,0000	0,0000
it31	-0,0012	-0,0232	0,0009	0,0002	-0,0002	-0,0001	0,0002
it32	-0,0049	-0,0807	0,0045	0,0009	-0,0012	0,0002	0,0007
it33	-0,0014	-0,0268	0,0024	0,0001	-0,0007	-0,0002	0,0002
it34	-0,0006	-0,0023	0,0006	0,0001	0,0002	0,0001	0,0000
it41	0,0017	0,0454	-0,0009	0,0000	-0,0002	-0,0001	0,0002
it42	0,0043	0,1807	-0,0027	-0,0002	-0,0014	0,0000	0,0006
it43	0,0012	0,0508	0,0020	-0,0002	-0,0012	0,0002	0,0002
it44	0,0003	0,0031	-0,0001	-0,0001	0,0003	0,0000	0,0000
it51	0,0001	-0,0426	0,0119	0,0003	-0,0018	0,0000	0,0003
it52	-0,0006	-0,1992	0,0441	0,0012	-0,0063	0,0002	0,0010
it53	0,0005	-0,0558	0,0137	0,0002	-0,0021	0,0001	0,0003
it54	0,0003	-0,0042	0,0009	-0,0001	-0,0003	-0,0002	0,0000
it61	-0,0004	0,0233	-0,0065	-0,0001	0,0010	0,0001	0,0000
it62	-0,0020	0,1085	-0,0234	-0,0004	0,0030	-0,0001	0,0000
it63	-0,0002	0,0425	-0,0078	-0,0003	0,0009	0,0001	0,0000
it64	0,0004	0,0024	-0,0004	-0,0001	0,0000	0,0001	0,0000
rti	-0,0012	-0,0068	-0,0079	0,0007	0,0007	-0,0005	0,0000
rentafam	0,0338	0,0007	0,0037	-0,0067	-0,0011	0,0012	0,0000
ninmuebles	-0,0002	-0,0424	-0,0336	-0,0017	0,0079	-0,0001	-0,0035
numfam	-0,0037	-0,0022	-0,0060	0,0051	0,0020	0,0001	0,0000
edad2	0,0003	0,0010	0,0000	0,0981	0,0069	-0,0004	0,0000
edad3	-0,0005	0,0019	0,0024	0,1189	0,0065	-0,0010	0,0000
edad4	-0,0010	0,0013	0,0019	0,1244	0,0065	-0,0014	0,0000
edad5	0,0005	0,0000	-0,0006	0,1090	0,0060	-0,0012	0,0000
edad6	0,0061	-0,0011	-0,0031	0,1031	0,0062	-0,0010	0,0000
edad7	-0,0009	0,0001	0,0001	0,0100	0,0001	0,0000	0,0000
BIECorrector1	0,0007	0,0007	-0,0009	-0,0007	-0,0002	0,0004	0,0000
pormedtit	0,0008	0,0261	0,1584	-0,0016	-0,0337	0,0005	0,0000
IP	-0,0012	-0,0002	-0,0024	0,0040	0,0005	0,1291	0,0000
RC	0,0252	0,0007	0,0015	0,0003	-0,0002	-0,0006	0,0000
Year	0,0002	0,0004	0,0009	0,0006	-0,0002	0,1301	0,0000

Tabla A.17: Correlaciones de las Dimensiones 49 a la 55 con las variables originales

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10	Dim.11	Dim.12	Dim.13	Dim.14
estcv2	0,318	0,550	-0,293	-0,286	-0,041	0,168	-0,054	-0,149	0,018	0,009	0,041	0,025	0,137	-0,052
estcv3	-0,001	-0,311	0,200	0,323	-0,064	0,065	-0,020	-0,052	-0,192	-0,078	0,040	-0,204	0,166	0,130
estcv4	-0,042	-0,186	0,043	0,176	0,082	-0,223	0,093	0,151	0,064	-0,017	-0,021	-0,007	-0,212	0,004
act11	0,007	0,055	0,055	-0,097	0,063	-0,099	0,043	0,029	0,575	0,268	0,128	-0,175	0,160	0,290
act12	0,100	0,075	-0,064	-0,055	0,098	-0,215	0,149	0,100	0,245	0,278	0,115	-0,212	0,102	-0,190
act14	0,027	0,007	0,021	-0,006	-0,001	-0,003	0,011	-0,001	0,015	0,057	0,036	-0,057	0,039	-0,084
act21	0,074	0,085	-0,004	-0,097	0,110	-0,216	0,120	0,085	0,602	0,383	0,165	-0,257	0,178	0,118
act22	0,042	0,034	-0,002	-0,027	0,038	-0,082	0,058	0,049	0,116	0,108	0,055	-0,095	0,047	-0,083
act24	0,009	0,001	0,002	-0,009	0,012	-0,027	0,020	0,013	0,030	0,043	0,039	-0,052	0,028	-0,068
tvh2	0,017	-0,092	0,065	0,100	-0,014	0,027	-0,010	-0,032	-0,076	-0,036	0,025	-0,056	0,155	0,072
tvh3	-0,148	-0,087	0,047	-0,071	0,035	-0,063	0,035	0,073	0,005	0,000	-0,031	0,029	-0,033	-0,025
tvh4	-0,308	-0,249	0,200	-0,059	0,009	-0,004	-0,010	0,087	0,223	0,144	-0,062	0,130	-0,126	0,005
tvh9	-0,177	-0,120	0,103	-0,034	-0,034	0,074	-0,014	-0,013	-0,050	-0,010	-0,062	0,095	-0,094	-0,104
ptvh-100	-0,012	-0,358	0,145	0,410	0,064	-0,264	0,132	0,186	-0,122	-0,104	0,089	-0,226	-0,036	0,040
ptvh-50	0,354	0,574	-0,332	-0,261	-0,065	0,230	-0,121	-0,295	-0,067	-0,033	-0,004	-0,112	0,004	0,066
ptvh-Ma50	0,022	-0,031	-0,010	0,024	0,036	-0,089	0,025	0,067	-0,031	0,009	-0,010	0,364	0,412	-0,090
ptvh-Me50	0,003	-0,070	0,023	0,040	0,018	-0,038	0,006	0,058	-0,001	0,034	0,002	0,456	0,447	-0,050
pctvh-100	0,024	-0,008	0,022	0,027	-0,013	0,016	0,023	0,105	0,067	0,022	0,049	0,011	0,047	-0,096
pctvh-50	0,361	0,599	-0,320	-0,283	-0,066	0,233	-0,108	-0,275	-0,046	-0,018	0,007	-0,101	0,018	0,037
pctvh-Ma50	0,009	-0,010	-0,014	0,003	0,023	-0,038	0,007	0,052	-0,007	0,009	0,012	0,247	0,240	-0,030
pctvh-Me50	0,023	-0,016	-0,005	0,010	0,030	-0,069	0,026	0,078	-0,021	0,036	0,000	0,522	0,560	-0,118
dvt-Dist0	0,097	0,140	-0,262	0,008	0,202	-0,397	-0,213	0,180	-0,210	-0,029	0,035	0,000	-0,012	0,260
it11	0,125	0,510	0,386	0,153	-0,003	-0,058	0,030	-0,110	0,062	-0,001	0,081	-0,092	0,123	-0,079
it12	0,679	-0,125	-0,242	0,487	-0,103	0,084	-0,010	-0,041	0,074	0,065	0,000	0,172	-0,132	0,098
it13	0,072	0,151	0,017	0,051	-0,063	0,139	0,011	0,242	0,087	0,006	0,192	-0,019	0,081	-0,035
it14	0,002	-0,003	0,009	0,066	0,461	0,206	0,005	-0,007	-0,003	0,000	0,000	0,002	0,004	-0,003
it21	0,148	0,496	0,509	0,094	0,026	-0,111	0,024	-0,070	0,002	0,021	-0,101	-0,058	0,082	-0,043
it22	0,740	-0,271	-0,064	0,136	-0,032	0,013	-0,038	-0,028	0,068	0,099	-0,399	-0,023	0,031	0,036
it23	0,167	0,179	-0,031	0,016	-0,120	0,276	-0,006	0,416	0,003	0,017	0,008	0,004	-0,012	0,022
it24	0,006	-0,005	0,015	0,085	0,648	0,293	0,004	-0,015	0,000	0,009	-0,001	-0,004	-0,003	-0,012
it31	0,157	0,464	0,582	0,031	0,050	-0,150	0,018	-0,039	-0,052	0,019	-0,122	0,023	-0,008	0,015
it32	0,763	-0,361	0,056	-0,113	-0,001	0,005	-0,034	-0,043	0,049	0,046	-0,217	-0,052	0,051	0,006
it33	0,170	0,132	0,056	-0,047	-0,090	0,237	-0,021	0,527	0,029	-0,065	-0,069	-0,060	0,040	0,003
it34	0,006	-0,006	0,013	0,078	0,605	0,273	0,004	-0,010	0,002	0,010	-0,006	-0,004	0,005	-0,007
it41	0,158	0,418	0,612	-0,025	0,066	-0,166	0,025	-0,029	-0,100	0,013	-0,062	0,126	-0,118	0,079
it42	0,732	-0,421	0,125	-0,244	0,024	-0,013	-0,041	-0,126	0,016	0,011	0,052	-0,013	0,017	-0,006
it43	0,199	0,111	0,053	-0,109	-0,112	0,303	0,010	0,622	0,004	-0,117	0,044	-0,018	-0,037	0,027
it44	0,008	-0,009	0,015	0,068	0,540	0,247	0,003	-0,010	-0,003	0,008	0,002	-0,002	0,004	-0,008
it51	0,143	0,354	0,573	-0,055	0,068	-0,169	0,027	-0,042	-0,118	0,014	-0,010	0,183	-0,172	0,114
it52	0,688	-0,432	0,157	-0,308	0,033	-0,023	-0,041	-0,179	-0,006	-0,015	0,259	0,021	-0,014	-0,020
it53	0,184	0,059	0,105	-0,155	-0,070	0,229	0,023	0,508	-0,001	-0,145	0,195	0,016	-0,049	0,015
it54	0,010	0,000	0,030	0,076	0,611	0,272	0,006	-0,018	-0,006	0,006	0,000	0,003	-0,010	-0,011
it61	0,116	0,301	0,494	-0,037	0,066	-0,148	0,027	-0,040	-0,102	0,011	0,019	0,180	-0,172	0,109
it62	0,608	-0,396	0,149	-0,272	0,038	-0,038	-0,040	-0,206	-0,009	-0,013	0,258	0,011	-0,007	-0,024
it63	0,183	0,048	0,091	-0,160	-0,080	0,233	0,031	0,490	-0,012	-0,142	0,216	0,029	-0,074	0,022
it64	0,007	-0,005	0,016	0,084	0,680	0,313	0,006	-0,007	0,000	0,007	0,003	-0,002	-0,005	-0,012
rti-Dist0	0,734	0,131	-0,040	0,506	-0,116	0,117	-0,006	-0,011	0,120	0,071	0,053	0,112	-0,046	0,050
renfam-1stQ	-0,285	-0,097	0,173	-0,080	-0,116	0,253	-0,090	-0,084	0,356	0,222	-0,099	0,325	-0,328	-0,151
renfam-masMedia	0,459	0,152	-0,318	0,041	0,233	-0,509	0,234	0,172	-0,130	-0,153	-0,014	-0,011	-0,091	-0,162
renfam-Me0	0,037	0,037	0,068	0,008	-0,007	0,009	-0,001	0,026	0,086	0,056	0,038	-0,002	0,008	0,042
renfam-MeMedia	-0,248	-0,082	0,166	-0,027	-0,112	0,259	-0,151	-0,104	-0,192	-0,044	0,090	-0,293	0,390	0,296
ninmuebles1	0,041	0,189	-0,210	0,528	-0,082	0,037	0,041	-0,143	0,086	-0,037	0,572	0,204	-0,128	0,037
ninmuebles2	0,185	0,161	-0,148	0,369	-0,075	0,075	-0,003	0,017	0,048	0,120	-0,267	0,013	0,008	0,025
ninmuebles3	0,244	0,075	-0,009	0,192	-0,024	-0,009	-0,016	0,059	0,081	0,083	-0,491	-0,124	0,137	-0,019
ninmuebles4	0,288	0,033	0,060	0,040	-0,037	0,070	-0,011	0,193	0,030	0,031	-0,347	-0,072	0,051	0,022
ninmuebles6	0,731	-0,289	0,329	-0,346	0,037	0,008	-0,024	-0,025	-0,037	-0,054	0,296	0,072	-0,073	0,016
numfam2	0,059	0,280	0,097	0,008	-0,103	0,233	-0,037	0,046	0,002	-0,024	0,069	-0,103	0,077	-0,231
numfam3	0,038	0,182	-0,246	-0,157	0,105	-0,150	0,031	-0,031	-0,004	-0,048	-0,045	0,135	-0,039	0,163

Continúa en la página siguiente...

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10	Dim.11	Dim.12	Dim.13	Dim.14
numfam4	0,005	0,186	-0,093	-0,140	0,070	-0,111	0,039	0,068	0,062	0,016	0,079	-0,047	0,116	0,141
numfam5	-0,007	0,068	-0,018	-0,066	0,034	-0,045	0,022	0,050	0,032	0,008	0,030	0,002	0,051	0,040
numfam6	-0,014	0,021	-0,007	-0,037	0,012	-0,015	0,012	0,021	0,018	0,004	0,011	-0,002	0,016	0,013
numfam7	-0,009	0,008	-0,003	-0,025	0,008	-0,011	0,004	0,010	0,000	-0,001	0,008	0,018	0,002	-0,001
numfam8	-0,003	0,005	-0,007	-0,005	0,003	-0,005	0,001	0,003	0,001	0,002	0,004	0,007	-0,004	0,005
numfam9	-0,001	0,005	-0,001	-0,001	0,001	-0,001	0,003	0,004	-0,001	0,004	-0,003	0,009	0,003	-0,007
numfam10	0,000	0,003	-0,006	-0,002	0,002	-0,003	0,007	0,000	0,003	0,001	0,003	0,001	-0,002	0,001
numfam11	0,001	0,005	-0,003	-0,003	0,002	-0,004	0,013	-0,002	0,001	0,005	0,001	-0,010	0,004	-0,010
numfam12	-0,001	0,003	-0,003	-0,003	0,002	-0,003	0,002	0,001	-0,001	0,001	-0,001	-0,003	0,000	-0,007
numfam13	0,000	0,002	-0,003	-0,003	0,001	-0,002	0,001	0,000	-0,007	0,003	0,000	0,000	-0,001	0,005
numfam16	0,003	0,001	-0,003	0,002	0,000	-0,001	0,000	-0,001	0,002	-0,004	-0,012	0,002	0,000	0,009
edad2	-0,238	-0,062	-0,026	-0,148	0,078	-0,132	0,022	0,068	-0,002	0,005	-0,009	0,044	-0,001	0,007
edad3	-0,035	0,126	-0,218	-0,122	0,130	-0,223	0,059	0,052	0,013	-0,055	0,030	0,094	0,033	0,286
edad4	0,174	0,132	-0,117	0,028	0,036	-0,069	0,044	0,028	0,097	0,035	-0,029	-0,002	-0,149	-0,218
edad5	0,236	0,045	0,094	0,084	-0,094	0,157	-0,024	-0,052	0,037	0,051	-0,011	-0,015	-0,015	-0,185
edad6	0,062	-0,152	0,224	0,256	-0,180	0,306	-0,112	-0,135	-0,215	-0,082	0,073	-0,209	0,225	0,135
edad7	-0,059	-0,055	0,049	-0,017	-0,011	0,023	0,009	0,002	0,033	0,022	-0,043	0,062	-0,062	-0,046
BIECorrector1	0,126	-0,081	0,040	0,022	-0,007	-0,005	0,029	-0,039	-0,009	-0,005	0,014	-0,045	0,042	-0,019
pormedtit-100	0,167	0,181	0,177	0,598	-0,053	-0,072	0,099	-0,070	0,106	0,000	0,316	-0,083	0,075	-0,059
pormedtit-50	0,351	0,114	-0,349	0,187	-0,116	0,219	-0,058	-0,041	0,014	0,046	-0,021	0,212	-0,222	0,107
pormedtit-Ma50	0,454	-0,093	0,191	-0,052	0,024	-0,029	-0,012	0,122	0,007	0,043	-0,296	-0,091	0,069	-0,002
pormedtit-Me50	0,284	-0,070	-0,086	-0,003	-0,003	0,020	-0,043	0,010	0,069	0,022	0,037	0,167	0,042	0,042
IP-2dQ	0,006	0,026	-0,023	0,061	0,077	-0,184	-0,762	0,146	-0,136	0,268	0,099	-0,054	-0,015	-0,214
IP-3thQ	0,011	-0,020	-0,049	-0,067	-0,073	0,153	0,555	0,026	-0,413	0,580	0,046	0,024	-0,037	0,102
IP-4thQ	0,005	0,015	0,030	-0,024	-0,037	0,061	0,333	-0,156	0,233	-0,486	-0,045	-0,109	0,079	-0,394
RentaCorrectora1	-0,437	-0,155	0,295	-0,015	-0,229	0,481	-0,215	-0,160	0,221	0,203	-0,023	0,123	-0,051	0,070
Year2009	0,011	-0,027	-0,021	-0,072	-0,087	0,182	0,635	-0,059	-0,180	0,202	0,010	-0,015	0,009	-0,173
Year2010	0,005	0,007	-0,059	-0,004	0,007	-0,019	-0,093	0,131	-0,369	0,546	0,072	0,005	-0,050	0,199
Year2011	0,006	0,026	0,000	0,063	0,073	-0,181	-0,702	0,083	-0,010	0,078	0,072	-0,055	0,008	-0,319
Year2012	-0,024	-0,027	0,046	0,023	0,027	-0,005	-0,050	-0,020	0,331	-0,378	-0,114	0,157	-0,034	0,596

Tabla A.18: Correlaciones de las Dimensiones 1 a la 14 con las variables originales discretizadas

	Dim.15	Dim.16	Dim.17	Dim.18	Dim.19	Dim.20	Dim.21	Dim.22	Dim.23	Dim.24	Dim.25	Dim.26	Dim.27	Dim.28
estcv2	-0,038	0,090	0,062	-0,073	0,024	-0,054	0,093	-0,095	-0,047	0,106	-0,027	0,123	0,044	-0,075
estcv3	0,170	-0,103	-0,064	0,347	-0,017	0,030	0,007	0,027	-0,009	0,160	0,126	-0,019	0,006	-0,075
estcv4	-0,143	-0,049	-0,025	0,067	0,002	0,038	-0,009	-0,040	-0,064	-0,417	-0,168	0,029	-0,024	0,145
act11	-0,040	-0,132	-0,045	0,055	-0,022	-0,092	-0,060	-0,045	-0,015	-0,091	0,084	0,098	-0,008	0,090
act12	0,236	-0,183	-0,077	0,037	0,008	0,096	-0,025	0,018	-0,045	0,141	-0,129	-0,069	-0,017	-0,075
act14	0,046	0,020	-0,007	-0,100	-0,019	0,595	0,407	0,125	0,009	-0,067	0,100	0,048	0,069	0,085
act21	0,101	-0,228	-0,091	0,094	-0,014	-0,057	-0,099	-0,037	-0,047	0,002	0,021	0,063	-0,019	0,066
act22	0,100	-0,048	-0,018	0,005	-0,009	0,141	0,065	0,014	-0,009	0,112	-0,093	-0,081	0,008	-0,134
act24	0,052	-0,014	-0,009	-0,107	-0,022	0,600	0,402	0,145	0,019	-0,029	0,103	0,048	0,077	0,082
tvh2	0,076	-0,073	-0,045	0,270	0,102	0,075	-0,020	-0,036	-0,024	0,148	0,168	0,038	0,027	-0,117
tvh3	-0,020	0,042	0,067	-0,204	0,016	-0,094	0,009	-0,109	-0,007	0,036	-0,151	0,313	0,618	0,170
tvh4	0,072	0,035	0,029	-0,288	0,001	0,007	0,021	0,205	0,185	-0,111	-0,030	-0,293	-0,214	-0,462
tvh9	0,092	0,073	0,028	0,053	-0,022	-0,057	-0,036	0,017	0,019	0,063	-0,019	-0,052	-0,068	0,468
ptvh-100	-0,069	-0,013	0,008	0,186	-0,028	0,033	-0,002	-0,117	-0,114	-0,005	0,061	0,105	-0,050	0,151
ptvh-50	-0,007	-0,073	-0,095	0,168	-0,030	0,058	-0,060	0,055	-0,016	-0,035	0,061	-0,064	-0,029	0,059
ptvh-Ma50	-0,032	-0,112	-0,042	0,067	-0,525	-0,150	0,191	0,087	-0,073	-0,026	-0,005	-0,021	-0,013	0,020
ptvh-Me50	-0,015	-0,048	-0,046	0,092	0,440	0,179	-0,202	-0,053	0,074	-0,042	-0,028	0,013	0,011	-0,002
pctvh-100	-0,028	0,123	0,137	-0,213	0,098	-0,146	0,239	-0,229	-0,124	0,307	-0,029	0,247	-0,076	-0,304
pctvh-50	-0,023	-0,032	-0,064	0,136	-0,020	0,046	-0,043	0,037	-0,010	-0,036	0,045	-0,051	-0,030	0,052
pctvh-Ma50	-0,002	-0,004	-0,020	0,038	0,557	0,223	-0,275	-0,072	0,114	0,004	-0,042	0,078	0,021	-0,058
pctvh-Me50	-0,045	-0,107	-0,042	0,075	-0,303	-0,078	0,107	0,042	-0,047	-0,065	-0,041	-0,007	-0,008	0,036

Continúa en la página siguiente...

	Dim.15	Dim.16	Dim.17	Dim.18	Dim.19	Dim.20	Dim.21	Dim.22	Dim.23	Dim.24	Dim.25	Dim.26	Dim.27	Dim.28
dvt-Dist0	-0,191	-0,032	-0,029	0,037	-0,056	0,067	-0,063	0,045	0,019	0,057	0,085	-0,068	-0,054	0,039
it11	-0,162	0,235	0,093	0,024	-0,064	0,015	-0,126	0,201	0,231	-0,053	-0,043	0,042	0,025	-0,002
it12	0,164	-0,084	-0,061	-0,117	-0,005	-0,009	-0,021	0,027	-0,022	-0,021	-0,030	-0,011	0,073	0,011
it13	-0,100	0,175	0,130	-0,088	0,135	-0,114	0,262	-0,211	-0,142	0,188	0,171	-0,112	-0,194	0,092
it14	-0,001	0,009	0,012	-0,001	0,011	0,010	-0,008	-0,007	0,001	-0,010	-0,006	0,003	-0,005	-0,002
it21	-0,096	0,138	0,121	0,017	-0,035	0,012	-0,099	0,173	0,187	-0,045	-0,020	0,041	0,015	0,002
it22	0,053	-0,042	0,138	-0,043	0,020	-0,009	-0,001	0,024	0,008	0,005	-0,002	0,004	0,008	0,057
it23	0,016	-0,136	0,288	-0,014	0,012	-0,017	0,072	-0,126	0,043	0,044	0,049	-0,069	-0,034	0,015
it24	0,000	0,003	0,013	-0,004	-0,004	-0,004	0,004	0,000	0,003	-0,008	-0,005	-0,003	0,001	-0,001
it31	0,017	0,069	-0,065	-0,013	0,006	-0,007	-0,035	0,095	0,010	-0,021	-0,011	0,012	0,033	-0,023
it32	0,006	0,160	-0,189	-0,048	0,017	-0,026	-0,013	0,075	-0,107	-0,009	-0,027	-0,020	0,067	-0,014
it33	0,001	0,122	-0,113	-0,002	0,028	-0,026	0,063	-0,057	-0,050	0,068	0,062	-0,044	-0,076	0,031
it34	0,004	0,004	-0,012	-0,003	-0,005	0,001	-0,005	-0,009	0,003	0,007	0,009	0,000	0,000	0,008
it41	0,137	-0,108	-0,099	-0,031	0,010	-0,001	0,048	-0,126	0,001	0,009	0,012	-0,002	-0,007	0,005
it42	-0,040	0,110	-0,107	-0,025	-0,059	0,008	0,017	-0,143	0,212	-0,001	-0,005	0,026	-0,035	0,025
it43	0,086	0,022	-0,170	0,040	-0,073	0,030	-0,064	-0,012	0,148	-0,038	-0,012	0,039	0,004	0,003
it44	0,005	-0,001	-0,007	0,002	-0,002	-0,001	-0,003	0,007	-0,004	0,012	-0,001	0,004	0,006	0,001
it51	0,189	-0,231	-0,038	-0,035	0,057	-0,023	0,088	-0,140	-0,172	0,034	0,021	-0,050	-0,004	0,016
it52	-0,079	0,023	0,128	0,002	-0,004	-0,016	0,002	0,030	0,037	0,002	-0,005	-0,008	-0,004	-0,005
it53	0,044	-0,018	0,049	0,073	0,012	0,015	-0,096	0,266	-0,066	-0,046	-0,032	0,057	0,031	0,019
it54	0,004	-0,003	-0,001	-0,003	-0,005	-0,004	0,006	0,001	-0,003	0,005	-0,001	-0,005	0,004	-0,007
it61	0,194	-0,230	-0,045	-0,043	0,066	-0,023	0,101	-0,149	-0,190	0,043	0,025	-0,053	-0,011	0,032
it62	-0,082	0,020	0,135	-0,001	-0,022	-0,010	0,000	0,021	0,048	0,017	-0,011	0,012	0,000	-0,043
it63	0,071	-0,067	0,048	0,081	-0,022	0,025	-0,141	0,297	-0,041	-0,090	-0,058	0,065	0,081	-0,013
it64	0,003	0,001	-0,005	-0,006	-0,009	-0,001	-0,005	0,004	-0,005	-0,001	-0,003	-0,001	0,003	0,006
rti-Dist0	0,053	0,065	0,033	-0,118	0,013	-0,035	0,006	0,046	0,034	0,011	0,005	-0,029	0,018	0,051
renfam-1stQ	-0,047	0,192	0,030	0,376	-0,089	0,084	-0,030	-0,002	-0,019	0,197	0,092	0,053	0,074	0,006
renfam-masMedia	-0,067	-0,044	-0,007	0,090	-0,027	0,033	-0,057	-0,021	0,017	0,094	0,079	-0,050	0,013	-0,125
renfam-Me0	0,017	0,005	0,015	-0,012	0,003	-0,006	0,020	0,024	-0,047	-0,023	0,030	0,039	-0,021	0,172
renfam-MeMedia	0,111	-0,133	-0,023	-0,413	0,107	-0,107	0,088	0,025	0,004	-0,267	-0,163	0,004	-0,076	0,105
ninmuebles1	0,056	0,137	-0,294	-0,122	-0,003	-0,034	0,011	0,025	-0,056	0,009	-0,009	-0,023	0,034	-0,020
ninmuebles2	0,017	-0,309	0,681	0,003	-0,017	0,026	-0,008	-0,046	0,264	-0,001	0,024	0,023	-0,059	0,097
ninmuebles3	-0,029	0,237	-0,122	-0,056	0,149	-0,082	-0,015	0,370	-0,522	0,026	-0,006	-0,099	0,123	-0,052
ninmuebles4	0,072	0,194	-0,493	-0,052	-0,151	0,053	0,018	-0,393	0,462	-0,015	0,003	0,068	-0,061	0,036
ninmuebles6	-0,010	-0,045	0,123	0,015	0,015	-0,016	-0,004	0,071	-0,031	-0,003	-0,008	-0,003	0,005	0,005
numfam2	-0,454	-0,164	-0,029	-0,065	-0,010	0,039	-0,007	-0,190	-0,115	-0,098	-0,087	-0,012	0,041	-0,136
numfam3	0,246	0,119	-0,005	-0,017	0,090	-0,108	0,116	0,273	0,067	0,096	0,075	0,523	-0,389	0,089
numfam4	0,162	0,248	0,177	0,033	-0,036	-0,012	0,017	-0,128	0,087	0,051	0,051	-0,525	0,337	0,065
numfam5	0,058	0,135	0,093	0,007	-0,007	-0,024	0,030	-0,051	0,030	-0,004	-0,079	0,055	0,186	-0,067
numfam6	0,023	0,074	0,047	0,002	0,011	-0,015	0,026	-0,043	0,003	-0,021	-0,070	0,092	0,172	-0,018
numfam7	0,020	0,033	0,025	0,015	0,035	0,006	0,020	-0,046	-0,019	-0,027	-0,056	0,051	0,122	-0,028
numfam8	-0,006	0,012	0,007	0,009	0,007	-0,008	0,017	-0,003	0,001	0,019	-0,043	-0,019	0,017	0,011
numfam9	0,010	-0,007	0,033	0,003	0,013	-0,016	0,027	-0,024	0,009	0,001	-0,022	-0,001	-0,010	0,008
numfam10	0,009	0,004	0,014	0,003	0,002	-0,005	0,008	-0,019	-0,005	0,000	-0,016	0,006	0,013	-0,018
numfam11	0,004	0,004	-0,003	0,017	-0,001	-0,002	-0,005	-0,003	0,017	-0,014	0,005	-0,002	-0,005	-0,021
numfam12	0,005	0,000	0,012	0,014	0,002	-0,003	0,001	-0,007	0,004	-0,006	0,000	-0,012	-0,015	0,038
numfam13	-0,003	0,004	0,001	0,006	-0,004	0,006	-0,004	-0,002	-0,003	0,005	-0,010	-0,006	-0,009	-0,006
numfam16	0,000	-0,002	0,022	0,004	-0,004	0,006	0,001	-0,005	0,003	-0,002	0,003	0,002	-0,004	-0,015
edad2	-0,069	-0,165	-0,094	-0,444	-0,057	-0,046	-0,215	0,230	0,170	0,384	0,359	-0,020	0,152	0,159
edad3	0,361	0,456	0,251	0,122	-0,094	0,104	-0,019	-0,183	-0,138	-0,235	-0,177	0,038	-0,055	-0,078
edad4	-0,268	-0,214	-0,163	0,262	0,319	-0,286	0,477	0,127	0,192	-0,141	-0,185	-0,029	0,056	0,029
edad5	-0,221	-0,055	0,019	-0,158	-0,205	0,300	-0,369	-0,252	-0,336	-0,082	0,073	0,110	-0,196	0,015
edad6	0,161	-0,069	-0,030	0,246	-0,003	-0,036	0,078	0,031	0,053	0,170	0,023	-0,010	0,095	-0,136
edad7	0,048	0,027	0,010	-0,017	0,003	0,003	-0,031	0,078	0,075	-0,041	-0,083	-0,250	-0,230	0,044
BIECorrector1	0,039	0,024	0,003	-0,078	-0,005	-0,004	0,034	-0,006	-0,024	-0,039	-0,077	0,046	-0,023	-0,011
pormedtit-100	-0,126	0,292	0,063	-0,052	-0,045	-0,034	-0,026	0,094	0,130	0,044	0,007	0,050	-0,035	0,026
pormedtit-50	0,267	-0,299	-0,099	-0,099	-0,090	0,034	-0,068	-0,003	0,019	-0,052	-0,095	0,033	0,160	-0,150
pormedtit-Ma50	-0,024	0,026	0,120	0,030	-0,019	-0,023	0,054	0,028	-0,055	0,041	0,007	0,074	-0,033	-0,059

Continúa en la página siguiente...

	Dim.15	Dim.16	Dim.17	Dim.18	Dim.19	Dim.20	Dim.21	Dim.22	Dim.23	Dim.24	Dim.25	Dim.26	Dim.27	Dim.28
pormedtit-Me50	-0,059	0,124	-0,037	-0,072	0,256	-0,052	0,095	-0,067	-0,104	-0,010	0,139	-0,272	-0,106	0,341
IP-2dQ	0,110	0,029	0,007	0,005	0,004	-0,043	-0,022	-0,003	-0,001	-0,036	-0,009	0,031	0,007	0,010
IP-3thQ	-0,090	0,081	-0,011	0,018	0,011	-0,042	-0,002	0,021	0,010	-0,094	0,130	0,025	0,020	-0,038
IP-4thQ	0,302	-0,069	0,005	-0,047	-0,009	0,019	-0,073	-0,006	-0,010	0,178	-0,301	-0,070	-0,094	0,161
RentaCorrectora1	0,055	0,086	0,004	0,095	-0,009	0,006	0,037	0,015	-0,024	0,006	-0,026	0,055	0,013	0,088
Year2009	0,142	0,046	0,014	-0,004	0,046	-0,134	-0,003	-0,019	0,004	-0,283	0,384	0,043	0,057	-0,051
Year2010	-0,185	0,036	-0,037	0,015	-0,046	0,113	-0,032	0,056	0,000	0,324	-0,489	-0,041	-0,090	0,098
Year2011	0,201	0,029	0,028	0,002	0,035	-0,116	-0,006	-0,032	0,005	-0,252	0,321	0,055	0,060	-0,059
Year2012	-0,383	-0,048	-0,006	0,025	-0,015	0,091	0,104	-0,006	-0,002	0,000	0,117	0,010	0,063	-0,130

Tabla A.19: Correlaciones de las Dimensiones 15 a la 28 con las variables originales discretizadas

	Dim.29	Dim.30	Dim.31	Dim.32	Dim.33	Dim.34	Dim.35	Dim.36	Dim.37	Dim.38	Dim.39	Dim.40	Dim.41	Dim.42
estcv2	-0,076	-0,004	-0,031	-0,016	-0,008	0,027	0,026	-0,002	0,001	0,001	-0,004	0,007	0,000	-0,050
estcv3	0,043	-0,002	0,129	0,089	-0,021	-0,050	-0,014	0,009	-0,001	0,003	-0,001	0,010	0,043	0,047
estcv4	0,088	-0,136	-0,153	-0,049	-0,003	0,005	-0,031	-0,024	0,000	0,001	0,000	-0,015	0,011	0,019
act11	-0,243	-0,032	0,099	-0,022	-0,028	-0,028	-0,023	0,009	0,009	-0,005	0,000	0,013	-0,005	0,010
act12	0,377	0,169	-0,143	0,029	0,061	0,040	0,032	-0,009	-0,013	-0,003	0,003	-0,012	-0,042	-0,068
act14	-0,127	-0,011	0,040	-0,005	-0,016	-0,003	0,001	-0,003	0,003	0,003	0,000	0,005	0,004	-0,011
act21	-0,104	0,030	0,125	0,049	0,050	0,048	0,034	-0,007	0,001	0,001	-0,001	0,005	-0,089	-0,076
act22	0,494	0,205	-0,425	-0,148	-0,087	-0,133	-0,082	0,028	-0,010	-0,023	0,011	-0,008	0,166	0,111
act24	-0,087	0,023	0,087	0,025	0,031	0,046	0,024	-0,007	0,002	0,002	0,000	-0,003	-0,063	-0,046
tvh2	0,023	-0,006	0,169	0,248	-0,188	-0,150	-0,050	0,008	0,011	0,041	-0,006	0,043	0,312	0,080
tvh3	0,223	-0,039	0,081	0,160	-0,021	-0,125	-0,036	0,016	0,003	0,014	0,011	-0,003	0,151	0,088
tvh4	0,035	-0,188	0,140	0,052	0,002	0,006	-0,007	0,001	-0,003	-0,016	-0,011	0,018	-0,009	0,076
tvh9	-0,134	0,598	-0,065	-0,063	0,054	-0,002	0,000	0,013	-0,007	-0,002	0,014	-0,027	-0,055	-0,078
ptvh-100	-0,067	-0,101	-0,107	-0,122	0,022	0,059	0,016	-0,010	-0,011	-0,006	0,000	-0,005	-0,089	-0,046
ptvh-50	0,059	-0,054	0,001	-0,025	0,003	-0,007	-0,018	0,006	0,003	0,002	0,002	-0,004	-0,009	0,038
ptvh-Ma50	0,026	-0,014	-0,011	0,008	0,027	-0,013	-0,002	0,007	0,002	-0,008	-0,001	-0,008	-0,017	0,015
ptvh-Me50	-0,035	-0,010	-0,024	-0,022	-0,020	0,009	-0,018	0,001	-0,002	0,005	0,001	-0,002	0,019	0,013
pctvh-100	-0,249	0,241	-0,075	0,137	-0,106	0,034	0,101	-0,038	0,033	0,030	-0,012	0,026	0,130	-0,234
pctvh-50	0,040	-0,049	-0,004	-0,030	0,005	-0,001	-0,013	0,007	0,001	0,001	0,002	-0,004	-0,013	0,029
pctvh-Ma50	-0,112	0,034	-0,108	-0,130	0,086	0,020	-0,025	0,015	-0,017	-0,027	0,004	-0,030	-0,094	0,051
pctvh-Me50	0,004	-0,029	-0,012	-0,005	-0,006	0,005	-0,002	-0,003	-0,001	0,002	0,000	0,000	0,002	-0,002
dvt-Dist0	-0,010	0,093	-0,023	-0,077	0,034	0,025	0,009	0,011	-0,006	-0,019	0,005	-0,014	-0,078	-0,002
it11	0,021	0,037	-0,008	0,046	-0,013	-0,020	0,007	-0,003	-0,005	-0,001	0,000	0,010	0,019	-0,001
it12	-0,016	0,011	-0,013	-0,008	-0,006	0,029	0,020	-0,006	0,011	0,006	-0,002	0,007	-0,009	-0,050
it13	0,103	-0,077	0,041	-0,058	0,102	-0,061	-0,097	0,023	-0,027	-0,024	0,010	-0,042	-0,049	0,167
it14	0,010	0,009	0,003	-0,114	0,049	0,023	-0,007	-0,004	-0,008	-0,011	0,000	-0,022	-0,080	0,017
it21	0,040	0,037	0,009	0,059	-0,022	-0,020	0,016	-0,008	-0,001	-0,001	-0,001	0,007	0,023	-0,029
it22	0,020	-0,032	-0,001	0,013	-0,017	0,029	0,035	-0,007	0,020	0,010	-0,002	0,004	0,008	-0,081
it23	0,016	-0,019	0,014	-0,059	0,057	-0,028	-0,071	0,008	-0,047	-0,040	0,006	-0,034	-0,066	0,150
it24	0,004	0,003	0,004	-0,057	0,024	0,011	-0,004	-0,003	-0,004	-0,006	0,000	-0,011	-0,039	0,009
it31	0,013	0,039	0,012	0,036	-0,017	-0,005	0,005	-0,008	-0,006	-0,003	-0,001	0,009	0,005	-0,011
it32	-0,026	0,014	-0,008	0,000	-0,007	0,018	0,019	-0,014	-0,001	0,000	0,000	0,005	-0,011	-0,020
it33	0,018	-0,045	0,000	-0,089	0,082	-0,031	-0,070	0,020	-0,019	-0,019	0,004	-0,031	-0,061	0,133
it34	-0,003	-0,004	0,000	0,030	-0,012	-0,005	0,002	0,002	0,002	0,003	0,000	0,006	0,018	-0,004
it41	0,003	-0,008	0,008	0,001	-0,006	0,016	0,007	-0,003	0,004	0,003	-0,002	0,005	-0,007	-0,031
it42	0,023	-0,007	-0,013	0,006	0,002	0,001	0,004	-0,003	0,000	-0,001	0,001	0,002	-0,002	-0,006
it43	-0,006	-0,001	0,001	0,012	-0,013	-0,002	0,008	0,000	0,002	0,001	-0,001	0,006	0,006	-0,009
it44	-0,011	-0,006	-0,004	0,090	-0,042	-0,019	0,006	0,003	0,007	0,010	0,000	0,018	0,068	-0,015
it51	-0,024	-0,041	0,002	-0,036	0,015	0,027	-0,007	0,001	0,003	0,002	-0,001	-0,008	-0,025	-0,006
it52	0,014	0,008	-0,010	-0,020	0,028	-0,011	-0,021	-0,004	-0,010	-0,012	0,001	-0,011	-0,025	0,038
it53	-0,005	0,006	0,001	0,073	-0,064	0,028	0,059	-0,003	0,026	0,030	-0,003	0,027	0,054	-0,121
it54	-0,001	-0,002	-0,003	0,012	-0,002	-0,001	0,000	0,001	0,000	0,000	0,000	0,001	0,006	0,001

Continúa en la página siguiente...

	Dim.29	Dim.30	Dim.31	Dim.32	Dim.33	Dim.34	Dim.35	Dim.36	Dim.37	Dim.38	Dim.39	Dim.40	Dim.41	Dim.42
it61	-0,013	-0,054	0,011	-0,038	0,015	0,027	-0,003	0,006	0,006	0,001	-0,001	-0,012	-0,017	-0,003
it62	-0,028	0,024	-0,024	-0,052	0,050	-0,025	-0,049	0,000	-0,022	-0,020	0,003	-0,021	-0,033	0,087
it63	-0,023	0,030	-0,004	0,091	-0,084	0,030	0,081	-0,017	0,027	0,028	-0,005	0,035	0,063	-0,146
it64	-0,005	-0,001	-0,006	0,021	-0,008	-0,002	0,003	0,001	0,001	0,002	0,000	0,004	0,012	-0,006
rti-Dist0	0,028	0,005	-0,004	-0,006	0,019	0,002	-0,004	-0,001	0,001	-0,001	0,002	-0,002	-0,019	-0,002
renfam-1stQ	0,014	-0,098	-0,057	-0,059	0,009	0,043	0,014	-0,004	0,001	-0,006	-0,002	-0,003	-0,053	-0,029
renfam-masMedia	-0,038	0,034	0,116	0,049	0,003	-0,019	-0,010	0,005	0,001	-0,003	-0,003	0,004	0,003	0,032
renfam-Me0	-0,127	-0,107	-0,056	-0,117	-0,083	-0,172	-0,058	0,056	-0,017	0,010	0,043	-0,034	0,287	0,511
renfam-MeMedia	0,028	0,062	-0,061	0,013	-0,008	-0,008	0,002	-0,005	-0,001	0,009	0,001	0,002	0,027	-0,040
ninmuebles1	-0,013	0,033	-0,014	-0,019	0,026	-0,008	-0,019	0,006	0,002	0,005	0,002	0,005	-0,009	0,031
ninmuebles2	0,072	-0,046	0,015	0,033	-0,020	0,007	0,019	0,002	0,010	0,000	-0,001	-0,007	0,018	-0,063
ninmuebles3	-0,051	0,030	0,007	-0,043	0,029	0,002	-0,016	-0,010	-0,016	-0,013	0,002	-0,012	-0,044	0,057
ninmuebles4	0,026	-0,014	-0,004	0,028	-0,023	0,001	0,021	-0,001	0,004	0,004	0,000	0,016	0,018	-0,028
ninmuebles6	0,005	-0,001	-0,008	-0,006	0,011	0,005	-0,004	-0,004	-0,001	-0,002	0,000	-0,004	-0,013	-0,002
numfam2	0,029	0,009	-0,041	-0,069	0,068	0,096	0,040	-0,015	0,018	-0,022	-0,005	-0,006	-0,133	-0,084
numfam3	0,105	-0,094	-0,078	0,070	0,015	-0,054	-0,020	0,005	-0,012	-0,011	0,000	-0,008	0,023	0,031
numfam4	-0,280	-0,075	-0,266	0,117	-0,068	-0,051	0,012	-0,018	0,008	0,024	-0,006	0,005	0,138	-0,048
numfam5	0,170	0,131	0,584	-0,605	-0,113	0,014	0,003	-0,063	0,034	0,054	-0,018	0,012	0,084	-0,031
numfam6	0,083	0,086	0,133	0,394	0,079	0,390	-0,482	0,068	0,010	-0,071	-0,012	-0,107	-0,301	0,287
numfam7	0,052	-0,034	0,105	0,175	0,491	-0,566	0,387	0,045	-0,027	-0,117	0,020	0,063	-0,250	0,071
numfam8	0,018	-0,001	0,053	0,006	-0,090	0,117	0,098	0,542	-0,750	-0,017	0,211	0,212	-0,018	-0,057
numfam9	0,007	0,114	-0,032	0,072	-0,376	0,030	0,267	-0,489	-0,141	0,220	0,141	0,402	-0,358	0,369
numfam10	0,012	0,047	-0,010	0,021	-0,439	-0,006	0,342	0,552	0,432	-0,067	-0,071	-0,202	-0,310	0,163
numfam11	0,027	0,034	0,029	0,069	0,225	0,315	0,421	-0,081	-0,112	0,431	0,204	-0,574	0,196	0,192
numfam12	0,012	0,081	0,023	0,019	0,203	0,427	0,367	-0,022	0,039	-0,488	-0,379	0,288	0,299	0,253
numfam13	-0,005	-0,002	0,023	-0,006	0,047	0,107	0,024	-0,032	0,270	-0,390	0,858	0,095	0,108	-0,013
numfam16	0,007	0,013	-0,036	-0,031	0,320	0,132	-0,061	0,322	0,354	0,576	0,060	0,542	0,091	0,056
edad2	0,008	-0,104	-0,065	-0,122	0,031	0,053	0,015	0,000	0,000	-0,007	0,008	-0,003	-0,090	0,019
edad3	0,037	-0,028	-0,039	-0,027	0,028	0,057	0,020	-0,019	0,000	-0,009	-0,005	-0,003	-0,064	-0,057
edad4	-0,094	0,010	-0,012	0,025	-0,054	-0,044	-0,020	0,002	0,012	0,015	-0,002	0,006	0,103	0,007
edad5	0,011	0,039	0,079	0,158	-0,096	-0,106	-0,022	0,013	-0,017	0,029	-0,001	0,023	0,171	0,046
edad6	0,050	-0,017	-0,010	-0,071	0,090	0,076	0,027	-0,006	0,002	-0,029	-0,001	-0,017	-0,152	-0,031
edad7	-0,070	0,434	0,259	0,126	0,067	-0,194	-0,120	0,079	0,011	-0,005	0,013	-0,053	0,060	0,132
BIECorrector1	-0,104	-0,074	-0,085	-0,153	-0,125	-0,030	0,008	0,143	-0,031	-0,010	0,017	-0,011	0,111	0,143
pormedtit-100	-0,007	-0,020	-0,029	-0,028	0,020	0,001	0,000	0,006	0,001	-0,005	0,002	-0,006	-0,019	0,013
pormedtit-50	-0,134	0,117	-0,014	-0,032	0,021	-0,035	-0,040	-0,009	-0,033	-0,022	0,004	-0,017	0,008	0,076
pormedtit-Ma50	-0,101	0,005	-0,132	-0,115	0,086	-0,014	-0,057	0,029	-0,007	-0,002	0,004	-0,016	-0,046	0,074
pormedtit-Me50	0,388	-0,150	0,207	0,211	-0,121	0,074	0,115	-0,029	0,063	0,043	-0,011	0,050	0,034	-0,225
IP-2dQ	-0,037	-0,048	0,022	0,014	-0,012	-0,004	-0,003	0,002	0,003	0,008	0,000	0,003	0,010	0,002
IP-3thQ	0,006	0,042	-0,013	-0,001	-0,002	-0,001	0,002	0,003	0,000	-0,002	0,000	0,005	0,001	-0,011
IP-4thQ	-0,100	-0,212	0,074	-0,024	0,004	0,024	-0,009	-0,015	0,016	0,003	-0,004	0,001	-0,037	0,010
RentaCorrectora1	0,049	-0,042	-0,104	-0,046	0,005	0,023	0,015	-0,009	0,001	0,002	0,002	-0,005	-0,013	-0,029
Year2009	0,007	0,050	-0,032	-0,029	0,007	0,001	-0,003	0,013	-0,015	-0,010	0,009	0,008	-0,022	-0,010
Year2010	-0,069	-0,136	0,070	0,037	-0,014	-0,001	0,000	-0,019	0,023	0,017	-0,017	0,001	0,018	0,008
Year2011	0,014	0,058	-0,025	-0,013	-0,002	0,000	-0,001	0,014	-0,010	-0,007	0,010	0,005	-0,001	-0,008
Year2012	0,132	0,213	-0,081	0,019	0,008	-0,021	0,009	0,008	-0,017	-0,005	0,004	-0,012	0,028	0,001

Tabla A.20: Correlaciones de las Dimensiones 29 a la 42 con las variables originales discretizadas

	Dim.43	Dim.44	Dim.45	Dim.46	Dim.47	Dim.48	Dim.49	Dim.50	Dim.51	Dim.52	Dim.53	Dim.54	Dim.55	Dim.56
estcv2	0,071	0,003	0,000	-0,014	0,085	-0,046	0,070	0,011	-0,056	-0,169	0,019	-0,007	-0,008	0,039
estcv3	-0,026	0,010	-0,083	-0,015	-0,016	-0,188	-0,192	-0,019	0,049	0,086	-0,033	0,005	0,055	-0,038
estcv4	-0,139	-0,002	0,058	0,057	-0,132	0,270	0,186	0,115	0,029	0,483	0,039	0,025	0,064	-0,096
act11	-0,022	0,081	0,003	-0,159	0,306	0,113	-0,216	-0,054	0,044	0,077	-0,067	0,169	0,004	-0,018
act12	0,008	-0,077	-0,002	0,222	-0,348	-0,129	0,203	0,047	-0,055	-0,097	0,038	-0,086	-0,012	0,014

Continúa en la página siguiente...

	Dim.43	Dim.44	Dim.45	Dim.46	Dim.47	Dim.48	Dim.49	Dim.50	Dim.51	Dim.52	Dim.53	Dim.54	Dim.55	Dim.56
act14	-0,032	0,017	0,001	-0,080	0,139	0,062	-0,082	-0,013	0,018	0,071	0,122	-0,571	-0,005	0,051
act21	-0,046	-0,022	-0,001	0,122	-0,088	-0,021	-0,018	-0,018	0,008	-0,002	0,016	-0,127	-0,008	0,012
act22	0,109	0,091	0,010	-0,301	0,355	0,103	-0,147	-0,007	0,003	0,054	-0,038	0,106	-0,005	-0,010
act24	-0,018	-0,037	0,005	0,110	-0,100	-0,036	0,046	0,005	-0,020	-0,042	-0,127	0,570	0,010	-0,058
tvh2	-0,104	0,056	0,022	-0,077	-0,135	0,545	0,276	-0,150	0,047	-0,181	0,095	0,030	0,078	0,076
tvh3	-0,289	-0,011	-0,066	0,234	0,127	-0,112	-0,093	-0,102	0,075	-0,015	0,021	-0,002	-0,002	-0,015
tvh4	0,009	-0,034	-0,005	-0,038	-0,080	-0,002	-0,124	-0,139	0,085	-0,060	0,042	0,000	0,007	-0,014
tvh9	0,035	-0,014	0,040	-0,101	-0,243	0,151	-0,245	-0,225	0,124	-0,053	0,056	0,004	-0,010	-0,045
ptvh-100	0,085	0,026	0,030	-0,081	0,151	-0,053	0,157	0,241	-0,185	-0,284	-0,074	-0,012	-0,032	0,173
ptvh-50	-0,010	0,035	0,005	0,019	-0,019	0,006	0,074	-0,001	0,046	0,183	0,024	0,010	0,023	-0,072
ptvh-Ma50	0,001	0,009	-0,018	0,008	-0,001	-0,030	0,006	-0,017	0,058	0,065	0,135	0,046	0,371	0,186
ptvh-Me50	-0,028	0,007	0,022	-0,008	-0,011	0,028	0,019	0,027	-0,043	0,014	-0,122	-0,038	-0,325	-0,206
pctvh-100	0,178	-0,077	-0,042	0,029	-0,028	0,099	0,055	0,112	-0,123	0,345	-0,059	0,000	0,071	-0,053
pctvh-50	0,004	0,038	0,003	0,007	0,001	-0,018	0,075	0,007	0,016	0,090	0,040	0,012	0,011	-0,057
pctvh-Ma50	-0,017	-0,004	0,003	0,020	0,026	-0,141	-0,028	0,004	0,088	0,063	0,186	0,063	0,423	0,233
pctvh-Me50	-0,011	0,008	0,001	-0,019	0,001	0,007	0,025	0,055	-0,056	-0,012	-0,048	-0,015	-0,219	-0,095
dvt-Dist0	0,088	0,012	0,047	0,051	0,027	0,072	0,034	0,024	0,020	0,015	0,000	-0,017	-0,023	0,035
it11	-0,033	-0,010	-0,006	-0,002	-0,047	0,032	-0,027	-0,049	-0,140	0,012	-0,038	-0,015	0,061	-0,060
it12	0,037	-0,004	0,011	-0,013	0,043	0,014	0,050	-0,072	-0,028	-0,042	-0,070	-0,022	0,067	-0,031
it13	-0,185	0,035	0,028	-0,002	-0,089	-0,018	-0,001	0,161	0,431	0,076	0,099	0,054	-0,236	0,195
it14	0,017	0,001	-0,664	0,026	0,007	0,124	0,039	-0,014	0,008	-0,020	-0,010	-0,001	-0,008	0,077
it21	-0,023	-0,019	-0,004	0,015	-0,068	0,062	-0,079	0,086	-0,055	0,050	-0,105	0,003	-0,056	0,201
it22	0,099	-0,007	-0,002	-0,010	0,064	-0,042	0,090	-0,120	0,024	0,052	0,169	0,050	-0,126	0,087
it23	-0,178	-0,004	0,043	0,020	-0,077	0,148	-0,186	0,319	-0,067	-0,184	-0,046	-0,053	0,264	-0,324
it24	0,008	0,000	-0,325	0,010	0,002	0,058	0,010	0,003	0,003	-0,006	0,004	0,001	-0,014	-0,057
it31	-0,025	-0,023	0,000	0,012	-0,055	0,080	-0,094	0,221	0,066	-0,015	-0,169	-0,019	0,002	0,146
it32	0,031	-0,044	0,005	-0,002	0,027	0,045	-0,012	0,145	0,091	-0,073	0,158	0,027	0,011	-0,117
it33	-0,181	0,055	0,015	-0,007	-0,049	-0,028	0,009	-0,347	-0,418	0,134	-0,042	0,010	-0,057	0,166
it34	-0,002	0,000	0,172	-0,003	0,004	-0,025	-0,011	0,010	0,020	0,011	-0,025	0,000	-0,003	0,051
it41	0,010	0,012	-0,001	-0,006	0,013	0,002	0,008	0,055	0,092	-0,008	-0,079	-0,017	0,004	0,048
it42	0,010	-0,036	-0,002	-0,002	-0,003	0,007	0,027	0,035	0,075	-0,011	0,018	0,010	-0,013	0,007
it43	-0,001	0,018	-0,010	0,007	-0,004	0,018	-0,040	-0,035	-0,211	0,000	0,092	0,010	0,006	-0,048
it44	-0,012	0,000	0,535	-0,024	-0,005	-0,097	-0,015	0,003	-0,007	-0,001	0,000	-0,001	0,022	0,051
it51	0,005	0,030	0,002	-0,015	0,052	-0,034	0,052	-0,087	-0,021	-0,015	0,076	0,010	0,020	-0,090
it52	-0,076	-0,029	0,013	0,004	-0,036	0,056	-0,047	0,043	-0,109	-0,017	-0,012	-0,008	0,028	0,001
it53	0,134	0,032	-0,025	-0,018	0,050	-0,097	0,110	-0,096	0,183	0,080	-0,025	0,022	-0,117	0,170
it54	-0,003	-0,002	0,053	0,001	-0,004	-0,009	-0,006	0,001	-0,009	-0,002	0,036	0,001	0,003	-0,137
it61	0,046	0,020	0,003	-0,007	0,070	-0,053	0,050	-0,129	-0,086	0,004	0,279	0,052	-0,036	-0,126
it62	-0,129	-0,013	0,023	-0,008	-0,057	0,048	-0,039	0,003	-0,136	0,019	0,010	0,011	-0,047	0,089
it63	0,167	-0,008	-0,041	-0,011	0,063	-0,024	0,032	0,143	0,236	-0,064	0,065	-0,003	0,026	-0,123
it64	0,001	0,001	0,133	-0,003	0,000	-0,035	-0,011	0,006	-0,018	0,014	-0,003	0,000	-0,002	0,037
rti-Dist0	-0,029	-0,002	0,011	-0,019	-0,009	0,023	0,032	-0,034	0,035	-0,021	-0,067	-0,015	0,024	0,014
renfam-1stQ	0,018	0,061	0,021	0,055	0,049	0,024	0,061	0,056	-0,021	-0,007	0,003	-0,001	0,002	0,018
renfam-masMedia	-0,014	-0,037	-0,016	0,045	0,017	-0,018	-0,170	-0,070	0,047	0,056	-0,007	0,004	-0,024	-0,044
renfam-Me0	0,449	-0,515	-0,012	0,156	-0,085	-0,053	-0,015	-0,026	-0,023	0,009	-0,042	-0,025	0,007	0,001
renfam-MeMedia	-0,030	0,014	-0,003	-0,109	-0,052	-0,001	0,113	0,026	-0,025	-0,052	0,013	0,001	0,017	0,033
ninmuebles1	-0,059	0,019	0,001	-0,022	-0,012	-0,015	0,065	-0,103	0,075	-0,014	-0,187	-0,047	0,072	-0,031
ninmuebles2	0,067	0,015	-0,001	0,007	0,019	-0,012	0,025	-0,057	-0,002	0,030	0,053	0,014	-0,044	0,051
ninmuebles3	-0,077	-0,018	0,017	0,001	-0,014	0,078	-0,082	0,107	-0,074	-0,032	0,093	0,025	-0,002	-0,020
ninmuebles4	0,061	-0,026	-0,007	0,005	0,007	0,002	0,017	0,078	0,112	-0,027	0,041	0,010	-0,008	-0,029
ninmuebles6	-0,029	-0,009	0,006	-0,005	-0,005	0,014	0,003	-0,012	-0,052	0,005	0,001	0,002	-0,004	0,026
numfam2	0,073	-0,170	0,102	0,230	0,170	0,258	-0,030	-0,184	0,080	-0,211	0,018	0,004	-0,003	0,041
numfam3	-0,139	0,020	-0,025	0,020	0,024	-0,029	-0,026	0,054	-0,050	-0,074	0,035	-0,007	0,020	-0,043
numfam4	-0,056	0,070	-0,076	-0,082	-0,058	-0,117	0,105	0,080	-0,053	0,004	0,015	0,006	0,020	0,000
numfam5	0,099	0,051	0,075	-0,196	-0,090	-0,043	0,134	0,073	-0,031	0,055	0,000	-0,010	0,038	-0,014
numfam6	0,315	0,117	0,017	-0,200	-0,032	0,030	0,082	0,032	-0,025	0,050	-0,008	-0,013	0,005	-0,007
numfam7	0,261	0,108	0,032	-0,172	-0,049	0,070	0,056	0,021	-0,025	0,035	-0,013	-0,005	-0,006	-0,022
numfam8	-0,029	-0,073	-0,005	-0,049	0,025	0,007	0,033	-0,001	-0,003	0,029	0,003	0,004	0,003	-0,001

Continúa en la página siguiente...

	Dim.43	Dim.44	Dim.45	Dim.46	Dim.47	Dim.48	Dim.49	Dim.50	Dim.51	Dim.52	Dim.53	Dim.54	Dim.55	Dim.56
numfam9	-0,043	0,055	-0,001	0,019	0,041	-0,022	0,041	-0,016	0,011	0,011	0,008	0,003	-0,002	0,009
numfam10	-0,112	-0,062	-0,004	-0,061	-0,015	-0,009	0,032	0,013	-0,001	0,005	0,005	0,001	-0,005	0,005
numfam11	-0,037	0,065	0,009	-0,064	0,050	0,026	-0,005	0,006	0,002	0,002	-0,008	0,004	0,003	0,003
numfam12	-0,073	0,042	-0,009	-0,027	0,044	-0,003	0,024	0,010	0,002	0,010	-0,003	0,003	-0,003	0,005
numfam13	-0,031	0,014	-0,006	-0,014	-0,005	-0,005	0,007	0,005	-0,003	-0,003	0,000	0,000	0,001	0,000
numfam16	-0,022	0,000	-0,007	-0,006	-0,004	0,023	-0,031	0,038	-0,019	-0,001	0,013	0,004	-0,002	0,005
edad2	0,052	0,032	0,060	-0,008	0,037	0,183	0,185	0,053	-0,016	0,058	-0,005	-0,011	0,019	-0,012
edad3	0,005	-0,110	0,064	0,105	0,055	0,158	-0,023	-0,103	0,037	-0,117	0,000	0,008	-0,006	0,043
edad4	0,004	0,064	-0,047	-0,151	-0,058	-0,103	0,027	0,023	-0,036	-0,107	0,021	0,010	0,014	0,030
edad5	-0,051	0,116	-0,113	-0,193	-0,116	-0,247	-0,030	0,064	-0,036	0,012	0,004	0,006	0,024	-0,021
edad6	0,022	-0,097	0,040	0,240	0,087	0,008	-0,129	0,019	0,027	0,226	-0,033	-0,018	-0,044	-0,060
edad7	-0,107	0,047	-0,002	0,343	0,460	-0,063	0,303	0,148	-0,074	0,029	-0,008	-0,005	0,013	0,006
BIECorrector1	0,287	0,725	0,032	0,418	-0,132	0,066	-0,141	-0,020	0,019	-0,042	-0,026	-0,017	-0,007	0,014
pormedtit-100	0,011	0,024	0,008	-0,011	0,036	-0,048	0,076	-0,064	-0,007	-0,027	0,370	0,062	-0,045	-0,182
pormedtit-50	-0,102	-0,047	0,016	-0,018	-0,083	0,128	-0,124	0,193	-0,049	0,008	-0,049	0,024	-0,135	0,268
pormedtit-Ma50	-0,110	0,049	0,021	-0,070	0,005	-0,116	0,287	-0,277	0,259	-0,029	-0,426	-0,115	0,158	-0,157
pormedtit-Me50	0,198	-0,014	-0,045	0,094	0,060	0,054	-0,159	0,001	-0,111	0,031	-0,187	-0,040	0,168	0,045
IP-2dQ	-0,015	0,011	-0,007	-0,013	0,036	0,009	-0,017	-0,002	0,011	0,001	0,006	0,022	0,003	-0,020
IP-3thQ	0,014	0,001	-0,001	0,011	0,006	-0,006	0,001	-0,021	-0,007	0,013	-0,001	0,005	-0,006	0,013
IP-4thQ	-0,016	-0,017	0,021	-0,064	0,070	0,099	0,002	0,016	0,016	0,027	-0,038	0,019	0,017	0,016
RentaCorrectora1	0,018	0,002	0,011	-0,009	-0,013	-0,001	0,145	0,065	-0,042	-0,026	0,004	-0,010	0,026	0,034
Year2009	0,015	-0,006	0,012	0,012	0,042	0,009	0,012	0,004	-0,007	0,008	-0,006	0,014	0,002	0,004
Year2010	-0,026	0,006	-0,012	-0,045	0,008	0,031	-0,028	-0,015	0,018	0,011	-0,005	0,015	0,004	-0,003
Year2011	0,012	0,002	0,003	0,020	0,021	-0,011	0,008	-0,003	-0,008	0,002	0,001	0,005	-0,004	0,001
Year2012	0,011	0,008	-0,017	0,066	-0,117	-0,106	0,009	0,008	-0,016	-0,045	0,040	-0,044	-0,014	-0,018

Tabla A.21: Correlaciones de las Dimensiones 43 a la 56 con las variables originales discretizadas

	Dim.57	Dim.58	Dim.59	Dim.60	Dim.61	Dim.62	Dim.63	Dim.64	Dim.65	Dim.66	Dim.67	Dim.68	Dim.69	Dim.70
estcv2	-0,014	0,011	-0,042	-0,193	-0,056	0,008	-0,033	0,006	0,022	0,269	0,026	0,030	0,169	-0,006
estcv3	0,020	-0,028	0,066	0,300	0,267	-0,066	0,181	-0,022	0,057	0,231	0,013	0,009	0,059	0,002
estcv4	0,017	0,017	-0,031	-0,095	0,035	-0,010	0,040	0,001	0,031	0,199	0,014	0,019	0,085	0,000
act11	-0,005	0,009	-0,001	-0,020	-0,015	0,005	-0,014	0,000	-0,004	-0,002	-0,009	-0,005	0,000	0,001
act12	0,002	-0,007	-0,001	0,021	0,003	-0,001	0,000	0,003	0,003	0,005	0,012	0,004	0,004	-0,001
act14	-0,008	-0,002	-0,001	0,004	0,006	-0,001	0,007	0,000	-0,002	-0,004	0,002	-0,002	-0,003	-0,002
act21	-0,005	0,005	-0,005	-0,010	-0,014	0,005	-0,015	0,000	-0,003	-0,007	0,002	-0,004	0,000	-0,002
act22	0,000	0,004	-0,003	-0,010	-0,006	0,002	-0,005	0,000	-0,002	0,001	0,001	0,000	-0,001	-0,003
act24	0,009	0,002	0,008	-0,001	-0,008	0,002	-0,007	0,000	0,002	0,004	-0,002	0,003	0,002	0,001
tvh2	-0,019	-0,001	-0,020	-0,070	-0,074	0,002	-0,045	0,000	-0,003	0,003	0,007	-0,016	-0,029	0,000
tvh3	0,009	-0,004	0,017	0,069	-0,074	0,019	-0,054	0,005	-0,002	0,024	0,011	-0,008	-0,006	-0,001
tvh4	0,011	0,005	-0,013	-0,033	-0,071	0,015	-0,037	0,004	0,012	0,138	0,025	0,005	0,043	-0,002
tvh9	0,006	0,011	-0,024	-0,084	0,002	-0,005	-0,005	0,000	0,010	0,094	0,013	0,006	0,037	-0,002
ptvh-100	-0,050	-0,001	-0,028	-0,120	-0,020	0,011	0,066	0,001	0,015	-0,010	-0,009	0,012	-0,010	0,007
ptvh-50	0,018	-0,002	0,022	0,076	0,085	-0,024	-0,007	-0,005	-0,017	-0,080	-0,011	-0,041	-0,088	-0,004
ptvh-Ma50	-0,031	-0,013	0,010	0,006	0,010	-0,001	0,002	0,004	-0,028	-0,177	-0,039	0,115	0,223	0,002
ptvh-Me50	0,037	0,004	0,010	0,055	0,042	-0,012	0,004	-0,002	-0,022	-0,155	-0,038	0,119	0,256	0,001
pctvh-100	0,015	-0,008	0,027	0,117	0,021	-0,001	0,017	-0,001	0,000	-0,046	0,003	-0,013	-0,048	0,000
pctvh-50	0,012	0,004	-0,002	-0,004	0,046	-0,015	-0,030	-0,002	-0,017	-0,048	-0,006	-0,014	-0,027	-0,006
pctvh-Ma50	-0,053	-0,014	0,002	-0,038	-0,020	0,000	0,001	0,003	0,011	0,073	0,019	-0,068	-0,149	0,001
pctvh-Me50	0,016	0,013	-0,019	-0,028	-0,008	0,002	-0,007	-0,003	0,033	0,200	0,045	-0,157	-0,324	-0,002
dvt-Dist0	0,012	-0,033	0,097	0,419	-0,365	0,105	-0,184	0,026	0,007	0,119	0,090	0,017	0,060	0,001
it11	0,013	-0,004	-0,054	0,030	-0,176	0,066	0,300	0,013	0,141	-0,032	-0,011	0,228	-0,115	0,149
it12	0,000	-0,001	-0,002	0,015	0,073	-0,031	-0,133	-0,004	-0,060	0,023	0,009	-0,103	0,048	-0,063
it13	-0,041	-0,016	0,138	-0,019	-0,034	0,015	0,093	-0,006	-0,003	-0,029	-0,009	0,026	-0,020	0,011
it14	0,322	0,268	0,025	-0,004	0,062	0,237	-0,017	-0,117	0,007	-0,003	-0,001	0,005	-0,001	-0,003

Continúa en la página siguiente...

	Dim.57	Dim.58	Dim.59	Dim.60	Dim.61	Dim.62	Dim.63	Dim.64	Dim.65	Dim.66	Dim.67	Dim.68	Dim.69	Dim.70
it21	-0,047	-0,018	0,080	-0,001	-0,028	0,026	0,109	-0,018	-0,070	-0,034	0,025	-0,262	0,110	-0,285
it22	-0,006	-0,012	0,062	-0,004	0,026	-0,008	-0,037	0,003	0,045	0,022	-0,004	0,111	-0,038	0,111
it23	0,070	0,044	-0,237	0,090	-0,006	-0,006	-0,034	-0,003	-0,015	-0,035	-0,014	-0,018	-0,018	-0,021
it24	-0,235	-0,096	-0,002	0,008	-0,058	-0,269	0,037	0,472	-0,044	0,001	-0,004	-0,018	0,004	0,009
it31	-0,038	0,004	0,024	0,001	0,133	-0,048	-0,223	-0,015	-0,168	0,040	-0,004	-0,063	0,026	0,238
it32	0,029	0,011	-0,022	0,029	-0,056	0,014	0,106	-0,002	-0,015	0,001	-0,004	0,060	-0,024	-0,103
it33	-0,032	-0,003	-0,071	0,028	0,043	-0,008	-0,098	0,032	0,256	-0,050	0,012	-0,107	0,045	0,022
it34	0,337	-0,590	-0,103	-0,046	0,060	0,196	-0,021	0,007	-0,011	0,003	0,002	0,004	-0,005	-0,001
it41	-0,012	0,016	-0,038	-0,001	0,118	-0,056	-0,247	0,027	0,177	0,011	-0,001	0,195	-0,068	0,023
it42	-0,005	0,016	-0,162	0,056	-0,035	0,012	0,068	0,001	0,085	-0,027	0,006	-0,119	0,048	-0,011
it43	0,017	-0,054	0,376	-0,070	0,000	0,002	0,025	-0,030	-0,335	0,057	-0,013	0,136	-0,064	-0,013
it44	0,185	0,438	0,073	0,013	0,063	0,157	0,018	0,255	-0,015	0,003	-0,004	-0,011	0,000	0,005
it51	0,050	-0,014	0,078	-0,020	-0,042	-0,009	0,044	0,027	0,088	0,002	-0,001	0,135	-0,062	-0,297
it52	0,012	-0,025	0,145	-0,011	0,017	-0,005	-0,027	-0,002	0,025	-0,004	0,003	-0,050	0,021	0,091
it53	-0,049	0,070	-0,465	0,100	0,007	-0,006	0,016	-0,007	-0,177	0,028	-0,006	0,081	-0,032	-0,054
it54	-0,608	0,000	0,008	0,013	0,078	0,302	-0,039	-0,227	0,030	0,002	-0,001	0,022	-0,005	-0,018
it61	0,024	0,004	-0,088	0,031	-0,110	0,063	0,224	-0,032	-0,147	-0,021	-0,002	-0,225	0,092	0,221
it62	-0,016	0,000	-0,026	0,038	0,064	-0,015	-0,083	0,004	-0,080	0,019	-0,003	0,091	-0,029	-0,041
it63	0,030	-0,041	0,230	-0,076	-0,008	0,010	-0,024	0,023	0,327	-0,062	0,000	-0,141	0,059	0,057
it64	0,104	0,086	0,013	0,007	-0,162	-0,472	0,009	-0,375	0,034	-0,008	0,004	-0,001	0,004	0,004
rti-Dist0	-0,006	-0,005	0,018	0,023	-0,022	0,010	0,044	-0,003	0,003	-0,004	0,001	0,011	-0,014	0,014
renfam-1stQ	0,002	-0,007	0,020	0,098	-0,042	0,012	-0,017	0,005	0,006	0,034	0,014	0,005	0,012	0,000
renfam-masMedia	0,003	0,017	-0,035	-0,148	0,042	-0,014	0,025	-0,004	-0,003	-0,020	-0,015	-0,004	-0,009	0,000
renfam-Me0	-0,002	0,001	0,004	-0,010	-0,001	-0,001	-0,014	0,001	0,000	0,006	0,003	0,007	-0,002	-0,007
renfam-MeMedia	-0,004	-0,011	0,019	0,077	-0,006	0,004	-0,007	-0,001	0,000	-0,002	0,004	0,000	-0,003	0,001
ninmuebles1	-0,003	-0,002	0,025	-0,007	-0,030	0,012	0,042	-0,004	-0,010	0,006	0,002	0,003	-0,007	0,015
ninmuebles2	-0,009	-0,008	0,038	0,001	0,010	-0,002	-0,006	-0,001	-0,001	-0,004	0,001	0,002	-0,004	0,003
ninmuebles3	0,012	-0,003	-0,022	0,022	0,004	-0,001	-0,006	0,003	0,005	-0,003	-0,002	0,008	-0,002	-0,007
ninmuebles4	0,002	0,019	-0,070	0,023	-0,006	0,001	0,011	0,005	0,030	-0,010	0,000	-0,026	0,012	0,001
ninmuebles6	-0,001	-0,006	0,016	0,015	0,008	-0,004	-0,009	-0,001	-0,007	0,005	0,001	0,013	-0,007	-0,010
numfam2	-0,005	-0,012	0,017	0,090	0,163	-0,046	0,101	-0,012	0,032	0,159	0,008	0,027	0,087	0,001
numfam3	0,009	0,006	-0,029	-0,045	0,030	-0,010	0,045	-0,001	0,031	0,174	0,011	0,041	0,109	-0,003
numfam4	0,001	0,004	-0,018	-0,046	0,033	-0,008	0,022	-0,001	0,022	0,157	0,014	0,027	0,100	0,001
numfam5	0,001	-0,001	-0,006	-0,025	0,022	-0,005	0,023	-0,006	0,011	0,069	0,009	0,014	0,047	-0,001
numfam6	0,001	0,001	0,003	-0,011	0,014	-0,003	0,009	0,000	0,007	0,030	0,003	0,004	0,019	-0,002
numfam7	0,004	0,002	-0,005	-0,007	0,014	-0,004	0,004	-0,001	0,001	0,014	-0,001	0,004	0,009	-0,002
numfam8	0,000	-0,001	0,005	-0,004	0,007	-0,002	0,003	0,000	0,001	0,006	-0,001	0,001	0,005	-0,001
numfam9	-0,002	-0,001	0,004	-0,003	0,001	0,000	0,003	0,000	-0,001	0,004	0,000	-0,001	0,005	0,001
numfam10	-0,001	0,000	0,001	-0,003	0,001	0,000	0,004	0,000	0,002	0,004	0,000	0,002	0,002	0,000
numfam11	-0,001	0,000	0,002	0,002	0,000	0,000	-0,002	0,000	0,000	0,005	0,000	0,001	0,002	0,001
numfam12	-0,001	0,000	0,002	0,000	0,002	-0,001	0,001	0,000	0,000	0,002	0,002	0,000	0,002	0,000
numfam13	0,000	0,000	-0,001	-0,003	0,002	-0,001	0,002	0,000	0,000	0,001	0,002	0,000	0,001	0,000
numfam16	0,000	0,000	-0,003	0,004	-0,003	0,001	0,000	0,001	0,002	0,003	0,000	0,001	0,002	-0,001
edad2	-0,001	0,000	0,003	0,012	0,179	-0,049	0,108	-0,012	0,017	0,038	-0,016	-0,001	0,002	0,002
edad3	-0,009	-0,016	0,029	0,111	0,136	-0,037	0,067	-0,011	0,001	-0,056	-0,017	-0,014	-0,043	0,002
edad4	0,006	-0,014	0,021	0,091	0,022	-0,003	0,001	-0,004	0,004	0,017	0,002	0,000	0,009	-0,002
edad5	0,001	0,002	-0,006	-0,019	-0,062	0,018	-0,034	0,007	0,006	0,092	0,018	0,015	0,062	-0,002
edad6	0,001	0,038	-0,069	-0,289	-0,217	0,053	-0,098	0,018	-0,007	0,056	0,016	0,021	0,061	-0,002
edad7	-0,005	0,001	-0,005	-0,020	0,042	-0,011	0,025	-0,002	0,006	0,032	-0,001	0,002	0,014	0,000
BIECorrector1	-0,005	-0,001	0,004	0,001	0,017	-0,003	0,024	-0,001	0,010	-0,004	-0,002	-0,001	-0,004	0,003
pormedtit-100	0,061	0,007	-0,036	0,046	0,091	-0,044	-0,198	0,005	-0,022	-0,022	0,000	-0,025	-0,002	-0,036
pormedtit-50	-0,054	-0,009	-0,007	0,027	-0,050	0,032	0,169	0,004	0,078	0,004	-0,003	0,046	0,004	0,027
pormedtit-Ma50	0,006	-0,007	0,090	-0,027	-0,053	0,017	0,063	-0,017	-0,074	0,004	0,003	-0,032	-0,004	0,024
pormedtit-Me50	-0,023	0,004	0,007	-0,043	-0,041	0,016	0,041	0,001	-0,004	0,026	0,005	0,014	-0,019	0,003
IP-2dQ	-0,003	0,009	-0,020	-0,084	0,092	-0,024	0,052	-0,004	-0,012	-0,072	0,360	0,024	-0,001	0,008
IP-3thQ	-0,002	-0,003	0,003	0,013	-0,025	0,007	-0,010	0,000	0,010	0,033	-0,230	-0,013	-0,006	-0,003
IP-4thQ	0,000	-0,011	0,034	0,130	-0,088	0,026	-0,048	0,006	0,000	0,028	0,021	0,001	0,016	0,001
RentaCorrectora1	-0,004	-0,013	0,024	0,086	-0,015	0,007	-0,011	-0,003	0,000	0,018	0,008	-0,001	0,010	0,000

Continúa en la página siguiente...

	Dim.57	Dim.58	Dim.59	Dim.60	Dim.61	Dim.62	Dim.63	Dim.64	Dim.65	Dim.66	Dim.67	Dim.68	Dim.69	Dim.70
Year2009	0,001	0,000	0,011	0,058	-0,026	0,010	-0,011	0,005	-0,010	-0,033	0,365	0,027	0,018	0,009
Year2010	-0,003	0,002	-0,006	-0,055	0,047	-0,013	0,026	-0,004	-0,003	-0,017	0,039	0,002	-0,002	0,002
Year2011	0,003	-0,001	0,000	0,017	-0,029	0,007	-0,015	-0,001	0,010	0,044	-0,284	-0,018	-0,004	-0,009
Year2012	0,001	0,008	-0,027	-0,100	0,069	-0,021	0,032	-0,002	-0,005	-0,036	0,100	0,003	-0,010	0,001

Tabla A.22: Correlaciones de las Dimensiones 57 a la 70 con las variables originales discretizadas

	Dim.71	Dim.72	Dim.73	Dim.74	Dim.75	Dim.76	Dim.77	Dim.78	Dim.79	Dim.80	Dim.81	Dim.82	Dim.83	Dim.84
estcv2	-0,001	0,018	-0,002	0,228	-0,092	-0,002	-0,002	-0,004	-0,047	-0,001	0,000	0,000	0,000	0,000
estcv3	-0,001	0,000	-0,001	0,091	0,004	-0,002	0,001	-0,002	-0,024	-0,002	0,000	0,000	0,000	0,000
estcv4	0,000	0,000	0,000	0,061	-0,003	-0,003	0,000	-0,004	-0,020	0,000	0,000	0,000	0,000	0,000
act11	0,000	-0,009	0,001	0,001	0,001	0,186	-0,001	-0,001	-0,001	-0,001	0,000	0,000	0,000	0,000
act12	0,000	-0,022	0,002	0,004	0,000	0,163	-0,002	-0,001	-0,001	-0,001	0,000	0,000	0,000	0,000
act14	0,004	-0,002	0,000	0,001	0,001	0,032	0,000	0,000	-0,001	0,000	0,000	0,000	0,000	0,000
act21	0,001	0,017	-0,001	0,001	-0,002	-0,225	0,001	0,000	0,001	-0,001	0,000	0,000	0,000	0,000
act22	0,001	0,007	-0,001	-0,001	0,000	-0,079	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
act24	-0,001	0,004	0,000	0,000	-0,001	-0,033	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
tvh2	-0,001	0,000	-0,001	-0,001	0,002	0,000	0,000	-0,001	0,001	0,000	0,000	0,000	0,000	0,000
tvh3	0,001	0,006	-0,001	0,020	0,043	0,001	0,002	-0,002	0,006	0,000	0,000	0,018	0,000	0,000
tvh4	0,000	0,005	-0,002	0,040	0,061	-0,002	0,003	-0,003	0,011	-0,001	0,000	0,030	0,000	0,000
tvh9	0,001	-0,004	-0,002	0,022	0,030	0,003	0,001	0,001	0,008	0,000	0,000	0,015	0,000	0,000
ptvh-100	-0,001	0,000	0,002	0,026	0,068	-0,001	0,002	0,001	-0,008	0,000	0,000	0,033	0,000	0,000
ptvh-50	-0,001	-0,003	0,003	-0,071	-0,151	0,001	-0,006	0,001	-0,007	0,000	0,000	0,040	0,000	0,000
ptvh-Ma50	0,004	-0,003	-0,002	0,006	0,006	0,000	0,000	0,000	-0,001	0,000	0,000	0,005	0,000	0,000
ptvh-Me50	0,005	0,001	-0,006	0,015	0,010	0,000	0,001	-0,001	-0,002	0,000	0,000	0,008	0,000	0,000
pctvh-100	0,001	-0,002	0,001	-0,002	0,044	0,000	0,001	-0,001	0,002	0,000	0,000	0,012	0,000	0,000
pctvh-50	0,002	-0,002	-0,006	0,084	0,294	-0,003	0,009	-0,002	0,009	-0,001	0,000	0,001	0,000	0,000
pctvh-Ma50	-0,003	-0,002	0,000	-0,006	0,006	0,000	0,000	0,000	0,001	0,000	0,000	0,000	0,000	0,000
pctvh-Me50	-0,006	0,000	0,001	-0,010	0,014	0,000	0,000	0,000	0,002	0,000	0,000	0,001	0,000	0,000
dvt-Dist0	-0,001	0,030	0,002	0,002	0,003	0,002	0,013	-0,003	0,005	0,000	-0,002	0,000	0,000	0,000
it11	0,029	0,004	0,006	0,022	-0,006	-0,002	0,007	0,041	-0,001	-0,001	0,000	0,000	0,000	0,001
it12	-0,013	-0,009	-0,005	-0,012	0,004	0,001	0,017	0,076	-0,003	0,000	0,000	0,000	0,000	0,002
it13	-0,004	0,002	-0,004	0,011	-0,001	0,000	0,006	0,027	-0,001	0,000	0,000	0,000	0,000	0,001
it14	0,000	0,000	0,000	0,001	-0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
it21	-0,127	0,001	-0,005	0,003	0,000	0,000	0,003	0,014	0,000	0,000	0,000	0,000	-0,002	-0,002
it22	0,052	-0,001	-0,015	-0,008	0,001	0,000	0,008	0,038	-0,001	-0,001	0,000	0,000	-0,004	-0,004
it23	0,002	0,003	-0,009	0,022	-0,007	0,001	0,005	0,016	-0,001	0,001	0,000	0,000	-0,002	-0,002
it24	-0,005	0,002	-0,001	0,000	0,000	-0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
it31	0,251	-0,003	-0,006	0,000	0,000	0,000	0,001	0,007	0,000	0,000	0,000	0,000	0,001	-0,001
it32	-0,092	0,005	-0,034	-0,003	-0,001	0,000	0,004	0,018	0,000	-0,001	0,000	0,000	0,004	-0,003
it33	0,029	0,000	-0,013	0,007	-0,002	0,000	0,001	0,006	0,000	0,001	0,000	0,000	0,001	-0,001
it34	-0,004	0,000	-0,002	0,000	0,000	-0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
it41	-0,322	0,006	-0,031	0,000	0,000	0,001	0,001	0,002	0,000	0,000	0,000	0,000	0,001	0,000
it42	0,133	0,004	-0,054	-0,004	-0,001	0,000	0,002	0,010	0,000	0,000	0,000	0,000	0,004	0,000
it43	-0,036	0,001	-0,028	0,007	-0,003	0,000	0,001	0,004	0,000	0,000	0,000	0,000	0,001	0,000
it44	0,000	0,002	-0,001	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
it51	0,271	-0,001	0,004	0,002	-0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
it52	-0,062	0,011	-0,139	-0,005	-0,001	0,000	0,001	0,006	0,000	0,000	0,000	0,000	-0,002	0,001
it53	-0,003	0,002	-0,022	0,003	-0,002	0,001	0,000	0,001	0,000	0,000	0,000	0,000	-0,001	0,000
it54	0,013	-0,001	-0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
it61	-0,106	-0,004	0,066	0,001	0,002	-0,001	0,001	0,002	0,000	0,000	0,000	0,000	0,000	0,000
it62	-0,022	-0,017	0,355	0,005	0,002	-0,003	0,002	0,003	0,000	-0,001	0,000	0,000	0,000	0,000
it63	0,025	-0,007	0,116	0,005	0,000	-0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
it64	-0,002	-0,001	0,007	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
rti-Dist0	-0,002	0,001	0,004	-0,006	0,002	-0,002	-0,056	-0,261	0,001	0,010	0,000	0,000	0,000	0,000

Continúa en la página siguiente...

	Dim.71	Dim.72	Dim.73	Dim.74	Dim.75	Dim.76	Dim.77	Dim.78	Dim.79	Dim.80	Dim.81	Dim.82	Dim.83	Dim.84
renfam-1stQ	-0,003	-0,207	-0,010	0,012	-0,002	-0,007	-0,001	0,008	0,000	0,086	0,000	0,000	0,000	0,000
renfam-masMedia	0,003	0,171	0,008	-0,010	0,005	0,013	0,002	0,008	-0,008	0,118	0,000	0,000	0,000	0,000
renfam-Me0	-0,003	-0,017	0,000	0,000	0,000	0,000	0,000	0,001	0,000	0,006	0,000	0,000	0,000	0,000
renfam-MeMedia	-0,002	-0,086	-0,004	0,011	-0,001	-0,009	0,002	0,010	0,000	0,104	0,000	0,000	0,000	0,000
ninmuebles1	-0,006	-0,009	0,017	0,000	0,004	0,000	0,017	0,076	-0,002	0,000	0,000	0,000	0,000	-0,002
ninmuebles2	-0,001	-0,004	0,019	0,003	-0,001	0,000	0,009	0,041	-0,001	0,000	0,000	0,000	0,003	0,001
ninmuebles3	-0,011	-0,001	0,033	0,001	0,000	0,000	0,004	0,019	0,000	0,000	0,000	0,000	0,000	0,002
ninmuebles4	0,006	-0,005	0,076	0,002	0,000	0,000	0,002	0,010	0,000	0,000	0,000	0,000	-0,003	0,002
ninmuebles6	0,013	0,010	-0,135	-0,003	-0,002	0,001	0,001	0,006	0,000	0,000	0,000	0,000	-0,002	0,001
numfam2	0,002	-0,006	0,003	-0,160	0,031	0,000	0,001	0,002	0,003	0,001	0,000	0,000	0,000	0,000
numfam3	0,001	-0,012	0,003	-0,184	0,032	-0,002	0,002	0,001	0,006	0,000	0,000	0,000	0,000	0,000
numfam4	0,000	-0,007	0,002	-0,159	0,025	-0,004	0,002	0,001	0,005	0,000	0,000	0,000	0,000	0,000
numfam5	0,002	-0,005	0,001	-0,079	0,014	-0,001	0,001	0,000	0,002	0,000	0,000	0,000	0,000	0,000
numfam6	-0,001	-0,001	0,000	-0,036	0,007	-0,001	0,000	0,000	0,001	0,000	0,000	0,000	0,000	0,000
numfam7	0,001	-0,002	0,000	-0,019	0,004	0,000	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
numfam8	-0,001	-0,001	0,001	-0,009	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
numfam9	-0,001	0,000	0,000	-0,005	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
numfam10	0,000	0,000	0,001	-0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
numfam11	0,000	0,000	0,000	-0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
numfam12	0,000	-0,001	0,000	-0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
numfam13	0,000	0,000	0,000	-0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
numfam16	0,000	0,000	0,000	-0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
edad2	0,000	0,012	-0,001	0,073	-0,015	0,000	-0,005	0,002	0,097	0,001	0,000	0,000	0,000	0,000
edad3	0,001	0,011	-0,001	0,089	-0,025	0,000	-0,006	0,002	0,117	0,001	-0,001	0,000	0,000	0,000
edad4	0,000	0,008	0,000	0,032	-0,021	-0,001	-0,003	0,004	0,123	0,001	-0,001	0,000	0,000	0,000
edad5	-0,001	0,001	0,000	-0,024	-0,011	0,001	-0,002	0,003	0,109	0,002	-0,001	0,000	0,000	0,000
edad6	0,001	-0,015	0,001	-0,036	-0,007	0,006	-0,003	0,002	0,103	0,002	-0,001	0,000	0,000	0,000
edad7	0,000	-0,001	0,000	0,009	-0,005	-0,001	-0,001	0,000	0,010	0,000	0,000	0,000	0,000	0,000
BIECorrector1	0,001	0,013	0,000	-0,003	0,000	0,000	-0,001	0,000	-0,001	0,001	0,000	0,000	0,000	0,000
pormedtit-100	0,015	0,001	-0,029	0,012	-0,008	-0,001	0,008	0,040	-0,004	0,000	0,000	0,000	0,000	0,001
pormedtit-50	0,005	-0,003	-0,017	-0,012	0,007	0,000	0,011	0,049	0,000	-0,001	0,000	0,000	0,000	0,001
pormedtit-Ma50	-0,026	-0,003	0,025	0,001	0,005	0,000	0,009	0,036	-0,001	-0,001	0,000	0,000	0,000	0,000
pormedtit-Me50	-0,007	-0,006	0,041	0,001	0,001	0,001	0,008	0,034	0,000	0,000	0,000	0,000	0,000	0,000
IP-2dQ	0,004	-0,010	-0,001	0,001	-0,009	0,001	0,144	-0,031	0,003	0,000	0,028	0,000	0,000	0,000
IP-3thQ	-0,001	-0,006	-0,001	-0,002	-0,008	0,001	0,157	-0,034	0,005	0,000	0,054	0,000	0,000	0,000
IP-4thQ	0,000	0,011	0,002	-0,004	0,001	0,000	0,038	-0,008	0,004	0,000	0,083	0,000	0,000	0,000
RentaCorrectora1	0,004	0,369	0,021	-0,018	0,004	0,014	0,010	0,006	0,000	0,018	-0,001	0,000	0,000	0,000
Year2009	0,004	0,011	0,003	0,000	0,004	-0,001	-0,117	0,025	-0,001	0,000	0,024	0,000	0,000	0,000
Year2010	-0,001	0,003	0,001	-0,002	0,008	-0,002	-0,146	0,032	-0,003	0,000	0,042	0,000	0,000	0,000
Year2011	-0,002	0,003	0,000	-0,003	0,006	0,000	-0,107	0,023	-0,002	0,000	0,063	0,000	0,000	0,000
Year2012	0,001	-0,006	-0,002	0,007	-0,002	0,000	0,001	0,000	0,000	0,000	0,083	0,000	0,000	0,000

Tabla A.23: Correlaciones de las Dimensiones 71 a la 84 con las variables originales discretizadas

	Dim.85	Dim.86	Dim.87	Dim.88
estcv2	0,0000	0,0000	0,0000	0,0000
estcv3	0,0000	0,0000	0,0000	0,0000
estcv4	0,0000	0,0000	0,0000	0,0000
act11	0,0000	0,0000	0,0000	0,0000
act12	0,0000	0,0000	0,0000	0,0000
act14	0,0000	0,0000	0,0000	0,0000
act21	0,0000	0,0000	0,0000	0,0000
act22	0,0000	0,0000	0,0000	0,0000
act24	0,0000	0,0000	0,0000	0,0000
tvh2	0,0000	0,0000	0,0000	0,0000

Continúa en la página siguiente...

	Dim.85	Dim.86	Dim.87	Dim.88
tvh3	0,0000	0,0000	0,0000	0,0000
tvh4	0,0000	0,0000	0,0000	0,0000
tvh9	0,0000	0,0000	0,0000	0,0000
ptvh-100	0,0000	0,0000	0,0000	0,0000
ptvh-50	0,0000	0,0000	0,0000	0,0000
ptvh-Ma50	0,0000	0,0000	0,0000	0,0000
ptvh-Me50	0,0000	0,0000	0,0000	0,0000
pctvh-100	0,0000	0,0000	0,0000	0,0000
pctvh-50	0,0000	0,0000	0,0000	0,0000
pctvh-Ma50	0,0000	0,0000	0,0000	0,0000
pctvh-Me50	0,0000	0,0000	0,0000	0,0000
dvt-Dist0	0,0000	0,0000	0,0000	0,0000
it11	-0,0002	-0,0012	0,0000	0,0000
it12	-0,0004	-0,0027	0,0000	0,0000
it13	-0,0001	-0,0008	0,0000	0,0000
it14	0,0000	0,0000	0,0000	0,0000
it21	0,0004	-0,0001	0,0000	0,0000
it22	0,0011	-0,0003	0,0000	0,0000
it23	0,0005	-0,0001	0,0000	0,0000
it24	0,0000	0,0000	0,0000	0,0000
it31	-0,0011	0,0000	0,0000	0,0000
it32	-0,0030	-0,0001	0,0000	0,0000
it33	-0,0010	0,0000	0,0000	0,0000
it34	0,0000	0,0000	0,0000	0,0000
it41	0,0012	0,0000	0,0000	0,0000
it42	0,0037	-0,0001	0,0000	0,0000
it43	0,0014	0,0000	0,0000	0,0000
it44	0,0001	0,0000	0,0000	0,0000
it51	-0,0002	0,0001	0,0000	0,0000
it52	-0,0007	0,0005	0,0000	0,0000
it53	-0,0002	0,0002	0,0000	0,0000
it54	0,0000	0,0000	0,0000	0,0000
it61	0,0000	0,0000	0,0000	0,0000
it62	0,0000	0,0000	0,0000	0,0000
it63	0,0000	0,0000	0,0000	0,0000
it64	0,0000	0,0000	0,0000	0,0000
rti-Dist0	0,0000	0,0000	0,0000	0,0000
renfam-1stQ	0,0000	0,0000	0,0000	0,0000
renfam-masMedia	0,0000	0,0000	0,0000	0,0000
renfam-Me0	0,0000	0,0000	0,0000	0,0000
renfam-MeMedia	0,0000	0,0000	0,0000	0,0000
ninmuebles1	0,0004	0,0010	0,0000	0,0000
ninmuebles2	-0,0005	0,0010	0,0000	0,0000
ninmuebles3	0,0016	0,0008	0,0000	0,0000
ninmuebles4	-0,0011	0,0007	0,0000	0,0000
ninmuebles6	-0,0008	0,0005	0,0000	0,0000
numfam2	0,0000	0,0000	0,0000	0,0000
numfam3	0,0000	0,0000	0,0000	0,0000
numfam4	0,0000	0,0000	0,0000	0,0000
numfam5	0,0000	0,0000	0,0000	0,0000
numfam6	0,0000	0,0000	0,0000	0,0000
numfam7	0,0000	0,0000	0,0000	0,0000
numfam8	0,0000	0,0000	0,0000	0,0000
numfam9	0,0000	0,0000	0,0000	0,0000
numfam10	0,0000	0,0000	0,0000	0,0000
numfam11	0,0000	0,0000	0,0000	0,0000
numfam12	0,0000	0,0000	0,0000	0,0000
numfam13	0,0000	0,0000	0,0000	0,0000

Continúa en la página siguiente...

	Dim.85	Dim.86	Dim.87	Dim.88
numfam16	0,0000	0,0000	0,0000	0,0000
edad2	0,0000	0,0000	0,0000	0,0000
edad3	0,0000	0,0000	0,0000	0,0000
edad4	0,0000	0,0000	0,0000	0,0000
edad5	0,0000	0,0000	0,0000	0,0000
edad6	0,0000	0,0000	0,0000	0,0000
edad7	0,0000	0,0000	0,0000	0,0000
BIECorrector1	0,0000	0,0000	0,0000	0,0000
pormedtit-100	-0,0001	0,0011	0,0000	0,0000
pormedtit-50	-0,0001	0,0011	0,0000	0,0000
pormedtit-Ma50	0,0000	0,0007	0,0000	0,0000
pormedtit-Me50	0,0000	0,0007	0,0000	0,0000
IP-2dQ	0,0000	0,0000	0,0000	0,0000
IP-3thQ	0,0000	0,0000	0,0000	0,0000
IP-4thQ	0,0000	0,0000	0,0000	0,0000
RentaCorrectora1	0,0000	0,0000	0,0000	0,0000
Year2009	0,0000	0,0000	0,0000	0,0000
Year2010	0,0000	0,0000	0,0000	0,0000
Year2011	0,0000	0,0000	0,0000	0,0000
Year2012	0,0000	0,0000	0,0000	0,0000

Tabla A.24: Correlaciones de las Dimensiones 85 a la 88 con las variables originales discretizadas

Bibliografía

Abdou, H.and Pointon, J. and El Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in egyptian banking. *Expert Systems with Applications*, 35(3):1275–1292.

Abell, D. (1980). *Defining the Business*. Prentice–Hall.

Abrahams, C. and Zhang, M. (2008). *Fair Lending Compliance: Intelligence and Implications for Credit Risk Management*. John Wiley and Sons, Inc., New Jersey.

Ahn, B. S., Cho, S. S., and Kim, C. Y. (2000). The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert Systems with Applications*, 18(2):65–74.

Almiñana, M., Escudero, L. F., Pérez-Martín, A., Rabasa, A., and Santamaría, L. (2014). A classification rule reduction algorithm based on significance domains. *TOP*, 22(1):397–418.

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4):589–609.

Altman, E., Marco, G., and Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks. *Journal of Banking and Finance*, 18(3):505–529.

Apilado, V. P., Warner, D. C., and Dauten, J. J. (1974). Evaluative techniques in consumer finance-experimental results and policy implications for financial institutions. *Journal of financial and Quantitative Analysis*, 9(02):275–283.

Artís, M., Guillén, M., and Martínez, J. M. (1994). A model for credit scoring: an application of discriminant anlysis. *QÜESTIIÓ*, 18(3):385–395.

Back, B., Laitinen, T., Sere, K., and Van Wezel, M. (1996). Choosing bankruptcy predictors using discriminant analysis, logit analysis, and genetic algorithms. *Turku Centre for Computer Science, Finlandia, Technical Report*, 40.

Baesens, B., Setiono, R., Mues, C., and Vanthienen, J. (2003a). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science*, 49(3):312–329.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003b). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635.

Bailey, M. (2004). *Consumer credit quality: underwriting, scoring, fraud prevention and collections*. White Box Publishing, Kingswood, Bristol.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 160(901):262–282.

Bates, D., Mächler, B., Bolker, B., and Walker, S. (2015a). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1).

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., and Grothendieck, G. (2015b). *Linear Mixed-Effects Models using 'Eigen' and S4*. 1.1-10.

Bessis, J. (2002). *Risk Management in Banking*. John Wiley and Sons, Inc., Inglaterra.

Beynon, M. J.and Peel, M. J. (2001). Variable precision rough set theory and data discretisation an application to corporate failure prediction. *OMEGA*, 29:561–576.

Blochlinger, A. and Leippold, M. (2006). Economic benefit of powerful credit scoring. *Journal of Banking and Finance*, 30(3):851–873.

Blum, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research*, 12(1):1–25.

BOE (06/05/2016). *Circular 4/2016, del Banco de de España*.

BOE (29/12/2007). *Real Decreto 1763/2007*.

BOE (30/12/2004). *Circular 4/2004, del Banco de de España*.

BOE (30/12/2008). *Real Decreto 2128/2008*.

BOE (31/05/2013). *Circular 1/2013, del Banco de de España*.

BOE (31/12/2009). *Real Decreto 2030/2009*.

BOE (31/12/2010). *Real Decreto 1795/2010*.

BOE (31/12/2011). *Real Decreto 1888/2011*.

BOE (31/12/2012). *Real Decreto 1717/2012*.

Boj, E., Claramunt, M., and Esteve, A.and Fortiana, J. (2009a). Criterios de selección de modelo en credit scoring, aplicación del análisis discriminante basado en distancias. *En Anales del Instituto de Actuarios Españoles*, 15:209–230.

Boj, E., Claramunt, M. M., Esteve, A., and Fortiana, J. (2009b). Credit scoring basado en distancias: coeficientes de influencia de los predictores. In Estudios, F. M., editor, *Investigaciones en Seguros y Gestión de riesgos: RIESGO 2009*, pages 15–22. Cuadernos de la Fundación MAPFRE, Madrid.

Bonilla, M., Olmeda, I., and Puertas, R. (2003). Modelos paramétricos y no paramétricos en problemas de credit scoring. *Revista Española de Financiación y Contabilidad*, 32(118):833–869.

Boyle, M., Crook, J., Hamilton, R., and Thomas, L. (1992). *Methods for credit scoring applied to slow payers in Credit scoring and Credit Control*. Oxford University Press, Oxford.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Chapman and Hall, Belmont: Wadsworth.

Cabrera Cruz, A. (2014). *Diseño de credit scoring para evaluar el riesgo crediticio en una entidad de ahorro y crédito popular*. PhD thesis, Universidad tecnológica de la Mixteca, Huajuapan de león, Oxaca, México.

Cao, L. (2002). Support vector machines experts for time series forecasting. *Neurocomputing*, 51:321–339.

Carter, C. and Catlett, J. (1987). Assessing credit card applications using machine learning. *IEEE Expert: intelligent systems and their applications*, 2(3):71–79.

Chen, M. C. and Huang, S. H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications*, 24(4):433–441.

Coats, P. K. and Fant, L. F. (1992). A neural network approach to forecasting financial distress. *The Journal of Business Forecasting*, Winter:9–12.

Coats, P. K. and Fant, L. F. (1993). Recognizing financial distress patterns using a neural network tool. *Financial Management*, 22(3):142–155.

Coffman, J. (1986). The proper role of tree analysis in forecasting the risk behaviour of borrowers. management decision systems. *MDS Reports*, pages 3–4–7–9.

Crook, J., Edelman, D., and Thomas, L. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3):1447–1465.

Davis, R. H., Edelman, D. B., and Gammerman, A. J. (1992). Machine-learning algorithms for credit-card applications. *Journal of Management Mathematics*, 4(1):43–51.

de Moivre, A. (1756). *The Doctrine of Chances: Or, A Method of Calculating the Probability of Events in Play*. W. Pearson.

Deakin, E. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, 10(1):161–179.

Death, G. and Fabricius, K. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology,Brooklin*, 81(11):3178–3192.

Demidenko, E. (2004). *Mixed models, Theory and Applications.* John Wiley, New York.

Desai, V., Crook, J., and Overstreet, G. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, 95(1):24–37.

Desai, V. S., Convay, D. G., Crook, J. N., and Overstreet, G. (1997). Credit scoring models in the credit union environment using neural networks and genetic algorithms. *IMA Journal of Mathematics Applied in Business and Industry*, 8(4):323–346.

Dietterich, T. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

DOUE (29/11/2016). *Reglamento (UE) 2016/2067 de la comisión.*

Dunteman, G. H. (1989). *Principal Components Analysis (Quantitative Applications in the Social Sciences) issue 69.* Quantitative Applications in the Social Sciences. Sage Publications, Inc.

Durand, D. (1941). *Risk Elements in Consumer Instalment Financing.* National Bureau of Economic Research, Massachusetts.

Eisenbeis, R. (1977). Pitfalls in the application of discriminant analysis in business, finance and economics. *The Journal of Finance*, 32(3):875–900.

Eisenbeis, R. (1978). Problems in applying discriminant analysis in credit scoring models. *The Journal of Banking and Finance*, 2(3):205–219.

Fan, A. and Palaniswami, M. (2000). Selecting bankruptcy predictors using a support vector machine approach. 6:354–359.

Fawcett, T. and Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, (1):291–316.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Foody, G. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80:185–201.

Friedman, J. (1977). A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions on Computers*, 26(4):404–408.

Frydman, H., Kallberg, J. G., and Kao, D. (1985). Testing the adequacy of markov chains and mover-stayer models as representations of credit behaviour. *Operations Research*, 33:1203–1214.

Galindo, J. and Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics*, 15(1):107–143.

Gomes Goncalves, E. (2009). *Estudio comparativo de técnicas de calificación creditícia*. PhD thesis, Universidad Simón Bolívar, Decanato de Estudios Postgrado Maestría en estadística.

González-Abril, L. (2003b). Modelos de clasificación basados en máquinas de vectores soporte. *Departamento de Economía Aplicada I, Universidad de Sevilla*, pages 1–19.

Grablowsky, B. and Talley, W. (1981). Probit and discriminant functions for classifying credit applicants: A comparison. *Journal of Economic Business*, 33(3):254–261.

Gunn, S. R. (1997). Support vectors machines for classification and regression. Technical report, University of Southampton.

Hall, M., Frank, E., Holmes, G., Pfahringer, P., Reutemann, P., and Witten, I. (2009). *The WEKA Data Mining Software: An Update.* SIGKDD Explorations, Volume 11, Issue 1.

Hammer, M. (1983). Failure prediction: sensitivity of classification accuracy to alternative statistical methods and variable sets. *Journal of Accounting and Public Policy*, 2(4):289–307.

Hand, D. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Royal Statistical Society*, 160(3):523–541.

Hand, D. J. (1981). *Discrimination and Classification.* John Wiley and Sons, Inc., Chichester, U.K.

Härdle, W., Moro, R., and Schäfer, D. (2004). Rating companies with support vector machines. Discussion paper, DIW Berlin, German Institute for Economic Research.

Härdle, W., Moro, R., and Schäfer, D. (2005). Predicting bankruptcy with support vector machines. Discussion paper, Deutsche Forschungsgemeinschaft.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning Data Mining, Inference, and Prediction.* Springer, Standford.

Henley, W. (1995). *Statistical aspects of credit scoring.* Open University Press.

Henley, W. E. and Hand, D. (1996). A k-nearest neigbour classifier for assessing consumer credit risk. *Statistician*, 45:77–95.

Hoffmann, F., Baesens, B., Mues, C., Gestel, T. V., and Vanthienen, J. (2007). Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. *European Journal of Operational Research*, 177(1):540–555.

Hsieh, N. (2005). Hybrid mining approach in the design of credit scoring models. *Expert Systems with Applications*, 28(4):655–665.

Huang, C., Chang, E., and Wu, H. (2009). A case study of applying data mining techniques in an outfitters customer value analysis. *Expert Systems with Applications*, 36(3):5909–5915.

Huang, Ch.and Chen, M. and Wang, C. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33(4):847–856.

Huang, W., Nakamori, Y., and Wang, S. (2005). Forecasting stock market movement direction with support vector machine. *Computers and Operations Research*, 32(10):2513–2522.

Huang, Z., Chen, H., Hsu, Ch.and Chen, W., and Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support System*, 37(4):543–558.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and their Applications*. Springer-Verlag, New York.

Juan Camilo Ochoa P., J., Wilinton Galeano M., W., and Agudelo V., L. (2010). Construcción de un modelo de scoring para el otorgamiento de crédito en una entidad financiera. *Perfil de conyuntura económica*, (16):191–222.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119–127.

Khashman, A. (2009). A neural network model for credit risk evaluation. *International Journal of Neural Systems*, 19(04):285–294.

Khattree, R. and Naik, D. (2000). *Multivariate Data Reduction and Discrimination with SAS Software*. Cary, NC: SAS Institute Inc., Berlin and New York SIGLO XII.

Koh, H. (1992). The sensitivity of optimal cutoff points to misclassification costs of type i and type ii errors in the going-concern prediction context. *Journal of Business Finance and Accounting*, 19(02):187–197.

Kolesar, P. and Showers, J. L. (1985). A robust credit screening model using categorical data. *Management Science*, 31(2):123–133.

Lachenbruch, P. A. (1975). *Discriminant analysis*. Computational Mechanics Publications, New York.

Lacher, R. C., Coats, P. K., Sharma, S. C., and Fant, L. F. (1995). A neural network for classifying the financial health of a firm. *European Journal of Operational Research*, 85(1):53–65.

Laitinen, T. and Kankaanpää, M. (1999). Comparative analysis of failure prediction methods: the finnish case. *European Accounting Review*, 8(1):67–92.

Lane, W. (1972). Submarginal credit risk classification. *Journal of Financial and Quantitive Analysis*, 7(2):1379–1385.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.

Lean Yu, L., Wang, S., Lai, K., and Zhou, L. (2008). *Bio-Inspired Credit Risk Analysis computational intelligence with support vector machines*. Springer-Verlag, Berlín.

Lee, T. and Chen, I. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, 28(4):743–752.

Lee, T., Chiu, C., Lu, C., and Chen, I. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, 23(3):245–254.

Leonard, K. J. (1993). Detecting credit card fraud using expert systems. *Computers and Industrial Engineering*, 25(1–4):103–106.

Li, H. and Sun, J. (2008). Ranking-order case-based reasoning for financial distress prediction. *Knowledge-Based Systems*, 21(8):868–878.

Li, H. and Sun, J. (2009a). Forecasting business failure in china using case-based reasoning with hybrid case representation. *Journal of forecasting*, 29(5):486–501.

Li, H. and Sun, J. (2009b). Gaussian case-based reasoning for business failure prediction with empirical data in china. *Information Sciences*, 179(1–2):89–108.

Li, H. and Sun, J. (2009c). Predicting business failure using multiple case-based reasoning combined with support vector machine. *Expert Systems with Applications*, 36(6):10085–10096.

Li, H. and Sun, J. (2010). Business failure prediction using hybrid 2 case-based reasoning. *Computers and Operations Research*, 37(1):137–151.

Li, H., Sun, J., and Sun, J. (2009). Financial distress prediction based on or-cbr in the principle of k-nearest neighbors. *Expert Systems with Applications*, 36(1):643–659.

Lichman, M. (2013). UCI machine learning repository.

Lincoln, M. (1982). *An empirical study of the usefulness of accounting ratios to describe levels of insolvency risk*. PhD thesis, University of Melbourne.

Liu, C.and Frazier, P. and Kumar, L. (2007). Comparative assessment of the measures of thematic classification accuracy. *Remote Sens. Environ*, 107(4):606–616.

López, J. and Martín, L. (1996). *La dirección estratégica de la empresa: teoría y aplicaciones*. Civitas.

Loterman, G., Brown, I., Martens, D., and Mues, C. Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, (28):161–170.

Lucas, A. (1992). *In Credit Scoring and Credit Control*, chapter Updating scorecards: removing the mystique, pages 180–197. Oxford University Press, Oxford.

Lucas, A. (2001). Statistical challenges in credit card issuing. *Applied Stochastic Models in Business and Industry*, 17(1):83–92.

Mahalanobis, P. C. (1936). *On Generalized Distance in Statistics*. Proceedings of the National Institute of Sciences, India.

Makowski, P. (1985). Credit scoring branches out: Decision tree - recent technology. *Credit World*, 75:30–37.

Malhotra, R. and Malhotra, D. (2002). Differentiating between good credits and bad credits using neuron-fuzzy systems. *European Journal of Operational Research*, 136:190–211.

Malhotra, R. and Malhotra, D. (2003). Evaluating consumer loans using neural networks. *Omega*, 31(2):83–96.

Marqués, A., García, V., and Sánchez, J. (2013). A literature review on the application of evolutionary computing to credit scoring. *Journal of the Operational Research Society*, 64(9):1384–1399.

Martín, D. (1977). Early warning of bank failure. *Journal of Banking and Finance*, 1(3):249–276.

Martinez-Murcia, F., Górriz, J., Ramírez, J., Illán, I., and Ortiz, A. (2014). Automatic detection of parkinsonism using significance measures and component analysis in datscan imaging. *Neurocomputing*, (126):58–70.

Masters, T. (1995). *Advanced Algorithms for Neural Networks: AC++ Sourcebook*. John Wiley and Sons, Inc., New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall/CRC.

Mehta, D. (1968). The formulation of credit policy models. *Managment Science*, 15:30–50.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2014). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-4.

Morales Gonzalez, D., Pérez-Martín, A., and Vaca, M. (2013). Monte carlo simulation study under regression models to estimate credit banking risk in home equity loan. *Data Management and Security Applications in Medicine, Science and Engineering*, 45:141–152.

Moreno Valencia, S. (2013). *El Modelo Logit Mixto para la construcción de un Scoring de Crédito*. PhD thesis, Universidad Nacional de Colombia, Universidad Nacional de Colombia. Facultad de Ciencias, Escuela de Estadistica Medellín. Colombia.

Moses, D. and Liao, S. (1987). On developing models for failure prediction. *Journal of Commercial Bank Lending*, 69(7):27–38.

Mures, M., García, A., and Vallejo, M. (2005). Aplicación del análisis discriminante y regresión logística en el estudio de la morosidad de las entidades financieras. comparación de resultados. *Pecnvia*, (1):175–199.

Myers, J. and Forgy, E. (1963). Journal of american statistical association. *The development of numerical credit evaluation systems*, 58:799–806.

Mylonakis, J. (2010). Evaluating the likelihood of using linear discriminant analysis as a commercial bank card owners credit scoring model. *International Business Research*, 3(2).

Njuho, P. and Milliken, G. (2005). Analysis of linear models with one factor having both fixed and random effects. *Communications in Statistics - Theory and Methods*, 34(9):1979–1989.

Ong, C. S., Huang, J. J., and Tzeng, G. H. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications*, 29(1):41–47.

Orgler, Y. (1970). A credit scoring model for commercial loans. *Journal of Money, Credit and Banking II*, 2(4):435–445.

Orgler, Y. (1971). Evaluation of bank consumer loans with credit scoring models. *Journal of Bank Research*, pages 31–37.

Orgler, Y. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109–131.

Overstreet, J. G. and Bradley, J. E. (1994). Applicability of generic linear scoring models in the usa credit union environment: Further analysis. Working paper, University of Virginia.

Overstreet, J. G., Bradley, J. E., and Kemp, R. (1992). The flat-maximum effect and generic linear scoring model: A test. *1MA Journal of Mathematics Applied in Business and Industry*, 4(1):97–109.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572.

Pérez López, C., Burgos Prieto, M., Huete Vázquez, S., and Gallego Vieco, C. (2012). La muestra de declarantes del irpf de 2009: Descripción general y principales magnitudes. Documentos de trabajo 11/2012, Instituto de Estudios Fiscales.

Pérez López, C., Burgos Prieto, M., Huete Vázquez, S., and Pradell Huete, E. (2013). La muestra de declarantes del irpf de 2010: Descripción general y principales magnitudes. Documentos de trabajo 22/2013, Instituto de Estudios Fiscales.

Pérez López, C., Villanueva Garcíia, J., Burgos Prieto, M., Bermejo Rubio, E., and Khalifi Chairi El Kammel, L. (2015). La muestra de declarantes del irpf de 2012: Descripción general y principales magnitudes. Documentos de trabajo 18/2015, Instituto de Estudios Fiscales.

Pérez López, C., Villanueva Garcíia, J., Burgos Prieto, M., Bermejo Rubio, E., and Khalifi Chairi El Kammel, L. (2016). La muestra de declarantes del irpf de 2013: Descripción general y principales magnitudes. Documentos de trabajo X/2016, Instituto de Estudios Fiscales.

Pérez López, C., Villanueva Garcíia, J., Burgos Prieto, M., Pradell Huete, E., and Moreno Pastor, A. (2014). La muestra de declarantes del irpf de 2011: Descripción general y principales magnitudes. Documentos de trabajo 17/2014, Instituto de Estudios Fiscales.

Pérez-Martín, A. (2008). *Estimación en áreas pequeñas bajo modelos lineales mixtos con dos factores aleatorios anidados*. PhD thesis, Universidad Miguel Hernández, Elche, Alicante.

Pérez-Martín, A. and Vaca, M. (2017). Computational experiment to compare techniques in large data sets to measure credit banking risk in home equity loans. *Data Management and Security Applications in Medicine, Science and Engineering*, 5(5):771–779.

Picos Sánchez, F., Pérez López, C., Gallego Vieco, C., and Huete Vázquez, S. (2011). La muestra de declarantes del irpf de 2008: Descripción general y principales magnitudes. Documentos de trabajo 14/2011, Instituto de Estudios Fiscales.

Piramuthu, S. (1999). Financial credit-risk evaluation with neural and neurofuzzy systems. *European Journal of Operational Research*, 112(2):310–321.

Porter, M. (1982). *Estratégica competitiva*. C.E.C.S.A.

Quinlan, J. (1983). *Learning efficient classification procedures, en Machine learning: an Artificial Intelligence approach*. Tioga Press, Palo Alto, EUA.

Quinlan, J. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, Inc., California, EUA.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabasa Dolado, A. (2009). *Método para la reducción de Sistemas de Reglas de Clasificación por dominios de significancia*. PhD thesis, Universidad Miguel Hernández, Elche, Alicante.

Reichert, A., Cho, C., and Wagner, G. (1983). An examination of conceptual issues involved in developing credit-scoring models. *Journal of Business and Economic Statistics*, 1(2):101–114.

Rodríguez Sala, J. J. (2014). *Método para generación y ordenación de reglas de clasificación. Diseño y estudio computacional. Aplicación a la Inteligencia de Negocio*. PhD thesis, Universidad Miguel Hernández, Elche, Alicante.

Rosenberg, E. and Gleit, A. (1994). Quantitative methods in credit management: a survey. *Operations Research*, 42(4):589–613.

Safavian, S. R. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on system, man and cybernetics*, 21(3):660–673.

Salchenberger, L.and Cinar, E. and Lash, N. (1992). Neural networks: a new tool for predicting thrift failures. *Decision Sciences*, 23(4):896–916.

Salinas Flores, J. (2005). Patrones de morosidad para un producto crediticio usando la técnica de árbol de clasificación cart. *Revista de la Facultad de Ingeniería Industrial UNMSM*, 8(1):29–36.

Schebesch, K. B. and Stecking, R. (2005). Support vector machines for classifying and describing credit applicants: detecting typical and critical regions. *Journal of the Operational Research Society*, 56:1082–1088.

Searle, S., Casella, G., and McCullogh, C. (1982). *Variance components*. John Wiley and Sons, Inc., New-York.

Shannon, C. E. (1951). Prediction and entropy of printed english. *The Bell System Technical Journal*, 30(1):50–64.

Shin, K. and Han, I. (2001). A case-based approach using inductive indexing for corporate bond rating. *Decision Support Systems*, 32(1):41–52.

Showers, S. R. and Chakrin, L. M. (1981). Reducing uncollectable revenue from residential telephone customers. *Interfaces*, 11(6):21–31.

Srinivasan, V. and Kim, Y. H. (1987a). The bierman–hausman credit granting model: a note. *Management Science*, 33(10):1361–1362.

Srinivasan, V. and Kim, Y. H. (1987b). Credit granting: a comparative analysis of classification procedures. *Journal of Finance*, 42(3):665–681.

Steenackers, A. and Goovaerts, M. (1989). A credit scoring model for personal loans. *Insurance: Mathematics and Economics*, 8(1):31–34.

Stehman, S. V. (1997a). Estimating standard errors of accuracy assessment statistics under cluster sampling. *Remote Sensing of Environment*, 60(3):258–269.

Stehman, S. V. (1997b). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1):77–89.

Sullivan, A. C. (1981). *Consumer Finance. In E. I. Altman, Financial Handbook*. John Wiley and Sons, Inc., New York.

Sun, J. and Li, H. (2009). Financial distress early warning based on group decision making. *Computers and Operations Research*, 36(3):885–906.

Tam, K. and Kiang, M. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science*, 38:926–947.

Tay, F. and Cao, L. (2002). Modified support vector machines in financial time series forecasting. *Neurocomputing*, 48(1–4):847–861.

Therneau, T., Atkinson, B., and Ripley, B. (2014). *Recursive Partitioning and Regression Trees*.

Thomas, L. C. (2000). A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2):149–172.

Thomas, L. C. and Edelman, D. B.and Crook, L. N. (2002). Credit scoring and its applications. Technical report, Society for Industrial and Applied Mathematics.

Trias, R., Carrascosa, F., Fernández, D., Parés, L., and Nebot, G. (2005). Riesgo de créditos: Conceptos para su medición, basilea ii, herramientas de apoyo a la gestión. *AIS Group - Financial Decisions*.

Trias, R., Carrascosa, F., Fernández, D., Parés, L., and Nebot, G. (2008). El método rdf (risk dynamics into the future). el nuevo estándar de stress testing de riesgo de crédito. *AIS Group - Financial Decisions*.

Van Gestel, T., Baesens, B., Garcia, J., and Van Dijcke, P. (2003). A support vector machine approach to credit scoring. *Bank en Financiewezen*, 2:73–82.

Van Gestel, T., Suykens, J.A.K.and Baestaens, D., Lambrechts, A., Lanckriet, G., Vandaele, B.and De Moor, B., and Vandewalle, J. (2001). Financial time series prediction using least squares support vector machines within the evidence framework. *IEEE Transactions on Neural Networks*, 12(4):809–821.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, USA.

Varetto, F. (1998). Genetic algorithms applications in the analysis of insolvency risk. *Journal of Banking Finance*, 22(10–11):1421–1439.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.

Wang, G., Hao, J., and Ma, J. e. a. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 38:223–230.

Wang, G. and Ma, J. (2011). Study of corporate credit risk prediction based on integrating boosting and random subspace. *Expert Systems with Applications*, 38(11):13871–13878.

Wang, Y., Wang, S., and Lai, K. (2005). A new fuzzy support vector machine to evaluate credit risk. *IEEE Transactions on Fuzzy Systems*, 13(6):820–831.

West, D. (2000). Neural network credit scoring models. *Computers and Operations Research*, 27:1131–1152.

Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, 15(03):757–770.

Yobas, M.B.and Crook, J. and Ross, P. (2000). Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics*, 11:111–125.

Yobas, M. B., Crook, J. N., and Ross, P. (1997). Credit scoring using neural and evolutionary techniques. Working Paper 2, University of Edinburgh.

Yu, L. (2014). Credit risk evaluation with a least squares fuzzy support vector machines classifier. *Hindawi Publishing Corporation Discrete Dynamics in Nature and Society*, page 9.

Yu, L., Wang, S., and Cao, J. (2009a). A modified least squares support vector machine classifier with application to credit risk analysis. *International Journal of Information Technology and Decision Making*, 8:697–710.

Yu, L., Wang, S., and Lai, K. (2009b). An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: the case of credit scoring. *European Journal of Operational Research*, 195:942–959.

Yu, L., Wang, S., Wen, F., Lai, K., and He, S. Y. (2008). Designing a hybrid intelligent mining system for credit risk evaluation. *Journal of Systems Science and Complexity*, 21:527–539.

Yu, L., Yao, X., Wang, S., and Lai, K. (2011). Credit risk evaluation using a weighted least squares svm classifier with design of experiment for parameter selection. *Expert Systems with Applications*, 38(12):15392–15399.

Yu, L., Yue, W., Wang, S., and Lai, K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Systems with Applications*, 37:1351–1360.

Yu, L. A., W. S. and Lai, K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications*, 34(2):1434–1444.

Zekic-Susac, M., Sarlija, N., and Bensic, M. (2004). Small business credit scoring: A comparison of logistic regression, neural networks, and decision tree models. In *26th International Conference on Information Technology Interfaces*, pages 265–270, Croatia.

Zhou, L. G., Lai, K. K., and Yu, L. (2009). Credit scoring using support vector machines with direct search for parameters selection. *Soft Computing*, 13:149–155.

Zhou, L. G., Lai, K. K., and Yu, L. (2010). Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications*, 37:127–133.

Índice de gráficos

UNIVERSITAS
Miguel
Hernández

Índice de tablas

UNIVERSITAS
Miguel
Hernández

UNIVERSITAS
Miguel
Hernández

UNIVERSITAS
Miguel
Hernández

UNIVERSITAS
Miguel
Hernández