UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Aprendizaje Estadístico en Educación

una propuesta de modelización para carreras de grado en Ingeniería

Daniel Eduardo Alessandrini López

Programa de Posgrado en Ingeniería Matemática
Facultad de Ingeniería
Universidad de la República

Montevideo – Uruguay
Noviembre de 2019

UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Aprendizaje Estadístico en Educación

una propuesta de modelización para carreras de grado en Ingeniería

Daniel Eduardo Alessandrini López

Tesis de Maestría presentada al Programa de Posgrado en Ingeniería Matemática, Facultad de Ingeniería de la Universidad de la República, como parte de los requisitos necesarios para la obtención del título de Magister en Ingeniería Matemática.

Director de tesis:
Ph.D. Prof. Paola Bermolen

Codirector:
Ph.D. Prof. Mathías Bourel

Director académico:
Ph.D. Prof. Paola Bermolen

Montevideo – Uruguay
Noviembre de 2019

Alessandrini López, Daniel Eduardo

Aprendizaje Estadístico en Educación / Daniel Eduardo Alessandrini López. - Montevideo: Universidad de la República, Facultad de Ingeniería, 2019.

XI, 130 p. 29, 7cm.

Director de tesis:

Paola Bermolen

Codirector:

Mathías Bourel

Director académico:

Paola Bermolen

Tesis de Maestría – Universidad de la República, Programa de Ingeniería Matemática, 2019.

Referencias bibliográficas: p. 95 – 100.

1. Herramienta diagnóstica, 2. Aprendizaje automático, 3. Modelos de consenso, 4. Correspondencia múltiple, 5. R. I. Bermolen, Paola *et al.* II. Universidad de la República, Programa de Posgrado en Ingeniería Matemática. III. Título.

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

Ph.D. Prof. Carolina Crisci

Mág. Prof. Ximena Otegui

Ph.D. Prof. Marco Scavino

Montevideo – Uruguay
Noviembre de 2019

A mi hija Delfina y a mi esposa Victoria, que fueron
mi sostén emocional durante todo este tiempo.
Y para mis abuelos, María Zunilda y Nené, que estuvieron en
alma y espíritu cuando este eterno estudiante lo necesitó.

Agradecimientos

Finalmente, llegó la hora tan esperada... la de cerrar una página que estuvo abierta mucho tiempo, durante el cual he aprendido innumerables lecciones, más allá de lo estrictamente académico, que me permiten arribar a este momento con serenidad.

Quiero en primer lugar agradecer a mis tutores, los doctores Paola Bermolen y Mathías Bourel por la paciencia, comprensión y disposición que cada uno de estos largos días han tenido conmigo. Seguidamente agradecer a todos mis compañeros de la Unidad de Enseñanza de la Facultad de Ingeniería, desde los actuales hasta aquellos que han seguido por diferentes rumbos, porque cada uno a su medida ha sido importante para ayudarme a llegar a la meta. Además agradecer al Dr. Marco Scavino, quién fue mi primer contacto con esta maestría que estoy culminando con este trabajo, y que sin su aporte no estaría escribiendo estas líneas. Extiendo este saludo a docentes y funcionarios de la Facultad y a mis compañeros de la maestría que de un modo u otro hicieron posible la finalización del mismo.

Finalmente, agradecer a mi esposa María Victoria, a mi hija Delfina y, por extensión, al resto de mi familia y amigos por el aguante y las largas jornadas que no pude disfrutar de la vida en compañía de los que uno más quiere.

RESUMEN

El objetivo del presente trabajo es intentar predecir, con el menor error posible, qué ocurre con estudiantes al ingreso de las carreras de grado de Facultad de Ingeniería, respecto a su progreso o eventual desvinculación. Para ello se combinan dos fuentes de datos: por un lado características sociodemográficas disponibles de cada alumno al ingreso y por otro una evaluación de conocimientos y habilidades utilizada desde hace más de una década: la Herramienta Diagnóstica al Ingreso, diseñada, mantenida y analizada por distintas entidades dentro de la Facultad.

Para lograr una mirada más general al problema, se utilizaron simultáneamente seis modelos muy usados en la práctica dentro del ambiente del Aprendizaje Automático, junto con tres tipos de Modelos de Consenso, que surgen de agregar de una cierta manera al resto de los modelos. Dentro de un mismo "loop" se generaron distintos conjuntos de entrenamiento y prueba, y se ajustaron los modelos a éstos, mediante relaciones aditivas entre las variables explicativas y las dicotómicas de interés (rendimiento y desvinculación). Esta -abundante- información fue resumida en distintas medidas que surgen de una matriz de confusión o evaluación, como p.ej. la sensibilidad o el error general cometido al predecir el resultado de cada variable. Con ello se construye un *ranking* de fórmulas para determinar cuáles son más significativas, no solo para comparar los resultados por modelo sino también para estudiar como aciertan o erran los mismos. Para este último punto se utiliza el Análisis de Correspondencia Múltiple, proyectando variables e individuos en espacios comunes. Para facilitar el trabajo, se creó un paquete en el software `R` que combina funcionalidades existentes con otras propias para generar, resumir y analizar toda la información necesaria.

Se destacaron como variables explicativas, independientemente de la variable a predecir, los resultados en Matemática de la HDI y la edad al ingreso, y en menor medida el lugar de origen y subsistema de educación preuniversitaria. Además, se identificaron grupos de acierto y error para los distintos individuos, ayudando así a una caracterización más afinada de los alumnos ingresantes. Se puede afirmar que, visto puramente desde lo académico, la herramienta diseñada (reuniendo datos de pruebas diagnósticas, datos sociodemográficos y resultados de modelos de predicción) puede ser vista como una "*prueba de tamizado*", en donde con altas chances se puede identificar a estudiantes en dificultades con sus estudios.

Palabras claves:
Herramienta diagnóstica, Aprendizaje automático, Modelos de consenso, Correspondencia múltiple, `R`.

ABSTRACT

This work aims to predict academic achievement and drop-out for first-year students at UdelaR's Faculty of Engineering, during their very first weeks inside this institution (so-called "freshers"), combining two main sources of information: a diagnostic, compulsory general aptitude test (called HDI) and sociodemographic information obtained for administrative purposes.

To tackle this challenge in a general way, nine different machine learning models were used: six of them ("simple models") taken from different previous works in this field, the remaining three are "aggregated heterogeneous models" (or simply *"consensus models"*), combining different forms of voting that take into account the simple models aforementioned. In order to assure a fair comparison, all these models were fit into a loop, using the same training and testing samples for each model, but with random rearranging of indexes in each loop count. The relations between the dicotomic variables of interests and those predicting them were modeled in a formula-based fashion, considering all the possible additive models for each dependent variable. To summarise the resulting bulk of information, different metrics from confusion matrices were obtained, then used to build a formula-based ranking to properly sort the most significant covariates. Finally, Multiple Correspondence Analysis was used to get some more information in terms of hits and misses of each model.

HDI's Math subtest and entry age were the most important independent variables regardless the performance variable used, followed by birthplace and public/private-funded school, interchangeably. Overall, models tend to have better success rates for the most common categories of predictors that are related to the dependent variables. It can be concluded that this 'ensemble tool' -using this mix of academic and sociodemographic data- could potentially be used as a *screening test* to detect students at risk in early stages of their careers.

Keywords:
Standardized test, Machine learning, Consensus methods, Multiple correspondence analysis, R.

Lista de siglas

Lista de siglas

ACM Análisis de Correspondencia Múltiple 51
ACS Análisis de Correspondencia Simple 51
CART Classification and Regression Tree 29, 36
CEAM Cuestionario de Estrategias de Aprendizaje y Motivación 11, 68
EM Algoritmo Expectation-Maximization 48
EOC Espacio de Orientación y Consulta 3
FIng Facultad de Ingeniería 1, 9
GAL1 Geometría y Álgebra Lineal 1 68
HDI Herramienta Diagnóstica al Ingreso 9
KDE Kernel Density Estimation 43
RF Random Forests 34
ROC Receiver Operator Characteristic 50
SGAE Sistema de Gestión Administrativa de la Enseñanza 67
SGB Sistema de Gestión de Bedelías 67
TDC Tabla Disyuntiva Completa 51
TIPE Taller de Introducción a la Planificación Estratégica 3
UEFI Unidad de Enseñanza de la Facultad de Ingeniería 1
UTU Universidad del Trabajo del Uruguay 74, 104
UdelaR Universidad de la República 3

Tabla de contenidos

Capítulo 1

Introducción

Predecir algunas características académicas de los estudiantes en un cierto momento del tiempo tiene infinidad de usos: desde administrativos hasta académicos, pasando por temas de otra índole (presupuestales, políticos,...). Lo anterior depende de la existencia de datos de fácil acceso, completos y confiables. En caso de fallar alguno de los requisitos anteriores, quien trabaje con estos datos tendrá que valerse de información aproximada, ya sea por fuentes diferentes a las pretendidas, o con estudios propios; incluso puede crear la información buscada, aunque a costos de trabajo y tiempos seguramente muy elevados.

En el caso particular de este trabajo, los problemas de información incompleta aparecen desde el vamos, al no disponer de elementos de juicio y académicos suficientes desde fuentes previas al ingreso de los estudiantes -como p.ej. la educación preuniversitaria-, por varios motivos. Afortunadamente, existe información que si bien es creada y mantenida desde la Universidad, es información que acompaña en -al menos- buena parte de su vida a cada estudiante: los datos sociodemográficos (fecha de nacimiento, sexo biológico, domicilio, datos de contacto, etc.). Para suplir a los datos académicos, aparece una herramienta que desde hace más de una década viene aplicándose de forma sistemática en todos los inicios de curso: la Herramienta Diagnóstica al Ingreso (HDI), coordinada por la Unidad de Enseñanza de la Facultad de Ingeniería (UEFI). Es una prueba de competencias y desempeños entendidas como básicas para quienes pretenden estudiar en la Facultad de Ingeniería (FIng), aplicable exclusivamente a los alumnos que ingresan por primera vez a dicha casa de estudios en un año determinado. Busca servir de diagnóstico para la institución y de autoevaluación para los alumnos, además de ser un potencial insumo para que la institución pueda adelantarse y mitigar ciertos problemas de alto impacto, como la desvinculación y el rezago académico.

Una pregunta pertinente es, con los datos actuales ¿qué metodología se utilizaría para predecir?, e incluso *¿qué* se puede predecir? Hay muchas respuestas posibles, todo depende del tipo y calidad de los datos, la intuición y el conocimiento del investigador. En el presente trabajo, el interés principal radica -para quienes trabajamos en la UEFI- en predecir rendimiento académico y desvinculación en el corto plazo dadas una serie de variables cualitativas y cuantitativas, con lo

cual tomando una serie acotada de paradigmas de modelización es más que suficiente.

Otra pregunta importante es ¿cómo implementar lo anterior? Dentro de la variada gama de software disponible, `R` es sin dudas una de las mejores opciones, y la finalmente utilizada en este trabajo. Se utilizarán varios paquetes con gran parte de la modelización pretendida, además de crear un nuevo paquete que agrega funcionalidades nuevas -a medida para este trabajo-, aprovechando así una de las ventajas de uso de `R`.

El trabajo se organiza de la siguiente manera: en el Capítulo 2 se realiza una puesta a punto sobre la predicción de fenómenos curriculares, haciendo un racconto de bibliografía de interés. El Capítulo 3 por su parte introduce todas las cuestiones metodológicas, desde las principales características de los modelos estadísticos utilizados hasta la implementación del paquete `uefi` (y su versión mejorada `uefi2`) que reúne funcionalidades existentes con nuevas creaciones para ajustar simultáneamente dichos modelos. El Capítulo 4 presenta los resultados obtenidos en el trabajo y el Capítulo 5 cierra el trabajo a modo de consideraciones finales y desafíos a futuro. Se incluyen apéndices específicos sobre información educativa, aspectos metodológicos e informáticos.

1.1. Objetivos

El objetivo del presente trabajo es predecir, en el corto plazo, qué ocurre con los estudiantes ingresantes a la Facultad de Ingeniería de la Universidad de la República respecto de su desempeño escolar e incluso con la posibilidad de abandonar sus estudios, utilizando métodos provenientes del aprendizaje automático.

Si bien es sabido que ambas dimensiones son medibles de distintas formas, en este trabajo se parte de tres definiciones operativas de dichos fenómenos, propuestas de antemano por distintos actores dentro de la Facultad. Sobre ellas se propone un conjunto de variables posiblemente explicativas y se preselecciona un conjunto reducido para predecir las variables de interés. Estas variables explicativas surgirán de dos fuentes: la Herramienta Diagnóstica al Ingreso (HDI) y datos sociodemográficos que se recogen al momento de la inscripción de cada estudiante.

Los objetivos específicos son:

- Ajustar simultáneamente varios modelos utilizados dentro del aprendizaje automático para un grupo de formulaciones aditivas entre las variables de interés y las posibles explicativas, permutando muestras de entrenamiento y prueba varias veces para introducir variabilidad adicional y evitar extraer conclusiones con una sola muestra disponible
- Describir exhaustivamente las herramientas predictivas, comparando los resultados de todos los modelos considerados mediante indicadores pertinentes e identificar, dentro de cada formulación para cada variable, un modelo con mayor poder de clasificación correcta
- Detectar patrones de acierto y error para las formulaciones anteriores, en base a técnicas de análisis multivariado

1.2. Alcance y limitaciones

Este trabajo apunta a obtener herramientas efectivas para intervenciones educativas lo más cercana posible al ingreso de los alumnos y servir de elementos adicionales a quiénes dentro de la institución los apoyan (p.ej.: Espacio de Orientación y Consulta (EOC), acciones específicas como el Taller de Introducción a la Planificación Estratégica (TIPE)). Los estudiantes son alumnos que ingresan a Facultad de Ingeniería de la Universidad de la República (UdelaR) provenientes de distintos subsistemas educativos, con una formación en matemática y ciencias básicas en general de las más fuertes dentro de los bachilleratos disponibles. A pesar de lo anterior, durante el primer año de las diferentes carreras se da una lógica de "cuello de botella" que genera inconvenientes para todos los involucrados. Conviene entonces posicionarse al comienzo de la historia para atacar el problema de forma más efectiva.

Poder determinar con la mayor exactitud posible si alguien está en franco riesgo de presentar inconvenientes en sus próximos meses ayudaría a retener e incluso aumentar las posibilidades de un fin exitoso (léase *egreso*). Se busca además que dicha información sea de fácil obtención y "barata" de producir. Por ello, los constructos empleados para medir el rendimiento y la desvinculación son de naturaleza unidimensionales, a sabiendas de que el problema es mucho más complejo[1]. Se trata de dar un primer paso para conocer ventajas y desventajas de esta propuesta y elaborar en un futuro indicadores que se ajusten más a la complejidad de los fenómenos abordados.

1.3. Justificación

¿Por qué investigar rendimiento y deserción en el corto plazo?

Ambos temas tienen consecuencias sociales, económicas y políticas importantes, y desde el presente trabajo se quiere dejar constancia que es posible aportar información con relativamente poco esfuerzo. El foco en el corto plazo radica en que la FIng es conocida como una "facultad de primer año", en donde se concentra una parte muy importante de los estudiantes de cada generación a lo largo de los años, generando problemas tanto para los estudiantes (masividad, despersonalización) como para la institución (cursos sobrecargados, docentes asignados masivamente a cursos iniciales dejando a otros desprovistos, salones con capacidad limitada) [MLO05].

Incluso varios estudios en distintos lugares, utilizando metodología y datos diferentes[2] coinciden en que el primer año de trayectoria es *buen predictor del resto* del desempeño. Promover esta iniciativa ayuda a posicionarse *antes* de que ocurra todo el primer año y así adelantarse a los hechos.

Además, se busca proponer herramientas de investigación en uso por la comunidad educativa alrededor del mundo y que incluso en Uruguay están teniendo relevancia en los últimos años.

[1] El enfoque se hará sobre resultados educativos inmediatos y buscando determinar si cada estudiante está activo al comenzar su segundo año lectivo en la institución

[2] Referirse a las secciones 2.3.3 a 2.3.5 del presente trabajo.

¿Por qué usar fuentes de datos con medidas *antes de ingresar* a la Facultad de Ingeniería?

Porque se pretende saber el alcance predictivo de la HDI junto con las variables sociodemográficas disponibles en ese momento, para determinar si son una alternativa adecuada a la falta en la actualidad de información académica previa, como pueden ser por ejemplo las calificaciones promedio de la educación secundaria.

¿Por qué usar modelos provenientes del aprendizaje automático?

Porque son ampliamente usados en la bibliografía, pues ya han probado dar buenos resultados en un sinfín de aplicaciones, ayudando a descubrir patrones en los datos que en un análisis simplemente descriptivo serían difíciles de detectar. Además, su versatilidad y disponibilidad actual en diferentes programas hace que sea más sencilla su utilización, particularmente porque pueden ser ajustados de manera simultánea sin recaudos adicionales.

Capítulo 2

Estado del Arte

En este capítulo se detalla el estado actual de la investigación predictiva en base a pruebas diagnósticas al inicio de la vida universitaria. Para ello se detallan definiciones, estudios empíricos y de modelización predictiva que sirven de preludio a los aspectos metodológicos de este trabajo.

2.1. La llegada a la universidad

Este no es un problema nuevo. Desde hace mucho tiempo se investiga los fenómenos de deserción y rendimiento en muchas universidades a lo largo y ancho del mundo, ya que ambos traen aparejados un sinfín de consecuencias no solo para el sistema educativo sino también para la sociedad en su conjunto. En el presente apartado se abordarán ambos temas por separado para una mejor comprensión.

2.1.1. Qué ocurre en la UdelaR

Hay una premisa bastante extendida que dice que el conocimiento (y en particular el que se genera en las universidades) es un bien caro de producir, almacenar y distribuir. Si se observa desde el lado de la oferta, tanto los presupuestos como el plantel docente y las locaciones no son ilimitadas. Por el lado de la demanda por su parte, si bien es una porción del total de posibles estudiantes los que llegan a la educación terciaria, presionan de igual modo a la oferta disponible en los recursos mencionados[1]. Concretamente, los alumnos que ingresan a la universidad son, según menciona D. Jolis (2000) (citada en [MLO05]), "sujetos adolescentes" mayoritariamente, sometidos a una sucesión de "cambio-pasaje-adaptación" que les genera una serie de conflictos internos que se agregan a los existentes y los resignifican. Esto sumado a las nuevas reglas de interacción universitarias -institucionalidad, masividad, exigencia académica- opera de forma más negativa sobre aquellos menos preparados para hacerle frente (Marrero (1996) también citada en

[1] En Uruguay, del prácticamente 100 % de estudiantes que accede a educación primaria, menos del 30 % finaliza estudios preuniversitarios, a pesar del sostenido incremento en los potenciales estudiantes de la educación superior durante las últimas 4 décadas [DGP17]

[MLO05]). Si a esto le agregamos algunas carencias presentes (formación específica distorsionada, estrategias de aprendizaje casi nulas, tendencia elevada a un lento avance o al abandono), se genera una combinación de factores propicia para desestimular al estudiante a continuar, afectando su rendimiento y, en el peor de los casos, acelerando su salida del sistema: de cada 10 ingresantes a la FIng por generación, solo un máximo de 4 egresará; de estos últimos tan solo el 1 % (sobre el total de ingresantes) realiza su trayecto en consonancia con lo estipulado con cada Plan de Estudios ([LBA+17, UEF18]).

Las implicancias directas e indirectas de estos fenómenos impactan tanto a los propios estudiantes como así también a las casas de estudio y a la sociedad, aunque en distinta forma: las universidades suelen tener grandes cuellos de botella en los primeros años, donde muchas veces las necesidades de infraestructura y capital humano formado son superadas por las demandas estudiantiles, los docentes más jóvenes y menos formados deben hacer frente muchas veces a grandes cantidades de alumnos desgastando su tarea; por su parte la baja titulación condiciona al desarrollo del país (tanto en el sector público como en el privado) por ser las necesidades de masa crítica -en docencia, investigación y producción- superiores a la oferta existente[2].

En la UdelaR concretamente, esto ocurre a pesar de ser, curiosamente, una universidad con un modelo de acceso y vinculación irrestricto[3] que prioriza la permanencia del estudiante y permite el desarrollo de trayectorias ajustadas a cada individuo y su entorno ([Seo15]). ¿Qué ocurre a grandes rasgos en el mundo? Eso se intentará explicar a continuación.

2.1.2. Transición entre sistemas

Por lo mencionado anteriormente, el ingreso a la vida universitaria en casi todo el mundo es restringido: a través de distintos mecanismos se le invita al futuro estudiante a elegir uno o varios posibles caminos, y luego mediante el resultado de una o más pruebas (que pueden ser durante su último año lectivo preuniversitario y/o poco antes de comenzar la universidad) las casas de estudios eligen, ordenan o asignan -dependiendo del caso- a los estudiantes en distintas carreras. Es común encontrar una entidad estatal que gestiona alguna de estas pruebas e incluso sirve de nexo entre los subsistemas preuniversitarios y las universidades compilando la información generada por los primeros. De esta forma, tanto unos como otros cuentan con información muy detallada no solo del presente, sino también de la trayectoria académica de cada estudiante[4].

Sobre las pruebas, se puede decir que existen dos enfoques: el de currículo o trayectoria y el de aptitud ([MLO05, pp.25]). En un comienzo las pruebas propuestas pertenecían a alguna de estas categorías; sin embargo al día de hoy es muy común observar instrumentos que mezclan de forma

[2]Según estimaciones, en Uruguay hay aproximadamente 15000 ingenieros en actividad ([Ins13]). Otro indicador en el mismo sentido es el Indice de Innovación Global (GII), que estima entorno al 15 % a los egresados de carreras relacionadas a Ciencia e Ingenierías en los últimos años ([GII18]).

[3]Salvo las formalidades requeridas, p.ej. acreditar haber terminado nivel preuniversitario.

[4]Incluso en algunas universidades que, por distintos motivos no están bajo un mismo programa gubernamental, exigen pruebas adicionales a los estudiantes para verificar sus conocimientos y aptitudes y decidir si le otorgan un lugar para continuar sus estudios allí.

disímil ambas tradiciones. Exámenes como el GaoKao[5] o el JEE-Adv[6] suelen ser considerados las pruebas más exigentes a nivel mundial, en particular por la gran cantidad de estudiantes que se presenta, la duración (más de un día) y las resultantes admisiones otorgadas (entre 1 y 10 % del total de estudiantes)[7]. En los países desarrollados, exámenes como el Abitur (Alemania), el *Bac* (Francia) el SAT (EEUU) o los denominados Matura (p.ej. en Suiza) suelen usarse como estándar para las universidades ([MLO05],[AG09]). En América Latina, varios países cuentan con pruebas como las mencionadas. Dependiendo de la universidad a la cual se quiera acceder, puede que las pruebas sean obligatorias u opcionales. En general, las universidades públicas o las que tienen alta demanda son las que proponen este tipo de exámenes para "seleccionar" a sus futuros alumnos. También pueden recurrir a información previa del estudiante, como pueden ser promedios en alguna etapa de su educación preuniversitaria para "afinar" su decisión final. En el Apéndice 1 se presenta una tabla con lo investigado para algunos países latinoamericanos.

Uruguay y Argentina han sido históricamente una excepción. En ambos países, y a pesar de los vaivenes políticos, se ha priorizado el derecho a la educación respecto a la libertad de enseñanza, entendido el primero como uno de los derechos fundamentales del ser humano[8]; en el caso de la libertad de enseñanza se le otorga potestad a las instituciones para decidir qué y cómo enseñan ([GR18]). Vale decir que hay actualmente excepciones a la regla anterior. Particularmente en Uruguay, algunas universidades privadas proponen pruebas estandarizadas para otorgar becas[9] o para conocer aptitudes en estudiantes de algunas carreras[10], mientras que en algunas carreras de la UdelaR hay pruebas específicas de ingreso[11] ([Rom10]).

También las hubo en el pasado: durante buena parte de la intervención a esta casa de estudios (1973-1985) se aplicaron gradualmente pruebas de admisión de estudiantes en algunas facultades, para luego llegar a 1980 con una suerte de examen de ingreso para toda la Universidad, restringiendo así el acceso de forma selectiva. Esto duró dos años: en 1982 se sustituyen estos exámenes por un sistema de cupos prefijados, dejando sin efecto lo anterior ([UCU07]).

2.2. Herramientas de diagnóstico

2.2.1. Generalidades

En muchas casas de estudio alrededor del mundo las pruebas de diagnóstico no son necesarias en cuanto ya existen otras pruebas que pueden dar una idea de la preparación de los bachilleres

[5] GaoKao: Examen general para el ingreso a la educación superior, China

[6] JEE-Adv: Examen de Admisión Conjunta - Avanzado, exclusivamente para acceder a los institutos tecnológicos nacionales (IIT), India

[7] Nota de prensa, Hindustan Times

[8] Art.71 de la Constitución de la República [con67].

[9] Prueba de Actitud Académica 'SAT en español' para presentarse a becas en universidades privadas, www.ort.edu.uy/30210/9/prueba-de-aptitud-academica.html

[10] http://www.admisionesum.uy/pasos-para-inscribirse/

[11] Traductorado: exámenes de lenguas para ingresar; Escuela Universitaria de Música: pruebas de admisión en Conocimientos Musicales y Específica para Licenciatura en Música o Instrumentos

al ingresar a la universidad, como se mencionó en la Subsección 2.1.2.

Las pruebas diagnósticas revisten mucha utilidad en aquellos lugares donde no se cuenta con ese tipo de instrumentos de evaluación[12], porque permiten conocer a los nuevos estudiantes en cuanto a su formación recibida y su contexto de procedencia, logrando acciones que acompañen su trayectoria y evitando así su desvinculación temprana ([ACF+18]).

2.2.2. Pruebas diagnósticas a nivel universitario en Uruguay

En Uruguay, a diferencia de muchos países de la región y del mundo, el pasaje de la educación preuniversitaria a la universitaria es con un certificado[13] que da fe de los conocimientos mínimos adquiridos por el alumno en referencia a su futura casa de estudios. En ningún caso hay necesidad de rendir una prueba -generalmente estandarizada- que indique específicamente qué conocimientos y habilidades ha obtenido cada estudiante. Esto tiene claras implicancias respecto a la información computarizada disponible en la actualidad: el sistema universitario sólo cuenta con datos sociodemográficos comunes (fecha de nacimiento, sexo biológico, lugar de procedencia, etc.) y -eventualmente- con información adicional proveniente de cuestionarios generales (como el Formulario 69A al ingreso) o específicos (cuestionarios de estrategias de aprendizaje, socioeconómicos, etc.).

¿Qué ocurre con la información curricular preuniversitaria? Si bien varios trabajos citan a este insumo como importante para predecir determinadas cuestiones en la universidad, hasta ahora no se han encontrado trabajos realizados en Uruguay con estas características. De hecho, las experiencias propias de la UEFI en este sentido no dieron frutos: durante parte del proyecto Moebius ([MBAP15]) fue prácticamente imposible obtener información académica de forma sistematizada y ordenada, dada la multiplicidad de criterios de almacenamiento y sistemas informáticos utilizados en los distintos liceos de la experiencia, incluso a pesar de ser relativamente pocos los estudiantes participantes.

En cuanto a la utilización de pruebas de diagnóstico, se pueden citar los ejemplos de las Facultades de Ciencias Económicas y Administración ([ACF+18]), de Química ([RA07],[RARD11]), de Medicina[14], de Ciencias ([Mí08]) y del Centro Universitario Regional Este ([RM17]) en la UdelaR, además de otros en el ámbito privado ([lVCI+07], [BDL+13]). Cada una de ellas tiene especificidades en función de los saberes necesarios en cada disciplina, salvo en el caso del CURE donde los conocimientos evaluados son más generales por tratarse de alumnos de un Ciclo Inicial Optativo[15].

[12] En varias universidades de América Latina existen en su lugar "cursos iniciales de nivelación" (denominados de maneras diferentes) con el objetivo de nivelar los conocimientos adquiridos en la educación preuniversitaria; dependen de cada carrera y están diseñados como un módulo inicial dentro de cada facultad. Una vez que el estudiante aprueba dicho módulo, pasa a ser un "estudiante efectivo" de la carrera escogida

[13] *Formulario 69A* expedido por liceos públicos y privados (ANEP), certificado de egreso de UTU, títulos de egreso de carreras técnicas, docentes u otras; todo esto dependiendo de la(s) carrera(s) escogida(s).

[14] La herramienta se denomina Prueba Inicial de Evaluación Diagnóstica (PIED), realizada a través de la plataforma EVA.

[15] Los CIOs son ciclos flexibles, administrados durante todo un año lectivo, que apuntan a alumnos con "débil

En el siguiente apartado se hará foco sobre la utilizada en la FIng.

2.2.3. La Herramienta Diagnóstica al Ingreso

Historia

Las pruebas de diagnóstico tienen un largo historial dentro de Facultad de Ingeniería. Se puede decir -a grandes rasgos- que hubo tres grandes etapas: en la primera (1992-2001) los institutos de Matemática (IMERL) y Física (IFFI) de la Facultad confeccionaban y procesaban pruebas por separado con conocimientos específicos de dichas disciplinas, para que luego la UEFI elaborara informes con los resultados y así poner en conocimiento al Consejo de Facultad. En una segunda etapa, previa a la HDI de la actualidad (2002-2004), comenzó un período de maduración de las pruebas, pasando de ser planteos concebidos y desarrollados por separado a ser una única unidad de medición. Además, en 2002 y 2004 se introducen pruebas de Lengua y Química respectivamente[16]. ([MLO05])

Finalmente, desde el año 2005 la Facultad de Ingeniería cuenta con la *Herramienta Diagnóstica al Ingreso (HDI)*, un instrumento que tiene como cometidos principales:

- A nivel institucional, servir de diagnóstico para cada generación ingresante, tanto de sus conocimientos a nivel académico como así también de hábitos de aprendizaje, de estrategias cognitivas, etc.
- A nivel estudiantil, ser una forma de autoevaluación, para que cada alumno sepa con qué conocimientos viene de la educación media
- Para los docentes, saber el nivel de preparación de sus futuros alumnos y adecuar sus cursos de forma acorde

Dada la importancia otorgada institucionalmente a este tema, el Consejo de Facultad resolvió, con el fin de obtener un "diagnóstico de la situación de sus ingresantes", que se realizará de ahora en más una prueba "de carácter obligatorio" antes de comenzar los cursos, sus resultados serán tenidos en cuenta "sólo en forma positiva"[17], en caso de aprobación se otorgarán "hasta un máximo de 5 puntos" para asignaturas de primero, y en caso de inasistencia sin justificación habrán sanciones ("imposibilidad de rendir exámenes... en julio/agosto...")[18]. Esto se mantiene -con algunas pequeñas modificaciones- hasta la actualidad. Una resumida línea de tiempo se aprecia en la Figura 2.1.

vocación específica", evitando que tengan que reiniciar sus estudios de bachillerato en caso de cambiar de vocación. Fuente: http://www.cure.edu.uy/?q=node/40

[16] Durante este período se da un crecimiento en el trabajo con equipos multidisciplinarios de distintos actores del sistema educativo.

[17] Resolución s/n del 20/12/2004

[18] Resolución s/n del 14/02/2005, ampliada luego por Resolución 1175 del 05/07/2007

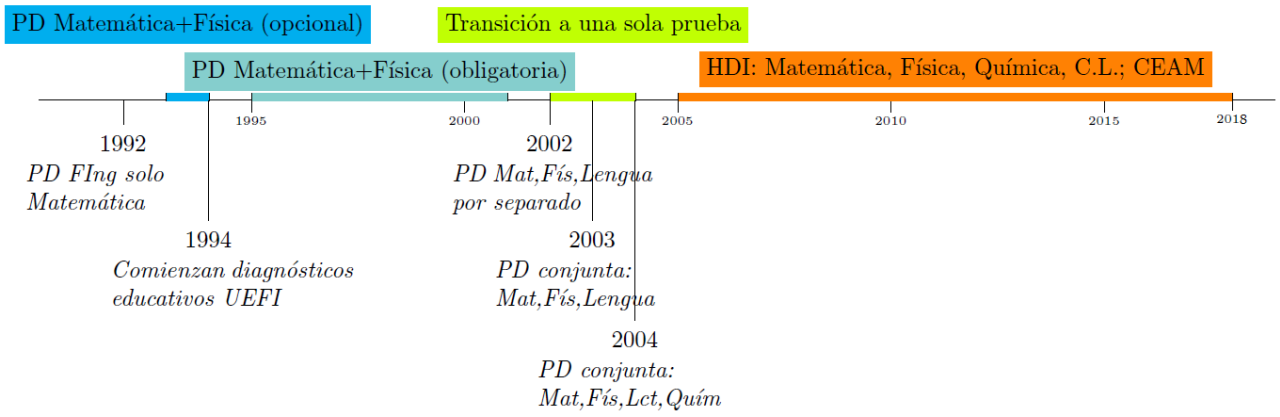

Figura 2.1: Pruebas diagnósticas al inicio en FIng: línea de tiempo 1992-2018

Características de la HDI

Es una prueba de *bajo impacto*[19] y de carácter obligatorio para los ingresantes a todas las carreras de grado y técnicas de FIng, enfocada principalmente en evaluar ciertos "niveles de desempeño"[20], con preguntas de fácil y rápida corrección pero que contribuyan con información relevante ([MLO05, pp.30-38]). Solo se debe rendir una única vez (aquellos alumnos que reingresan a FIng y/o rindieron la prueba en otro momento están exentos de realizarla) y se efectúa durante una jornada, inmediatamente antes del comienzo de los cursos de cada semestre. La duración en la actualidad es de 3 horas, aunque ha llegado a ser de hasta 4 horas en algunas ediciones anteriores, en función de la cantidad de componentes y la longitud de las preguntas en cada uno. En cualquier caso, no se puede usar material de consulta ni calculadora durante toda la duración de la prueba.

Si bien el formato de la prueba ha tenido cambios a lo largo del tiempo, en líneas generales la misma ha contado con:

- Preguntas:
  - *Preguntas cerradas*: la mayoría son de múltiple opción, con una sola opción correcta y entre 3 y 4 distractores (dependiendo de la edición y el componente), valiendo 1 punto la respuesta correcta y cero la incorrecta. A través de escáners se generan las bases de datos que luego la UEFI utilizará para corregir la prueba
  - *Preguntas abiertas*: los componentes de Matemática y Comprensión Lectora han contado en distintos momentos del tiempo con preguntas de este tipo, generalmente propuestas de desarrollo (de cálculos en caso de Matemática o de ideas en caso de Comprensión

[19] Evaluaciones de bajo impacto o riesgo: son aquellas que no tienen consecuencias para quienes las realizan, en base a su performance o resultado final. Están en contraposición con las evaluaciones de alto impacto o riesgo, que sí tienen consecuencias en función del resultado obtenido. Fuente: http://uis.unesco.org/en/glossary-term/high-stake-assessment.

[20] Se decide de antemano una opción teórica sobre modelos de aprendizaje en la enseñanza de las ciencias.

Lectora). La corrección de estos componentes ha estado a cargo de algunos grupos de trabajo (en el IMERL para Matemática, en UEFI[21] para Comprensión Lectora)

- Actividades: las preguntas se agrupan en *Componentes* (Matemática, Física, Química, Comprensión Lectora), éstos se agrupan en *Actividades*

  - La Actividad 1 hace referencia a los componentes de Matemática, Química y Física de la prueba; se posicionan en bloques de preguntas con numeración sucesiva[22]. En el caso de Matemática, las preguntas se basan en saberes esperables de los alumnos al ingresar: uso de herramientas del cálculo (derivadas, límites, extremos), lógica y teoría de conjuntos, geometría y trigonometría, sistemas lineales y cálculo de porcentajes
  - La Actividad 2 agrupa a las preguntas de Comprensión Lectora, con un texto principal y preguntas relacionadas al mismo
  - Las actividades adicionales (Actividad 3, Actividad 4) hacían referencia a componentes adicionales que no formaban parte de la prueba puntuable pero sí brindaban información complementaria sobre los estudiantes, como el Cuestionario de Estrategias de Aprendizaje y Motivación (CEAM), incluido desde 2005 hasta 2017[23] [MCC+07].

Para aquellos alumnos que superan un umbral de suficiencia (60 % del puntaje total de la prueba), se otorgan gradualmente, dependiendo de la calificación final de la prueba, hasta 5 puntos[24] para ser utilizados en las primeras tres asignaturas durante el primer año: Cálculo 1, Geometría y Álgebra Lineal 1 y Física 1, siendo válidos solamente durante el primer año de cursada.

Componentes o dimensiones La cantidad y tipo de componentes ha variado a lo largo de la historia de las pruebas diagnósticas en FIng, aunque para la HDI hubo 3 momentos de quiebre: entre 2005 y 2015 se utilizaron los cuatro componentes originales –Matemática, Física, Química y Comprensión Lectora con distintos ítems, cantidad de posibilidades e incluso preguntas abiertas en Comprensión Lectora y/o Matemática–; desde 2016 se quitó la componente de Química de la prueba por sugerencia de la UEFI, y finalmente desde el segundo semestre de 2017 se dejó de utilizar la componente de Física, quedando en la prueba solo preguntas de Matemática y Comprensión Lectora.

[21] Durante algunas ediciones de HDI se realizaron contrataciones externas para corregir este componente

[22] En la actualidad la Actividad 1 solo agrupa preguntas de Matemática

[23] El CEAM fue variando desde su primera versión, teniendo 37 (2013-2015), 50 (2017), 60 (2016) hasta 68 (2005-2012) afirmaciones de las que los alumnos debían mostrar su grado de acuerdo -en una escala Likert de 4 puntos-, todas referidas a temas motivacionales y de estrategias de estudio.

[24] Estos puntos son en una escala de 0 a 100, coincidente en general con la escala al interior de cada asignatura. Al finalizar cada curso, los docentes transforman este puntaje en la escala de *notas* de UdelaR, que va de 0 a 12.

Alumnos que rinden la prueba Según datos disponibles entre 2006 y 2016, los alumnos que se presentan a la HDI tienen estas características:

- Casi tres cuartas partes de los ingresantes (72,5 %)[25] hizo la HDI. Como se muestra en la tercer columna de la Tabla 2.1, el promedio anual se sitúa en el entorno del 70 %; desde 2011 se logró aumentar ese guarismo gracias a tener dos instancias de HDI, una antes del inicio de *cada semestre*
  - Más del 98 % de los alumnos que efectivamente hizo la HDI pertenece a su generación; aquellos que rinden la prueba en un momento diferente lo hacen como máximo un año antes
  - A pesar de que la HDI se puede hacer sólo una vez, casi el 1 % la hizo *más de una vez* (con una mediana de 1 año entre la primera y segunda vez)
- Más del 95 % de alumnos que cada año hacen HDI y cursan su(s) primera(s) unidad(es) curricular(es) (UC), lo hacen en el mismo año[26]

Con estos datos se puede decir que la porción de alumnos que rinde la prueba es alta (aunque no extrema, sobre todo dependiendo qué carrera escoge al iniciar su tránsito en FIng), de éstos la mayoría tiene actividad durante el mismo año en el que realiza la prueba, aunque no en las asignaturas *esperables* para muchos de los casos, algo ya constatado en [UEF09b].

Algunas medidas psicométricas En la Tabla 2.1[27] se muestra, para cada generación que hizo HDI, total de ingresantes, cobertura e "indicadores HDI", que hacen referencia a *promedios* en los respectivos valores de *consistencia interna* (α *de Cronbach*), dificultad (*Índice de Dificultad*) y discriminación entre rendimientos estudiantiles (*Índice de Discriminación*) para cada una de las ediciones de la prueba[28]. En líneas generales se trata de una prueba de dificultad media-alta, con un nivel medio a bajo de discriminación y una consistencia interna adecuada[29].

Resultados En líneas generales, y a pesar de algunos cambios dentro de cada componente de la prueba, se observa un descenso paulatino en los niveles de aprobación, ya sea por componentes o considerando resultados globales (como aprobar simultáneamente cada uno de las subpruebas -suficiencia global- o llegar al mínimo de puntos totales en la prueba -suficiencia en la HDI). Los niveles de suficiencia están en sintonía con lo encontrado en otros centros de estudio con similares características ([RM17]). En la Figura 2.2 se resumen estos guarismos.

[25] El total de alumnos considerados es de 16140, entre 2006 y 2016

[26] El total en este caso es sobre 12175 alumnos que hacen HDI y además cursan al menos 1 UC, también entre 2006 y 2016

[27] Para la columna de preguntas de Matemática (*Pregs. Mat.*): 'c':cerradas, 'a':abiertas (corregidas por docentes); las preguntas de CL fueron 5 cerradas, salvo en 2014 que se agregó 1 abierta.

[28] Desde 2011 se toman los "indicadores HDI" solo del primer semestre de cada edición.

[29] Para valores de referencia, referirse a https://www.statisticshowto.datasciencecentral.com/cronbachs-alpha-spss/.

Generacion de ingreso	Total inscripción	%Cobertura HDI	INDICADORES HDI			Pregs. Mat.
			α-Cronbach	IDif	IDis	
2006	1305	67,7 %	0,783	0,519	0,235	13c;2a
2007	1284	68,1 %	0,779	0,544	0,248	13c;2a
2008	**1318**	**70,4 %**	**0,796**	**0,494**	**0,244**	**13c;1a**
2009	1317	67,8 %	0,817	0,519	0,283	13c;2a
2010	1501	68,4 %	0,792	0,48	0,345	13c;2a
2011	1446	78,0 %	0,795	0,562	0,273	13c;2a
2012	1478	75,3 %	0,748	0,569	0,365	13c;1a
2013	1628	78,4 %	0,76	0,542	0,34	13c;1a
2014	1567	76,8 %	0,759	0,599	0,262	13c;1a
2015	1651	71,8 %	0,705	0,44	0,25	15c;1a
2016	**1645**	**73,4 %**	**0,838**	**0,442**	**0,134**	**16c**

Tabla 2.1: HDI: cobertura e indicadores de consistencia, 2006-16

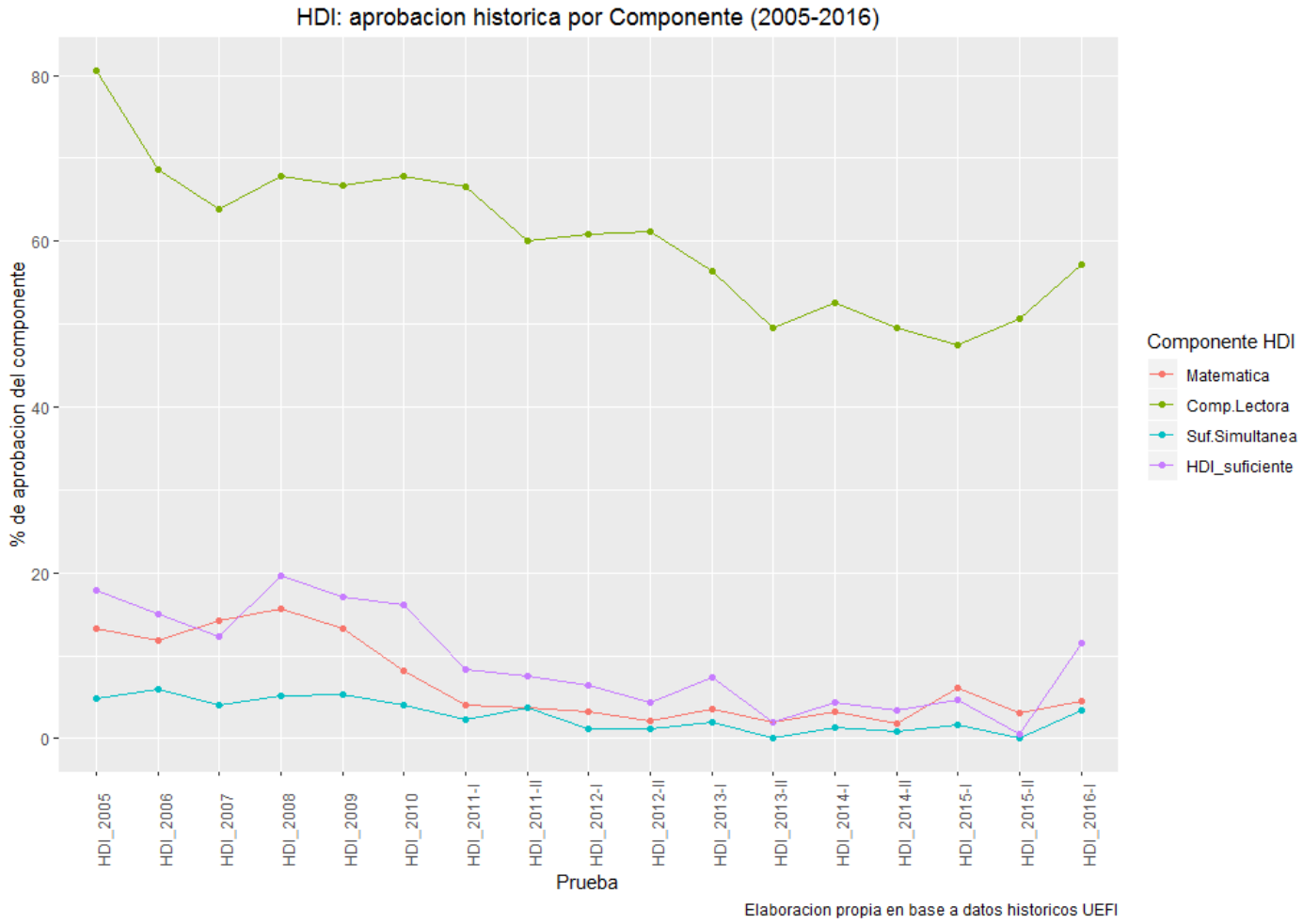

Figura 2.2: Niveles históricos de aprobación en HDI (2005-2016)

2.3. Investigación en Educación en la Universidad

La investigación en educación superior abarca muchísimos temas y miradas, ya sea por sobre qué actores se considera (alumnos, instituciones, sistemas educativos) como también las definiciones utilizadas (parciales, absolutas, totales...). Las formas metodológicas y las aproximaciones a cada fenómeno también lo son (estudios cualitativos, cuantitativos, de reflexión, históricos...), y con ello los datos utilizados (transversales, longitudinales o de panel). A la infinidad de trabajos existentes se puede acceder por distintas fuentes, ya sean revistas especializadas[30] o también en publicaciones que no son principalmente de enseñanza pero que toman a este tema como dentro de sus cometidos[31]; incluso las jornadas de investigación o grupos de trabajo sobre variadas temáticas son cada vez más diversas[32].

En este trabajo se pretende hacer foco solamente en estudios cuantitativos que busquen saber qué ocurrirá con el rendimiento o la vinculación de los alumnos en el *corto plazo*, con información al ingreso a la universidad. Para ello es necesario hacer una breve mención sobre algunos trabajos seleccionados.

2.3.1. Un tema "de larga data"

Si de datos se trata, la educación está viviendo -como muchas otras áreas del conocimiento- una verdadera explosión en cuanto a la disponibilidad de información en la actualidad, aunque es sabido que estos temas tienen cabida hace mucho: tanto la importancia del rendimiento académico previo de los estudiantes universitarios como el impacto del abandono en los sistemas educativos vienen siendo estudiados en varias partes del mundo, desde por lo menos la segunda mitad del siglo XX ([RA07],[Dor93][33]).

Esta historia acumulada hace que, cuando se requiere realizar revisiones sistemáticas para generar nueva información, sea muchas veces necesario realizar otra investigación (véase p.ej.: [PA14b], [BY09], [RV07]) o citar minuciosamente a quiénes las hayan realizado (p.ej. [RA07, págs.30-39,142-160]). Lo que está ocurriendo en estos últimos años es que muchas veces la gran cantidad de información educativa que se genera no puede ser examinada por medios más tradicionales; de ahí el surgimiento de nuevas miradas presentadas a continuación.

2.3.2. Las disciplinas emergentes en base a datos de gran tamaño

En la actualidad, la información educativa no se basa solamente en resultados provenientes de actas de curso o en registros cerrados en una base de datos. Con la irrupción de las tecnologías de la información y sus consecuencias a nivel de disponibilidad de fuentes (registros administrativos,

[30] Ejemplo de revista especializada: https://revistas.uam.es/riee

[31] Ejemplo de artículo en revista de Ingeniería Industrial: http://revistas.ubiobio.cl/index.php/RI/article/view/56

[32] Ejemplo de Jornadas de Investigación en Educación Superior: http://jies.cse.udelar.edu.uy/

[33] El autor agrega -25 años después- una sección de "Métodos" aparte junto con otras interesantes aclaraciones; ver http://shermandorn.com/wordpress/?p=8519

redes sociales, cursos masivos en línea, textos, juegos con perfil educativo...) y dada la naturaleza de esa información (almacenada en bases de gran tamaño, con estructuras disímiles, con diferentes formas y permisos de acceso...) surgen dos nuevos paradigmas de recolección y análisis de datos: los denominados "Minería de Datos Educativos" (Educational Data Mining, EDM) y "Analítica del Aprendizaje" (o Learning Analytics, LeAn) que si bien difieren en sus objetivos, comparten muchas cosas. Ambos términos consideran como parte fundamental a la llamada *minería de datos*.

Esta disciplina tiene su origen fuera de la ciencia estadística; de hecho surge dentro de las ciencias informáticas y en particular dentro de la comunidad de administradores de bases de datos. Hacia fines de la década de 1980 aflora el interés por usar los sistemas de manejo o administración de base de datos (DBMS) con un fin diferente al usado hasta el momento (almacenar y recuperar información rápidamente con fines administrativos); emerge el enfoque de "apoyo a la toma de decisiones": obtener información pertinente en base a consultas o preguntas que surgen desde distintos interesados. Como lo anterior podía ser realizado solamente por usuarios con experticia en base de datos, surgen aplicaciones informáticas que facilitan dichas consultas para no expertos, conocidas como de minería de datos en el ámbito comercial[34]. La combinación de automatización de registros, arquitecturas paralelas y máquinas con mejor hardware ha generado la proliferación de bases de datos de gran tamaño, y con ello la necesidad de explotar comercialmente la información oculta en ellas; de ahí la popularidad del término ([Fri97]).

Entonces, ¿cuál es la relación de lo anterior con la educación? La gama de definiciones posibles para Minería de Datos Educativos y Analítica del Aprendizaje es amplia, sin embargo se pueden citar las siguientes:

EDM *Minería de Datos Educativos*: Paradigma creado para el diseño de modelos, tareas, métodos y algoritmos para extraer conocimiento de datos provenientes de sitios, aplicaciones o lo que sea relacionado con la educación ([PA14b])

LeAn *Analítica del Aprendizaje*: Es la medición, recolección y análisis de datos respecto a los alumnos y su contexto, con el propósito de comprender y mejorar el aprendizaje y el ambiente en el que éste se desarrolla, incluso en tiempo real ([SB12])

Las diferencias entre ambas disciplinas son sutiles y puede que algunas investigaciones de una disciplina utilicen características de la otra ([CMP+15]). En la Figura 2.3 se muestra la interacción entre estas disciplinas y otras ciencias.

Este nuevo enfoque comienza tímidamente sobre el final de la década de 1990[35], expandiéndose notablemente bien comenzado el siglo XXI en consonancia con los avances tecnológicos, la cada vez mayor disposición de estos elementos (computadoras personales, celulares, tabletas,...) y la consecuente abundancia de información durante todo el ciclo educativo. El EDM en particular busca identificar patrones y predecir el comportamiento y logros de los estudiantes, además de

[34]De hecho, en el ámbito académico surge el concepto de "Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases, KDD)", pero nunca llegó a tener la popularidad del término "Data Mining", a pesar que hablan de lo mismo.

[35]El trabajo de Baker ([BY09]) menciona incluso un trabajo pionero publicado en 1973.

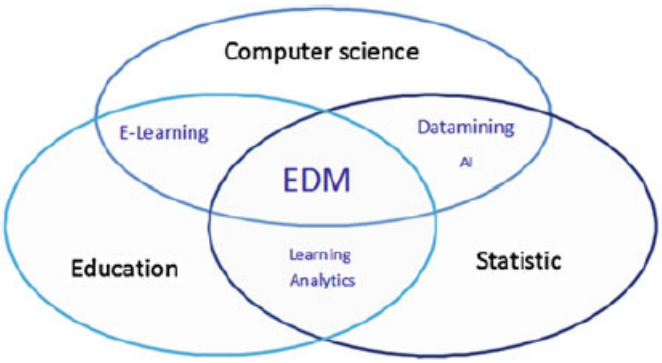

Figura 2.3: Diagrama de Venn explicando relaciones entre EDM y LeAn. Fuente: [PA14a], pp.6

investigar disciplinas y contenidos específicos, evaluaciones, funcionalidades educativas y aplicaciones utilizadas en educación ([RV07]). Estos paradigmas incorporan además una nueva escala temporal a los procesos de aprendizaje, ya que se permite observar prácticamente momento a momento que ocurre con cada estudiante, como por ejemplo la evolución en las capacidades de aprendizaje, cambios en la concentración y el estado emocional, además de poder devolver resultados prácticamente en tiempo real[36] ([CMP+15]).

El trabajo de Peña ([PA14b]) realiza una exhaustiva descripción de una muestra de trabajos contemporáneos (2010-2014) y de años anteriores (citando otros trabajos), identificando modelos, tareas, métodos y algoritmos en cada uno y agrupándolos en atributos específicos, mediante diferentes técnicas de minería de datos. Se muestra allí que la modelización y evaluación del rendimiento, comportamiento y aprendizaje de los estudiantes y la evaluación de funcionalidades bajo disciplinas específicas son las aplicaciones más comunes de este paradigma. Los datos provienen tanto de sistemas informáticos (p.ej. sistemas de administración del aprendizaje (LMS) como Moodle, o sistemas de tutoría inteligente) como de fuentes tradicionales (p.ej. actas de cursos). Además, la modelización predictiva se destaca sobre la descriptiva (utilizando herramientas probabilísticas, estadísticas o de minería de datos); las tareas asociadas son la clasificación y el agrupamiento (clustering); los métodos más usados son los bayesianos, árboles de decisión y regresión logística, implementados con algoritmos k-means, EM, J48 y naïve Bayes principalmente[37].

En ese mismo trabajo, Peña identifica las fortalezas, debilidades, oportunidades y amenazas del paradigma en su conjunto, destacando que el mismo está en su "adolescencia", que poco a poco se ha organizado en distintos eventos, publicaciones y conferencias a nivel mundial, presentando alternativas para la evaluación de la educación en su conjunto -en particular pensando en el enfoque pedagógico que busca centrar la educación en cada estudiante- y con mucho campo para seguir creciendo.

[36] Se han creado con este fin un sinnúmero de aplicaciones informáticas que sirven para ayudar a los docentes en el aula, taller o laboratorio, acompañando juicios humanos con evaluaciones y otras características de cada alumno.

[37] Algunos de estos métodos serán desarrollados en la Sección 3.1.

Si bien estas nuevas iniciativas están teniendo una preponderancia muy marcada en estos últimos años, desde mucho antes se ha investigado sobre dos temas que las universidades a lo largo y ancho del mundo han identificado como sus principales problemas: la deserción o desvinculación de estudiantes y el rendimiento académico de los mismos[38]. Se detallará cada uno de estos temas por separado para facilitar su comprensión.

2.3.3. Desvinculación o Deserción Universitaria

Uno de los principales problemas constatados a nivel terciario es el relacionado al abandono o deserción[39] de los alumnos. En modo general se define a la deserción como "el abandono prematuro de un programa de estudios antes de alcanzar el título o grado, considerando un tiempo suficientemente largo como para descartar la posibilidad de que el estudiante se reincorpore al mismo" (Himmel E., citado en [MLO05]).

Varios autores han trabajado este fenómeno desde un punto de vista teórico. Por nombrar algunos, Vincent Tinto diferencia el abandono por exclusión académica respecto a la denominada deserción voluntaria, Margarita Latiesa explica la deserción como un fenómeno multidimensional donde influyen tanto factores institucionales como también del contexto social e individual de cada estudiante, Pierre Bourdieu y Jean Claude Passeron atan el rendimiento académico y la deserción con características sociodemográficas como sexo, tipo de familia, origen social y/o cultural (todos citados en [FR13]).

A nivel empírico se han desarrollado innumerables investigaciones en este sentido. En distintas partes del mundo se han creado foros de debate, revistas especializadas en la temática, grupos de trabajo y un sinfín de cuestiones alrededor de este tema. Al ser la UdelaR una universidad diferente en su funcionamiento respecto a otras en el resto del mundo, en lo que sigue se mencionarán trabajos aplicados a su realidad.

Dependiendo de los datos trabajados y los análisis realizados, se han obtenido diferentes miradas del problema. Boado (en [Boa11]) investigó según las visiones anteriores, que ocurre en particular con la deserción en varias carreras, concluyendo que para las denominadas "tradicionales" -como lo son consideradas las distintas ramas de la Ingeniería- son "la desmotivación frente a los estudios realizados, la dificultad de trabajar y estudiar al mismo tiempo y los bajos rendimientos académicos", además de la gran distancia entre lo que los alumnos esperaban y lo que pudieron vivir dentro de la institución educativa, vinculados principalmente con la enseñanza y el aprendizaje (docentes, pruebas, exigencias).

Custodio por su parte retoma los resultados principales encontrados por Boado e intenta generar perfiles de deserción ([Cus09]). Primeramente caracteriza a una muestra representativa de desertores en función de variables de corte hipotetizadas en trabajos precedentes, para luego

[38]En FIng se viene reclamando el "potenciar la investigación en enseñanza", "profundizar los estudios sobre desafiliación y rezago" y "detectar a los estudiantes cuando dan indicios" desde hace tiempo [Enr14].

[39]Dejando de lado -por simplicidad- la corrección política de lado, se hablará de deserción, abandono o desvinculación indistintamente, como fue sugerido en la presentación de [Boa11].

agruparlos en tres grupos: de estudiantes *inversores*, *consumidores* y *atípicos*. Para los primeros, definidos por la autora como "tipicamente universitarios", los factores más incidentes en la deserción son bajos rendimientos y cambios en factores vocacionales. Por su parte, para el segundo grupo las dificultades institucionales y externas combinadas con su característica de ingreso extra edad inciden mayoritariamente. En el último caso, los estudiantes "atípicos" tienen mayor influencia del trabajo sobre su decisión de abandonar los estudios.

El trabajo de Fiori y Ramírez utiliza datos de censos estudiantiles y de registros académicos, concluyendo con postulados evidenciados en Tinto y Latiesa: desafiliación dependiente del nivel de avance (más marcado cerca del inicio de las carreras), e influencia de variables académicas (grado de avance, asistencia a clases) ([FR13]). Los tres trabajos mencionados realizan una pormenorizada descripción de los estudiantes que abandonan, sin diferenciar que ocurre en cada facultad, escuela o instituto.

Mirando por servicios de la UdelaR, los trabajos de Goyeneche, Urrestarazu y Zoppolo por una parte, y Barros y Míguez por otra aportan luz sobre este tema utilizando modelización para datos de series temporales: en el primero de ellos, se estima en función de los créditos obtenidos para la carrera de Contador Público entre 1990-98, que pasará con los *estudiantes activos*[40] de esas generaciones, separándolos en desertores y egresados y calculando mediante estimadores de funciones de densidad por núcleos qué pasará con cada uno de estos grupos. Concluyen que el 75 % de los activos abandonarán los estudios en un lapso promedio de 8,5 años ([GUZ01]). El segundo trabajo también observa qué sucede con créditos obtenidos por los Contadores Públicos entre 1998 y 2001, agregando en su análisis nuevos indicadores (como tasa de aprobación de asignaturas y velocidad de avance por alumno). Con esto, generan 4 grupos de estudiantes y estiman en base a los datos observados una matriz de transición de estados para luego simular qué ocurre a 12 años con una cadena de Markov[41] no homogénea. Concluyen entre otras cosas que la permanencia de los estudiantes en la universidad (más allá de su rendimiento) hace caer la probabilidad de abandono de sus estudios ([BM03]).

En el caso de la Facultad de Ingeniería, la UEFI observó la desvinculación de los estudiantes entre 1997 y 2009, siendo ésta un 48,2 % en promedio. Se concluye que los resultados obtenidos están en sintonía con lo encontrado por Boado: sobre la deserción inciden factores propios del estudiante (motivación, inteligencia, vocación, etc.), factores "endógenos" a la institución (masividad, horarios, esfuerzo necesario, etc.) y "exógenos" (características sociodemográficas, inserción laboral). Se encontró particularmente que las mujeres, los alumnos provenientes de liceos públicos y con edad al ingreso mayor a 22 años tenían mayor nivel de deserción que sus contrapartes ([UEF13]).

A nivel internacional, en el caso de grupos de trabajo y conferencias, no se puede dejar de mencionar a la Conferencia Latinoamericana sobre el Abandono en la Educación Superior (CLABES), que busca reunir a los distintos actores de la vida universitaria para determinar factores

[40] Definición en Glosario (pp.101).
[41] Una definición simplificada: https://economipedia.com/definiciones/cadena-de-markov.html

asociados, efectos individuales y sociales de la desvinculación y políticas a adoptar para lograr permanencia y egreso de los estudiantes en riesgo[42].

En otras partes del mundo, la mirada aportada por las nuevas disciplinas mencionadas ha dado diferentes frutos, aunque en proporciones mucho menores al rendimiento estudiantil como se verá más adelante. Trabajos como los de Bayer ([BOP+12]) y Al-Shargabi y Nursari ([ASN10]) muestran cómo la combinación de distintas técnicas matemático-estadísticas en función de la naturaleza disímil de los datos manejados, ayudan a descubrir información oculta[43] o a predecir la posibilidad de abandonar la universidad[44].

2.3.4. Rendimiento en la Universidad

El rendimiento, éxito académico o estudiantil es algo más complicado de definir, ya que existen infinidad de posibilidades. Es además uno de los temas más estudiados a nivel universitario[45]. Como mencionan Zimmermann y cols. ([ZBHB15]), cualquier medida creada para medir el éxito estudiantil será en realidad una variable *proxy* del verdadero éxito. Se han realizado a lo largo de los años diferentes estudios sobre este cometido, utilizando diferentes enfoques y técnicas, y con variados resultados ([PA14b]).

La noción de "éxito" o "fracaso" escolar puede estar centrada en distintos elementos, ya sea sobre el propio alumno o también en su entorno (instituciones, sistema educativo) o combinando total o parcialmente estas dimensiones. Destaca Rodríguez-Ayán que todos los trabajos presentan en común la multidimensionalidad y complejidad del concepto. La investigadora señala además que es vital distinguir los indicadores centrados en las instituciones (p.ej.: tasa de graduación) de los centrados en los alumnos (p.ej.: calificaciones, créditos acumulados durante un período) [RA07]. En el presente trabajo se pretende -como ya se mencionó- una mirada exclusivamente sobre los estudiantes.

Como indica Rodríguez-Ayán, los trabajos que miden rendimiento utilizan frecuentemente resultados a pruebas (por ejemplo: promedios de calificaciones, resultados en pruebas estandarizadas, rendimiento según alguna medida de posición -media, mediana-, etc.), aunque también pueden utilizarse medidas alternativas, como lo son la cantidad de créditos acumulados a cierto momento ([UEF09b],[BM03]), la aprobación de un umbral de asignaturas ([IM10]) o incluso la correlación entre créditos acumulados y puntajes en pruebas diagnósticas ([Enr15]).

Otros estudios buscan una descripción más completa de los estudiantes, no solamente en base a un solo atributo, tanto sea por su desempeño, su vinculación con las casas de estudios, o ambas. Ejemplos de ello -en la UdelaR- son los trabajos de Seoane, Míguez, nuevamente Rodríguez-Ayán

[42] https://revistas.utp.ac.pa/index.php/clabes

[43] Utilizando distintas técnicas de modelado y agrupación de datos descubren patrones de logro académico. Uno de los factores que incide en el mismo es la propensión a desertar de los alumnos estudiados.

[44] Concluyen que la performance e, indirectamente, la posibilidad de abandonar estudios, está relacionada con los hábitos sociales, particularmente con la comunicación frecuente entre pares.

[45] Ocurre que en muchos estudios tratan las problemáticas de deserción y rendimiento en el mismo trabajo; ver p.ej. [GUZ01], [BM03], [AOD13].

y Serna.

En el primero, la autora realizó, en base a información de actas de cursos, de exámenes y escolaridades junto con datos sociodemográficos de los estudiantes, un seguimiento a la generación ingresante en 2009 a Odontología durante 2 años (denominado "tramo inicial"). Concluye entre otras cosas que el resultado del primer examen rendido es un indicador del rendimiento posterior, que la desvinculación parece ser algo "esperable dentro del modelo de acceso a la UdelaR" y que las variables sociodemográficas que forman parte del Formulario Estadístico de la UdelaR no parecen tener potencia predictora de las trayectorias definidas. Además, habla de los estudiantes rezagados o en tránsito ajustado como aquellos que efectivamente constituyen la norma, ya que en una parte importante de los servicios de toda la universidad son, dentro de los vinculados, los mayoritarios, siendo excepcionales los que tienen un tránsito esperable según los distintos planes de estudio ([Seo15]).

Por su parte, la tesis de Míguez ([Mí08]) aporta una descripción somera de los estudiantes ingresantes a la Facultad de Ingeniería: los define con una orientación motivacional que, desde su interior, los impulsa a realizar una carrera universitaria, aunque con carencias en sus estrategias de aprendizaje para favorecer "un aprendizaje autorregulado y significativo". Esto concuerda en cierto modo con otro estudio realizado para alumnos de Ingeniería pero de una universidad privada uruguaya ([Pag11]). Míguez define así al estudiante "exitoso" como aquél que puede distanciarse del clima institucional, utilizando la idealización como mecanismo de autodefensa, revalorizándose a sí mismo y actuando de manera tenaz ante la adversidad. Además, en función del resultado de la HDI, puede separar a los ingresantes en dos grupos: aquellos que llegan a la suficiencia tienen en general buen rendimiento a posteriori, mientras que los de puntajes más bajos están en riesgo de fracaso académico.

El tercer trabajo propone, además de una completa síntesis de diferentes estudios empíricos sobre rendimiento y deserción académicos, la utilización de créditos como indicador de desempeño, y lo compara con uno tradicional -calificaciones promedio- obteniendo similares resultados, incluso utilizando modelización diferente (regresión lineal jerárquica (HLM), regresión logística, modelos de ecuaciones estructurales) [RA07].

A nivel agregado de la UdelaR, el trabajo de Serna y cols. investiga a fondo a la generación ingresante en 1995 para ocho carreras en distintos servicios de la UdelaR, observando que ocurre con el ingreso, la permanencia y el egreso de cada una de ellas ([SMN+05]). De lo anterior surge una clasificación en 18 grupos que son producto del cruce según modalidades de aprobación de asignaturas[46], escolaridad[47] y avance[48]. Junto con lo anterior, propone una batería de indicadores que servirán de base para el actual Sistema de Indicadores para la Evaluación Universitaria, SIEU ([DGP16]).

De todas maneras, aclara Rodríguez-Ayán que al momento falta "un cuerpo de investigaciones

[46]Por Curso, Examen o Curso y Examen
[47]Según si su escolaridad era menor o mayor a la mediana general
[48]Tres grupos definidos: Avanzados, Rezagados y Vulnerables

sobre el rendimiento académico estudiantil", con énfasis en la construcción y validación de modelos predictivos ([RA07, pp.16]). A continuación se presentan algunas investigaciones relacionadas con la predicción de los fenómenos mencionados.

2.3.5. Predicción de Rendimiento y Deserción Universitarios

Como ya se mencionó, la diversidad de definiciones, posturas y modelos utilizados hace que las formas de predicción del rendimiento y la deserción sean muy variadas.

En una primera instancia, y como ya se hizo en las Secciones 2.3.1 y 2.3.2, se puede hacer -siempre desde el punto de vista de las fuentes de información y en menor medida de lo que considera cada investigador- una primera división entre un enfoque más tradicional y otro más moderno.

Para el caso de los trabajos de corte tradicional, podemos citar como ejemplos a Ibarra ([IM10]), Rovira ([RPI17]), Arias ([AOD13]) y Míguez ([MLO05]). Respecto a los datos utilizados, los dos primeros tienen un enfoque transversal, mientras que los siguientes realizan estudios en base a datos longitudinales.

Si se observan las tareas realizadas, estas pueden ser de clasificación ([IM10], [RPI17]) o predicción de determinados valores ([MLO05], incluso [RPI17]) utilizando modelización diferente: el trabajo de Ibarra y Michalus ([IM10]) busca saber qué incidencia tiene diferentes tipos de variables en el rendimiento académico (medido como promedio de asignaturas aprobadas anualmente) mediante regresión logística; Míguez y cols. ([MLO05]) por su parte utiliza un modelo temporal de tendencia lineal para estimar, en base a los créditos acumulados por año, qué podría ocurrir con los alumnos estudiados en el futuro, particularmente con el objetivo de saber en cuanto tiempo egresaría el alumno en cuestión. Este modelo partía de ciertos supuestos fuertes[49], pero ha servido como una forma rápida y sencilla de comprender el fenómeno dentro de FIng.

El trabajo de Arias utiliza modelos de tiempo discreto bajo el paradigma de modelos competentes dentro del análisis de supervivencia[50] para identificar el momento exacto en que los eventos 'deserción' y 'egreso' ocurren para dos cohortes completas de alumnos universitarios, mediante funciones de riesgo específicas. Finalmente el trabajo de Rovira tiene dos objetivos: predecir calificaciones y deserción para estudiantes de grado mediante la creación de una aplicación para tutores de cursos en diferentes carreras; para ello utilizan una serie de modelos de aprendizaje automático -apoyada con visualización adecuada- para cumplir ambos planteos[51]: regresión logística, clasificador bayesiano ingenuo gaussiano, SVM, Random Forest y Boosting (versión AdaBoost).

Por su parte, para el enfoque moderno la variedad de disciplinas, tareas y técnicas utilizadas es amplia; dependiendo del problema y el origen de los datos se pueden configurar un sinfín de

[49]Por ejemplo linealidad en tasa de aprobación por año: si el alumno/a llegaba a 90 créditos en el primer año, egresaría en los 5 años que el plan estipula (ya que el mínimo exigido para el egreso es 450 créditos).

[50]Bajo el análisis de supervivencia estándar los individuos están expuestos a un solo evento, mientras que en la vida real hay varios eventos que compiten entre sí; con esto surge este paradigma de "modelos competentes".

[51]Curiosamente, este trabajo podría ser perfectamente mencionado como de Learning Analytics o EDM; sin embargo los autores no hacen este tipo de aclaraciones en ninguna parte del mismo.

combinaciones, como ya se mencionó en la Subsección 2.3.2 ([PA14b]).

El trabajo de Celis y cols. ([CMP+15]) es sin dudas uno de los más cercanos a la presente propuesta. Utiliza herramientas del paradigma Learning Analytics para construir un modelo (GLM-logit con ajuste *stepwise*, utilizando *validación cruzada* que predice la 'causal de eliminación' (o doble reprobación) por motivos académicos de estudiantes del primer año de un plan común en una facultad de ingeniería, utilizando información curricular hasta antes del inicio del segundo semestre. Generan así un modelo con alto poder predictivo, particularmente para aquellos alumnos que caen en la doble reprobación; además muestra que las mujeres exhiben mejor rendimiento académico, en consonancia con otros estudios similares.

El trabajo de Oñate por su parte busca construir un modelo de clasificación para identificar y predecir a estudiantes con bajo rendimiento académico[52], agrupando previamente a los estudiantes según perfiles socioeconómicos y resultados de una prueba de egreso del sistema preuniversitario ([OB16]). Los modelos utilizados son el clasificador bayesiano ingenuo y árboles de decisión (mediante el algoritmo C4.5), considerando como variable de interés una indicadora de bloqueo académico para distintos períodos[53]. En este trabajo se aprecia el uso de una prueba general denominada "Saber11", tanto para agrupar a los estudiantes según su desempeño como también para usar alguno de sus resultados como variables predictivas; esto es lo más cercano a una prueba de diagnóstico como la utilizada en este trabajo. El autor concluye que la predicción del bloqueo académico es factible si se considera tanto el historial académico de cada estudiante (prueba diagnóstica, información universitaria) junto con información socioeconómica.

A nivel extrarregional se citan los trabajos de Saarela y cols. ([SK15]) y Zimmermann y cols. ([ZBHB15]), en relación a este nuevo paradigma de análisis. El primero busca ajustar la currícula a estudiantes de ciencias informáticas mediante una triangulación multifase, evaluando los efectos de los denominados cursos principales en el rendimiento estudiantil con distintas técnicas entrelazadas (análisis de correlaciones, clustering robusto, modelización mediante redes neuronales) y proponiendo mejoras. Concluyen en líneas generales que las capacidades generales de aprendizaje -y no las habilidades computacionales aprendidas durante los cursos- son las que mejor predicen el éxito del estudiante. El segundo por su parte busca determinar qué información proveniente de estudiantes aspirantes a estudios de posgrado puede ser determinante para saber si el estudiante en cuestión podrá terminar la especialización a la que aspira. Descubrieron que, inesperadamente, las calificaciones promedio del tercer año son las que mejor predicen si un estudiante puede egresar de alguno de los posgrados ofrecidos.

[52] En este estudio le denominan "bloqueo académico", que es una condición administrativa que se impone cuando un/a estudiante no cumple con determinadas condiciones impuestas, en este caso de rendimiento (p.ej. reprobar una misma asignatura en un período determinado).

[53] Oñate se refiere a *matrículas*, que son cuotas que los alumnos deben pagar para continuar sus estudios en determinados períodos (pueden ser semestrales o anuales); la predicción se hace para las matrículas 2, 3 y 4 aunque no especifica más detalles.

Definiciones y datos utilizados

Las definiciones de *rendimiento* y *deserción* varían en función de las fuentes de datos disponibles. Además, como los nuevos enfoques de análisis pueden hacer más cortos los períodos de trabajo, muchas veces se estudian estos fenómenos para cursos puntuales, haciéndose énfasis en la falta de interacción humana en los grandes cursos en línea, necesaria para establecer con mayor claridad que ocurre con cada estudiante ([HZZA18]).

Variables predictoras En una recopilación realizada por Rodríguez-Ayán, la autora cita al rendimiento previo como clave para predecir el desempeño futuro; mientras más cercano en el tiempo sea ese rendimiento previo, mayor será el poder predictivo del modelo (p.ej: notas preuniversitarias o resultados de pruebas estandarizadas al ingreso)[54]. En el mismo trabajo, un metaanálisis realizado por Sirin descubre que en 58 artículos analizados se utilizaron como indicadores de rendimiento calificaciones en áreas específicas. Otros autores señalan los promedios por año (GPA), medidas de "inteligencia" (p.ej. la estimación del coeficiente intelectual) e incluso otras variables como determinantes, aunque en menor medida que las anteriores ([RA07]).

Variables a predecir Respecto a como definir el rendimiento *per sé* hay muchísimas formas, tanto sea utilizando variables continuas (p.ej.: puntajes en una escala) como no continuas, tanto cuantitativas (p.ej.: cantidad de cursos o exámenes aprobados en un cierto período) como cualitativas (p.ej.: llegar o no a determinado umbral). En lo anterior pesa mucho la técnica utilizada: de regresión para las primeras variables, de clasificación para las últimas. A continuación se muestran algunas definiciones utilizadas en la bibliografía consultada.

- *Dedicación o compromiso con el curso (Bajo,Alto)*: [HZZA18]
- *Indicadora de alcanzar o superar nivel medio de calificaciones*: [IM10]
- *Indicadora 'activo/no activo'*: [RPI17], [SMG+18]
- *Ranking de calificaciones (4 categorías)*: [RPI17]
- *Indicador de "causal de eliminación"*: [CMP+15]

Como se observa, varios de los trabajos mencionados utilizan variables de tipo dicotómico, dadas su fácil implementación e interpretación, además de su versatilidad a la hora de ser utilizada en modelos de naturaleza diferente. En base a lo anterior, y a la forma de funcionamiento que será comentada en la Subsección 4.1.2 y siguientes, se considerarán variables dependientes de tipo dicotómico, tanto para referirse a rendimiento académico como a desvinculación.

[54] No obstante, la varianza explicada tiene un techo de aproximadamente 33 % si se mira el rendimiento dentro del primer año universitario, datos coincidentes con [AG09]. Incluso en FIng se ha demostrado correlación significativa entre el puntaje de HDI y los créditos acumulados durante el primer año, pero ésta tiene porcentajes de varianza explicada entorno al 25 % ([UEF09b],[Enr14])

Herramientas de modelización

Ambos enfoques -tradicional y moderno- han utilizado y utilizan en la mayoría de los casos herramientas estadísticas y de aprendizaje automático, buscando predecir distintos fenómenos vinculados con el rendimiento estudiantil. Para el caso del enfoque moderno aparecen otras técnicas más relacionadas con la informática, como por ejemplo la minería de asociación o la minería de textos ([PA14b]).

Algunos ejemplos en la bibliografía consultada:

- Modelos lineales generalizados (en particular regresión logística): ([IM10], [RPI17], [CMP+15], [UEF09b])
- Árboles de clasificación o decisión: ([OB16], [HZZA18])
- Métodos de ensamble:
  - Boosting ([HZZA18], [RPI17])
  - Random Forests ([RPI17])
- SVM: [RPI17]
- Clasificadores bayesianos: ([OB16], [HZZA18], [RPI17])

Históricamente, y tal como lo señalaban Rodríguez-Ayán y Peña, la regresión logística ha sido de los métodos más populares, por su versatilidad y fácil interpretación. Con los cambios mencionados en la cantidad y calidad de datos disponibles, otras herramientas comienzan a tener más peso en la comunidad de investigadores educativos, como son las redes neuronales, el aprendizaje profundo y tantos otros ([RA07], [PA14b]).

Hasta el momento, ninguno de los trabajos mencionados ha apuntado a la predicción de rendimiento o desvinculación en el corto plazo utilizando información académica desde una herramienta de diagnóstico y otros datos de tipo sociodemográfico, como se propone este trabajo. El ajuste simultáneo de los modelos propuestos además de agregar en dicho proceso modelos de consenso es otro aporte inédito, al menos en este campo de conocimiento.

Capítulo 3

Metodología

En este capítulo se presenta una reseña de los modelos estadísticos utilizados -todos en el marco del *aprendizaje supervisado*-, las medidas de desempeño para realizar comparaciones entre modelos (Sección 3.2) y la implementación de lo anterior para predecir rendimiento y desvinculación en el corto plazo, mediante la creación y desarrollo de un paquete original en el software `R` (Sección 3.3), junto con otras consideraciones metodológicas. En el Glosario (pp.101) se amplían las definiciones que se describen con fuente *inclinada*.

3.1. Modelización estadística

Las técnicas elegidas fueron escogidas en base a diferentes criterios, desde haber sido usadas en antecedentes directos de este trabajo hasta la actual popularidad en variadas disciplinas del saber. Como algunas de ellas son más conocidas y presentes en el ámbito de la investigación educativa, se hará una introducción más breve (subsecciones 3.1.2, 3.1.3, 3.1.7) que para los métodos en los cuales se necesita un mayor desarrollo (como p.ej. la Subsección 3.1.5, en particular por su implementación informática). Buena parte de la modelización es extraída del libro de James y otros ([JWHT13]), salvo mención de otras fuentes.

3.1.1. Planteo del Problema

Sea $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$ una muestra aleatoria iid de vectores con $\mathbf{x}_i \in \mathbb{R}^p$ (es decir $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$), provenientes de una población con distribución multivariada $\mathbb{P}(\mathbf{X} = \mathbf{x})$, desconocida y etiquetas de pertenencia a una clase o grupo $y_i \in \{0, 1\}$ (dicotómicas, como ya se mencionó en la Sección 2.3.5), se busca predecir a qué clase o grupo pertenece un nuevo dato, $\mathbf{x}$.

Para predecir la variable de interés Y se utilizará un clasificador general h, una función $h : \mathbb{R}^p \to \{0, 1\}$ que buscará acercarse lo más posible al denominado *clasificador de Bayes*, que asigna a la nueva observación $\mathbf{x}$ para la clase con mayor probabilidad a posteriori, es decir:

$$h(\mathbf{x}) = \underset{k \in \{0,1\}}{\operatorname{argmax}} \mathbb{P}(Y = k | \mathbf{X} = \mathbf{x}) \tag{3.1}$$

Este clasificador minimiza -en promedio- al denominado "error de testeo"[1] ([BCM17, pp.48], [JWHT13, pp.37-38]).

Para definir dicho error, se partirá a la muestra original en dos submuestras: una de entrenamiento, $S_{ME} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_m, y_m)\}$ y otra de prueba o testeo $S_{MT} = \{(\mathbf{x}_{m+1}, y_{m+1}), (\mathbf{x}_{m+2}, y_{m+2}), ..., (\mathbf{x}_n, y_n)\}$, con $m < n$, utilizando muestreo aleatorio simple para decidir qué elemento va para cada conjunto. Se entrenarán los modelos que definirán un clasificador h utilizando S_{ME} y se verificará el error de clasificación sobre S_{MT}, definido como

$$\text{err}(h, S_{MT}) = \frac{1}{n-m} \sum_{i=m+1}^{n} \mathbb{I}_{\{h(\boldsymbol{x}_i) \neq y_i\}}, \tag{3.2}$$

siendo $\mathbb{I}_{\{condición\}}$ la función indicatriz que vale 1 si la condición es verdadera y 0 si es falsa. Como el error dependerá de la partición resultante, es necesario realizar este procedimiento varias veces para reducir la varianza del mismo.

Se usarán, para construir dicho clasificador, diferentes modelos utilizados en el aprendizaje estadístico, buscando en todos los casos predecir con la mayor exactitud (o el menor error posible) qué ocurre con la variable de interés (en este caso, la dicotómica $Y \in \{0, 1\}$) dadas las variables explicativas X_i.

Otro de los desafíos en este trabajo es comparar la predicción de los modelos utilizados, para decidir por alguno que se destaque en términos de performance predictiva. Como la medida de error en (3.2) hace referencia al error general (es decir, no separa el error de predicción por una u otra clase), y teniendo en cuenta que es fundamental para el trabajo realizado por la UEFI comprender sobre qué poblaciones se acierta o se erra más para tomar las decisiones correctas[2], se calcularán otras medidas de aciertos y errores comúnmente utilizadas con modelos predictivos. Todo esto será visto en detalle en la Sección 3.2.

3.1.2. Modelos Lineales Generalizados

Sean dos variables aleatorias (v.a.) X y Y, y suponiendo que Y es binaria (más precisamente una v.a. Bernoulli[3], que toma sólo dos valores: 0 y 1). ¿Cómo modelizar la relación entre $p(x) = \mathbf{P}(Y = 1|X = x)$ y X? Si se utiliza una regresión lineal del tipo $p(x) = \beta_0 + \beta_1 x$ se pueden generar valores -dependiendo de los que pueda tomar X- fuera del intervalo $[0, 1]$. Será necesario realizar una transformación que asegure que los valores obtenidos sean *realmente* probabilidades, es decir entre 0 y 1. Una posibilidad es utilizar la *función logística*,

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{3.3}$$

[1] También esto es equivalente al llamado "riesgo empírico", pero en este caso es necesario introducir una función de pérdida o ajuste del modelo a los datos $L(\mathbf{X}, \mathbf{y}, \beta)$.

[2] Según [CMP+15] los clasificadores tienden a predecir mejor a la *ausencia del atributo* pero con condiciones de ingreso diferentes a las encontradas en la UdelaR; en este trabajo se busca primero conocer las características de varias de estas cantidades empíricas para luego decidir cuál modelo es el que predice mejor.

[3] Las v.a de tipo Bernoulli tienen esperanza $E(X) = p$ y varianza $Var(X) = p(1 - p)$.

que ayuda a que los valores resultantes cumplan con la propiedad deseada. Manipulando (3.3) se llega a

$$\exp\left(\beta_0 + \beta_1 x\right) = \frac{p(x)}{1 - p(x)}, \tag{3.4}$$

cociente denominado "razón de momios" u *odds ratio*, que permiten comparar las chances de cumplirse uno u otro valor de Y (es decir, $Y = 0$ o $Y = 1$) dados valores determinados de X. Tomando el logaritmo en ambos lados de (3.4), se obtiene

$$\ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x \tag{3.5}$$

que es la transformación denominada *logit*, que permite linealizar distribuciones de tipo sigmoide como la propuesta en (3.3).

Regresión Logística

El modelo de regresión logística está enmarcado en los modelos lineales generalizados (GLM). Estos son utilizados para modelizar variables provenientes de la familia exponencial, de forma genérica $f(y, \theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\}$. Estas v.a. pueden ser tanto continuas (normal, integrantes de familia gamma -exponencial, weibull, beta-, etc.) como discretas (binomial, poisson, binomial negativa entre otras). En la práctica tienen un nivel de uso elevado dada su versatilidad: no requiren los rígidos supuestos de normalidad de los datos ni homocedasticidad (igualdad de varianza) de los residuos, aplicándose así a una mayor variedad de problemas, tanto de regresión como de clasificación.

Características Todos los modelos lineales generalizados constan de tres componentes:

- *Predictor(es) lineal(es)*: componente sistemático del modelo, representado por $\beta_0 + \beta_1 x$
- *Variable (aleatoria) dependiente proveniente de familia exponencial*: Y, que a su vez es representada por el parámetro $\mu = E(Y|X)$
- *Función de enlace (link)*: debe ser monótona y diferenciable $g : \mathbb{R} \to \mathbb{R}$ que relaciona la media con el predictor, es decir $g(\mu) = \beta_0 + \beta_1 x$

Así, si v.a. dependiente es de tipo Bernoulli, $\mu = E(Y|X = x) = \mathbf{P}(Y = 1|X = x) = p(x)$, con lo cual usando (3.5) se tiene

$$g(p(x)) = \beta_0 + \beta_1 x \overset{(3.5)}{=} \ln\left(\frac{p(x)}{1 - p(x)}\right),$$

de donde $g(p) = \ln\left(\frac{p}{1-p}\right)$.

De este modo, la función de verosimilitud[4] a maximizar (para hallar los parámetros β_j) es

$$L(\beta_0, \beta_1 | Y = y) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}, \tag{3.6}$$

con $p_i = p(x_i) = \mathbf{P}(Y|X = x_i)$. Utilizaremos como estimadores de β_0 y β_1 a aquellos valores que maximicen la verosimilitud[5]:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \underset{\boldsymbol{\beta_0}, \boldsymbol{\beta_1}}{\text{máx}} \ln(L(\beta_0, \beta_1 | Y = y)). \tag{3.7}$$

Todo lo anterior puede generalizarse bajo la denominada *regresión logística multivariada*, considerando (en lugar de una sola variable) p variables aleatorias independientes agrupadas en una matriz $\underset{n\times(p+1)}{\mathbf{X}}$ y con un vector de parámetros β compatible con $\mathbf{X}$. Por ejemplo, la ecuación (3.4) en este caso es, considerando $\ln(p(\mathbf{x})) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$:

$$p(\mathbf{x}) = \frac{\exp(\mathbf{x}^\mathbf{T}\beta)}{1 + \exp(\mathbf{x}^\mathbf{T}\beta)} \tag{3.8}$$

Bondad del ajuste Para estimar la *bondad de ajuste*, se utiliza como referencia un modelo llamado "saturado": aquel que se ajusta perfectamente a los datos, ya que es la solución a un sistema compatible determinado (es decir, tiene tantas incógnitas -parámetros- como ecuaciones -observaciones- diferentes). Se asume en este caso que $p(1) = \frac{\#\{y_i=1\}}{n}$ y $p(0) = 1 - p(1)$.

Para comparar modelos entre sí se realizan hipótesis utilizando como estadístico la 'devianza' de Wilks D, que es un cociente de verosimilitudes. Para la comparación del modelo saturado respecto al completo (el que incluye todas las variables -sin interacciones- para una cantidad de variables $p < n$):

$$D = -2 \ln \frac{L_{comp}(\theta|y)}{L_{sat}(\theta|y)}, \quad D \sim \chi^2_{n-(p+1)} \text{ (bajo } H_0),$$

con H_0 la hipótesis nula que indica en este caso buen ajuste del modelo a los datos, $\theta = (\beta_0, \beta_1, ..., \beta_p)$ el vector de parámetros y χ^2 la distribución chi-cuadrado con $n-(p+1)$ grados de libertad, partiendo del modelo multivariado expresado en (3.8).

La lectura de este estadístico se hace en la tabla ANOVA[6] del modelo, reportada en general en el software utilizado. Mientras más pequeña sea, mejor es el ajuste del modelo a los datos. Para el caso de modelos M_j anidados (es decir, con $M_1 \subset ... \subset M_{comp}$) la prueba de hipótesis es similar a la anterior, con un estadístico $\Lambda_{reduc} = \frac{L_{reduc}}{L_{comp}}$ con distribución asintótica bajo la hipótesis

[4] La función de verosimilitud L es igual a la distribución conjunta de probabilidad de una muestra aleatoria $X_1, .., X_n$ evaluada en los puntos $x_1, .., x_n$; simbolizada en general como $L(\theta|x)$, siendo θ el vector de parámetros poblacionales que caracterizan dicha distribución.

[5] En la práctica la maximización se realiza mediante métodos iterativos, p.ej.: Newton-Raphson.

[6] El Análisis de Varianzas (ANOVA) compara cuán diferentes son las medias de dos o más grupos definidos, observando variabilidad entre e intra grupos. Una tabla ANOVA organiza los distintos componentes de la varianza total en las posibles fuentes de la variabilidad, suma de cuadrados, grados de libertad, etc.

nula $-2\ln(\Lambda_{reduc}) \sim \chi^2_{p-q}$, siendo q la cantidad de variables del modelo reducido, que es aquel modelo ajustado más parsimonioso (o con menos variables independientes) en su fórmula, pero con un nivel de ajuste similar a modelos más complejos como el completo o el saturado. Con una lógica similar, se pueden comparar modelos reducidos entre sí. Para seleccionar el mejor modelo de un grupo de modelos anidados, se utiliza -entre otros- el estadístico AIC (*Akaike Information Criterion*)[7], que estima la pérdida relativa de información de un modelo, como balance entre el ajuste a los datos y su simplicidad:

$$AIC = 2k - 2\ln \hat{L}_{reduc}(\theta),$$

siendo k los parámetros estimados del modelo y $\hat{L}$ el valor de la función de verosimilitud para el modelo reducido a ser comparado.

Como se mencionó antes, el clasificador resultante será aquel que asignará la etiqueta que maximice la probabilidad que el individuo i-ésimo tome el valor k para $k \in \{0, 1\}$:

$$h_{GLM}(\mathbf{x}_i) = \underset{k\in\{0,1\}}{\arg\max}\, \mathbb{P}(Y = k|\mathbf{X} = \mathbf{x}_i) \tag{3.9}$$

3.1.3. Árboles y Métodos de Ensamble

Se presentan a continuación métodos basados en *árboles de decisión*, algoritmos de clasificación que segmentan el espacio de las variables independientes $\mathbf{X}$ en formas simples, y predicen el valor de clase Y como el valor más frecuente para esa región.

Classification and Regression Tree (CART)

Este método predictivo, propuesto por Breiman en 1984, divide primeramente al espacio en formas sencillas en base a reglas de verdadero/falso anidadas entre sí. Es muy utilizado por ser muy intuitivo, de fácil interpretación (particularmente para no expertos ya que simplifica relaciones complejas entre los datos), y cuenta con la ventaja de lidiar con datos faltantes sin inconvenientes.

Componentes Un gráfico con una típica estructura de árbol se muestra en la Figura 3.1. En el mismo se aprecian los principales elementos de los modelos con estructura arbórea que se describe a continuación:

Nodos Todas las observaciones parten de un *nodo raíz*, el nodo inicial del árbol. Utilizando un criterio de partición se divide al nodo raíz en dos "nodos hijos" de tal forma que estos hijos sean más homogéneos (o que presenten menor impureza) en relación a la variable a predecir, Y. Si se logra lo anterior, se vuelve a repetir el proceso (nodos hijos que se vuelven padres continuando

[7] En realidad, para comparar modelos, importa la *diferencia* entre los AIC para cada uno. Si ésta supera 10, será mucho mejor el modelo con el menor AIC.

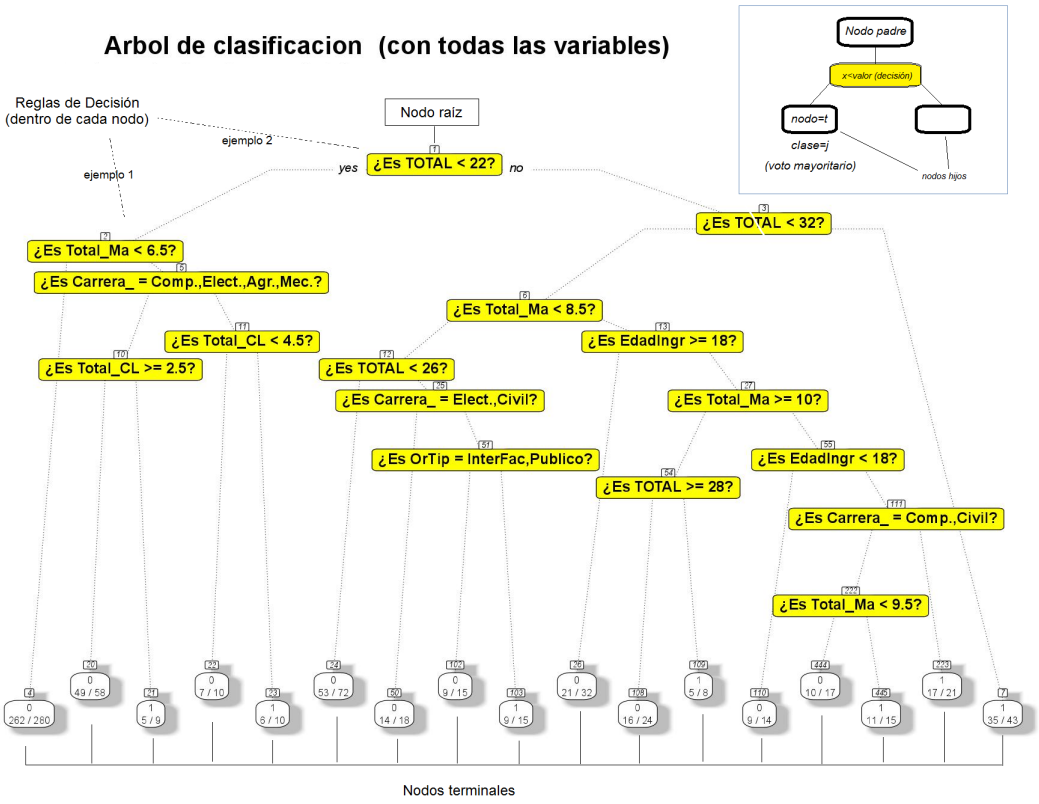

Figura 3.1: CART: ejemplo de gráfico de árbol, aplicado los datos disponibles de la generación 2008 (elaboración propia)

con las particiones) hasta llegar -en el caso extremo- a nodos terminales u hojas donde cada uno de ellos es una observación.

Criterios de decisión Son condiciones binarias (verdadero o falso) en base a un valor específico para una de las variables independientes que se utilizan en el modelo. En cada partición se forman dos nuevos nodos hijos, uno izquierdo y otro derecho, asociados a si la regla es verdadera o falsa, respectivamente.

Ramas Segmentos de los árboles que conectan los nodos entre sí. En general, si la condición es verdadera, habrá una rama con el nodo izquierdo, caso contrario será con el derecho. En cualquier caso se asigna un valor predicho de la variable Y a cada uno de los nodos, que será el valor de la clase más frecuente para todos los datos que cumplan esa condición. En general, mientras más largas sean las ramas del árbol en el gráfico, mejor será la discriminación entre clases.

Hojas Son simplemente los nodos terminales, como se observan en el extremo inferior de la Figura 3.1.

Construcción del árbol El método usado para construir el árbol requiere definir tres criterios:

- *Criterio de Partición (como subdividir)*: como dividir recursivamente, buscando la mejor regla que parta en dos al nodo padre, de la manera más homogénea posible
- *Criterio de Parada*: decidir cuando un nodo es terminal y finalizar el proceso
- *Criterio de Asignación (como "etiquetar")*: cómo asignar la etiqueta a cada hoja

Luego de que se logran los tres cometidos anteriores, hay que decidir dónde se poda al árbol y, eventualmente, con qué sub-árbol es mejor quedarse

Partición Para particionar correctamente, debe introducirse una función que indique la homogeneidad de los nodos hijos. Esta se conoce como *función de impureza*, simbolizada por ϕ. Sean p_k las proporciones de las observaciones de la muestra de entrenamiento que pertenecen a la clase k ($k \in \{0, 1\}$). La función de impureza se define como

$$\phi : \{p = (p_0, p_1) \in \mathbb{R}^2 : 0 \leq p_0, p_1 \leq 1 \ , p_0 + p_1 = 1\} \longrightarrow \mathbb{R} \tag{3.10}$$

Esta debe, para ser útil a estos propósitos, cumplir estas propiedades:

- Simetría
- Tener mínimos en la base canónica, en este caso $\mathbb{R}^2$ $(((1, 0), (0, 1)))$
- Tener un único máximo en $\left(\frac{1}{2}, \frac{1}{2}\right)$

Las siguientes funciones cumplen con las propiedades anteriores:

- Impureza de Gini: mide la varianza total sobre las 2 clases posibles:

$$\phi(p) = \sum_{k \in \{0,1\}} \hat{p}_k(1 - \hat{p}_k)$$

- Entropía (o entropía cruzada): similar al índice de impureza de Gini; calcula la entropía para cada nodo así:

$$\phi(p) = -\sum_{k \in \{0,1\}} \hat{p}_k \log \hat{p}_k$$

- Resustitución o tasa de error de clasificación, similar a (3.2)

Se define entonces la *impureza del nodo t* como

$$i(t) = \phi(p_0(t), p_1(t)), \tag{3.11}$$

siendo $N(\cdot)$ frecuencias absolutas -con $N(t)$ elementos de la muestra de entrenamiento S_{ME} en el nodo t y $N_k(t)$ elementos de la misma muestra pero pertenecientes a la k-ésima clase-, donde $p_k(t) = \frac{N_k(t)}{N(t)}$ es la proporción de las clases en el nodo t. Las funciones de impureza tomarán valores pequeños si el t-ésimo nodo es más puro (o menos heterogéneo) que el nodo padre correspondiente.

Es necesario conocer también la *variación de la impureza* del nodo t respecto de sus hijos t_L y t_R, luego de realizar la partición s:

$$\Delta i(t,s) = i(t) - p_L i(t_L) - p_R i(t_R), \tag{3.12}$$

siendo los subíndices L y R indicadores de particiones izquierda y derecha, respectivamente, tanto para los nodos hijos (t_L, t_R) como para las proporciones (p_L, p_R).

La idea en todos los casos es obtener, para todas las particiones posibles del nodo t, aquellas que verifiquen: $s^*(t) = \arg\max_{s \in S} (\Delta i(t,s))$.

Finalmente, la impureza total del árbol T se define como

$$I(T) = \sum_{t \in \tilde{T}} p(t) i(t), \tag{3.13}$$

siendo $\tilde{T}$ el conjunto de hojas del árbol T. Breiman demostró[8] que maximizar la diferencia de impureza en cada nodo equivale a minimizar la impureza global del árbol T.

Parada El criterio de parada es elegido por el usuario antes de comenzar a crecer el árbol, de forma tal de no obtener un árbol demasiado grande pero también que éste no sobreajuste a la muestra de entrenamiento. Hay dos criterios posibles:

- Umbral de impureza: elegir umbral a partir del cual un nodo es "puro"
- Mínimo de observaciones por nodo: decidir que no se particiona más allá de un nodo que contiene m observaciones (valor arbitrario)

Asignación En este caso, como la variable es cualitativa, se realiza un "voto mayoritario simple"; es decir se asigna el valor de Y que sea el de la clase mayoritaria en cada nodo terminal, T_0. Si $\mathbf{x} \in T_0$,

$$h_{CART}(\mathbf{x}) = \arg\max_{k \in \{0,1\}} p_k(T_0) \tag{3.14}$$

Poda del árbol Sea t un nodo del árbol T, una *rama* (proveniente de t) será el subárbol T_t de T que tiene como nodo raíz al nodo t. Podar el árbol consiste, en este contexto, en suprimir todos los nodos descendientes de t, exceptuando a t; el árbol que se obtiene se denota como $T - T_t$.

La poda se realiza utilizando el algoritmo de "mínimo costo-complejidad" para lograr el mejor árbol posible evitando el sobreajuste de los datos. La medida de costo complejidad se calcula utilizando

$$C_\alpha(T) = R(T) + \alpha |\tilde{T}|, \tag{3.15}$$

[8] En el libro Classification and Regression Trees (Breiman, L.,Friedman, J. H.,Olshen, R. A. and Stone, C. J. (1984)).

siendo R el costo de clasificación errónea, definido como

$$R(T) = \frac{\sum_{t\in\tilde{T}} \#\text{obs muestra } S_{ME} \text{ mal clasificadas en t}}{|S_{ME}|} = \sum_{t\in\tilde{T}} R(t),$$

siendo $\tilde{T}$ la complejidad del árbol, medida por la cantidad total de hojas. El parámetro α controla el balance entre la complejidad del subárbol T' y el ajuste de éste a los datos de entrenamiento: valores grandes de α penalizan aquellos árboles con muchas hojas, compensando el crecimiento de $R(T)$ en (3.15). De este modo, se compara un subconjunto de todos los posibles árboles en función del penalizador α y se retiene el mejor utilizando los métodos de particiones ya vistos[9].

Elección del mejor sub-árbol El dilema entre elegir árboles grandes pero con problemas de sobreajuste -y, eventualmente, de generalización de los resultados a la muestra de prueba- o árboles más pequeños pero con resultados sesgados, se soluciona con la siguiente estrategia ([JWHT13, pp.309], válida para muestras pequeñas):

1. Hacer crecer un gran árbol en base a particiones recursivas, terminando solamente si cada nodo terminal tiene menos de un mínimo prefijado de observaciones
2. Aplicar poda por mínimo costo-complejidad (3.15), para obtener una secuencia de subárboles como función del párámetro α
3. Utilizar validación cruzada para elegir un valor de α (se elige aquel α^* que minimiza el error de clasificación general)
4. Quedarse con el subárbol de 2. correspondiente al valor α^*

Ventajas y desventajas Si bien CART tiene ventajas destacables sobre otras técnicas, el algoritmo de construcción de estos árboles es de tipo voraz (greedy), es decir que utiliza la mejor partición en cada momento, con lo cual puede dejar de lado algunas otras variables que podrían llegar a ser importantes para explicar la variabilidad de los datos. Otro inconveniente de CART es la inestabilidad resultante ante pequeños cambios en las observaciones. Es necesario entonces presentar alternativas razonables, por ello se presentan tres métodos de agregación: por un lado el método de agregación de árboles remuestreados (bootstrap aggregating o *bagging*), luego el más popular en la actualidad, bosques aleatorios (o *random forest*) y finalmente *boosting*[10].

[9] En general, se separa a los datos en conjuntos de *entrenamiento-prueba* para datos de gran tamaño, mientras que se usa validación cruzada para conjuntos pequeños.

[10] En el caso de los métodos de *bagging* o *boosting* ambos pueden aplicarse a modelos generales de clasificación o regresión.

Bagging

Bagging es un método que realiza una operación "intuitiva" sobre los árboles originales: para estabilizarlos los agrega y los promedia. Esto está inspirado en p.ej. tener un conjunto de observaciones iid $Z_1, ..., Z_n$ muy inestable -o sea cada v.a. Z_i con varianza σ^2 muy alta- y que se decide promediar a estas observaciones. Como resultado, la varianza del estimador será $Var(\bar{Z}) = \frac{\sigma^2}{n}$, que es más pequeña que σ^2, cumpliendo el objetivo deseado.

Para lograr lo anterior se generan B conjuntos de entrenamiento utilizando la técnica del *bootstrap*[11], luego se entrenan los datos remuestreados en cada uno de los B árboles (a partir de los B conjuntos remuestreados) -sin podar en ningún caso- y finalmente se utiliza el voto mayoritario para decidir a que clase se asigna a la observación X. Para lograr un valor adecuado de B es necesario probar con valores altos (p.ej. superiores a 100).

Una peculiaridad de este método es que las observaciones no sorteadas en el bootstrap -que aproximadamente son, para muestras grandes, $\frac{1}{3}$ del total de datos de entrenamiento-, denominadas OOB (observaciones *out of bag*) se utilizan para probar la performance del modelo; esto es: se usan como muestra de testeo para medir el error de clasificación cometido por cada árbol. Este método tiene el respaldo de ser asintóticamente equivalente al denominado *leave-one-out cross-validation* (LOO-CV) que sortea una observación y la deja fuera del entrenamiento, para usarla como prueba del modelo, repitiendo el proceso n veces (tanto como datos haya). Este método surge de otra familia de técnicas de remuestreo denominadas *jackknife*, muy utilizadas en la estadística clásica.

Si bien este método es una mejora sustancial respecto a los árboles "clásicos", se pierde muchísimo en interpretabilidad del modelo, siendo éste el principal problema del bagging.

Random Forests (RF)

Los *random forests* surgen como respuesta a un problema que tienen los modelos bagging: la correlación alta entre muestras y la consecuente baja reducción en la varianza, a pesar de emplear muchos de estos árboles. Este método decorrelaciona a los árboles utilizando un principio sencillo: en cada partición se le prohíbe al algoritmo considerar siquiera una mayoría de las variables en existencia: solo una porción del total -elegida al azar- será tenida en cuenta. Un valor bastante común de este subconjunto es $m = \sqrt{p}$, siendo p la cantidad de variables predictoras (con ello, tan solo $\frac{p-m}{p}$ de las particiones considerarán -en caso de existir- un predictor "fuerte" dentro del conjunto de prueba, dándole más chance a otras variables a ser escogidas). [JWHT13, pp.320]. En este caso tampoco se poda a los árboles resultantes, se gana en tiempos menores de entrenamiento y se evita el sobreajuste.

Para el presente problema, el clasificador resultante de Random Forests será similar al de CART (3.14).

[11] Referirse al Glosario

Algoritmo 1 Seudocódigo: boosting para árboles de clasificación (AdaBoost)

Datos: datos $\mathbf{x}_i$, pesos iniciales $w_0(i) = \frac{1}{n}$, $i = 1, ..., n$, cantidad de iteraciones B
Resultado: clasificador mejorado, $h_{Bstg}(\mathbf{x}_i)$
para $b = 1, 2, ...B$ **hacer**

1. Ajustar clasificador $h_b(\mathbf{x}_i) = \{0, 1\}$ con pesos $w_b(i)$ sobre muestra entrenamiento S_{ME}
2. Calcular errores $e_b = \sum_{i=1}^{n} w_b(i) \cdot \mathbb{I}_{\{h_b(\mathbf{x}_i) \neq y_i\}}$ y $\alpha_b = \frac{1}{2}\left(\frac{1-e_b}{e_b}\right)$ (factor de actualización de pesos)
3. Actualizar nuevos pesos: $w_{b+1}(i) = \sum_{i=1}^{n} w_b(i) \cdot \exp\left(\alpha_b \cdot \mathbb{I}_{\{h_b(\mathbf{x}_i) \neq y_i\}}\right)$, normalizar pesos $(\sum_{i=1}^{n} w_{b+1}(i) = 1)$

fin
Clasificador final:

$$h_{Bstg}(\mathbf{x}_i) = \underset{k \in \{0,1\}}{\arg\max} \sum_{b=1}^{B} \alpha_b \cdot \mathbb{I}_{\{h_b(\mathbf{x}_i) = k\}} \tag{3.16}$$

Boosting

A diferencia del bagging o random forests, boosting hace crecer a los árboles de forma *secuencial*, usando información de los árboles ajustados previamente sobre versiones modificadas de los datos originales. Este es catalogado como un algoritmo de aprendizaje lento; este tipo de procedimiento en general presenta mejoras respecto a algoritmos más ágiles. Una aplicación bastante utilizada de esta familia es el AdaBoost o Boosting Adaptivo, *meta-algoritmo*, creado por Freund y Schapire en 1995.

Este clasificador se aplica repetidas veces sobre los datos de entrenamiento; en cada iteración el foco se centra en algunos elementos de ese conjunto utilizando pesos flexibles, $w_b(i)$, que cambiarán en cada iteración b: los individuos mal clasificados tendrán sus pesos incrementados, mientras que los que están correctamente clasificados tendrán pesos menores, para que el clasificador se vea forzado en la próxima iteración $(b + 1)$ a focalizarse en aquellos individuos mal clasificados. La diferencia entre los pesos actualizados será mayor en caso de que el error de clasificación sea bajo, ya que en este caso los errores tendrán más importancia y así se buscará minimizarlos. Una vez culminado el proceso se combinan todos los clasificadores en uno final, mucho más preciso que los intermedios. El método se presenta en versión de seudocódigo en el algoritmo 1 (tomado de [AGG13, pp.4], con algunas modificaciones menores).

Para el presente problema, el clasificador resultante de Boosting será el definido en la ecuación (3.16).

Otras consideraciones

Importancia de variables Las formas tradicionales de particiones recursivas (como p.ej. en CART) se basan en medidas empíricas de reducción de impureza, como es p.ej. el índice de impureza de Gini. Breiman fue claro al decir que, usando el índice de Gini para particionar, "la selección está sesgada a variables que tengan más valores (y así más particiones) posibles". Esto es una desventaja de estos algoritmos; investigaciones como la de Strobl ([Str08]) proponen alternativas tanto a criterios de partición como a formas de muestreo para solucionarlos, pero esto escapa a los objetivos de este trabajo.

Manejo de datos "perdidos" Para el caso de los bosques aleatorios, el algoritmo original no funciona con datos faltantes, aunque diferentes implementaciones proveen herramientas para imputar (ver Sección 3.3). Según Breiman[12], por la aleatoriedad y la cantidad de árboles ajustados la forma de imputación no debería afectar la precisión del algoritmo. El autor destaca dos formas de lidiar con datos perdidos: una (la "simple") es imputar por la mediana (para variables continuas) o la moda (para variables categóricas). La otra posibilidad es reemplazar los datos perdidos sólo en el conjunto de entrenamiento usado, imputando iterativamente comenzando con valores "simples", ajustando un árbol y nuevamente imputando hasta lograr valores estables[13].

3.1.4. Support Vector Machines (SVM)

Las Máquinas de Soporte Vectorial (o *Support Vector Machines*, SVM) provienen de una lógica un tanto diferente a lo visto anteriormente: la idea es colocar una "frontera separadora" -en general, hiperplanos- entre distintos grupos de datos utilizando aquellas observaciones que estén "cerca" de esa frontera, e ignorando al resto[14] ([JWHT13, pp.337-353]).

El problema en general se presenta explicando qué ocurre para variables de interés binarias y en tres casos particulares, en grado creciente de complejidad: el caso de separabilidad total entre clases, la existencia de un "margen blando" que permite cierto grado de error al clasificar y el caso de no separabilidad, en donde es necesario enviar a los datos a un espacio de dimensión mayor para encontrar hiperplanos que los separen. El *margen* en este contexto es la suma de las distancias del hiperplano hacia los datos más cercanos a un lado y al otro del mismo.

Los mencionados arriba son en definitiva problemas de optimización convexa, con ciertas peculiaridades en cada caso. Los más complejos son los de margen blando y el de no separabilidad.

- *Margen blando*: en este caso se introducen variables de holgura en el problema (marcadas como ξ_i), para que cierta cantidad de observaciones puedan estar del lado equivocado del

[12] Como consta en *Manual on Setting Up, Using and Understanding Random Forests*, https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf pp.8, versión 3.1 del algoritmo.

[13] Para el software `R` ambas formas de imputación están programadas en la función `rfInmpute()` disponible en el paquete `randomForest`.

[14] Se las denomina *máquinas* porque hacen la separación no lineal *automáticamente*, agrandando el espacio original de variables usando el kernel trick, descrito más adelante.

margen o del hiperplano separador.

- *No separabilidad entre clases*: en este caso se "envían" los datos a un espacio de mayor dimensión que el actual, utilizando una función de mapeo llamada *núcleo* (no necesariamente lineal) aplicando el denominado "truco de los núcleos" (kernel trick).

Es común que a los tres modelos que se verán en este capítulo se los ponga en el mismo rótulo de Máquinas de Soporte Vectorial, aunque sean tres enfoques con distintos grados de complejidad.

El Clasificador de Margen Maximal

Clasificación usando hiperplanos Un hiperplano $\{\boldsymbol{x}, \beta \in \mathbb{R}^p : \langle \beta, \boldsymbol{x} \rangle + \beta_0 = 0,\ \beta_0 \in \mathbb{R}\}$ es un subespacio vectorial que tiene dimensión $(p-1)$ y que divide en dos al espacio. Por ejemplo, suponiendo datos que sean separables linealmente, $\langle \beta, \boldsymbol{x} \rangle + \beta_0 < 0$ define aquellos elementos a un lado del hiperplano (negativos) y su opuesto $\langle \beta, \boldsymbol{x} \rangle + \beta_0 > 0$ a los que están del otro lado (positivos), con $\beta = (\beta_0, \beta_1, ..., \beta_p)$. Esto equivale a decir

$$\begin{aligned} \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} > 0, \quad \text{si } y_i = 1 \\ \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} < 0, \quad \text{si } y_i = -1, \end{aligned} \tag{3.17}$$

o lo que es lo mismo, alcanza con calcular el signo de $(\langle \beta, \boldsymbol{x} \rangle + \beta_0)$ para saber de qué lado del hiperplano se encuentra un punto cualquiera. De esta manera, una nueva observación estará bien clasificada si $y_i\,(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) > 0, \quad \forall\ i = 1, ..., n$, y así será sencillo construir un clasificador: se asignará una nueva observación a una de las clases dependiendo a qué lado del hiperplano se encuentra.

En el gráfico (a) de la Figura 3.2, se presentan tres posibles soluciones para un ejemplo de espacios linealmente separables en $\mathbb{R}^2$, todas válidas ya que ninguna toca a los puntos en cuestión. De este modo, queda claro que este problema no tiene solución única. Es necesario restringir las soluciones a una condición de optimalidad, que será la que propicie como resultado el gráfico (b) de dicha figura.

Margen maximal Una primera opción es elegir aquel hiperplano separador que presente el menor margen. Sean $\beta_0, ..., \beta_p$ los parámetros del hiperplano separador antedicho, este nuevo clasificador asignará a una nueva observación $\mathbf{x}$ a uno de los lados del hiperplano, según el $\text{sg}(h(\mathbf{x})) = \text{sg}(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$. Ahora es necesario determinar cual de todos los hiperplanos que cumplen con lo anterior es el mejor. Formalmente, es necesario encontrar la solución del siguiente problema de optimización, siendo M el margen -máximo- a encontrar:

$$\begin{cases} \underset{\beta_0, \beta_1, ..., \beta_p}{\text{máx}} & M(\beta_0, ..., \beta_p) = \frac{2}{\|\beta\|},\ \beta = (\beta_0, ..., \beta_p) \\ s.a. & y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) > M(\beta_0, ..., \beta_p), \quad \forall i = 1, ..., n \\ & \|\beta\| = 1, \beta_0 \in \mathbb{R} \end{cases} \tag{3.18}$$

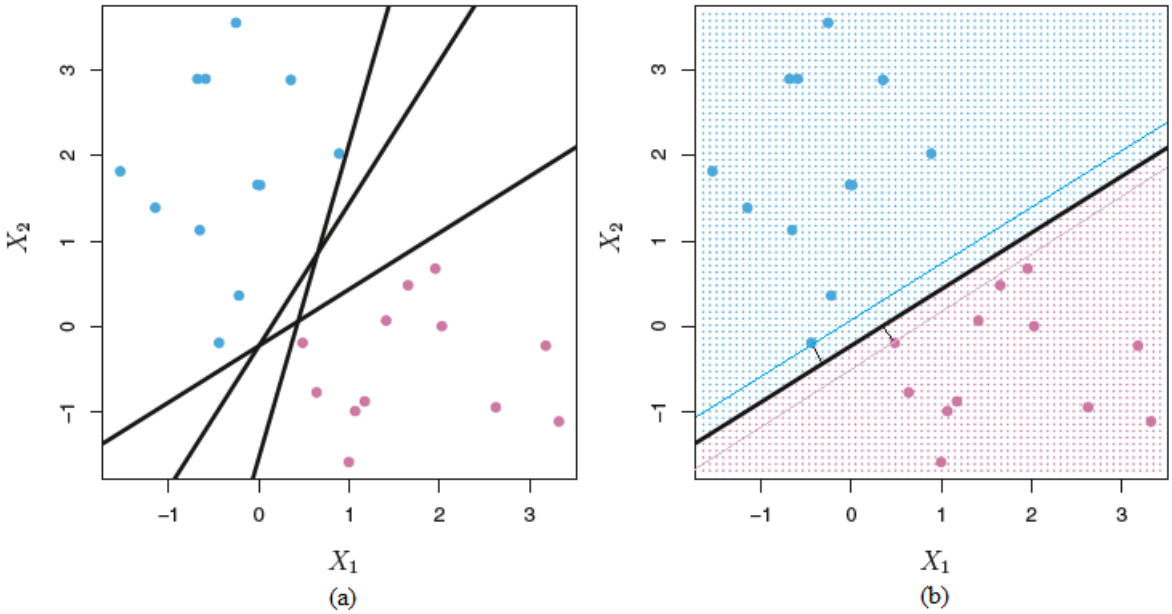

Figura 3.2: Problema con datos separables: (a) Distintos hiperplanos separadores; (b) Solución óptima para esos datos (modificado de [JWHT13, pp.340])

Utilizando herramientas de optimización como las relajaciones de Lagrange[15] y solucionando el problema dual resultante, se arriba al clasificador óptimo, que es aquel que maximiza el margen entre todos los hiperplanos posibles: $h(\mathbf{x}) = \text{sg}(\beta_0^* + \beta_1^* x_1 + \cdots + \beta_p^* x_p)$, siendo los $\beta_i^*,\ i = 0, 1, ..., p$ los valores hallados en la solución del problema (3.18). En definitiva, la solución queda expresada en términos de los datos de entrenamiento con multiplicadores de Lagrange diferentes de cero. Estos son los denominados "vectores de soporte".

El problema de este clasificador de "margen duro" es que no siempre se encuentran hiperplanos separadores, como por ejemplo cuando la frontera entre las clases es no lineal. Además, agregando pocas observaciones adicionales a los datos pueden hacer cambiar drásticamente la forma del clasificador de margen maximal con consecuente problema de sobreajuste. Para poder considerar estos casos es necesario generalizar el problema, introduciendo un "margen blando" que permita -hasta cierto umbral- que algunas observaciones queden del lado incorrecto (tanto del margen como del hiperplano). Esta generalización se denomina Clasificador de Soporte Vectorial.

El Clasificador de Soporte Vectorial

Estos nuevos clasificadores cumplen dos condiciones fundamentales: son más robustos respecto a cada observación y clasifican correctamente *a la mayoría* de las observaciones de entrenamiento. Estos clasificadores "de margen blando" son permisivos con algunas observaciones para que en

[15] Utilizando los *multiplicadores de Lagrange*: estrategia para encontrar extremos de una función f sujeta a restricciones g que definen la región factible del problema (qué valores posibles de las variables pueden ser escogidos para encontrar el extremo deseado), permitiendo pasar a resolver sistemas de ecuaciones para encontrar los puntos críticos de la función denominada Lagrangiano, $f_\lambda = f + \lambda g$. Bajo ciertas condiciones, esos puntos críticos son extremos del Lagrangiano y, a su vez, de f.

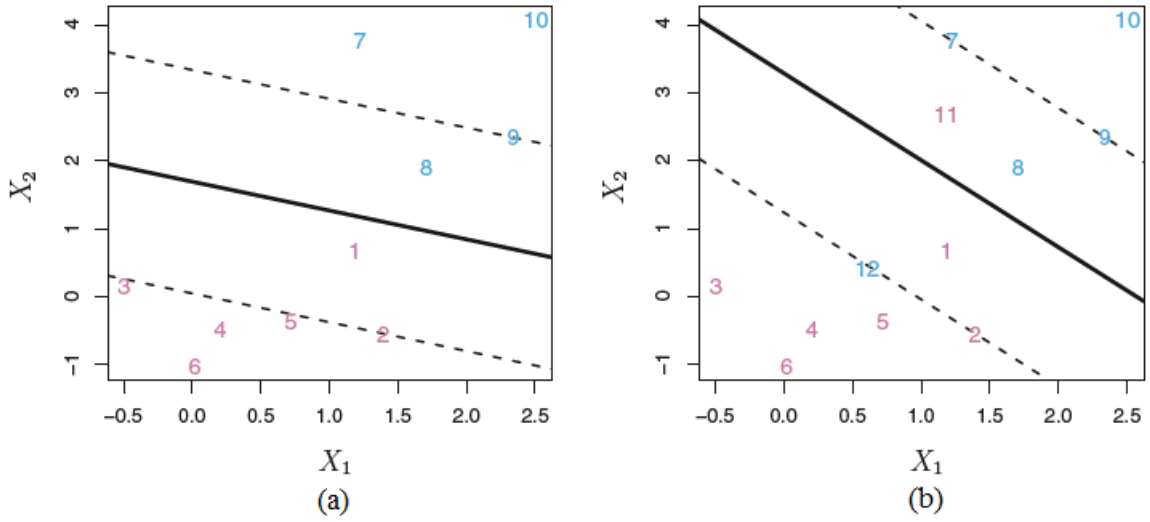

Figura 3.3: Clasificador de margen blando: ejemplo (fuente: [JWHT13, pp.346])

conjunto el resultado obtenido sea el mejor posible. El problema a resolver en este caso es:

$$\begin{cases} \underset{\beta_0,\beta_1,...,\beta_p}{\text{máx}} & M(\beta_0,...,\beta_p) \\ s.a. & y_i(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}) \geq M(\beta_0,...,\beta_p)(1-\xi_i), \quad \forall i=1,...,n \\ & \xi_i \geq 0 \; i=1,...,n; \quad \sum_{i=1}^{n}\xi_i \leq C \\ & \|\beta\|=1, \beta_0 \in \mathbb{R} \end{cases} \tag{3.19}$$

siendo ξ_j variables de holgura, $C \geq 0$ un parámetro de ajuste, M el ancho de margen. ¿Qué significan dichos valores?

- ξ_i: son las denominadas *variables de holgura*: dicen en qué lugar se sitúa cada observación, en relación al margen y al hiperplano. Si $\xi_i = 0$, la observación i está del lado correcto del margen, si $\xi_i > 0$ i estará del lado incorrecto del margen y si $\xi_i > 1$ i estará del lado incorrecto del hiperplano
- C: es el parámetro que determina la tolerancia permitida de violaciones a las condiciones de margen e hiperplano
  - C grande: más tolerancia a estar del lado incorrecto del margen, por esto el margen se ensancha[16]
  - C chico: menos tolerancia a estar del lado incorrecto, se achica el margen[17]

En la práctica, C es el parámetro que controla el posible sobreajuste del modelo (o balance entre sesgo y varianza), y es escogido por validación cruzada.

[16] Si $C > 0$, se permiten hasta C observaciones del lado incorrecto del hiperplano
[17] Caso extremo $C = 0$: optimización de margen maximal, ya que se cumple $\xi_i = 0, \;\; \forall i = 1,...,n$

El gráfico en la Figura 3.3 muestra que ocurre con un ejemplo de datos de entrenamiento "de juguete". El hiperplano -dibujado en línea continua negra- y los márgenes -en líneas punteadas- separan a las dos clases definidas en colores azul y violeta, respectivamente. En el gráfico (a) ninguna de las observaciones está incorrectamente clasificada[18], mientras que en el (b) se introducen dos observaciones adicionales (11 en violeta, 12 en azul), ambas erróneamente clasificadas por el modelo (ya que están del lado incorrecto del hiperplano).

El clasificador en este caso es, al igual que en el problema anterior, $h(\mathbf{x}) = \text{sg}(\beta_0^* + \beta_1^* x_1 + \cdots + \beta_p^* x_p)$, siendo los β_i^*, $i = 0, 1, ..., p$ en este caso los valores de solución correspondientes al problema (3.19).

Máquinas de Soporte Vectorial

Para problemas con fronteras no lineales entre las clases o grupos, se puede agrandar el espacio original de las variables adicionando términos polinomiales de mayor orden (ej: cuadráticos, cúbicos, de orden 4,...)[19], como es realizado en otras técnicas tradicionales de modelización. Sin embargo, este procedimiento puede dar lugar a un crecimiento desmedido de la cantidad de variables predictoras y, consecuentemente, a cálculos computacionales muy demandantes. Es necesario evitar el sobreajuste, además de aprovechar las propiedades de los SVM para obtener clasificadores eficientes.

La solución al problema de optimización en (3.19) requiere solamente de *productos internos*, definido éste entre dos puntos de esta manera:

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}, \tag{3.20}$$

Considerando esta definición, se puede reescribir el clasificador lineal de soporte vectorial así:

$$\text{sg}\left(h(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \right)$$

siendo α_i multiplicadores de Lagrange. Como además son pocos los valores de $\alpha_i \neq 0$ (solo para aquellos vectores de soporte), se puede reescribir lo anterior a un conjunto $\mathcal{S}$ más pequeño, donde solo están aquellos α_i diferentes de cero:

$$\text{sg}\left(h(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle x, x_i \rangle \right) \tag{3.21}$$

En definitiva, para generar al clasificador h y sus coeficientes β_i se necesitan productos internos como el de (3.21).

[18] Clase violeta: observaciones correctas 3, 4 ,5 y 6; sobre el margen: obs. 2; del lado incorrecto del margen obs. 1. Clase azul: observaciones correctas 7 y 10; sobre el margen: obs. 9; del lado incorrecto del margen obs. 8

[19] ¿Por qué esto da resultado? Porque en ese espacio "aumentado" de variables, las fronteras entre clases son lineales, aunque en el espacio original forman una frontera no lineal. Un ejemplo partiendo de un problema en $\mathbb{R}^2$, sería crear un nuevo espacio $[x_1, x_2, (x_1 \cdot x_2), x_1^2, x_2^2]$ usando las variables originales $[x_1, x_2]$.

Kernel	Fórmula	Parámetros	Usos
Lineal	$\mathbf{u}^\top\mathbf{v}$	(no tiene)	datos dispersos
Polinómico	$\gamma\left(\mathbf{u}^\top\mathbf{v}+c_0\right)^p$	γ, d, c_0	regresión/imágenes
RBF (Gaussiano)	$\exp\left\{-\gamma\cdot\lvert\mathbf{u}-\mathbf{v}\rvert^2\right\}$	γ	si no hay información
RBF (Laplace)	$\exp\left\{-\gamma\cdot\lvert\mathbf{u}-\mathbf{v}\rvert\right\}$	γ	a priori de los datos
Sigmoide	$\tanh\left\{\gamma\mathbf{u}^\top\mathbf{v}+c_0\right\}$	γ, c_0	Redes Neuronales

Tabla 3.1: Ejemplos de estimadores por núcleo para SVM

Kernels Lo anterior puede ser generalizado si se utilizan funciones llamadas *kernels*, simbolizadas por $K(x,y)$, que cuantifican la similitud entre dos observaciones. Supongamos que los datos pueden ser transformados por un mapa $\phi:\mathbb{R}^p\longrightarrow\mathscr{H}$, con $\mathscr{H}$ un espacio de mayor dimensión. La solución equivalente a (3.21) estará dada en términos del producto interno $\langle\phi(x_i),\phi(x_{i'})\rangle$, que se define como $K(x_i,x_{i'})$ que cumple las propiedades de kernel[20]. Esto equivale a plantear un hiperplano en un espacio de mayor dimensión.

Sustituyendo a $\langle x,x_i\rangle$ por $K(x,x_i)$ en (3.21) y considerando al clasificador h, el mismo determinará -como los anteriores- para qué lado del hiperplano irá la nueva observación $\mathbf{x}$ en función del signo obtenido:

$$h(x)=\text{sg}\left(\beta_0+\sum_{i\in\mathcal{S}}\alpha_iK(x,x_i)\right) \tag{3.22}$$

Sin perder generalidad, y suponiendo a $\mathbf{u},\mathbf{v}$ vectores reales, se resumen los principales estimadores por núcleo en la Tabla 3.1.

Respecto al clasificador utilizado en el presente trabajo, para hacerlo compatible con los restantes -esto es, que indiquen con el valor 1 la presencia del atributo que se está midiendo- se hará una transformación sencilla de (3.22), valiendo 1 si $h(x)=1$ y cero en cualquier otro caso:

$$h_{SVM}(x)=\mathbb{I}\{\text{sg}(\beta_0+\sum_{i\in\mathcal{S}}\alpha_iK(x,x_i))>0\} \tag{3.23}$$

3.1.5. Clasificador Bayesiano

El clasificador bayesiano que se introduce en esta subsección es en realidad una versión modificada del mismo, ya que no se asumirá ninguna distribución de probabilidad para las subpoblaciones determinadas por cada clase. A raíz de lo anterior, es necesario resolver dos problemas: por un lado la estimación de la distribución de las variables independientes, y por otro cómo estimar las proporciones o probabilidades a priori para las dos clases que serán definidas para cada variable Y en el Capítulo 4. Se introducen así los estimadores por la estimación de densidades por núcleo, complementado esto con el algoritmo EM para estimar las proporciones buscadas.

[20]El denominado *teorema de Mercer* asegura que, dada una función kernel definida positiva, existirá un espacio de Hilbert (o de características) donde el mismo es igual al producto interno de las funciones ϕ definidas en ese espacio.

Algoritmo 2 Seudocódigo: cálculo del clasificador discriminante bayesiano (CDB)

Datos: datos de entrenamiento $\mathbf{X}$

Resultado: Decisión h

1. Escoger función kernel de (3.2), determinar ancho de banda $\hat{b}$
2. Estimar por separado las densidades $\hat{f}_k(x)$, $k = 0, 1$ por (KDE)
3. Dadas densidades en 2., estimar "probabilidades mezcla", $\hat{\alpha_k}$ con EM mediante algoritmo 4
4. Dados $\hat{\alpha_k}$ y $\hat{f}_k(x)$ calcular el cociente de mezcla de densidades

$$\mathbb{P}(k|\mathbf{x}) = \frac{\hat{\alpha_k} \cdot \hat{f}_k(\mathbf{x})}{\sum_{k\in\{0,1\}} \hat{\alpha_k} \cdot \hat{f}_k(\mathbf{x})} \tag{3.24}$$

5. Con 4. construir el *clasificador discriminante bayesiano* (CDB):

$$h(\mathbf{x}) = \underset{k\in\{0,1\}}{\arg\max}\, \mathbb{P}(k|\mathbf{x}) \tag{3.25}$$

Estimación no paramétrica de Funciones de Probabilidad

Los estimadores no paramétricos son técnicas estadísticas que no requieren asumir ninguna forma funcional para estimar los objetos de interés. Han aumentado su notoriedad en los últimos años debido tanto a los avances teóricos como también la gran disponibilidad de librerías en distintos programas informáticos.

En este apartado el centro se hará sobre los denominados estimadores por núcleo (kernel estimation)[21]. Todos los kernels son en realidad funciones de suavizado, que para cada observación realizan una ponderación en base a una cierta regla, dependiente ésta de la distancia entre observaciones. Algunos ejemplos son el histograma (que divide al recorrido de los datos en "bandas" de igual longitud y pondera por igual a cada observación que cae dentro de cada banda) y el denominado "núcleo caja" (que centra cada uno de las ventanas usadas en cada uno de los datos, pero utilizando funciones discontinuas)[22]. Se presentará en particular el denominado método de estimación por núcleos con ancho de banda fijo (KMWFB, según notación usada por Elamin [Ela13]), con el cual se estiman funciones de probabilidad suavizando directamente a las variables a trabajar[23].

Estimación de densidad por núcleos (Kernel) para variables de tipo numérico Para introducirnos en tema, será necesario diferenciar la notación para variables continuas (tendrán un supraíndice c, como p.ej. X^c, x^c) y para las discretas (con supraíndice d: X^d, x^d). De este modo, sea $X_i^c, i = 1, ..., n$ una variable continua extraída con una muestra iid de una población cuya

[21] Suele haber confusión en esto: si bien las funciones kernel tanto de este apartado como las utilizadas en SVM *transforman* a los datos, cada una lo hace de modo diferente; por esto las de SVM se simbolizaron $K(\cdot)$ y las de este apartado con $k(\cdot)$.

[22] Un ejemplo propuesto explicando las diferencias en detalle se encuentra en http://www.mvstat.net/tduong/.

[23] Existe una contraparte con anchos de banda flexibles, similar al método de los vecinos más cercanos, que es más complejo y escapa al objetivo de este trabajo.

distribución de probabilidad $f(x^c)$ es desconocida. El objetivo es estimar la densidad de X^c.

Para solucionar los problemas de dependencia de los extremos de las bandas y para suavizar las distribuciones de los histogramas, Rosenblatt sugiere[24] el denominado método de estimación de densidad por núcleos (*Kernel Density Estimation (KDE)*). A diferencia del histograma, el KDE pondera a las observaciones en un intervalo de radio h alrededor de x^c.

De este modo, x^c es el punto de estimación o suavizado, los puntos X_i^c son ponderados en relación a x^c en un intervalo fijo denominado *ancho de banda* o ancho de ventana, y los pesos que la función kernel utiliza para ponderar a las observaciones son denominados pesos locales. Dependiendo qué función kernel se utilice, cambiará la importancia dada a cada punto alrededor de x^c y con ello los pesos correspondientes[25]. En la Tabla 3.2 se presentan algunas funciones KDE usadas habitualmente.

Nombre	Fórmula	Características
Uniforme o "ingenuo"	$\frac{1}{2}\cdot\mathbb{I}_{\{\lvert z\rvert<1\}}$	ventana rectangular
Triangular	$(1-\lvert z\rvert)\cdot\mathbb{I}_{\{\lvert z\rvert<1\}}$	ventana triangular
Epanechnikov	$\frac{3}{4}(1-z^2)\cdot\mathbb{I}_{\{\lvert z\rvert<1\}}$	ventana parabólica
Normal	$\frac{1}{\sqrt{2\pi}}\cdot\exp\{-\frac{z^2}{2}\}$	ventana "gaussiana"

Tabla 3.2: Funciones KDE para variables continuas

Notación Para una función univariada de una variable aleatoria continua, $f(x^c)$, el estimador kernel de ancho de banda fijo en el valor x^c usando los elementos de la muestra $\{X^c\}_{i=1}^n$ es el promedio muestral de los pesos locales, para todas las observaciones en la muestra considerada:

$$\hat{f}(x^c) = \frac{1}{n\hat{b}}\sum_{i=1}^{n} k\left(\frac{X_i^c - x^c}{\hat{b}}\right) \tag{3.26}$$

$$= \frac{1}{n\hat{b}}\sum_{i=1}^{n} k(z_i) \tag{3.27}$$

siendo $b \geq 0$ el ancho de banda de la KDE; este valor se mantiene constante sin importar donde esté centrado el punto x^c.

Propiedades Todas las KDE anteriores cumplen con ser simétricas entorno a x^c y además satisfacen las siguientes condiciones:

1. $\int k(z)dz = 1$ (integra a 1)
2. $\int z^s k(z)dz = 0$ (para asegurar simetría)

[24] Rosenblatt, M.: "Remarks on Some Nonparametric Estimates of a Density Function", The Annals of Mathematical Statistics 27(3), pp.832-837, 1956. Disponible en https://projecteuclid.org/euclid.aoms/1177728190

[25] La regla general es: mientras más cercana sea la observación a x^c mayor peso se le dará a la misma. Por ejemplo el KDE uniforme pondera a cada observación con el valor $\frac{1}{2}$ dentro del intervalo $[x^c-1, x^c+1)$ y le asigna cero al resto.

3. $\int z^r k(z)dz = \tau_r \neq 0$ (KDE de orden alto o mayor, $r > 2$)

Varias investigaciones en Elamin ([Ela13]) sugieren que los kernels de orden alto son más adecuados para estimar funciones de densidad a partir de muestras pequeñas.

Estimación del ancho de banda Como se apreció anteriormente, el ancho de banda b es un parámetro clave en la estimación de funciones de densidad por kernels, tanto o más importante que la función $k(\cdot)$ que se pueda escoger. Una decisión errónea en estos influenciará en la precisión de estimación (mediante los errores estándar y la velocidad de convergencia). Es importante considerar a la vez el costo computacional de estas estimaciones, en particular según los métodos usados, el tamaño de muestra considerado y la cantidad de variables utilizadas.

Respecto a los métodos, se pueden considerar dos clases: aquellos que aproximan $\hat{b}$ al valor teórico del ancho de banda, b, y por otro lado los métodos "basados en datos" (*data-driven methods*), que estiman los anchos de banda a través de la optimización de una función objetivo que realiza un balance entre exactitud (sesgo) y precisión (varianza) del estimador kernel.

Aproximación al ancho de banda teórico El sesgo y la varianza del estimador en (3.26) es:

$$Sesgo(\hat{f}(x^c)) \approx \frac{b^2}{2} f^{(\prime\prime)}(x^c) \int z^2 k(z)dz \tag{3.28}$$

siendo $f^{(\prime\prime)}(x) = \frac{\partial^2 f}{\partial x^2}$ la derivada segunda de f y

$$Var(\hat{f}(x^c)) \approx \frac{f(x^c)}{nb} \int k^2(z)dz \tag{3.29}$$

Cuando las condiciones de consistencia mencionadas en la Sección 3.1.5 se cumplen, el sesgo en la estimación anterior desaparece. El ancho de banda óptimo, $\hat{b}_{opt}$, es aquel valor que minimiza el error cuadrático medio integrado (IMSE):

$$IMSE(\hat{f}) = \int E\left[\hat{f}(x^c) - f(x^c)\right]^2 dx \tag{3.30}$$

que da como resultado:

$$\hat{b}_{opt} = \left\{ \frac{k^2(z)dz}{\left[\int z^2 k(z)dz\right]^2 \cdot \left[\int f^{(\prime\prime)}(x)dx\right]^2} \right\}^{-\frac{1}{5}} \cdot n^{-\frac{1}{5}} = c_0 \cdot n^{-\frac{1}{5}}. \tag{3.31}$$

Entonces, el valor óptimo del ancho de banda es en realidad una función de la derivada segunda de la –desconocida– densidad verdadera, $f(x^c)$. Es necesario hacer supuestos sobre esa función desconocida para poder avanzar, lo cual hace que los métodos que aproximan al ancho de banda teórico sean semiparamétricos. Algunos de los métodos más utilizados son:

- Tanteo o selección gráfica: es la más sencilla de las implementaciones, pero válido solamente para muestras pequeñas y/o pocas variables (hasta dos en general). Se toman varias estimaciones de b y el usuario elige la mejor "a ojo"
- Método *plug-in*: se asume una cierta distribución de la variable X^c y luego utiliza la fórmula (3.31) para obtener un "valor inicial" de b
- Regla general (*rule of thumb*): método más popular (y más antiguo) para estimar el ancho de banda, creado por Silverman en 1986. Usa una densidad normal para aproximar $\int f^{(\prime\prime)}(x)dx$ en (3.31). Si la densidad verdadera es unimodal, aproximadamente simétrica y sin colas pesadas, este estimador funcionará mejor que los anteriores, que pueden usarse como valores iniciales. Es particularmente útil si los cálculos son complejos, aunque puede suavizar "demasiado" a los datos en algún caso.

Métodos basados en datos Estos métodos tienen como base la validación cruzada, estimando anchos de banda a través de la optimización de una función objetivo que balancea sesgo y varianza del estimador. Esta función puede tener muchas formas; entre las más usuales se destacan las de error cuadrático integrado (ISE), el promedio del error cuadrático integrado (MISE), y versiones ponderadas (WIMSE) y asintóticas (AIMSE) de la función IMSE mencionada en (3.30).

- Validación cruzada mediante Mínimos Cuadrados (LSCV): este método es recomendado si el modelo no tiene muchas variables y/o el tamaño de muestra no es muy grande, además de ser adecuado para la estimación de densidades en variables mixtas. Los anchos de banda estimados convergen a los valores teóricos, y no tiene el problema de las "colas pesadas" como su alternativa, el MLCV
- Validación cruzada mediante Máxima Verosimilitud (MLCV): en este caso se halla $\hat{b}_{opt}$ mediante la maximización de una función objetivo basada en la divergencia de Kullback-Leibler[26]

En el presente trabajo se utilizan mayoritariamente los métodos basados en datos. En el Apéndice 3 se muestra los distintos métodos de estimación de ancho de banda para las diferentes funciones que utiliza `asm2()`, la función de "ajuste simultáneo de modelos" introducida en la Subsección 3.3.3.

Variables de tipo categórico

Estimador de cuantía univariada Este estimador simplemente separa los datos en celdas para luego estimar la función de cuantía, al mejor estilo del histograma en el caso continuo. Sea $X_i^d,\ i = 1, ..., n$ una variable discreta extraída con una muestra iid de una población con cuantía $p(x^d)$ desconocida, con recorrido finito en $\{0, 1, ..., c-1\}$. El estimador se define como:

[26] La (*pseudo*)distancia o divergencia de Kullback-Leibler también es llamada entropía relativa; mide la divergencia entre dos distribuciones de probabilidad.

$$\hat{p}(x^d) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i^d = x^d\}}. \tag{3.32}$$

Este método produce resultados insesgados y asintóticamente coherentes si la función de cuantía tiene un recorrido con pocos valores posibles, aunque necesita de un tamaño de muestra grande para funcionar adecuadamente. Por el contrario, si la variable discreta toma muchos valores su precisión cae rápidamente. Si hay más celdas que observaciones, el método es poco eficiente. Tampoco sirve para realizar estimaciones multivariadas de probabilidad ([Ela13, pp.25]).

Estimador de Aitchinson-Aitken La extensión del método KMWFB para estimar funciones de cuantía como (3.32) fue propuesta por Aitchison y Aitken (1976), denominado de ahora en más (kernel) AAk:

$$\hat{p}(x^d) = \frac{1}{n}\sum_{i=1}^{n} l(X_i^d, x^d, \lambda), \tag{3.33}$$

con $\lambda \in [0, 1]$ un parámetro de suavizado que depende de c (cantidad de categorías diferentes que puede tomar la variable X) y un kernel $l(\cdot)$. Se puede determinar -como en Chi-Hang y otros ([CHP15])- que el sesgo y la varianza de (3.33) son respectivamente:

$$Sesgo(\hat{p}(x^d)) = \lambda \frac{1 - c \cdot \hat{p}(x^d)}{c - 1} \tag{3.34}$$

$$Var(\hat{p}(x^d)) = \lambda \frac{\hat{p}(x^d)[1 - \hat{p}(x^d)]}{n} \cdot \left(1 - \frac{\lambda c}{c - 1}\right)^2. \tag{3.35}$$

Si bien este estimador introduce sesgo, con un adecuado tamaño de muestra se reduce significativamente la varianza, y así el error cuadrático medio del mismo.

Las propiedades de $l(X_i^d, x^d, \lambda)$ para asegurar consistencia son (asumiendo $\lambda \in [0, 1]$):

1. $l(X_i^d, x^d, \lambda) \geq 0$
2. $\sum_{x^d=0}^{c-1} l(X_i^d, x^d, \lambda) = 1, \ i = 1, ..., n$
3. Propiedades asintóticas:
   - *a)* $\lambda \underset{n\to\infty}{\to} 0$, para asegurarse que $\hat{p}(x^d) \xrightarrow{p} p(x^d)$
   - *b)* $\frac{\lambda}{\sqrt{n}} \underset{n\to\infty}{\longrightarrow} 0$ y $l(X_i^d, x^d, \lambda)$ tiene derivada primera continua en 0 ($l'(X_i^d, 0, \lambda)$), para asegurar distribución asintótica Normal estándar

En esencia, la estructura de los estimadores para el caso de variables nominales y ordinales es la misma, cambiando solo la forma de ponderar la distancia entre pares (X_i^d, x^d): mientras que para las variables nominales se obtienen solamente dos pesos diferentes -uno para $X_i^d - x^d = 0$ y

el otro para $X_i^d - x^d \neq 0$-, para las ordinales estos pesos serán el total de categorías c, graduados por la distancia $|X_i^d - x^d|$.

A continuación se muestran dos ejemplos, un estimador kernel para datos "sin orden" (nominales) y otro para datos "con orden" (ordinales), tal como son mostrados en [Ela13, pp.26].

Nombre	Fórmula	Características
Aitchison-Aitken	$\begin{cases} 1-\lambda, & si\ X_i^d = x^d \\ \frac{\lambda}{c-1} & si\ X_i^d \neq x^d \end{cases}$	Para variables nominales
Wang-van Ryzin	$\begin{cases} 1-\lambda, & si\ X_i^d = x^d \\ \frac{1}{2}(1-\lambda)\lambda^{\lvert X_i - x\rvert} & si\ X_i^d \neq x^d \end{cases}$	Para variables ordinales

Tabla 3.3: Kernels: datos categóricos (nominales y ordinales)

Caso multivariado con variables de ambos tipos Es particularmente importante introducir este apartado debido a la necesidad de lidiar con este tipo de variables en este trabajo.

Estimador kernel multivariado para variables de tipo mixto Para estimar kernels en el caso multivariado, se utiliza el producto de kernels univariados. Por un lado, para el caso de variables continuas:

$$\hat{f}(\mathsf{x}^c) = \hat{f}(x_1^c, \ldots, x_q^c) = \frac{1}{n\hat{b}_1 \cdots \hat{b}_q} \sum_{i=1}^{n} \prod_{s=1}^{q} k\left(\frac{X_{is}^c - x_{is}^c}{\hat{b}_s}\right) \tag{3.36}$$

y por otro lado para variables categóricas o discretas:

$$\hat{p}(\mathsf{x}^d) = \hat{p}(x_1^d, \ldots, x_p^d) = \frac{1}{n} \sum_{i=1}^{n} \prod_{r=1}^{p} l(X_{ir}^d, x_{ir}^d, \hat{\lambda}_r), \tag{3.37}$$

siendo p y q las cantidades de variables continuas y discretas, respectivamente.

Por otra parte, para manipular datos provenientes de variables aleatorias mixtas -esto es, que toman valores categóricos y numéricos al mismo tiempo- Racine y Li ([RQ03]) introducen el siguiente estimador:

$$\begin{aligned} \hat{f}(\mathsf{x}) &= \hat{f}(\mathsf{x}^c, \mathsf{x}^d) = \hat{f}(x_1^c, \ldots, x_q^c, x_1^d, \ldots, x_p^d) && (3.38) \\ &= \frac{1}{n\hat{b}_1 \cdots \hat{b}_q} \sum_{i=1}^{n} \left[\prod_{s=1}^{q} k\left(\frac{X_{is}^c - x_{is}^c}{\hat{b}_s}\right) \prod_{r=1}^{p} l(X_{ir}^d, x_{ir}^d, \hat{\lambda}_r) \right] && (3.39) \\ &= \frac{1}{n\hat{b}_1 \cdots \hat{b}_q} \sum_{i=1}^{n} \mathbf{W}(\mathbf{x}, \mathbf{X_i}, \hat{\mathbf{h}}) && (3.40) \end{aligned}$$

siendo $\mathbf{W}(\mathbf{x}, \mathbf{X}_i, \hat{\mathbf{b}}) = \prod_{s=1}^{q} k\left(\frac{X_{is}^c - x_{is}^c}{\hat{b}_s}\right) \prod_{r=1}^{p} l(X_{ir}^d, x_{ir}^d, \hat{\lambda}_r)$ el producto de los kernels univa-

riados de las variables mixtas para la i-ésima observación de la muestra y $\hat{\mathbf{b}}$ el vector de anchos de banda estimados para ambos tipos de variable: $\hat{\mathbf{b}} = [\hat{b}_1^c, \ldots, \hat{b}_p^c, \hat{\lambda}_1^d, \ldots, \hat{\lambda}_q^d]$. Se prueba que la tasa de convergencia multivariada está dominada por la contraparte continua, $O_p(n^{-\frac{1}{q+4}})$ ([Ela13, pp.29]).

Caso particular: estimador kernel multivariado mixto usando método frecuencista para variables nominales Racine y Li muestran la performance de los estimadores frecuencistas y el AAk para distintos escenarios simulados y un ejemplo concreto ([RQ03, pp.9-13]). Como propiedad fundamental el estimador frecuencista es insesgado, pero se muestra que el error cuadrático medio de las estimaciones realizadas sobre los mismos datos para el estimador AAk están entre $\frac{1}{3}$ y $\frac{1}{2}$ del valor del estimador frecuencista.

En el presente trabajo se opta por utilizar un estimador multivariado como en (3.39), asumiendo $l(X_{ir}^d, x_{ir}^d, \hat{\lambda}_r) = l(X_{ir}^d, x_{ir}^d, 0) = \hat{p}(x_{ir}^d)$, ya que fue la manera más sencilla de implementarlo en `R`. Se obtendrán así estimadores sesgados -aunque con menor varianza- para construir el clasificador Bayesiano (3.24).

Estimación de las proporciones

Para estimar las probabilidades a priori desconocidas (α_j) se utilizó el *Algoritmo Expectation-Maximization (EM)*, que es un método iterativo utilizado para estimar parámetros en modelos estadísticos cuando éstos no pueden ser resueltos de forma directa. Este método tratará de hallar los estimadores máximo verosímiles[27] de los parámetros dados los datos *realmente* observados.

En el Apéndice 2 se proporcionan más detalles sobre este método, en particular en el denominado *"Caso 2"* -el que se utilizó para las estimaciones mencionadas-, en la página 117.

Clasificador Bayesiano *ad-hoc*

Origen El clasificador bayesiano –como su nombre lo indica– está estrechamente relacionado con el teorema de Bayes y la inferencia Bayesiana. Su expresión es:

$$\mathbb{P}(Y = j|X = x) = \mathbb{P}(j|x) = \frac{\alpha_j \cdot \mathbb{P}(x|j)}{\sum_{l=1}^{K} \alpha_l \cdot \mathbb{P}(x|l)},$$

siendo $\alpha_j = \mathbb{P}(j)$ la probabilidad a priori de pertenecer a la clase j-ésima.

Para el caso particular de $K = 2$ (dos clases o grupos para asignar a cada dato) existen varios métodos populares para clasificar a los individuos, como pueden ser el Análisis Discriminante o la Regresión Logística.

Problemas encontrados En primer lugar, la forma de la función de densidad multivariada de los datos $f(\cdot)$ es completamente desconocida. Seguido a ésto no se conoce de antemano la

[27] La técnica de Máxima Verosimilitud busca aquel valor de parámetros que *maximiza* la función de verosimilitud en una muestra.

probabilidad de pertenencia a priori a cada clase, ya que $f(\cdot)$ es una mezcla de 2 densidades, siendo sus proporciones desconocidas. Para afrontar el primer escollo, se propone como solución aproximar esta densidad multivariada utilizando -como ya se mencionó- las técnicas de estimación por kernel, considerando que coexisten variables predictoras tanto cuantitativas como cualitativas, y que ésto será tenido en cuenta al realizar los cálculos.

Para sortear el segundo problema es necesario aproximar las probabilidades de pertenencia a una u otra clase con las restricciones mencionadas, caso muy similar al denominado "densidades mezcla". Es por eso que se decide estimar a estas proporciones α_j utilizando el algoritmo EM, previa estimación de las densidades f_j.

Asumiendo que es posible estimar α_j y f_j se propone la solución a continuación.

Solución propuesta El clasificador bayesiano *ad-hoc* es la propuesta que reúne todo lo anteriormente mencionado, resumido en el seudocódigo del algoritmo 2. El clasificador en (3.25) será denominado *Clasificador Discriminante Bayesiano* (CDB) de aquí en adelante.

3.1.6. Métodos de Consenso

Como indica el artículo de Bourel, Crisci y Martínez ([BCM17]), los métodos de agregación consisten en combinar las predicciones realizadas por modelos base de forma separada, buscando reducir varianza y generando predicciones más estables y precisas que las obtenidas por los modelos originales. La lógica que sostiene este razonamiento se basa en una idea simple: si cada clasificador individual comete errores diferentes, combinar varios de ellos de una forma "inteligente" produciría una reducción del error general y así se podría mejorar las tasas correspondientes a los modelos individuales.

Estos modelos de ensamble se clasifican de dos formas: homogéneos (clasificadores de naturaleza similar, p.ej. bagging, random forest o boosting, para el caso de árboles de clasificación) y no homogéneos (también llamados *métodos de consenso*); en estos últimos se combinan clasificadores diferentes en pos de obtener mejores resultados. La forma de mezclarlos toma como base el denominado "voto mayoritario bayesiano", en donde se asigna un determinado peso a cada una de las hipótesis propuestas por los modelos individuales.

Sea S la muestra aleatoria definida en la Subsección 3.1.1 y $w_{h_m,S}$ una serie de ponderadores dependientes de los clasificadores

$$h(\mathbf{x}) = \underset{k\in\{0,1\}}{\arg\max} \sum_{m=1}^{M} w_{h_m,S}\mathbb{P}_{h_m}(Y_j = k|\mathbf{X} = \mathbf{x}) \tag{3.41}$$

siendo $S_{ME} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ un conjunto de datos supervisado. Todos los métodos de consenso tienen dos componentes: varios algoritmos que generan los clasificadores individuales y una forma de agregar éstos para generar el método de ensamble. Para cada clasificador individual, las "hipótesis de base" h generan vectores del tipo $\left(p_0^h(\mathbf{x}), p_1^h(\mathbf{x})\right)$, siendo $\mathbb{P}_h(Y_j = 0|\mathbf{X} = \mathbf{x}) =$

$p_0^h(\mathbf{x})$ y $\mathbb{P}_h(Y_j = 1|\mathbf{X} = \mathbf{x}) = p_1^h(\mathbf{x})$ las probabilidades a posteriori que $\mathbf{x}$ pertenezca a la clase 0 o 1, respectivamente.

Los tres métodos de consenso utilizados en este trabajo son los que se detallan a continuación.

Voto mayoritario

Dada una observación $(\mathbf{x}_i, y_i) \in S_{ME}$, se considerará como salida el valor de la clase que se repita más veces, para los M modelos considerados ([Bou12]):

$$h_{\text{VM}}(\mathbf{x}) = \underset{k\in\{0,1\}}{\text{arg máx}} \left\{ \sum_{m=1}^{M} \mathbb{I}_{\{h_m(\mathbf{x})=k\}} \right\} \tag{3.42}$$

En caso de empate (es decir, que la cantidad de etiquetas '0' o '1' sean iguales), se asigna aleatoriamente la etiqueta de clase a uno de esos dos valores.

Probabilidad media

Dada una observación $(\mathbf{x}_i, y_i) \in S_{ME}$, se generarán dos vectores de largo M (con elementos respectivos $(p_0^{h_m})_m$ y $(p_1^{h_m})_m$), cada uno describiendo la probabilidad de que cada observación pertenezca a cada una de las clases o grupos definidos para los distintos clasificadores, promediándose luego para cada uno de los modelos en cada clase:

$$h_{\text{MP}}(\mathbf{x}) = \underset{k\in\{0,1\}}{\text{arg máx}} \left\{ \frac{1}{M} \sum_{m=1}^{M} p_k^{h_m}(\mathbf{x}) \right\} \tag{3.43}$$

siendo las probabilidades a posteriori $p_k^{h_m}$ las definidas anteriormente.

Promedio ponderado del AUC

Se separa la muestra en dos partes: una para entrenar los modelos individuales y la otra para los pesos. Dada una observación cualquiera, se consideran las medias ponderadas de las probabilidades a posteriori de cada uno de los modelos por separado; estos pesos surgen de calcular el área debajo de la curva característica[28] para cada método, obtenida con la muestra de prueba. El clasificador resultante será:

$$h_{\text{WA-AUC}}(\mathbf{x}) = \underset{k\in\{0,1\}}{\text{arg máx}} \left\{ \sum_{m=1}^{M} \text{AUC}_m \times p_k^{h_m}(\mathbf{x}) \right\} \tag{3.44}$$

[28]Receiver Operator Characteristic (ROC); al área debajo de la curva se la denomina *AUC-ROC* o *AUROC*; más detalles en la Subsección 3.2.3.

3.1.7. Herramientas auxiliares utilizadas

Análisis de Correspondencia Múltiple

Esta es una técnica de análisis factorial que busca explicar el grado de asociación entre múltiples variables de tipo categórico. Permite además transformar estos datos a espacios continuos ([Bla06]). Se busca eliminar información redundante, trabajar en dimensiones más fáciles de interpretar, además de maximizar la separación de individuos, siempre manteniendo la forma de la nube de puntos original. El análisis de correspondencia múltiple presenta algunas características que lo diferencian de otros métodos factoriales[29]: puede utilizar variables cualitativas; individuos, variables y modalidades de éstas serán los objetos que formarán parte del análisis y sobre ellos se buscará determinar qué relaciones existen, además de permitir el estudio de elementos "suplementarios" que por distintos motivos no pueden ser utilizados con el resto de los datos.

Enfoque formal El Análisis de Correspondencia Múltiple (ACM) es una generalización del Análisis de Correspondencia Simple (ACS), que trabaja con tablas de datos cualitativos, tratando a filas o columnas de forma equivalente a través de los "perfiles fila" o "perfiles columna", y que sirve para resumir datos en dimensiones bajas, manteniendo la mayor cantidad de información original posible. Se parte de una tabla de datos original, con I filas o individuos y J columnas o variables, llamada $\mathbf{X}$:

$$\mathbf{X}_{I,J} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1J} \\ x_{21} & x_{22} & \cdots & x_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ x_{I1} & x_{I2} & \cdots & x_{IJ} \end{pmatrix},$$

representando x_{ij} el valor de la variable j-ésima para el individuo i-ésimo. De esta matriz se pueden obtener dos tipos de tablas que permiten realizar los cálculos necesarios para trabajar con ACM:

- *Matriz o Tabla Disyuntiva Completa (TDC)*: es una matriz que separa a las variables en sus correspondientes categorías y que para cada entrada vale 1 si el individuo cumple en esa modalidad y cero en caso contrario. Es muy útil para ver relaciones entre individuos. Simbolizada por $\mathbf{Z}$, tendrá la misma cantidad de filas que $\mathbf{X}$ pero tantas columnas como modalidades o categorías tengan todas las variables.
- *Tabla de Burt*: surge del producto de una TDC por su traspuesta ($\mathbf{Z}^T\mathbf{Z}$); resultando en una matriz simétrica, análoga a la matriz de covarianza para variables continuas. Por su forma de cálculo, existirá en su interior un grupo de submatrices diagonales, mientras que las subma-

[29] Como por ejemplo el Análisis de Componentes Principales (ACP). La principal diferencia entre ACM y ACP es que al ser las matrices a diagonalizar diferentes (ACP trabaja exclusivamente con datos numéricos), debe utilizarse otro tipo de distancia y pesos por filas y columnas. Estas técnicas factoriales comparten los objetivos de reducción de dimensionalidad con mínima pérdida de información.

trices restantes mostrarán la interacción entre las distintas categorías. Como contrapartida, se pierde la información de las filas.

Sobre cualquiera de estas tablas se calculan los valores y vectores propios correspondientes[30], obteniendo así un listado descendente en magnitud de las contribuciones a la *inercia* (o variación) de cada eje[31]. Para poder dotar de interpretación a cada eje factorial obtenido, es necesario observar las contribuciones de inercia a cada eje, la calidad de representación y qué contenido trae cada modalidad por separado.

Elementos de base Los principios del ACM son (al igual que para el ACS) tres: transformación de datos originales en perfiles fila o columna, ajuste o ponderación de puntos a los perfiles marginales, utilización de la distancia χ^2 para saber proximidad entre elementos. Más detalles al respecto se pueden encontrar en los libros de Blanco y cols. ([Bla06]) y Lebart, Morineau y Piron ([LMP95]).

Ayudas a la interpretación factorial Hay dos series de coeficientes que aportan información adicional: las "contribuciones" a la inercia (o varianza) explicada por cada factor y los "cosenos cuadrados". La contribución a la inercia de cada eje factorial dice cuanto aporta un individuo o modalidad a la formación de ese eje, mientras que el coseno cuadrado mide la asociación entre ese objeto (individuos o modalidades) y el eje factorial, ayudando ambos a interpretar mejor los significados de los ejes y las posiciones relativas de los objetos representados en los planos factoriales.

Variables activas y suplementarias Los elementos suplementarios -tanto modalidades como individuos- pueden ser proyectados en los factores obtenidos. Esto es muy útil para enriquecer el análisis, bien por ser éstas dejadas de lado a propósito por el investigador (p.ej. para no introducir ruido adicional, como variables de menor interés en el análisis) o también por ser éstas variables de naturaleza especial y que por ello no pueden ser parte del proceso (p.ej. variables numéricas). En la Figura 3.4 se muestra un esquema de cómo se visualizan cada una de las variables e individuos según su naturaleza, según la descripción de Lebart y cols.

3.2. Medidas de Desempeño de Clasificadores

Como es de interés en este trabajo cuantificar la performance de cada modelo utilizado en función de diferentes perspectivas, se introducirán otros indicadores promedio de interés -además

[30]Como indican [HJS16, pp.3-4] hay tres formas de presentar al ACM; en este trabajo se utiliza la propuesta más utilizada: aplicar ACS a una TDC.

[31]Matemáticamente, tanto ACM como otros métodos factoriales (p.ej. ACP) son una transformación lineal ortogonal de los datos originales a un nuevo espacio de coordenadas, de forma tal que la dispersión (p.ej. inercia en el caso de ACM, varianza en ACP) se mantenga lo más cercana a los datos originales, y que además las nuevas coordenadas ordenen esa dispersión en forma decreciente; esto es: en primer lugar tener la dirección con mayor porción de la varianza, en segundo lugar la segunda mayor porción, etc.

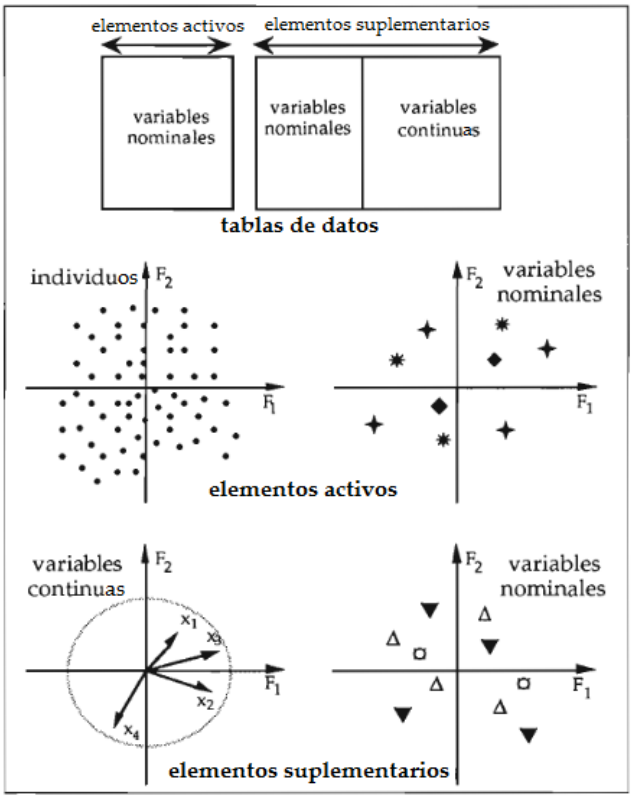

Figura 3.4: Esquema de representación de individuos y variables suplementarias (fuente: [LMP95, pp.123], con modificaciones respecto del original)

del ya mencionado *error de generalización* visto en (3.2)- que, como esta cantidad, también pueden ser calculados utilizando las denominadas *matrices de confusión*.

En la primera parte de esta sección se introduce este concepto, para luego desarrollar los diferentes indicadores que surgen de esta matriz y finalmente se introducirán otras herramientas de comparación de modelos, como las denominadas "curvas ROC".

3.2.1. Matriz de confusión

La *matriz de confusión* o matriz de error es una tabla que muestra la performance predictiva de un determinado algoritmo, respecto de la verdadera etiqueta o clase y en función de un umbral de clasificación previamente escogido[32]. Para desarrollar este apartado se presentará a los elementos de la matriz de esta manera:

- En filas: se muestran los valores observados o la "realidad"
- En columnas: se muestran los valores que el modelo propuesto predice

En el caso de un problema de clasificación, se tendrá tantas filas o columnas como valores pueda tomar la variable estudiada. Simplificando, se usará de aquí en más una variable de interés Y con

[32] Un ejemplo gráfico se muestra en la Subsección 3.2.3.

solo dos valores, $\{0, 1\}$, indicando ausencia o presencia del atributo que se intenta cuantificar o medir, respectivamente.

		Y^{pred} 0	1
Y^{obs}	0	VN	FP
	1	FN	VP

Tabla 3.4: Matriz de Confusión

De este modo, y suponiendo que se trabajará con más de una variable de interés y que cada una de ellas toma siempre los valores 0 o 1, la matriz de confusión tiene el formato presentado en la Tabla 3.4. De esta tabla, surgen cuatro indicadores de vital importancia, todos en base a las *frecuencias absolutas* en cada una de sus celdas:

- *Verdaderos Negativos (VN)*: total de individuos que el modelo asignó como negativos y además no tenían la condición (intersección de celdas $[Y^{obs} = 0, Y^{pred} = 0]$)
- *Falsos Negativos (FN)*: total de individuos que el modelo asignó como negativos pero que en realidad sí tenían la condición (intersección de celdas $[Y^{obs} = 1, Y^{pred} = 0]$)
- *Verdaderos Positivos (VP)*: total de individuos que el modelo clasificó correctamente con la condición (intersección de celdas $[Y^{obs} = 1, Y^{pred} = 1]$)
- *Falsos Positivos (FP)*: total de individuos clasificados incorrectamente con la condición (intersección de celdas $[Y^{obs} = 0, Y^{pred} = 1]$)

3.2.2. Indicadores generados por una matriz de confusión

Mirando las tablas por filas o columnas se muestran dos lecturas diferentes. En el primer caso se pueden identificar a aquellos individuos según su etiqueta o clase "real" u observada, mientras que para la segunda lectura se puede determinar qué ha decidido el modelo para cada una de las observaciones.

- Etiquetas reales u observadas
  - Condición Positiva: total de individuos que realmente tienen la característica $CRP = VP + FN$
  - Condición Negativa: total de individuos que realmente no tienen la característica $CRN = FP + VN$
- Indicadores según la realidad observada
  - Del total de individuos que tienen la característica...
    - ¿qué proporción son correctos?: Sensibilidad o $TVP = \frac{VP}{VP+FN}$
    - ¿qué proporción son incorrectos?: Falsos Negativos $TFN = \frac{FN}{VP+FN}$

- Del total de individuos que no tienen la característica...
  - ¿qué proporción son correctos?: Especificidad o $TVN = \frac{VN}{VN+FP}$
  - ¿qué proporción son incorrectos?: Falsos Positivos $TFP = \frac{FP}{VN+FP}$

- Etiquetas predichas por el modelo
  - Condición Predicha Positiva: total de individuos asignados como "tienen la característica" por el modelo $CPP = VP + FP$
  - Condición Predicha Negativa: total de individuos asignados como "no tienen la característica" por el modelo $CPN = FN + VN$

- Según la predicción del modelo
  - Del total de individuos asignados como "tienen la característica" por el modelo...
    - ¿qué proporción son correctos?: Valor Predictivo Positivo $VPP = \frac{VP}{VP+FP}$
    - ¿qué proporción son incorrectos?: Falso Descubrimiento $TFD = \frac{FP}{VP+FP}$
  - Del total de individuos asignados como que "no tienen la característica" por el modelo...
    - ¿qué proporción son correctos?: Valor Predictivo Negativo $VPN = \frac{VN}{VN+FN}$
    - ¿qué proporción son incorrectos?: Falsa Omisión $TFO = \frac{FN}{VN+FN}$

El error general visto en (3.2) es, según esta notación, $erg = 1 - \frac{VN+VP}{N}$, siendo N el total de datos.

En función del problema se buscará maximizar una o varias de las tasas o indicadores presentados. Es necesario tener cuidado en estos casos porque al maximizar una tasa puede que otras caigan a valores no aceptables; un ejemplo es la denominada "paradoja de la precisión": si la cantidad de falsos positivos supera a la de verdaderos positivos ($FP > VP$) la exactitud aumentará si se usa un modelo que asigne siempre la etiqueta "0" o de ausencia de característica de interés. Del mismo modo, si los falsos negativos superan a los verdaderos negativos ($FN > VN$) la exactitud aumentará si ese modelo asigna siempre la etiqueta "1". Esto hace que la exactitud -definida como el total de casos clasificados correctamente sobre el total de observaciones- no sea adecuada para comparar diferentes modelos; en su lugar se utilizan por ejemplo la tasa de valores predictivos positivos o la sensibilidad.

Otras medidas de interés usadas en este trabajo

- Kappa de Cohen: en este caso, mide cuan bueno es el clasificador respecto de asignar aleatoriamente la "etiqueta verdadera" a la nueva observación; sirve para corregir al valor obtenido en la precisión por aquellas asignaciones "aleatorias"[33]

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{3.45}$$

[33]Muy utilizado cuando las variables a predecir son desbalanceadas. A partir de un $\kappa >= 0.4$ se habla de un ajuste adecuado.

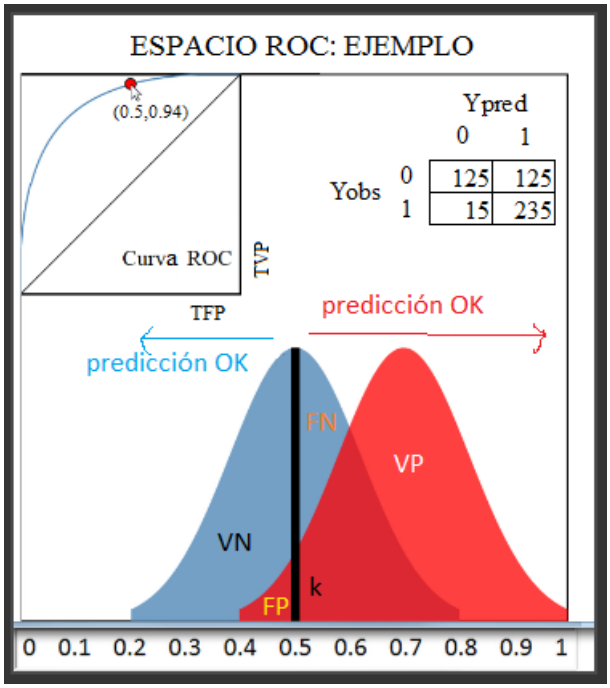

Figura 3.5: Espacio de curva ROC: ejemplo

con p_o el nivel de acuerdo observado y p_e el esperado entre el clasificador y los valores reales, como indican Solís y cols. ([SMG+18])

3.2.3. Herramientas para comparar modelos predictivos

Curvas ROC

Dado un clasificador como cualquiera de los descriptos en la Sección 3.1 y suponiendo solo dos clases como se mencionó en la Subsección 3.2.1, el cruce de "realidad" con "predicción" resulta en *una* matriz de confusión como la de la Tabla 3.4. Esta tabla única genera un único punto en el espacio de tasas de falsos y verdaderos positivos, como se muestra en la Figura 3.5[34]. En este ejemplo, se supone una población de 500 individuos, repartidos en dos clases (azul y roja) por igual. El umbral k determina el valor del punto de corte utilizado para decidir cuando una observación se clasifica de una u otra forma[35]. Si éste se fija en $k = 0.5$ como en la imagen, la matriz de confusión resultante es la que figura en el ángulo superior derecho de la figura y ésta dará como resultado el punto (0.5,0.94) señalado en la curva ROC que está a la izquierda del gráfico. Si el umbral cambia, lo harán también la matriz de confusión y el punto resultante en el espacio de valores de la curva ROC.

La curva ROC entonces surge de explorar todos los posibles puntos de corte -y las matrices de confusión resultantes- en un gráfico escalera de la Sensibilidad (o TVP) expresada en relación a la tasa de falsos positivos (TFP) del (o los) modelo(s) predictivo(s).

[34]Ejemplo extraído con modificaciones de *ROC curves and AUC explained*.

[35]Como se observa, hay superposición entre las dos clases, con lo cual el error cometido será necesariamente distinto de cero.

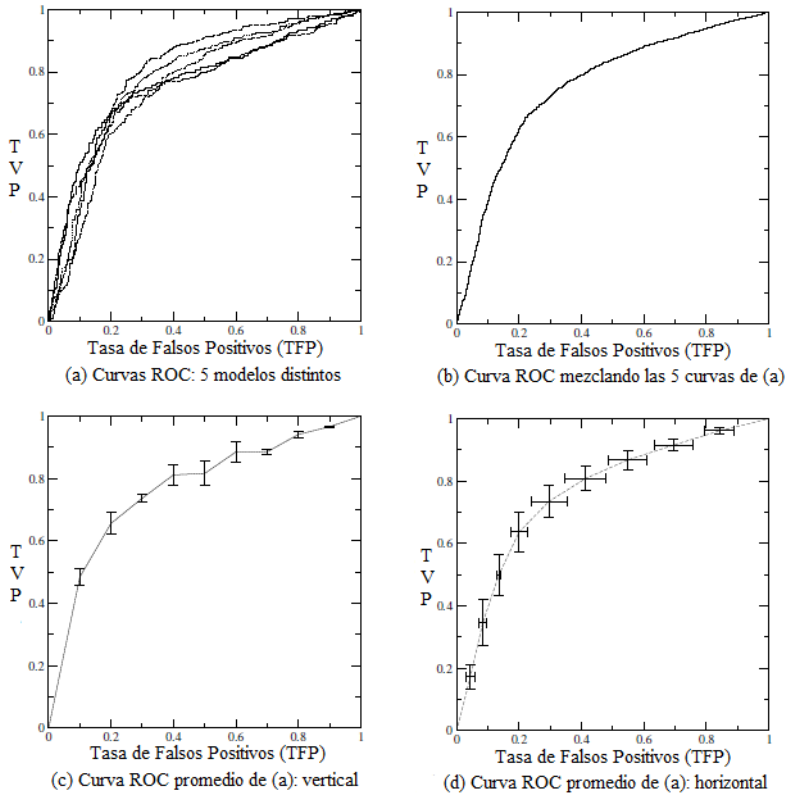

Figura 3.6: Promediado de curvas ROC, en base a [Faw04, pp.15]

Una pregunta recurrente es, ante muchos clasificadores de un mismo tipo, ¿es posible *promediar* estas curvas? Fawcett sugiere en su artículo sobre el tema ([Faw04, pp.14-17]) estas dos opciones:

- *Promedio vertical*: útil cuando se busca dejar fija la TFP. Para cada valor de TFP se elige el máximo de Sensibilidad o TVP, o se interpola entre puntos si es necesario. De este modo, se considera a cada curva ROC como una función R_i tal que $TVP = R_i(TFP), \quad i = 1, ..., k$ con k puntos de corte, equiespaciados entre sí. La curva promediada es $\overline{R_i(TFP)} = \bar{R}(TFP)$. Si se asume distribución Binomial, se pueden calcular intervalos de confianza para $\overline{TFP}$
- *Promedio horizontal*: también llamado "promedio por umbrales", este método busca para cada punto de corte el punto en la curva ROC y luego promedia a todos estos puntos

En la Figura 3.6 se muestran las diferentes posibilidades: cinco curvas ROC en (a), una curva que resulta de mezclar las anteriores en (b) y los promedios de éstas ya sea de modo vertical (gráfico (c)) como horizontal (gráfico (d)).

Área bajo la curva ROC

La *AUC* (o *AUROC*) es el área debajo de la curva ROC, que es una medida relativa de precisión entre modelos predictivos. Los valores más plausibles de este indicador van de 0.5 (el mínimo aceptable, indicando que el modelo no discrimina entre individuos de diferentes clases)

y 1.0 (el máximo, que indica discriminación perfecta), aunque puede pasar que caigan entre 0 y 0.5, indicando discriminación inversa. Valores por encima de 0.7 indican una discriminación aceptable[36].

Este indicador tiene una interesante propiedad: es equivalente a la probabilidad de que un clasificador asigne -a una observación aleatoria- una etiqueta "positiva" sea mayor que asignar una "negativa". Esto es equivalente a la prueba no paramétrica de suma de rangos de Wilcoxon. Incluso se relaciona con el índice de Gini ([Faw04, pp.13]). El punto de "clasificación perfecta" es el (0,1) en cualquiera de los gráficos de la Figura 3.6; mientras más cerca se esté de éste mejor será el modelo predictivo ya que su AUROC será cercana a 1.

A pesar de no considerar los valores de probabilidad predicha ni la bondad de ajuste de cada modelo, de resumir la performance predictiva de los modelos en regiones donde muchas veces no se tienen valores concretos y ponderar de la misma manera a los errores por acción (se toma decisión equivocada) o por omisión (no se hace nada, [BCM17]), es muy utilizado en la práctica por su versatilidad ([Faw04, pp.13]) y por esta razón es considerado en este trabajo.

Curva de Sensibilidad-Especificidad

Esta curva es similar a la curva ROC descrita en la Sección 3.2.3, pero cambiando el eje de abscisas por la Especificidad en lugar de la "no especificidad" o TFP.

3.3. Desarrollo de un paquete en `R` para el estudio de los datos

En lo que sigue se introduce el paquete `uefi2`, parte fundamental del desarrollo de este trabajo, creado y mantenido por este autor[37].

3.3.1. Forma de trabajo: "ajuste simultáneo"

El "ajuste simultáneo de modelos" consiste simplemente en, dado un grupo de datos previamente separados en muestras de entrenamiento y prueba, ajustarlos a modelos provenientes de distintos enfoques del aprendizaje automático (sobre los *mismos datos de entrenamiento* para *todos los modelos*), predecir con los mismos datos de prueba *sobre cada uno* de esos modelos ajustados y observar su performance predictiva.

Para lograr una mayor generalización de los resultados predictivos, este ajuste y prueba se realiza sorteando *varias veces* qué observaciones son de entrenamiento o de prueba, evitando así el obtener resultados de un sólo posible ajuste y generando variabilidad que ponga a prueba la adecuación de cada modelo a la misma. Esto se conoce como *remuestreo*.

[36] Estos valores son generalmente resultado de investigaciones empíricas. Este indicador fue tomado del sitio `http://gim.unmc.edu/dxtests/roc3.htm`.

[37] El código utilizado está libremente disponible en este sitio de GitHub.

Algoritmo 3 Seudocódigo: mejor subconjunto

Datos: datos $\mathbf{x}_i$
Resultado: $\mathcal{M}_p^*$ (mejor modelo de todos)
Inicio: Sea $\mathcal{M}_0$ el modelo nulo, aquel sin predictores **para** $j = 1, ..., p$ **hacer**

1. Ajustar todos los posibles $\binom{p}{j}$ modelos que tienen exactamente j variables predictivas
2. Elegir el mejor de los modelos en a), llamado $\mathcal{M}_j$, con algún criterio válido

fin
Modelo final: Elegir de todos los posibles "buenos modelos" $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_p$ al mejor de todos $(\mathcal{M}_p^*)$ utilizando algún criterio válido

Partiendo de un cierto conjunto de datos y de modelos predictivos, se buscará algo similar al algoritmo *best subset selection* modificando el algoritmo 3; esto será detallado en la Sección 3.3.3.

3.3.2. Parametrización de modelos

Buscando optimizar los parámetros de algunos de los modelos utilizados en este trabajo, se creó una función interna[38] llamada `ajp()` (por "ajuste de parámetros") que, partiendo de una fórmula general y datos de entrenamiento muestra los valores óptimos encontrados por cada método. Esta función es un *wrapper*[39] de otras funciones de calibración existentes en `R` (ejs.: `e1071::tune()`, `MASS::stepAIC()`).

Considerando los parámetros γ y C de SVM[40], m de Random Forests[41] y α en CART[42], se sortean distintas permutaciones de datos de entrenamiento y prueba y con los datos de entrenamiento para cada permutación se corre la función `ajp()`. Al final del proceso se toma la *mediana* de cada uno de los parámetros evaluados y con esa información se continúa trabajando.

3.3.3. "Loop" de Ajuste Simultáneo: la función `asm2()`

Esta función es la principal en `uefi2`, ya que ajusta a todos los modelos con los mismos datos de entrenamiento, además de verificar el ajuste en los datos de prueba y generar una batería de indicadores y salidas anexas para profundizar el análisis.

[38] Función interna es una función "privada", o sea aquella que no es accesible directamente al cargar un paquete, pero que sí está en su 'namespace'; para poder hacer uso de ellas en `R` debe usarse el operador `':::'`.
[39] Función de envoltura o wrapper es aquella subrutina que llama a otras en su interior, con pocas o nulas órdenes adicionales.
[40] Parámetros de "balance sesgo-varianza" y costo de clasificación errónea, como figuran en la Tabla 3.1 y en la ecuación (3.19), respectivamente.
[41] Tamaño del subconjunto de variables explicativas consideradas para realizar las particiones en Random Forests, con $m < p$; en la función `randomForest()` este se conoce como `mtry`.
[42] Parámetro de costo-complejidad, como en la ecuación (3.15).

Paquetes de `R` utilizados

Ajuste de modelos dentro de `asm2()` Se presentan a continuación aquellos paquetes utilizados -fuera de la distribución canónica o base- en orden alfabético, que implementan los modelos mencionados previamente:

`adabag` **Boosting**: este paquete implementa las versiones AdaBoost (*adapting* boosting) original (clasificación a 2 clases) y las generalizaciones conocidas como M1 y SAMME (Zhu) [AGG13]. Las funciones usadas son `boosting()` y algunos de sus métodos. Tanto los clasificadores "débiles" como los parámetros utilizados fueron los proporcionados por defecto: árboles de clasificación CART, para el algoritmo AdaBoost.M1[43].

`e1071` **SVM**: se utilizan para este modelo las funciones `svm()` y `tune.svm()`, además de `tuneRF()` para optimizar los parámetros de Random Forests [MDH+18]

`randomForest` **Random Forests**: de este paquete se usa la función `randomForest()` y varias funciones gráficas y métodos específicos [LW02]

`rpart` **CART**: se utiliza principalmente la función `rpart()`, junto con métodos específicos y complementos, como `prp()` para gráficos [TA18]

Otras funcionalidades El paquete `MASS` es de uso general, creado principalmente como soporte del libro [VR02]. En este trabajo se utiliza `stepAIC()` para estudiar la bondad de ajuste de modelos lineales generalizados. Para ACM por su parte, se utilizó el paquete `FactoMineR`, mientras que para generar distintos gráficos se usaron los paquetes `ggplot2`, `reshape2`[44] y `factoextra` ([LJH08, Wic16, Wic07, KM17]).

Uso de fórmulas Los cinco modelos de aprendizaje automático ya existentes en `R` tienen funciones o incluso métodos de los cuales uno de sus parámetros es un objeto de clase 'fórmula', es decir una descripción simbólica del tipo $Y \sim \mathbf{X}$, con Y la variable dependiente y $\mathbf{X}$ la(s) variable(s) independiente(s) o predictoras. Para cada caso, dentro del código de las funciones principales utilizadas, los métodos fórmula implementados llaman a la función genérica `model.frame()`, que devuelve siempre un `data.frame`[45] con el orden de sus columnas según la fórmula dada por el usuario[46].

Para el caso del clasificador Bayesiano, como no hay una función de ajuste definida se construye un `model.frame` desde la fórmula proporcionada.

[43] En la configuración estándar de la función, deja crecer a estos "clasificadores débiles" al máximo (asigna un $C_p = -1$) y pide que al menos tengan un nodo (lo que se conoce en términos de aprendizaje automático como 'decision stump').

[44] Para reordenar internamente tablas de datos bajo el paradigma *"split-apply-combine"*, que luego son graficadas mediante la función `plot.asm()`, del paquete `uefi2`.

[45] Un `data.frame` es una tabla de datos, que almacena la información por individuo en sus filas y por variable en sus columnas; éstas últimas pueden ser *de cualquier tipo*.

[46] Es decir que si p.ej. se tiene la fórmula `'y~x1+x2'` para el `data.frame` $[X_1, X_3, Y, X_2]$, `model.frame()` devolverá un `data.frame` $[Y, X_1, X_2]$.

Esquema de funcionamiento de `asm2()`

Para cada una de las variables dependientes definidas en la Subsección 4.1.3, se enumeran los pasos de funcionamiento de `asm2()`:

1. Controles previos: de datos perdidos, manejo de excepciones durante la ejecución del ajuste de los modelos propuestos
2. Loop: separando previamente -y de forma aleatoria- muestras de entrenamiento (ME) y de prueba o testeo (MT) utilizando validación cruzada una vez *(1-fold CV)*, y resorteando los índices para cada iteración[47]:
   a) Se ajusta cada modelo (CART, GLM, SVM, RF, Boosting, CDB) a datos ME
   b) Con objeto anterior se crea objeto `predict` (correspondiente a cada clase) que será usado para predecir Y usando MT
   c) Se generan matrices de confusión e indicadores respectivos
3. Se guardan `data.frame`s de valores predichos ($\{0, 1\}$) y probabilidades a posteriori para cada una de las dos categorías de cada Y, para las diferentes variables dependientes consideradas
4. Con lo anterior se calculan los valores predichos para los modelos de consenso citados en la Subsección 3.1.6
5. Salida final: información sobre variables, metadatos, matrices de confusión e indicadores para cada modelo y permutación de datos

Las salidas de la función `asm2()` son el principal insumo para observar los resultados del ajuste de los modelos; ésta genera un objeto de clase `asm2` que es en definitiva una lista con metadatos del ajuste y resultados para ser utilizados en otras partes del trabajo. Durante todas las corridas -y para cada iteración- se utilizó el 75 % del total de los datos para entrenamiento y 25 % para prueba.

Particularidades por modelo implementado

Implementación informática KDE En la función principal elaborada para este trabajo, `dk()`, se utilizan según la naturaleza de las variables:

- Para datos numéricos: se puede estimar densidades por núcleo mediante cualquiera de las funciones `stats::density()`, `np::npudens()` o `ks::kde()`[48] [Duo18, WLLR05, Wan15]
- Para datos categóricos: se crearon las funciones `nif()`, `predict.nif()` y `predict.nifM()`, en el paquete `uefi2`, mediante el modo "frecuencista" descripto en la Sección 3.1.5. Aquí

[47] Con este método se introduce variabilidad adicional buscando evaluar la *estabilidad* de las predicciones, buscando así mayor generalización en los resultados.

[48] Es necesario tener cargados los paquetes `KernSmooth` y `klaR` ya que `ks` utiliza internamente funciones de éstos. Incluso fue necesario modificar levemente algunas de esas funciones para concluir con alguno de los cálculos

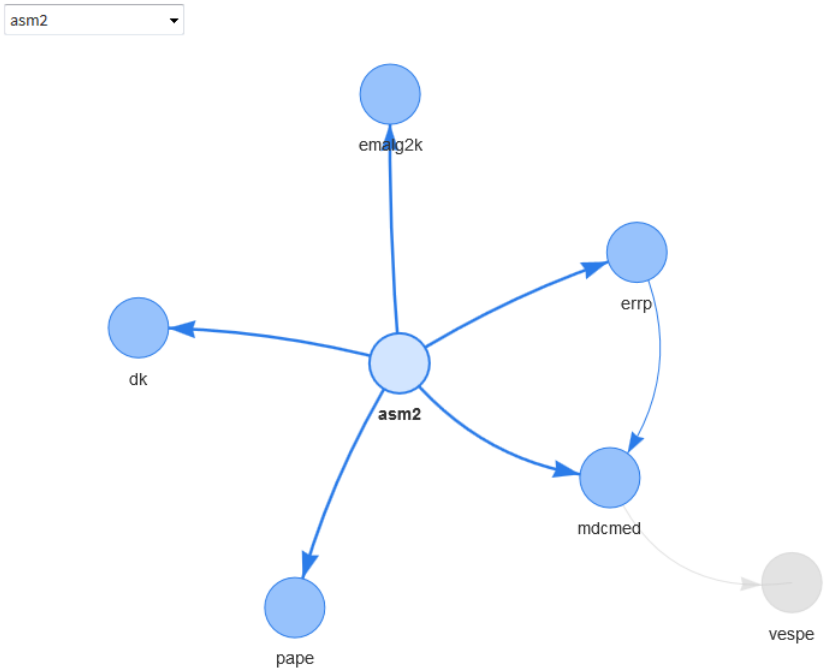

Figura 3.7: Paquete `uefi2`: dependencias entre funciones. Se muestra la principal, `asm2()` cuyo código está en el Apéndice 3.

`predict.nifM()`, versión multivariada de `predict.nif()`, asigna como predicción la misma frecuencia relativa a la categoría i-ésima de una variable categórica X^d, en base a lo hallado previamente por las otras dos funciones

- Para datos mixtos: primero se realiza el producto de variables categóricas, y con el resultado anterior se hace el producto con variables numéricas. Los tres resultados son devueltos en una lista

Implementación informática SVM Se probaron algunas alternativas durante la etapa inicial mencionada en la Subsección 4.1.2, proporcionadas por el paquete `e1071` de `R`, para finalmente decidir por usar el kernel de base radial (RBF) de tipo normal o gaussiano, ya que es el mejor cuando no se conoce con mucho detalle a la población a priori, como lo indica la Tabla 3.1. La implementación de SVM en `e1071` calcula además probabilidades a posteriori para cada una de las clases definidas, mediante la denominada "aproximación de Platt", utilizando una función sigmoide con parámetros desconocidos, los cuales se obtienen mediante un problema de optimización máximo verosímil regularizado ([HCR01, pp.1-2]). Este paquete utiliza la librería LibSVM que implementa dicha aproximación ([KMH06, pp.4]).

En cuanto a los parámetros definitivos, se utilizó la función `ajp()` para cada grupo de datos y variable como se explica al comienzo de la Subsección 3.3.2.

Implementación informática CDB Se decidió crear una implementación propia para esta modelización. En espíritu es similar a lo hecho por los otros modelos "de base" en el punto 2. del

esquema anterior, pero con estas peculiaridades:

Dadas las muestras de entrenamiento y prueba por separado:

1. Se estiman las densidades por núcleos $\hat{f^j}_{KDE}(x_{obs,i})$ mediante la función `dk()` con los datos de entrenamiento, por separado para cada clase
2. Dadas las densidades $\hat{f^j}_{KDE}(x_{obs,i})$ se estiman mediante el algoritmo EM (en algoritmo 5) las proporciones $\alpha_l^{(m)}$ para cada grupo, utilizando la función `emalg2k(f,...)`
3. Dados $\alpha_l^{(m)}$ y $\hat{f^j}_{KDE}(x_{obs,i})$ para cada una de las clases se calculan las probabilidades a posteriori mediante la función `pape()`, como lo indica la fórmula (3.24). En caso de no estimar los valores de $\alpha_l^{(m)}$, `pape()` asigna por defecto las proporciones por clase a priori

3.3.4. Estudio de resultados obtenidos con `asm2()`

Selección de "mejores fórmulas" para cada Y mediante *ranking*

Ordenar para reducir información Para minimizar la cantidad de información a utilizar, era imperioso jerarquizar para usar solo la "parte más importante". Se pensó en consecuencia en un ordenamiento sencillo, que fuera relativamente fácil de calcular e implementar y que pueda ser mejorable en caso de que sea necesario. De esto surge un ordenamiento de las fórmulas, basado en la mediana de las posiciones de uno o más estadísticos provenientes de las matrices de confusión obtenidas en cada validación cruzada, sin ponderadores ni penalizadores[49]. Estos indicadores pueden ser considerados como "positivos" (si se busca maximizarlos, p.ej: Sensibilidad) o "negativos" (si se busca lo contrario, p.ej: Falsos Positivos).

Implementación en `uefi2` La función `resblq()` recibe un objeto de tipo `asm2` -generalmente una lista con muchas fórmulas ajustadas según los datos proporcionados-, y con los parámetros `ord.mdcmed` y `mxr` se le dice qué indicador(es) de la matriz de confusión se considerarán para ordenar los resultados y cuántas fórmulas (las "n mejores") serán las devueltas por ese ranking, respectivamente[50].

Esta función `resblq()` puede verse como un "jurado": evalúa a todas las fórmulas en todos los modelos según el valor de su *mediana* de posición en los distintos indicadores (definidos en la Subsección 3.2.2); así emergerá como "ganadora" aquella fórmula de tipo $Y \sim \mathbf{X}$ que tenga más primeros lugares[51]. Luego de varias pruebas se decidió por el valor `mxr=5`, lo que equivale a contar la cantidad de veces que la i-ésima fórmula aparece entre los primeros 5 ordenamientos de la tabla generada.

[49] Esto se hace separado por generación y variable predictora.

[50] Esta función se apoya en dos internas, por un lado `bestinfo()` para extraer información del objeto `asm2` y por otro `ord_df_()` para generar un ranking ordinal (sin empates) utilizando los valores de los distintos indicadores de la matriz de confusión de modo adecuado a su "signo".

[51] Para esta evaluación se utiliza una matriz que en las filas tiene la posiciones (1°, 2°, etc.), en las columnas los diferentes modelos ajustados y en cada celda el número de fórmula que ocupa cada posición.

Respecto a qué estadístico(s) usar para evaluar lo anterior, y después de varias pruebas con los posibles indicadores, se optó por usar como referencia al estadístico κ definido en la ecuación (3.45), ya que aquellos modelos con valores altos de este indicador mostraban el mejor desempeño predictivo de todos.

Finalmente, una vez obtenido los resultados de la cantidad de veces que cada fórmula aparece entre los 5 primeros lugares (según el estadístico κ), se utiliza como "puntaje" la frecuencia absoluta de aparición en la tabla mencionada anteriormente, utilizando de aquí en más solamente las primeras cinco de ellas, reduciendo significativamente la información a utilizarse para el resto del trabajo.

Análisis de la performance predictiva

A continuación se muestran los pasos para realizar las curvas ROC de los modelos predictivos en este trabajo.

1. Para cada iteración, se extrae un `data.frame` del objeto `asm2` que consta de las probabilidades a posteriori para cada valor de la variable Y, mediante la función wrapper `uefi2:::furoc()`[52]
2. Para cada modelo se agrupan todas las iteraciones en `data.frame`s
3. Para cada fórmula seleccionada se agregan todos los `data.frame`s generados en 2.
4. Se resume por separado el resultado de cada fórmula, comparando la performance de los modelos mediante curvas ROC *promediadas verticalmente*, utilizando una función creada para estos propósitos -denominada `prom_ROCs()`- en base a lo descrito en la Sección 3.2.3

Análisis de aciertos y errores mediante ACM

En primera instancia se extrae toda la información necesaria, para cada una de las "mejores fórmulas" definidas previamente y separando por Y y generación, para luego realizar el ACM siguiendo los siguientes pasos:

1. Se extrae aciertos *por individuo* y se guardan en un `data.frame` mediante el wrapper `uefi2:::acm_err()`[53]. Acierto en este contexto quiere decir, dados fórmula y datos, para qué individuo(s) se acierta el valor de la variable Y en cada una de las iteraciones del loop mencionado en la Sección 3.3.3[54]
2. Se agrega la información en 1. mediante otro `data.frame` junto con los datos de cada individuo

[52] La función `furoc()` llama a su vez a `auroc()`, ambas wrappers de las funciones `prediction()` y `performance()` del paquete `ROCR` [SSBL05], adaptadas al formato de la salida de los datos en `uefi2`.

[53] La función wrapper `acm_err()` extrae datos para realizar ACM, además de llamar a funciones analíticas y gráficas de los paquetes `FactoMineR` y `factoextra`.

[54] Puede ocurrir que para un mismo individuo, algún modelo erre su predicción en tan solo una iteración; en este caso se considerará un *error*, por más que ocurran aciertos en otras iteraciones sobre el mismo individuo.

3. Se decide qué variables son activas y/o suplementarias a la formación de los ejes factoriales
4. Se llama a la función `MCA()` del paquete `FactoMineR`; este objeto será el punto de partida del análisis factorial
5. Mediante el paquete `factoextra` se analizan contribución y calidad de representación a cada modalidad, utilizando tablas y gráficos adecuados, agregando a las variables numéricas como suplementarias.

La presentación de los gráficos correspondientes al punto 5. se hace separada: por un lado estarán las variables activas y, para el mismo plano pero en un gráfico contiguo, estarán representadas las variables suplementarias, al estilo de la Figura 3.4.

El procedimiento de todo el trabajo se resume en la Figura 3.8.

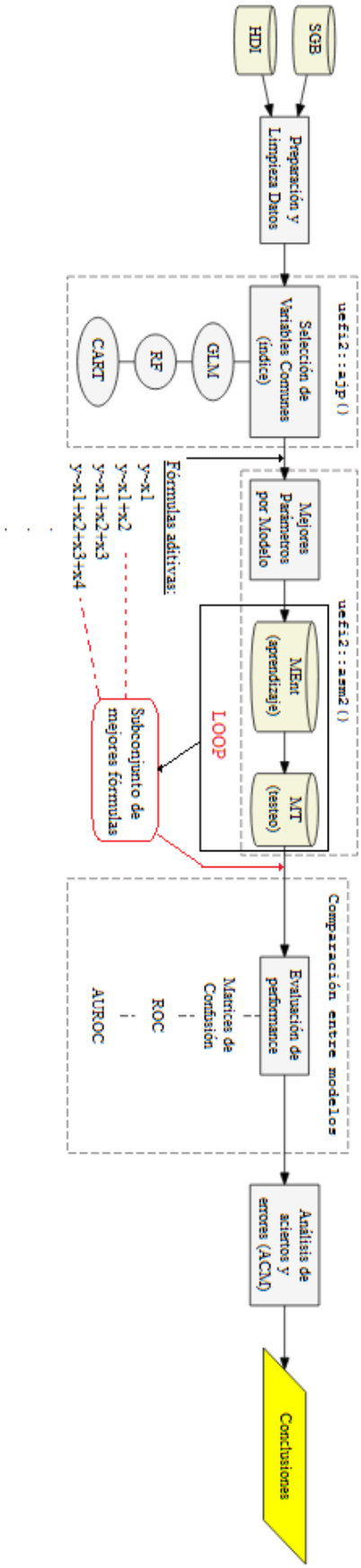

Figura 3.8: Diagrama de flujo del trabajo

Capítulo 4

Resultados

En este capítulo se presentan los resultados de la investigación. Primeramente se describen los pasos previos al análisis -Sección 4.1-, para luego mostrar una descripción somera de los datos trabajados -Sección 4.2- y finalmente los resultados principales, tanto para el rendimiento -Subsección 4.3.1- como para la desvinculación de los estudiantes -Subsección 4.3.2.

4.1. Preparación para el análisis

4.1.1. Información académica en FIng

Hasta el momento coexisten dos sistemas horizontales de gestión académica, ambos desarrollados por el Servicio Central de Informática Universitaria (SeCIU): por un lado el Sistema de Gestión de Bedelías (SGB), existente desde hace más de veinte años y con serias limitantes en la actualidad, y por otro el reciente Sistema de Gestión Administrativa de la Enseñanza (SGAE). Este nuevo sistema acompaña decisiones políticas importantes tomadas en los últimos años -flexibilización de planes de estudio, movilidad horizontal, creditización-, incorporando la noción de "estudiante único"[1]. La transición del SGB al SGAE comenzó por etapas en 2015, esperándose finalizar entre 2018 y 2019[2]. En este trabajo se utilizó el viejo sistema para extraer datos académicos y sociodemográficos de interés, como se viene haciendo hasta ahora.

Para extraer los datos pertenecientes a los estudiantes de FIng, es necesario acceder a una terminal conectada al SGB central, donde se accede con un "usuario FIng". Una vez dentro del sistema, se extraen datos de ciertas tablas de interés (con datos de cursos, alumnos, titulos, etc.) mediante consultas realizadas en lenguaje Oracle SQL, y con dicha información se actualiza una base de datos relacional armada a medida en la Unidad -utilizando SQLite, un gestor de base de datos mucho más liviano y adecuado para los propósitos específicos de la UEFI- para realizar consultas requeridas por distintos organismos dentro de Facultad (o incluso externos), tales como

[1] El estudiante queda identificado de manera única por su documento de identidad, sin importar a qué servicio(s) se anote.

[2] http://universidad.edu.uy/prensa/renderItem/itemId/40297/refererPageId/12

el avance por carreras, la detección de puntos críticos, la evaluación y seguimiento de nuevas modalidades de enseñanza, entre otros. Esta base de datos "local" es de donde proviene gran parte de la información de los alumnos ingresantes evaluados en este trabajo.

Por otra parte, los registros de las pruebas y cuestionarios (p.ej. HDI) figuran en tablas de datos aparte del sistema. Ambos conjuntos de datos son unidos mediante un identificador único por alumno, como el documento de identidad. En todos los casos se asegura que la información de una u otra fuente sea fidedigna realizando controles específicos.

4.1.2. Estudios preliminares

La idea principal del trabajo en sus comienzos se centró en responder *"¿Cómo predecir algún atributo de los estudiantes en el corto plazo?"* Para ello, se trabajó con datos históricos de estudiantes ingresantes a FIng, luego se ajustaron diferentes modelos a distintas propuestas de variables dependientes que fueron recabadas de estudios similares realizados por la Unidad y se crearon otras que podrían ser adecuadas, en función de la bibliografía consultada.

Esto añadió una gran complejidad a la hora de decidir tanto por modelos de predicción como por variables independientes y dependientes, ya que lo anterior no estaba implementado de manera simultánea[3]. Por esto se decidió en una segunda etapa del trabajo utilizar variables independientes comunes a todos los datos utilizados, además de elegir variables dependientes de tipo dicotómico -marcando presencia/ausencia de cierto atributo, numérico o categórico- para facilitar no solo la comparación entre distintos modelos sino también la puesta en práctica de los mismos en `R`.

Respecto a los datos utilizados, se decidió trabajar con dos generaciones: 2008 y 2016. La primera de ellas por algunos cambios importantes experimentados ese año (comienza la bisemestralización de cursos de primero, inician los primeros cursos de Cálculo 1 y Geometría y Álgebra Lineal 1 (GAL1) anuales), y la segunda por ser la más cercana sobre la cual se tenía información completa al comenzar este trabajo.

La información utilizada para ambos conjuntos de datos sería sociodemográfica y académica, como se describió en la Sección 1.1. En cuanto al CEAM –aplicado junto con la HDI–, la importancia de las variables dimensionales[4] fue marginal en prácticamente todos los casos previos consultados, decidiéndose dejarlas de lado para evitar introducir ruido a las predicciones[5].

Durante todo ese proceso previo se observó que los modelos tendían a predecir mejor a aquellos individuos con *ausencia* del atributo principal, logrando tasas de verdaderos negativos bastante más altas que de otras categorías. Esto, junto a bibliografía consultada que obtenía resultados

[3]Se llegó a trabajar con más de 100 variables independientes.

[4]Dimensiones definidas según la bibliografía especializada en Aprendizaje y Motivación. En la última versión de este cuestionario (2012) se constataron las dimensiones de Motivación (dividida en 4 subdimensiones), Estrategias de Aprendizaje y Lugar de Control (externo). Fuente: [Mí08]

[5]Durante parte del proceso previo se generaron fórmulas aditivas en base a un conjunto de tamaño medio de variables explicativas, generando así muchísimas fórmulas para probar su nivel predictivo, resultado esto en un proceso muy ineficiente, tanto en el tiempo insumido como en la memoria utilizada (durante la ejecución y luego de la misma).

similares ([CMP+15]) dio el puntapié para organizar de manera diferente tanto a los datos como a los modelos a tratar.

4.1.3. Variables de interés

Todas las variables de interés son binarias, indicando en cada caso la *presencia* del atributo que se mide; es decir si la variable toma el valor $Y_j = 1$ es porque el individuo posee la característica investigada.

La variable denominada "aprueba el 50 % (o más) de Cursos durante su primer año lectivo" (Y_1) surge de una propuesta realizada en un informe previo de la UEFI ([UEF09b]), en donde se buscó establecer la relación existente entre el resultado obtenido en la HDI y el rendimiento de los cursos de primer año, mediante modelos lineales, tanto simples como generalizados, mostrando así la utilidad de la HDI a la hora de predecir el comportamiento de los estudiantes en el primer año de su carrera[6]. Una de las variables a predecir utilizadas fue una indicadora denominada "Llegó50", tomando valor 1 si el estudiante aprueba el 50 % de los cursos a los cuales se anota durante su primer año de cursada. Si bien la creación de dicha variable fue arbitraria, se logró establecer una definición operativa del rendimiento en el corto plazo[7].

Por su parte, la variable denominada "llega al 50 % de Créditos durante su primer año lectivo" (Y_2) se crea de un modo similar a la variable "Llegó50", definida en este caso como la indicadora de si el estudiante llega a ganar el 50 % de los créditos acumulados de las seis asignaturas de primer año comunes a las carreras de grado de Agrimensura, Ingenierías Civil, en Computación, Eléctrica, Mecánica, Naval, en Producción, Química[8], y Licenciatura en Ciencias de la Atmósfera. Todas estas carreras tienen en sus currículas sugeridas a las asignaturas: Cálculo 1, Geometría y Álgebra Lineal 1, Física 1 (primer semestre) y Cálculo 2, Geometría y Álgebra Lineal 2 y Física 2 o Matemática Discreta 1[9] (segundo semestre). Se excluyen los tecnólogos en Cartografía, en Informática y en Mecánica, además de la Licenciatura en Ingeniería Biológica. Se consideró esta nueva variable al buscar un grupo de asignaturas que aglutine a la mayoría de estudiantes, dadas las dificultades encontradas por el comportamiento un tanto errático de las inscripciones de los alumnos, algo ya mencionado en [UEF09b].

Finalmente, la variable denominada *"activo al comenzar el segundo año"* (Y_3) es una indicatriz

[6] Para la mayoría de las asignaturas en los primeros años de cada carrera en FIng, la aprobación de las mismas puede ser al finalizar el curso si se llega a una calificación igual o mayor a 6 (en escala de 0 al 12), o bien realizando un examen, previa aprobación del curso de la asignatura (calificaciones entre 3 y 5).

[7] En el informe se menciona que definir una variable de rendimiento como ésta "no es fácil", ya que se desconocen "los motivos por los cuales (los estudiantes) se anotaron a pocas asignaturas" [UEF09b, pp.2]. La operatividad viene dada por el hecho de que, en general, más del 90 % de cada generación tiene actividad al ingresar a FIng, es decir que 9 de cada 10 alumnos que ingresan cursa al menos una asignatura durante su primer año lectivo.

[8] Sólo seis estudiantes que cursan Ingeniería en Alimentos fueron seleccionados; esto debido a que cursan más de una carrera dentro de FIng (ej. Computación); las asignaturas que se relevan son justamente de esas otras carreras

[9] Matemática Discreta 1 sustituye a Física 2 para los estudiantes de Computación; para el resto se utiliza Física 2.

que define si el estudiante declara actividad al comienzo del segundo año luego de su ingreso. Se optó por trabajar con esta definición ya que se corría riesgo de que la variable no sirviera como pasó previamente con otras opciones exploradas[10], quedando dentro de la restricción impuesta de trabajar en el corto plazo.

4.1.4. Preselección de variables explicativas

Criterio utilizado

Como la idea es buscar qué modelo se destaca del resto, es necesario realizar una preselección de variables acorde a todos ellos -y no a cada uno por separado. De este modo se procede:

1. *Variables comunes 2008-16*: solo se tomarán en cuenta aquellas variables explicativas presentes en *ambos* conjuntos de datos -y para cada variable de interés Y_j-, teniendo en cuenta que por los resultados previos algunas variables tenían un destaque errático (p.ej. el cuestionario CEAM como se mencionó más arriba); además fueron eliminadas:
   a) *Total de puntos en HDI*, porque se pretende observar el efecto del puntaje *por separado* de Comprensión Lectora y Matemática
   b) *Carrera*, porque la variable sufre cambios entre 2008 y 2016[11]
2. Con todas las variables de 1. se realiza una preselección de las 5 más importantes[12], mediante el ajuste a cada uno de los conjuntos de datos de fórmulas *aditivas*[13] `y~x1+x2+...` utilizando tres modelos (árboles a través de CART, Random Forests (RF), modelos lineales generalizados (GLM)), que tuvieran en sus implementaciones en `R` formas directas (CART, RF) o indirectas (GLM) del cálculo de la importancia de cada variable propuesta
   a) *Selección con GLM*: se parte de una fórmula del tipo $Y \sim \cdot$ (usando todas las variables explicativas), luego se observa el ajuste paso a paso (stepwise) hasta encontrar el mínimo valor del estadístico AIC, usando la función `stepAIC()` del paquete `MASS`
   b) *Selección con RF*: partiendo de un modelo saturado, se toman las 5 primeras variables según caída promedio en precisión (`MeanDecreaseAccuracy`), utilizando la función `varImpPlot()` del paquete `randomForest`
   c) *Selección con CART*: partiendo de un modelo completo, se contabiliza la información

[10] Por ejemplo, se trabajó con una indicatriz de aprobación de las asignaturas en Matemática hacia el fin del segundo año.

[11] Se *duplican* la cantidad de categorías posibles de una generación a otra

[12] Es importante recordar que este paso intermedio permite generar todas las combinaciones aditivas posibles de variables predictoras. Si se agrega una más al proceso, la cantidad de fórmulas posibles (2^n-1) pasa de 31 a 63; sin embargo los tiempos de cómputo se multiplican por 10 o más (ver Subsección 4.3.2), y las salidas de la función `asm2()` crecen de a razón de 4MB por cada fórmula ajustada

[13] Las fórmulas *aditivas* son un supuesto implícito sobre la relación funcional entre (X_i,Y_j); cada una de las variables independientes del modelo afectan de forma aditiva a la variable dependiente, y en particular sin interacción entre ellas.

sobre la "importancia por variable[14]", observando el contenido del modelo ajustado mediante el argumento `objeto$variable.importance`

Juntando cada una de las formas disponibles de decidir cuál variable es importante en su contexto, se crea un "índice" donde simplemente se cuenta la cantidad de veces que dicha variable aparece como importante para alguno de las combinaciones [variable dependiente, modelo ajustado, datos]:

Variable	#VecesImportante
HDIm	17
EdadIngreso	17
OrTip	8
HDIcl	8
OrLug	5

Estas cinco variables, puntaje en componente Matemática (*HDIm*) y de Comprensión Lectora (*HDIcl*) de la prueba, edad al ingreso, subsistema en donde terminó estudios de enseñanza media (*OrTip*) y lugar geográfico en donde se encuentra dicho centro (*OrLug*), serán las que formarán parte del resto del trabajo[15].

Datos faltantes e imputación

Se encontraron datos faltantes para alguna de las variables dependientes utilizadas. Como random forests es un modelo que no trabaja con variables que tengan datos perdidos, éstas fueron imputadas mediante el mecanismo "simple" descrito en la página 36. Comparando los resultados de la preselección de variables anterior no existen cambios sustanciales, con lo cual se continuó utilizando a estos datos imputados para el resto del trabajo, y para cada uno de los modelos citados.

4.2. Estadísticas Descriptivas

Si bien las propuestas de ambas pruebas fue diferente para los dos períodos considerados, los desempeños evaluados son los mismos en ambos casos. En suma: cambian algunos ítems, mas no así el *espíritu* de la prueba. En la Subsección 2.2.3 se mencionaron características generales; en lo que sigue se mostrarán algunos guarismos de interés sobre la prueba para las dos generaciones mencionadas, además de describir a nivel sociodemográfico a cada grupo.

[14] La idea general es, para cada nodo, observar qué variable fue usada para particionar; se guarda esa información y luego se "promedia" de algún modo, p.ej. mirando la función de pérdida y sus reducciones promedio

[15] Se decidió dejar fuera a la variable *Sexo* porque tenía el mismo nivel de importancia que *OrLug*, y por lo observado durante la etapa previa el sexo de cada estudiante no parece aportar mucha información predictiva sobre las variables a estudiar.

Procedencia de alumnos Comparando variables sociodemográficas de aquellos que realizan la prueba y los que no, se constatan diferencias en particular según dónde se realizaron los estudios preuniversitarios: los alumnos del subsistema Público son los que presentan menores tasas de participación histórica (63,7 %), mientras que los de UTU y de liceos privados tienen mayor participación (80,9 % y 88,8 %, respectivamente).

- *Diferencias 2008-16*: si bien las proporciones entre los que hacen y no hacen la prueba según lugar de origen se mantiene estable en el tiempo, para el caso de la generación 2008 hubo una proporción levemente menor de alumnos del interior que no realizaron la HDI
- *Similitudes 2008-16*: los porcentajes de no cobertura por generaciones son similares (29,6 % para 2008, 25,6 % para 2016), y en ambas generaciones hay patrones que se repiten: hay más mujeres (tanto sea del subsistema público como el privado), y hombres del interior que *no realizan la prueba* en estos dos períodos; esto parece estar en consonancia con la elección de carreras de estos alumnos

Cobertura por carrera Para las carreras "tradicionales" de FIng la proporción de estudiantes que hizo HDI supera el 80 % en ambos años (se considera solo el primer semestre de 2016), mientras que las carreras compartidas "tradicionales" (Ing. Química, Alimentos) y los tecnólogos tienen cifras bastante menores de cobertura[16]. Posibles razones de estas diferencias radican en los lugares de inscripción de dichos alumnos, en qué momento realizan su inscripción (si antes o después de la fecha de la prueba) e incluso la edad al ingreso: varias de las carreras no tradicionales mencionadas cuentan con estudiantes que han cursado otras carreras fuera de facultad, en general con un promedio mayor de edad que los ingresantes primarios.

Como diferencia a destacar, se observa para 2016 una caída importante para los ingresantes a Alimentos (de 9 % a 4,4 %) y un crecimiento importante de carreras no tradicionales respecto a 2008 (de 5,4 % a 14,7 %), explicado en gran medida por el aumento de oferta de carreras en los 8 años posteriores.

Descripción de variables predictoras: Para las variables numéricas escogidas -esto es, las tres que componen el resultado de HDI más la edad al ingreso- se observan diferencias puntuales, tal cual se aprecia en las tablas 4.2 y 4.3. La edad al ingreso presenta una relativa estabilidad entre ambos períodos, confirmando ciertamente un comportamiento similar en los últimos 20 años.

Por su parte, las variables provenientes de la HDI presentan matices. Por un lado, el puntaje de Comprensión Lectora es -sorprendentemente- bastante similar en ambas pruebas, mientras que para el Total la diferencia es marcada; esto se justifica por la diferencia en la cantidad de ítems y componentes: en 2008 había preguntas tanto de Física como de Química, mientras que en 2016 la componente de Física tenía más preguntas, que "sustituían" de algún modo a las de Química, que dejó de usarse desde ese año.

[16]La Licenciatura en Ingeniería Biológica es una excepción, pero incide también la baja cantidad de estudiantes que se inscribe (menos de 20 por año).

	Cobertura	
Carreras	HDI 2008	HDI 2016-I
Ing.I.Mecánica	92,4 %	85,9 %
Ing.Civil	91,5 %	84,4 %
Ing.Computación	90,2 %	83,7 %
Ing.Eléctrica	90,3 %	81,7 %
Agrimensura	81,8 %	87,0 %
Ing.Naval	71,4 %	86,7 %
Ing.Producción	0,0 %	84,8 %
Ing.Química	24,1 %	49,3 %
Lic.Ciencias Atmósfera	16,7 %	61,9 %
Tecnólogo Informático	11,8 %	52,5 %
Tecnólogo Mecánico	10,0 %	76,7 %
Ing.Alimentos	0,0 %	9,7 %
Lic.Ing.Biológica	----	100,0 %
Tecnólogo Cartografía	----	56,2 %
Total cobertura HDI	**70,4 %**	**73,4 %**

Tabla 4.1: Porcentaje de cobertura de HDI por carrera, para pruebas 2008 y 2016-I

Finalmente, el componente de Matemática presenta una suerte de "baja" en 2016 en relación a 2008: salvo los extremos, los valores de las cuantilas de 2016 están todos *por debajo* de los valores de 2008, a pesar de tener más ítems de evaluación -16 contra 14-, tal como se mencionó en la Tabla 2.1.

Generación 2008	Min.	p_{25}	p_{50}	Media	p_{75}	Max.	SD	n_{total}
HDI-Matemática	0,00	5,00	6,00	6,57	9,00	15,00	2,87	821
HDI-Comp.Lectora	0,00	3,00	4,00	3,87	5,00	5,00	1,05	821
HDI-Total	2,00	17,00	21,00	21,15	26,00	37,00	6,27	821
Edad al Ingreso	17,00	18,00	18,00	18,71	19,00	54,00	2,38	821

Tabla 4.2: Resumen estadístico generación 2008: variables numéricas

Generación 2016	Min.	p_{25}	p_{50}	Media	p_{75}	Max.	SD	n_{total}
HDI-Matemática	0,00	4,00	5,00	5,48	7,00	15,00	2,65	1208
HDI-Comp.Lectora	0,00	3,00	4,00	3,54	5,00	5,00	1,23	1208
HDI-Total	0,00	15,00	20,00	21,24	26,00	47,00	7,70	1208
Edad al Ingreso	17,40	18,20	18,60	20,07	19,92	51,20	3,85	1208

Tabla 4.3: Resumen estadístico generación 2016: variables numéricas

Por su parte, para las variables categóricas también se presentan algunas diferencias puntuales: por un lado aumenta la cantidad relativa de estudiantes mujeres en 2016; en cuanto a los subsistemas de origen de educación preuniversitaria los alumnos de liceos públicos caen casi en un 10 %, compensado esto en parte por el aumento de estudiantes de UTU y del exterior; final-

mente los alumnos del subsistema privado o mirados por lugar geográfico no presentan diferencias sustanciales, tal como lo muestra la Tabla 4.4.

		Gen 2008		Gen 2016	
	Categorías	f_i	%	f_i	%
Sexo	Masculino	643	78,3 %	917	75,9 %
	Femenino	178	21,7 %	291	24,1 %
Lugar de Origen	Montevideo	483	58,8 %	682	56,5 %
	Interior	313	38,1 %	459	38,0 %
	Otros	25	3,1 %	67	5,5 %
Sub-sis-te-ma	Público	453	55,2 %	529	43,8 %
	Privado	310	37,8 %	433	35,8 %
	UTU	52	6,3 %	184	15,2 %
	Otros	*6*	*0,7 %*	*62*	*5,1 %*

Tabla 4.4: Resumen estadístico: variables categóricas, ambas generaciones

Respecto a variables dependientes: Es importante mencionar que tanto para Y_1 como para Y_2 se excluyeron aquellos alumnos en carreras sin asignaturas comunes durante el primer año, como el caso de los Tecnólogos. Para Y_3 se consideraron todos los casos, ya que a priori la carrera escogida no influye sobre la posibilidad de abandono temprano. Si se considera solamente las seis asignaturas comunes a la absoluta mayoría de las carreras de grado durante el primer año (definidas en la Subsección 4.1.3), éstas son cursadas por entre el 57 % (2016) y el 73 % (2008) de los que realizan HDI.

En cuanto a diferencias sociodemográficas para cada variable dependiente, en base a distintos análisis se concluye:

- Que para Y_1:: *"estudiante aprueba el 50 % (o más) de Cursos durante su primer año lectivo"*, en ambos períodos se observan más alumnos de la Universidad del Trabajo del Uruguay (UTU) que no cumplen la condición respecto a las otras categorías
- Para Y_2:: *"estudiante llega al 50 % de Créditos durante su primer año lectivo"* por su parte, aquellos que sí presentan la característica son en proporción más alumnos provenientes del sector privado, con mayores calificaciones en el componente Matemática de HDI (HDIm), mientras que la ausencia del atributo en ambos períodos se presenta mayor medida para los alumnos provenientes de UTU

Se observó también que las correlaciones más altas entre todas las variables (siendo las Y_j consideradas como indicatrices numéricas) fueron del orden de 0,45 entre el componente Matemática de HDI y Y_2 para ambas generaciones. Si bien son correlaciones intermedias (ni altas ni bajas), dan claramente un indicio: parece que un buen desempeño en Matemática para la prueba inicial otorga chances importantes de lograr reunir al menos la mitad de los créditos -de las seis asignaturas mencionadas en la Subsección 4.1.3- durante el primer año.

	Generación 2008	Generación 2016
Iteraciones	30	30
Proporción ME	75 %	75 %
CART	$C_p = 0,01$	$C_p = 0,01827$
Random Forest	$m = 3, n_{arb} = 100$	$m = 2, n_{arb} = 100$
Boosting ($Bstg$)	$B = 100$	$B = 100$
Clasif.Bayesiano	$\hat{\alpha}_j^{(0)} = 0,5,\ j = 0;1$	$\hat{\alpha}_j^{(0)} = 0,5,\ j = 0;1$
SVM (RBF)	$\gamma = 0,5,\ C = 1$	$\gamma = 2,\ C = 0,5$
Casos considerados	$n_{2008} = 814$	$n_{2016} = 694$

Tabla 4.5: Parámetros de corridas para Y_1 "Aprueba al menos 50 % cursos de primer año", obtenidos con optimización utilizando la función `ajp()`

Es importante también aclarar que al no existir grandes desbalances en las posibles etiquetas por variables de interés (la menor proporción ronda el 20 %), no es necesario hacer uso de herramientas específicas para balancear proporciones[17].

4.3. Análisis

El proceso común será -como ya se mencionó- la selección de las cinco mejores fórmulas por variable dependiente y generación según el ranking establecido en la Sección 3.3.4, de éstas se estudiarán los valores de los estadísticos "positivos" de las matrices de confusión[18] y se calcularán las curvas ROC y el área debajo de ella (AUROC). Finalmente, se estudiarán aciertos y errores de los modelos utilizados en estas fórmulas mediante ACM, también explicado en la Sección 3.3.4.

4.3.1. Predicción del Rendimiento

Variable "Llegar al 50 % de los cursos de primer año" (Y_1)

Durante toda esta sección, y considerando todos los modelos ajustados, los parámetros obtenidos por optimización mediante la función `ajp()` y algunos de los parámetros iniciales fueron los que figuran en la Tabla 4.5[19].

Ranking de mejores fórmulas para Y_1 Según se observa en la Tabla 4.6 las cinco fórmulas con mejor desempeño (como se fundamentó en la Sección 3.3.4, utilizando la mediana de las posiciones

[17] Como por ejemplo técnicas de remuestreo para aumentar la cantidad de casos de la clase menos frecuente (SMOTE), como proponen Rovira y cols. ([RPI17]) para balancear los datos usados.

[18] Al ser las variables de interés dicotómicas en todos los casos, los valores "negativos" se obtienen restando 1 al valor correspondiente.

[19] Para el Clasificador Bayesiano los de la tabla fueron los parámetros de inicio en una primera instancia; como luego los predictores con dichos valores tendían a generar mucho más falsos positivos que el resto de los modelos, en una segunda etapa se utilizaron como parámetros las probabilidades *a priori* para cada clase, en cada una de las muestras de testeo.

de especificidad, sensibilidad, precisión, etc. para todas las ejecuciones) tienen en común -y para ambas generaciones- a las variables "componente Matemática de HDI" (HDIm) y la "edad al ingreso" (EdIng) en todos los casos; alternan en su presencia las variables "subsistema o tipo de centro educativo donde culminó estudios secundarios" (OrTip) y "lugar geográfico donde culminó estudios secundarios" (OrLug), mientras que el "componente de Comprensión Lectora" (HDIcl) aparece en tan sola una de las fórmulas[20].

Datos	Posición	Fórmula	HDIm	HDIcl	EdIng	OrLug	OrTip	Frec.Abs.
2008	1	20	√		√		√	8
	2	27	√		√	√	√	8
	3	14	√		√			7
	4	23	√		√	√		6
	5	31	√	√	√	√	√	6
2016	1	26	√		√	√	√	6
	2	31	√	√	√	√	√	6
	3	22	√		√	√		6
	4	19	√		√		√	5
	5	13	√		√			4

Tabla 4.6: Ranking Y_1 "Aprueba al menos 50 % cursos de primer año": cinco mejores fórmulas

La Tabla 4.6 -al igual que la Tabla 4.10- simplifican la notación de fórmula: se leen horizontalmente, y dada una variable dependiente Y_j se indica mediante √ que variable(s) independiente(s) la acompañan. Así, la fórmula 20 para Y_1 usando datos de la generación 2008 en la Tabla 4.6 es `y1~HDIm+EdIng+OrTip`.

Performance predictiva Para no generar confusión, se separarán los resultados por generación.

Para generación 2008 Si se observan los resultados de resumen de los distintos indicadores de las matrices de confusión generadas -recordemos que son promedios respecto a las 30 muestras (ME, MT) utilizadas para la validación cruzada de cada modelo-, la mediana de la sensibilidad de los modelos con mejor nivel (los de consenso) está entorno de 0,83, siendo muy parejos los resultados entre ellos. Por su parte, la tasa de valores predictivos positivos (VPP) para los modelos con mejor nivel es de 0,70, destacándose GLM y SVM. Se puede decir que los modelos de consenso son más precisos para detectar a aquellos individuos que llegan a aprobar la mitad de sus cursos, mientras que GLM y SVM asignan de forma más certera a sus predicciones de presencia del atributo.

[20]En dicha tabla, *Frec.Abs* es la cantidad de veces que i-ésima fórmula aparece entre los primeros 5 ordenamientos, como se explicó en la Sección 3.3.4.

		Modelos simples						Modelos de consenso		
		Bstg	CART	CDB	GLM	RF	SVM	MP	VM	WA
Formula-20	Especificidad	0,621	0,59	0,34	**0,667**	0,601	0,638	0,593	0,598	0,598
	VPP	0,696	0,686	0,606	0,708	0,688	**0,71**	0,696	0,694	0,694
	Sensibilidad	0,774	0,808	**0,946**	0,756	0,762	0,783	0,797	0,816	0,816
	VPN	0,722	0,738	**0,848**	0,713	0,7	0,736	0,739	0,755	0,757
	Precisión	0,7	0,7	0,644	0,713	0,693	**0,718**	0,71	0,713	0,713
	AUROC	0,763	0,732	0,766	**0,781**	0,741	0,768	0,75	0,75	0,773
Formula-27	Especificidad	0,628	0,59	0,33	**0,657**	0,599	0,651	0,591	0,616	0,612
	VPP	0,7	0,686	0,609	0,701	0,692	**0,708**	0,69	0,702	0,704
	Sensibilidad	0,778	0,806	**0,947**	0,758	0,795	0,781	0,827	0,829	0,829
	VPN	0,72	0,738	**0,846**	0,715	0,722	0,724	0,756	0,762	0,762
	Precisión	0,71	0,703	0,644	0,706	0,7	0,718	0,71	**0,72**	**0,72**
	AUROC	0,761	0,732	0,766	**0,781**	0,754	0,769	0,752	0,752	0,779
Formula-14	Especificidad	0,593	0,59	0,308	**0,662**	0,58	0,61	0,568	0,59	0,582
	VPP	0,688	0,686	0,605	**0,702**	0,68	0,691	0,682	0,692	0,692
	Sensibilidad	0,81	0,807	**0,962**	0,752	0,812	0,82	0,835	0,822	0,822
	VPN	0,748	0,734	**0,868**	0,713	0,748	0,749	0,76	0,747	0,747
	Precisión	0,706	0,703	0,639	0,708	0,686	**0,713**	0,706	0,703	0,703
	AUROC	0,763	0,732	0,765	**0,779**	0,742	0,769	0,746	0,746	0,772

Tabla 4.7: Tabla de comparación Y_1: mediana de indicadores seleccionados, para tres mejores fórmulas aditivas para la generación 2008, según Tabla 4.6

Por otra parte, la mediana de especificidad en los mejores casos -GLM-logit y SVM- es de 0,65, y la de valor predictivo negativo es de 0,75 -para los modelos de consenso-; cabe destacar que el clasificador bayesiano (CDB) obtuvo una mediana más alta (0,84), pero con performances alejadas de los mejores modelos para el resto de los indicadores mencionados. Consecuentemente, se puede decir que (al igual que en el párrafo anterior) los modelos GLM y SVM detectan más a los alumnos que no llegan a aprobar la mitad de sus UCs cursadas, mientras que los modelos de consenso asignan con más precisión a esos individuos a la categoría correcta. Curiosamente, ninguno de los métodos basados en árboles de clasificación pudo siquiera destacarse en alguno de los guarismos anteriores.

Observando lo que dice el estadístico AUROC promedio para cada una de las fórmulas mencionadas, y lo que muestran las curvas ROC para la fórmula con mayor destaque en este sentido (mostrado en la Figura 4.1), sin dudas el modelo GLM-Logit es el que tiene mejor desempeño, dejando en claro que es el modelo que mejor discrimina (esto es, clasifica correctamente a quiénes poseen o no la característica) para estos datos.

Para generación 2016 Para estos datos en cambio, la sensibilidad aumenta notoriamente a valores casi de perfección (una mediana de 0,97 para SVM en el mejor de los casos) y valores predictivos positivos más moderados (en el entorno de 0,77 para los modelos de consenso), aunque con valores de especificidad y predictivos negativos mucho más bajos que para 2008 (0,26 para Boosting y el Clasificador Bayesiano en el primer caso, 0,54 para el Clasificador Bayesiano en

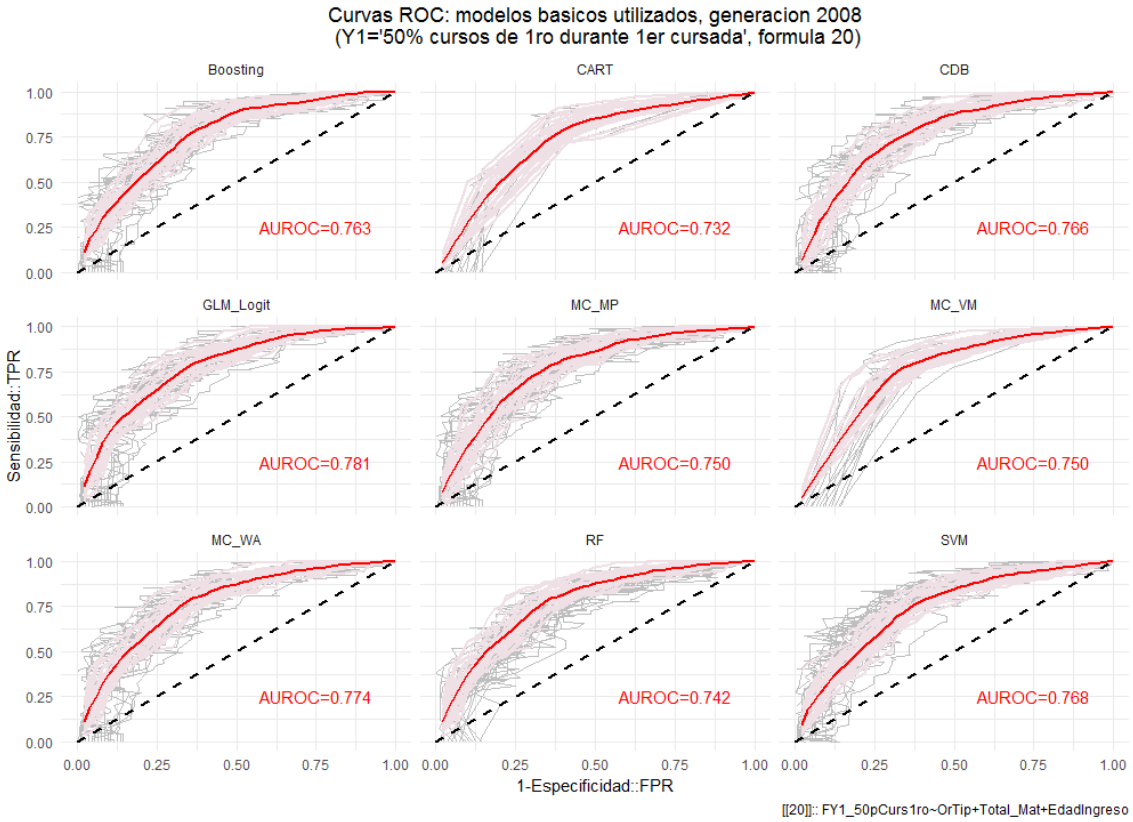

Figura 4.1: Curvas ROC (todas, promedio), generación 2008, fórmula 20 para Y_1 , todos los modelos utilizados

el segundo). La precisión general con mejor mediana la logra el modelo CART con 0,74, aunque superando apenas (aumento de $0,6\,\%$ de precisión promedio) al denominado *clasificador nulo*. Puede decirse que hay una mayor precisión en mediana para esta variable y esta generación, pero a costa de empeorar otros indicadores; en este caso *apenas* se logran mejoras respecto a utilizar el clasificador nulo. Una posible explicación es que efectivamente las generaciones son diferentes en cuanto a comportamientos relacionados con qué y cuándo cursar determinadas asignaturas.

Nuevamente como en el caso anterior, los modelos basados en árboles de clasificación presentan los valores más bajos en los estadísticos tomados de las matrices de confusión. Particularmente para estos datos, Random Forest presenta los valores más magros, aunque SVM tiene un comportamiento errático para la especificidad y los VPNs.

También como para los datos de 2008, GLM logra el mejor resultado en el estadístico AUROC promedio, como se observa en la Tabla 4.8.

En definitiva, para ambos grupos de datos y modelos ajustados se observa una alta sensibilidad (superior al $80\,\%$) y valores predictivos positivos altos (superiores al $75\,\%$), lo cual se corresponde en cierta manera con la prevalencia de estos grupos. Ahora bien, para el caso de aquellos estudiantes que no presentan la característica relevada, los guarismos predictivos son bastante diferentes entre las generaciones estudiadas: en el entorno del $70\,\%$ para 2008, y muy por debajo de este valor para 2016. La diferencia más notoria se da en la mejora de los modelos aplicados respecto

		Modelos simples						Modelos de consenso		
		Bstg	CART	CDB	GLM	RF	SVM	MP	VM	WA
Formula-26	Especificidad	**0,266**	0,172	0,248	0,112	0,228	0,042	0,183	0,23	0,202
	VPP	0,769	0,77	0,774	0,758	0,766	0,747	0,772	**0,776**	0,772
	Sensibilidad	0,863	0,937	0,905	0,962	0,894	**0,978**	0,938	0,912	0,929
	VPN	0,393	**0,5**	0,463	0,487	0,41	**0,5**	**0,5**	0,474	**0,5**
	Precisión	0,71	**0,744**	0,724	0,736	0,716	0,741	0,742	0,738	0,741
	AUROC	0,644	0,666	0,689	**0,7**	0,657	0,39	0,605	0,605	**0,7**
Formula-31	Especificidad	**0,311**	0,16	0,108	0,115	0,24	0,022	0,125	0,185	0,16
	VPP	**0,78**	0,764	0,756	0,758	0,768	0,744	0,758	0,768	0,765
	Sensibilidad	0,838	0,938	0,964	0,952	0,89	**0,989**	0,96	0,938	0,952
	VPN	0,405	0,484	**0,542**	0,406	0,434	0,4	0,5	0,514	0,536
	Precisión	0,701	0,741	0,741	0,736	0,721	0,733	0,741	**0,747**	**0,747**
	AUROC	0,644	0,662	0,694	0,699	0,666	0,383	0,61	0,61	**0,701**
Formula-13	Especificidad	0,194	0,134	0,194	0,07	**0,248**	0,063	0,136	0,198	0,176
	VPP	0,765	0,755	**0,77**	0,748	0,762	0,744	0,76	**0,77**	0,767
	Sensibilidad	0,897	0,96	0,934	0,976	0,862	**0,977**	0,946	0,931	0,938
	VPN	0,384	0,5	0,526	0,472	0,386	0,5	0,464	0,489	**0,512**
	Precisión	0,716	0,736	**0,747**	0,73	0,701	0,741	0,738	0,738	**0,747**
	AUROC	0,652	0,651	0,694	**0,704**	0,626	0,357	0,595	0,595	0,689

Tabla 4.8: Tabla de comparación Y_1: mediana de indicadores seleccionados, para tres mejores fórmulas aditivas para la generación 2016, según Tabla 4.6

a un clasificador nulo (alta para 2008, apenas perceptible para 2016).

Aciertos y errores generales Como se aprecia en la Figura 4.2 donde se representa el plano factorial principal (de modo similar como lo muestra la tercera fila de la Figura 3.4), el primer eje separa errores de aciertos de las distintas fórmulas. Por su parte, el segundo eje separa las categorías sociodemográficas que se asocian más comúnmente a ciertos comportamientos: 'Montevideo' y 'Privado' se asocia a llegar al umbral de la mitad de cursos realizados durante el primer año, y en mucho menor medida con los resultados en HDI (tanto en Matemática como en Lectura). Este segundo eje presenta la peculiaridad de agrupar a casi todas las modalidades mencionadas sobre el valor cero para el eje de abscisas, lo cual manifiesta una casi nula relación entre los errores y aciertos que realizan los modelos con estas otras variables. Respecto a las variables consideradas como suplementarias en este análisis, se observa una baja correlación entre los ejes factoriales y las variables numéricas, aunque el sentido es el mismo que puede ser visto para el resto de los análisis de correspondencia en este trabajo.

De esto -y de los resultados que se verán más adelante- se puede observar como para 2008 el comportamiento entre la variable a predecir y las independientes fue al menos "extraño" en relación a los restantes análisis.

Para 2016 en cambio, hay algunas diferencias con el caso anterior: si bien el primer eje separa nuevamente aciertos de errores, esta vez el sentido es contrario que para lo observado en 2008. El segundo eje aleja solamente a las categorías 'Exterior' y 'Sin Datos' relacionadas con los estudios

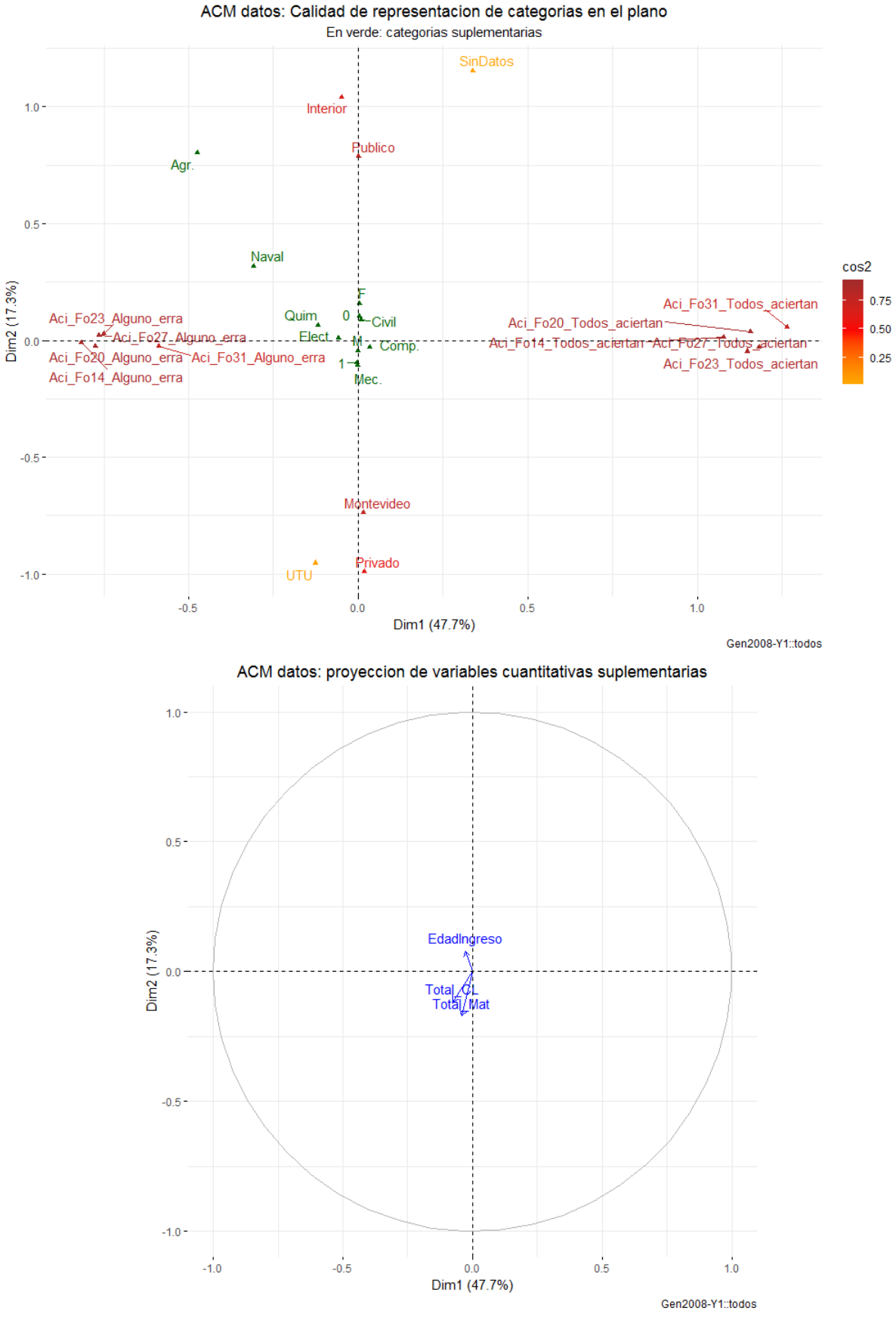

Figura 4.2: ACM para Y_1: resultados para generación 2008

	Generación 2008	Generación 2016
Iteraciones	30	30
Proporción ME	75 %	75 %
CART	$C_p = 0,01092$	$C_p = 0,01076$
Random Forest	$m = 3, n_{arb} = 100$	$m = 2, n_{arb} = 100$
Boosting ($Bstg$)	$B = 100$	$B = 100$
Clasif.Bayesiano	$\hat{\alpha}_j^{(0)} = 0,5,\ j = 0;1$	$\hat{\alpha}_j^{(0)} = 0,5,\ j = 0;1$
SVM (RBF)	$\gamma = 0,5,\ C = 2$	$\gamma = 0,5,\ C = 0,5$
Casos considerados	$n_{2008} = 658$	$n_{2016} = 694$

Tabla 4.9: Parámetros de corridas para Y_2 "Llegar al 50 % de los créditos de las seis asignaturas comunes en primer año", obtenidos con optimización utilizando la función `ajp()`

preuniversitarios de los estudiantes; esto parece tener relación con la edad al ingreso de dichos estudiantes (levemente mayores que el promedio), como muestra la Figura 4.3.

A diferencia del caso anterior, para 2016 la asociación entre aciertos de los modelos, buenos resultados en HDI (particularmente en Matemática) y algunas categorías de los estudios preuniversitarios (como 'Privado' y 'Montevideo') es destacable, aunque la calidad de representación en el plano para estas últimas es baja.

Se puede decir que existen diferencias entre las dos generaciones consideradas, y consecuentemente en los resultados de las predicciones de la variable de interés, poniendo en duda la utilidad de esta variable como una posible forma de medir el rendimiento en el corto plazo para los alumnos de FIng.

Variable "Llegar al 50 % de los créditos de las seis asignaturas comunes en primer año" (Y_2)

Para lo que sigue, y considerando todos los modelos ajustados, los parámetros sugeridos por la función `ajp()` y algunos de los parámetros iniciales fueron los que indica la Tabla 4.9[21]:

Ranking de mejores fórmulas para Y_2 La Tabla 4.10 muestra las cinco "mejores fórmulas" como ya se explicó al comienzo de este capítulo. A diferencia de lo ocurrido para la variable Y_1, y si bien las variables 'HDIm' y 'EdIng' son prácticamente omnipresentes, la presencia de las restantes es un tanto diferente: 'HDIcl' aparece solo una vez y 'EdIng' es sustituida por 'OrLug' en una de las fórmulas de los datos de 2016, 'OrTip' aparece solo en dos de las fórmulas y 'OrLug' tiene una presencia más en 2016.

[21] Nuevamente, para el Clasificador Bayesiano los de la tabla fueron los parámetros de inicio en una primera instancia, detalle ya comentado para la Tabla 4.5.

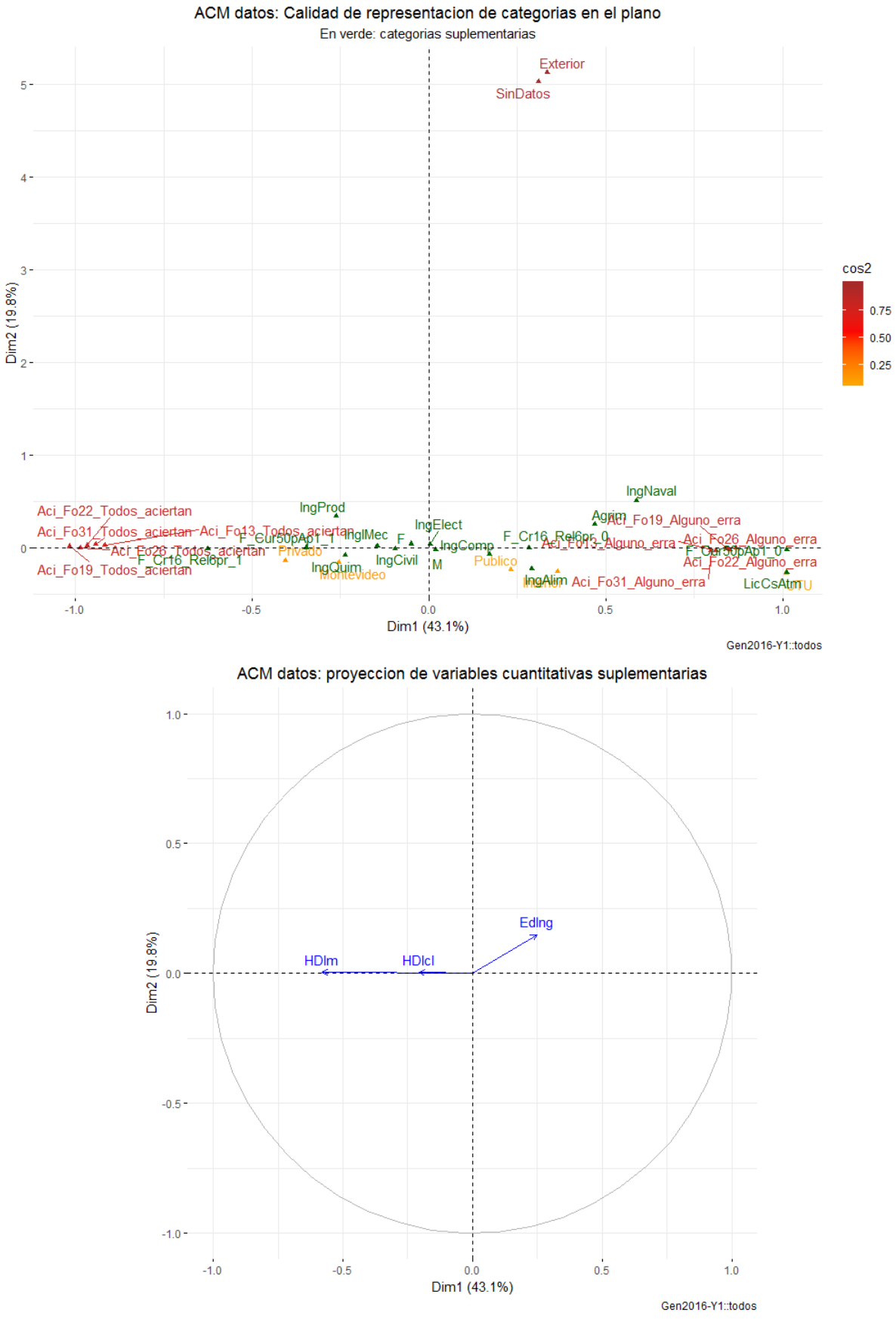

Figura 4.3: ACM para Y_1: resultados para generación 2016

Datos	Posición	Fórmula	HDIm	HDIcl	EdIng	OrLug	OrTip	Frec.Abs.
2008	1	22	√		√	√		7
	2	26	√		√	√	√	7
	3	13	√		√			4
	4	19	√		√		√	4
	5	25	√	√	√			4
2016	1	19	√		√		√	7
	2	22	√		√	√		7
	3	26	√		√	√	√	5
	4	11	√			√		4
	5	13	√		√			4

Tabla 4.10: Ranking Y_2 "Aprueba al menos 50 % créditos de las 6 'asignaturas comunes' en primer año": cinco mejores fórmulas

Performance predictiva Nuevamente para evitar confusiones, se separarán los resultados por generación.

Para generación 2008 Para estos datos la sensibilidad es baja (mediana de 0,48), destacándose los modelos de consenso de forma pareja. Los valores predictivos positivos por su parte logran resultados más altos (mediana de 0,63 para los modelos de consenso). En cambio, es notable el nivel predictivo para los valores negativos: la especificidad alcanza en mediana 0,94 para SVM, seguido muy de cerca por GLM y los modelos de consenso WAAUC y MP, mientras que los valores predictivos negativos están en el entorno de 0,83 para los modelos de consenso (agregando 'OrTip' mejora levemente VPN; vale decir que el CDB presenta una mediana de 0,864 pero tiene un peor comportamiento en otros indicadores, seguramente por tener valores más dispersos). La precisión general está en el entorno de 0,78 para los modelos de consenso, lo cual implica una mejora predictiva de 4,6 % en promedio respecto al clasificador nulo.

Respecto a los resultados en las curvas ROC y los promedios de AUROC verificados, nuevamente es GLM-Logit el que tiene el valor más alto de todos, seguido de cerca por el modelo de consenso WAAUC, mostrando que ambos son los modelos que mejor discriminan a ambas poblaciones para estos datos.

Para generación 2016 Nuevamente se repiten patrones observados en el apartado anterior: la sensibilidad nuevamente es baja (mediana de 0,51, con destaque para Boosting, GLM y VM) y los valores predictivos positivos se encuentran en el entorno de 0,66 para SVM y los modelos de consenso. También para la generación 2016, se da que los modelos predicen mejor a aquellos con ausencia del atributo ($Y_2 = 0$): la especificidad mediana es de 0,92 para SVM y los modelos de consenso WAAUC y MP, mientras que los valores predictivos negativos están en el entorno de 0,80 para VM y GLM (ocurre de modo similar que el Clasificador Bayesiano tiene valores más altos pero una performance peor para el resto de los indicadores). La precisión en este caso mejora un 7 % en promedio respecto al "clasificador nulo", resultando destacable los modelos de

		Modelos simples						Modelos de consenso		
		Bstg	CART	CDB	GLM	RF	SVM	MP	VM	WA
Formula-22	Especificidad	0,906	0,904	0,752	**0,919**	0,9	0,94	0,906	0,889	0,91
	VPP	0,628	0,629	0,504	0,59	0,61	0,632	**0,636**	0,62	**0,636**
	Sensibilidad	0,431	0,438	**0,639**	0,359	0,424	0,281	0,452	0,488	0,45
	VPN	0,819	0,826	**0,864**	0,804	0,824	0,796	0,828	0,832	0,828
	Precisión	0,776	0,779	0,73	0,77	0,776	0,776	**0,788**	**0,788**	**0,788**
	AUROC	0,791	0,72	0,803	**0,807**	0,762	0,753	0,733	0,733	0,803
Formula-13	Especificidad	0,913	0,918	0,791	0,921	0,92	**0,934**	0,923	0,906	0,92
	VPP	0,612	0,64	0,516	0,621	0,612	**0,642**	0,636	0,612	0,631
	Sensibilidad	0,375	0,409	**0,628**	0,39	0,35	0,322	0,398	0,428	0,404
	VPN	0,81	0,821	**0,86**	0,812	0,811	0,804	0,82	0,821	0,82
	Precisión	0,767	**0,776**	0,742	0,773	0,761	0,773	**0,776**	**0,776**	**0,776**
	AUROC	0,796	0,719	0,803	**0,805**	0,735	0,753	0,729	0,729	0,804
Formula-25	Especificidad	0,899	0,904	0,848	0,922	0,894	**0,937**	0,914	0,902	0,912
	VPP	0,596	0,619	0,558	0,623	0,546	0,604	**0,638**	0,621	0,633
	Sensibilidad	0,438	0,436	**0,542**	0,384	0,398	0,27	0,434	0,444	0,436
	VPN	0,827	0,824	**0,841**	0,812	0,814	0,795	0,826	0,83	0,826
	Precisión	0,767	0,773	0,761	0,77	0,764	0,755	**0,782**	**0,782**	**0,782**
	AUROC	0,779	0,718	0,785	**0,802**	0,748	0,738	0,721	0,721	0,798

Tabla 4.11: Tabla de comparación Y_2: mediana de indicadores seleccionados, para tres mejores fórmulas aditivas para la generación 2008, según Tabla 4.10

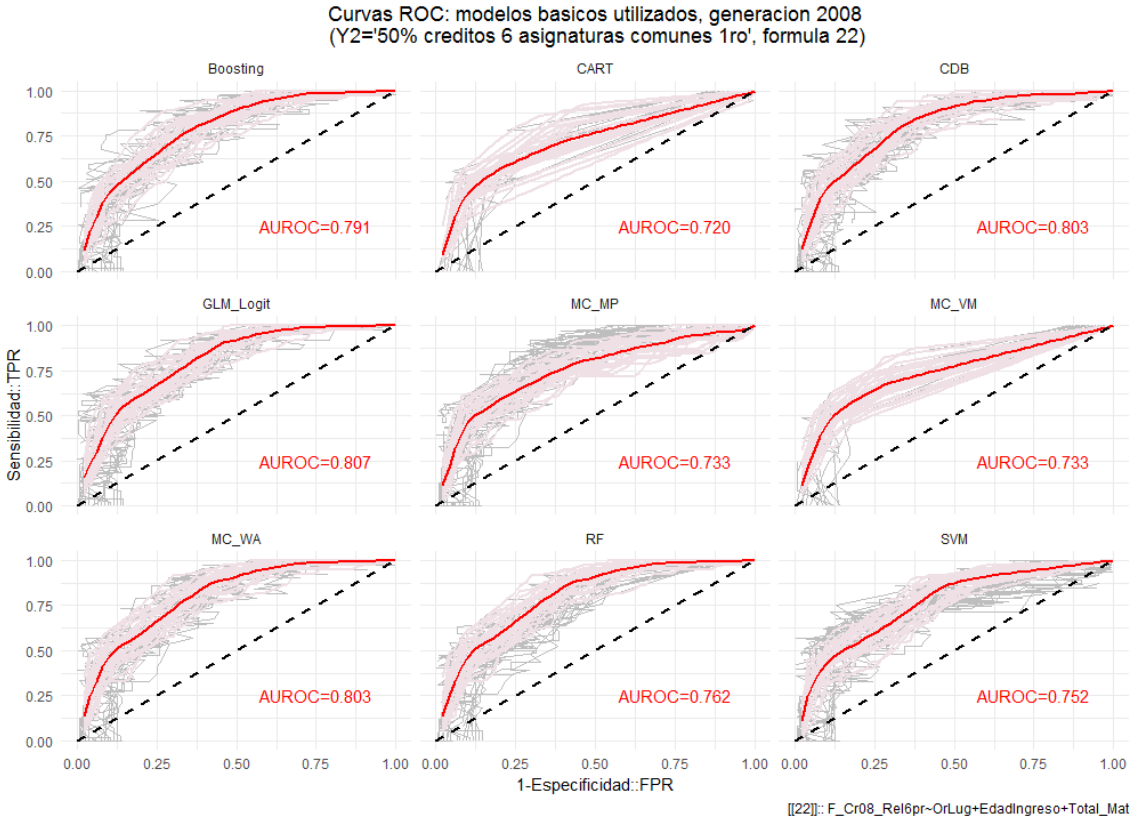

Figura 4.4: Curvas ROC (todas, promedio), generación 2008, fórmula 22, todos los modelos utilizados

		Modelos simples						Modelos de consenso		
		Bstg	CART	CDB	GLM	RF	SVM	MP	VM	WA
Formula-19	Especificidad	0,883	0,917	0,883	0,892	0,874	0,874	**0,923**	0,911	0,921
	VPP	0,637	0,663	0,654	0,646	0,613	0,613	0,69	0,683	**0,7**
	Sensibilidad	0,482	0,424	0,462	**0,5**	0,466	0,466	0,436	0,491	0,458
	VPN	0,795	0,788	0,8	0,804	0,799	0,799	0,79	**0,806**	0,794
	Precisión	0,764	0,764	0,762	0,775	0,749	0,749	0,772	**0,778**	0,775
	AUROC	0,754	0,699	0,771	0,795	0,742	0,742	0,74	0,746	**0,8**
Formula-22	Especificidad	0,895	0,923	0,899	0,893	0,889	0,889	**0,933**	0,913	0,928
	VPP	0,637	0,671	0,648	0,654	0,639	0,639	**0,71**	0,691	0,7
	Sensibilidad	0,474	0,404	0,456	**0,5**	0,492	0,492	0,405	0,469	0,434
	VPN	0,79	0,789	0,793	0,798	0,799	0,799	0,786	**0,8**	0,791
	Precisión	0,764	0,767	0,767	0,772	0,767	0,767	0,772	**0,78**	0,778
	AUROC	0,752	0,701	0,764	**0,789**	0,735	0,735	0,739	0,731	0,78
Formula-26	Especificidad	0,873	0,921	0,876	0,885	0,901	0,901	**0,923**	0,888	0,918
	VPP	0,62	0,667	0,624	0,626	0,653	0,653	**0,691**	0,645	0,682
	Sensibilidad	0,481	0,405	0,496	0,5	0,466	0,466	0,419	**0,524**	0,45
	VPN	0,8	0,788	0,806	0,804	0,8	0,8	0,786	**0,808**	0,79
	Precisión	0,759	0,764	0,764	0,764	0,772	0,772	0,77	**0,778**	0,77
	AUROC	0,74	0,698	0,765	**0,794**	0,739	0,739	0,733	0,732	0,779

Tabla 4.12: Tabla de comparación Y_2: mediana de indicadores seleccionados, para tres mejores fórmulas aditivas para la generación 2016, según Tabla 4.10

consenso VM y MP.

La curva ROC por su parte -en la Figura 4.5 se muestra la correspondiente a la fórmula 19- muestra al AUROC más alto, obtenido por el modelo de consenso WAAUC. Es destacable igualmente el comportamiento -nuevamente- del modelo Logit, apreciado en la Tabla 4.12 ocupando el segundo lugar (o incluso el primero) para este estadístico en varias de las fórmulas mencionadas.

Se puede concluir entonces que, para ambas generaciones, los resultados para la variable "llega al 50 % de los créditos de las 6 UCs comunes en primer año" se parecen mucho: altos valores de especificidad y valores predictivos negativos, precisión por encima del 75 % y una ganancia de entre 4 y 7 puntos porcentuales de precisión respecto al clasificador nulo. Se destacan en prácticamente todos los casos los modelos SVM, GLM y los de consenso. Como problemas se observan la baja sensibilidad (problemas para detectar a aquellos que poseen la característica) y mayor variabilidad para varios de los indicadores utilizados para ambas generaciones.

Aciertos y errores generales Para la generación 2008, se observa en la Figura 4.6 que el primer eje factorial separa aciertos de errores como en los casos anteriores. Por su parte, el segundo eje separa las categorías 'Privado' y 'Montevideo' de 'Publico' e 'Interior', que vuelven a estar cercanas en esta proyección. En este caso parece haber más asociación entre los aciertos y errores y alguna de las variables de estudios preuniversitarios, particularmente por la posición de las proyecciones de estas modalidades algo más alejadas del eje de ordenadas del plano, a

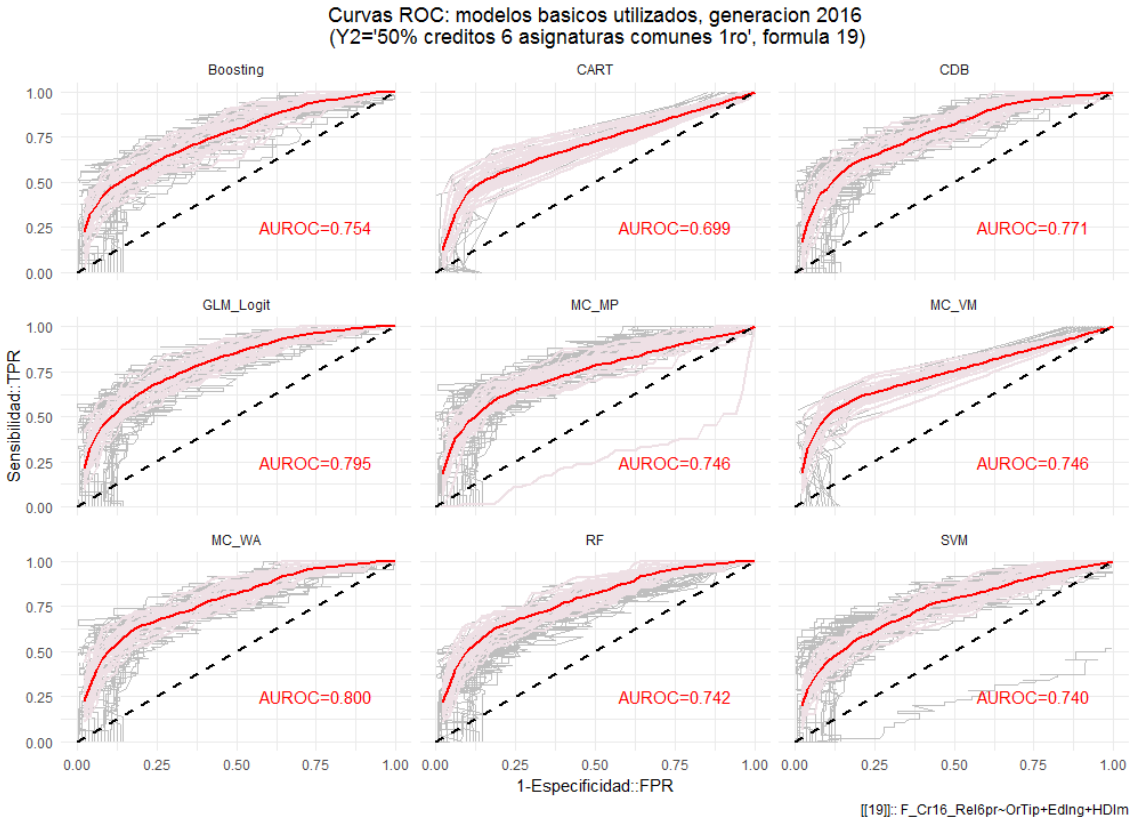

Figura 4.5: Curvas ROC (todas, promedio), generación 2016, fórmula 19, todos los modelos utilizados

diferencia de lo observado en la Figura 4.2.

En cuanto a la variable de interés, el acierto general está claramente asociado a la categoría de ausencia del atributo, es decir a los que no llegan a acumular la mitad de los créditos en esas seis asignaturas comunes, además de a resultados bajos en HDI (particularmente en Matemática). Aparecen también algunas modalidades poco frecuentes (como las carreras 'Agrimensura' e 'Ingeniería Naval') pero esto es parte de cómo las modalidades raras juegan papeles más preponderantes en los primeros ejes factoriales.

En el caso de la generación 2016, el primer eje separa -como es esperable- errores de aciertos. Nuevamente como se observó para Y_1, el segundo eje aleja a las categorías 'Exterior' y 'Sin Datos' relacionadas con los estudios preuniversitarios de los estudiantes, asociado esto de algún modo a la edad a la que ingresan, como muestra la Figura 4.7.

En comparación con lo que sucede para los cursos aprobados, con los créditos de las seis asignaturas comunes el camino es inverso: los aciertos se asocian en mayor medida a los alumnos con menor performance en HDI (más que nada en Matemática), además de con algunas categorías de la etapa previa a la universidad (como 'UTU', 'Público' o 'Interior'), que son en definitiva las categorías más comunes donde aparecen los resultados negativos en la variable estudiada, como se mencionó al comienzo de este capítulo.

Lo que parece común a todos los datos es la asociación de los aciertos totales de los modelos a características más frecuentes para cada variable estudiada. Esto particularmente es mucho más

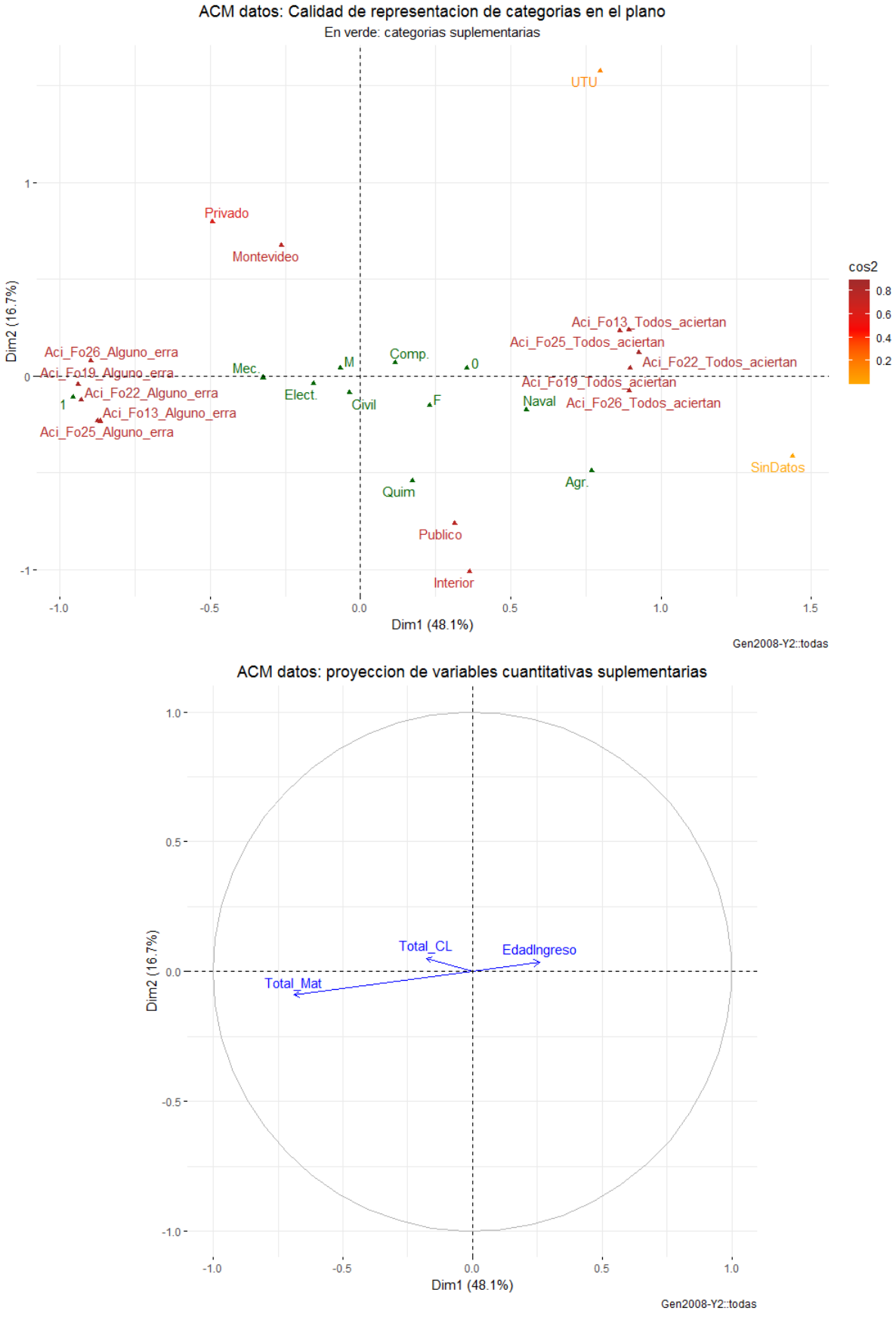

Figura 4.6: ACM para Y_2: resultados para generación 2008

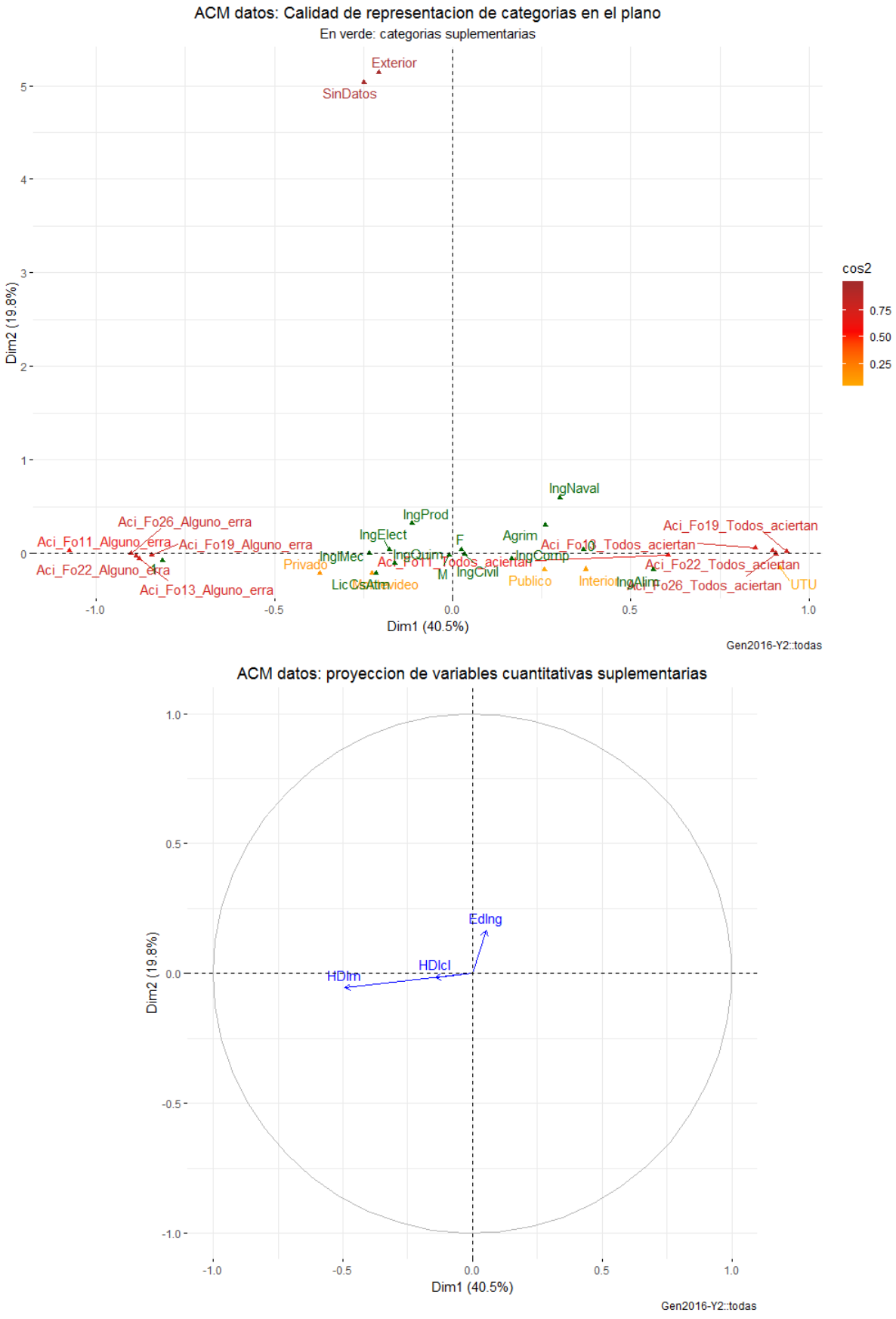

Figura 4.7: ACM para Y_2: resultados para generación 2016

marcado y consistente con la variable denominada Y_2: aquellos grupos que presentan mayores tasas de ausencia del atributo son las que los diferentes modelos tienden a acertar con mayor precisión, sin importar cuál de ellos lo haga. La discrepancia parecería venir desde los datos "raros": aquellos alumnos que "salen de la norma" -para cada variable estudiada- son los que los modelos asignan con la etiqueta errónea con mayor frecuencia.

4.3.2. Resultados adicionales

Predicción de la Desvinculación en el corto plazo

Ninguno de los alumnos que realiza HDI en ambas generaciones consideradas presenta "abandono inmediato" (es decir, se anota a la facultad y luego no realiza ninguna actividad más); todos ellos cursan al menos un semestre en la institución. Teniendo en cuenta además que son muy pocos los que hacen solo un semestre (para 2016 10 %, y para 2008 menos del 1 %), se considerará exclusivamente como variable de actividad la que indica si el estudiante está activo *al comienzo de su segundo año* en la institución.

Predicción para Y_3 Los parámetros utilizados por los modelos para las corridas de esta variable fueron prácticamente los mismos que los vistos en la Tabla 4.5. Luego de observar qué ocurría para algunas de las fórmulas sugeridas para las otras variables dependientes consideradas -que contenían variables predictoras comunes, como p.ej. el puntaje en Matemática de la HDI y la edad al ingreso-, y observando que los cambios eran mínimos entre las diferentes combinaciones de fórmulas aditivas, se decidió no proseguir con el resto del protocolo de análisis utilizado para las variables Y_1 y Y_2.

Para esta variable y para ambas generaciones, las predicciones en mediana rara vez superan a la clase mayoritaria, sin importar el modelo utilizado. En prácticamente todos los casos, estos modelos predicen mal a los que no tienen condición -o sea, ser inactivo antes de comenzar el segundo año. Esto puede deberse entre otras cosas a que la información que proporciona la prueba y los datos sociodemográficos complementarios no son suficientes para discriminar a las poblaciones de activos y no activos al cabo del primer año.

Reafirmando este razonamiento, todos los autores consultados en la Subsección 2.3.3 convergen en afirmar que las variables que parecen tener incidencia sobre la decisión de desvincularse son de mediano y largo plazo en su mayoría, no centradas en el corto plazo como las variables independientes manejadas en este contexto, o bien que aquellas que son efectivamente de corto plazo son mucho más difíciles de recolectar y almacenar, al menos en un sistema de gestión educativa como los actuales.

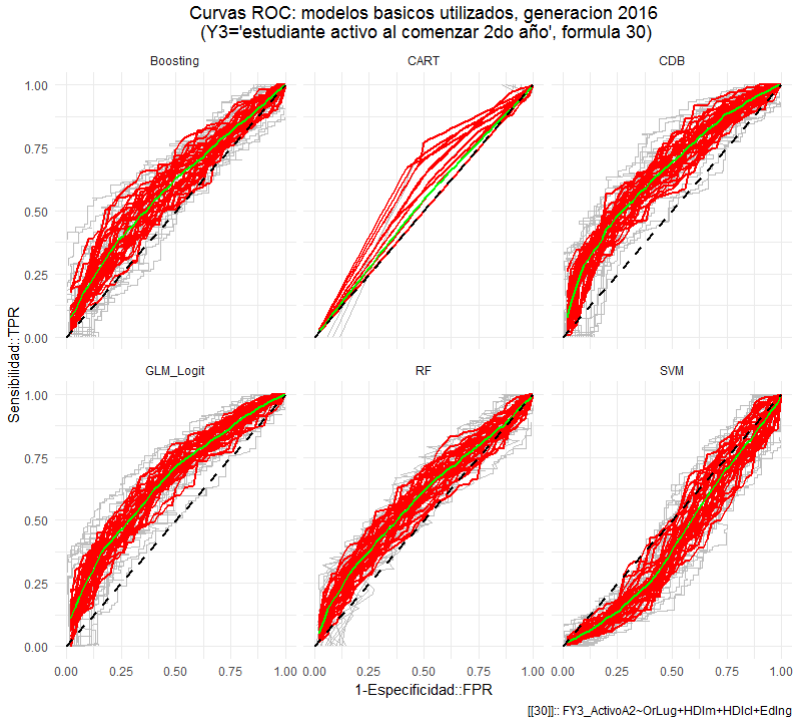

Figura 4.8: Ejemplo de curvas ROC promedio para Y_3, modelos no agregados, generación 2016

Lo anterior puede apreciarse en la Figura 4.8, donde se muestra claramente que la curva ROC promedio apenas se despega -e incluso para alguno de los modelos pasa *por debajo*- de la recta de clasificación aleatoria, marcada con una línea punteada.

Tiempos de ejecución A pesar de no ser un tema central en este trabajo, sobre los tiempos de ejecución o corrida se pudo observar:

- *Tiempos de ejecución generales*: correr uno de estos *loops* ha implicado tiempos de entre 4 a 6 horas, siempre en función de la cantidad de iteraciones y configuración de algunos de los modelos utilizados
- *Tiempos de ejecución por modelo*: Boosting resultó por lejos el más lento de los algoritmos utilizados en este problema; agregó tiempos de hasta un factor de 12 al resto de las corridas con otros modelos

Capítulo 5

Consideraciones finales

5.1. Conclusiones

Este trabajo fue concebido con la idea de retomar distintos esfuerzos realizados dentro y fuera de la Facultad de Ingeniería, para luego presentar otras técnicas que permitan echar luz sobre afirmaciones realizadas en otros trabajos de la Unidad respecto al *poder predictivo* de la HDI, en cuanto al desempeño futuro inmediato de los alumnos ingresantes.

Podemos concluir en primer lugar que efectivamente las pruebas diagnósticas de FIng pueden ser utilizadas, combinadas con otra información de interés y de fácil obtención, como herramienta para que la institución pueda anticiparse a ciertos eventos que se busca minimizar, aunque con algunos reparos: la mejora obtenida en los resultados con datos adicionados a la prueba se demostró para un cierto tipo de "trayectoria común" entre estudiantes de distintas carreras. Para considerar otros subgrupos, será necesario profundizar en estudios de trayectorias y cuantificar su impacto, construyendo así variables dependientes que tomen en cuenta estos aspectos.

Para la predicción del "rendimiento" a corto plazo se destacan claramente las variables "resultado en el componente Matemática de la (HDIm)" y "edad al ingreso" como las más importantes para los modelos con mejores niveles predictivos, aunque el "lugar de origen" y el "subsistema" tienen también relativa importancia. Por su parte, no es posible predecir la "deserción" con información disponible solo al ingreso de cada alumno; es necesario incluir más variables (de rendimiento, de cuestiones más personales) a plazos más largos para verificar si realmente es posible predecir este fenómeno como los anteriores *desde el inicio* de la vida universitaria.

Respecto a los modelos implementados, se pueden observar varias cosas. Antes que nada, es importante aclarar que fueron elegidos ciertos modelos de clasificación por popularidad, curiosidad o eficiencia *a priori*. Esto no quiere decir que no puedan existir otros que sean aún mejores que los propuestos; se remitirán las conclusiones exclusivamente a la realidad estudiada.

En primer lugar, destacar al modelo de regresión logística por su performance: es el que mejor discrimina presencia de ausencia del atributo para prácticamente todas las formulaciones [datos, variable de interés], mostrado ésto a través de los valores promedio de AUROC (se posicionó en el

tope en la amplia mayoría de formulaciones); incluso reflejó mayor porcentaje de aciertos respecto de lo observado para un mismo individuo en diferentes iteraciones. Juntando esto con el hecho de ser un modelo muy utilizado en el ámbito académico por su interpretabilidad y su amplia disponibilidad en diferentes programas, es sin dudas el "ganador" de entre todos los modelos considerados.

Siguen en orden (con disparidades, aunque en muchos casos juntos) los modelos de consenso. De hecho su performance es buena en algunas combinaciones [datos, variable de interés] e indicadores puntuales, sin embargo se observa que para un mismo individuo en distintas iteraciones pueden equivocarse en la predicción. Esto seguramente sea fruto de cómo están concebidos: los errores se dan con mayor frecuencia al existir discrepancias entre tres o más de los seis modelos individuales utilizados. En caso de empate se define la etiqueta resultante de manera aleatoria, con lo cual es posible que los primeros lugares no hayan sido ocupados por estos modelos dadas estas diferencias entre los modelos no agregados.

Los modelos SVM y el clasificador Bayesiano implementado estuvieron bastante lejos de lo esperado. El primero es conocido por adaptarse muy bien ante la existencia de fronteras no lineales en los datos, mientras que el segundo es un modelo clásico en sentido estricto y se basa en principios básicos de probabilidad para designar etiquetas. Sin embargo, fueron bastante oscilantes en sus resultados aunque con destaques puntuales en alguna fórmula, seguramente por existir solapamiento entre las clases consideradas para las diferentes variables. La performance de SVM quedó ensombrecida respecto a la del modelo logístico.

Finalmente, los modelos basados en árboles de clasificación fueron los que más lejos estuvieron de lo esperado. Este trabajo comenzó con la premisa que dichos modelos -en particular Boosting y Random Forest por ser mejoras de p.ej. CART- deberían estar en el tope en cuanto a los resultados predictivos, pero ocurrió todo lo contrario. Algunas posibles explicaciones pueden ser la falta de un ajuste o 'tuning' adicional de parámetros, corridas más largas (ej. con mayor cantidad de árboles para Boosting, aunque esto hubiera significado tiempos de corridas muy superiores a los actuales), incluso estos datos quizá sean *muy* particulares (Boosting parece captar mejor mucha de la variabilidad existente, al menos comparando algunos indicadores de las matrices de confusión).

De todos modos, para todos las variables y datos utilizados se pudo observar un patrón en común: los modelos parecen acertar a aquellos individuos que caen en las categorías de las variables predictoras más frecuentes en relación a la variable dependiente; en particular para la indicadora de llegar a la mitad de los créditos acumulados en las seis asignaturas comunes del primer año (denominada Y_2).

Al cierre de este trabajo ya se están aplicando las nuevas pruebas diagnósticas, que son diferentes -en contenido y espíritu- a la HDI. Los resultados presentados pueden tomarse como insumo para complementar lo que determinen estas pruebas con algunos datos adicionales de los alumnos, para afinar así la predicción sobre una posible variable de rendimiento a determinar, p.ej. culminar con éxito las UCs correspondientes a la asignatura Matemática. Claro que esto requiere como ya se mencionó un estudio *muy profundo* de las trayectorias iniciales de los alumnos,

para construir variables dependientes que puedan captar comportamientos estudiantiles de mejor manera posible (tanto continuas como discretas).

Limitaciones y alcance La principal limitante encontrada es respecto a quiénes realizan la prueba: se ha observado que las carreras compartidas son las que tienen menor nivel de "prevalencia" dada la multiplicidad de "ventanas de ingreso" a esas carreras y las exigencias institucionales respectivas. Esto es un punto importante para posibles trabajos futuros relacionados con pruebas diagnósticas, ya que la movilidad intra e inter facultades será cada vez mayor, generando nuevos desafíos para quienes trabajan con datos curriculares de alumnos en la Universidad de la República. Resta esperar el impacto del nuevo sistema de gestión estudiantil (SGAE) en estos fenómenos.

Otra limitante importante es que no hay suficiente evidencia como para asegurar que, respecto a las variables sociodemográficas no asociadas a las carreras, los datos faltantes son perdidos completamente al azar (MCAR)[1]. Es necesario investigar más a fondo esta cuestión para encontrar patrones de pérdida de datos y eventualmente eliminar los sesgos resultantes.

Algo no menor es qué información brinda la HDI sobre el *verdadero* rendimiento: al ser ésta una prueba de bajo impacto, el estudiante no es perjudicado en caso de un mal resultado, con lo cual no alcanzar los mínimos en esta prueba no es concluyente por sí de un futuro mal desempeño[2]; solo se puede observar una tendencia aunque nada definitivo.

También es importante destacar las limitantes durante la implementación informática: se buscó siempre la inmediata operatividad informática de este trabajo, dejando en un segundo plano la optimización del código generado, la creación de funciones más eficientes, y la mejora en tiempos de cálculo; todo esto es posible con una exhaustiva depuración y algo de paciencia.

5.2. Comentarios finales

Recomendaciones

Sería deseable que entre las instituciones preuniversitarias y la Universidad de la República exista intercambio de información -más allá del "Formulario 69A"- que podría ser de mucha utilidad para solucionar problemas como los mencionados en este trabajo. Un ejemplo de esto pueden ser las calificaciones de los alumnos que ingresan a la Universidad, que ha mostrado ser mucho mejor para predecir resultados en el corto y mediano plazo sobre estos alumnos dentro de la universidad, aunque esta hipótesis faltaría contrastarla en Uruguay.

[1] Considerando una variable independiente (X) y otra dependiente (Y), si los datos cumplen con esta característica, el hecho de que hayan datos perdidos p.ej. en X no generará sesgos en Y si se eliminan esos casos del análisis.

[2] De hecho, es *imposible* saber cuántos de esos estudiantes -que no alcanzan la suficiencia- se toman la prueba en serio.

Pensando en que en los próximos años se irán generando pruebas de diagnóstico más generales para ser aplicadas en distintas facultades de la UdelaR (y en base a otros trabajos), sería importante para lograr mejores resultados:

- Obtener datos de por lo menos dos años de trayectoria y realizar predicciones con más información académica, incluyendo en las mismas costos de clasificación errónea
- Estudiar la posibilidad de generar variables a predecir *dinámicas*, como sugiere el trabajo de Solís y cols. ([SMG+18])

Trabajo futuro

Creemos que es importante implementar otras variables -dependientes e independientes- y formas de predecir las mismas, como puede ser la predicción separada por carreras y la predicción del rendimiento a dos pasos. La primer propuesta podría ayudar a captar comportamientos que no pudieron ser observados en este trabajo, mientras que la segunda ayudaría a determinar si es posible predecir a plazos mayores, p.ej. al cabo del quinto año si se llega a cierta cantidad de créditos o si se desiste. Esto podría complementarse con análisis de supervivencia, siempre buscando una forma adecuada de aplicar a la realidad en FIng (el trabajo de Arias [AOD13] incluye estudiantes de ingeniería en sus análisis).

Estudiar a fondo que ha ocurrido con las poblaciones de activos, egresados y estudiantes desvinculados de la institución, según sus resultados en las diferentes pruebas diagnósticas disponibles, echaría luz sobre si la información de estas pruebas puede tener una proyección mayor a la del corto plazo.

Finalmente, sería ideal incluir otros paradigmas de clasificación, como por ejemplo redes neuronales o técnicas de aprendizaje profundo, dentro de las funciones creadas para comparar resultados. Podría ocurrir que otros clasificadores no tan "sofisticados" puedan incluso tener mejor poder predictivo (como menciona Donoho [Don17, 34-37]).

Referencias bibliográficas

[ACF+18] Laura Aspirot, Fedora Carbajal, Mercedes Fernández, Alina Machado, Andrea Vigorito, and Andrea Mesa. Pruebas diagnósticas realizadas a las generaciones de ingreso en la Facultad de Ciencias Económicas y de Administración. Informe de implementación y análisis de resultados 2017-2018, 2018.

[AG09] Richard Atkinson and Saul Geiser. Reflections on a century of college admissions tests. *Educational Researcher*, 38:665–676, 12 2009.

[AGG13] Esteban Alfaro, Matías Gámez, and Noelia García. adabag: An R package for classification with boosting and bagging. *Journal of Statistical Software*, 54(2):1–35, 2013.

[AOD13] Elena Arias Ortiz and Catherine Dehon. Roads to success in the Belgian French community's higher education system: predictors of dropout and degree completion at the Université Libre de Bruxelles. *Research in Higher Education*, 54(6):693–723, September 2013.

[ASN10] Asma A Al-Shargabi and Ali N Nusari. Discovering vital patterns from UST students data by applying data mining techniques. pages 547–551, Singapore, February 2010. IEEE.

[BCM17] M. Bourel, C. Crisci, and A. Martínez. Consensus methods based on machine learning techniques for marine phytoplankton presence–absence prediction. *Ecological Informatics*, 42:46–54, November 2017.

[BDL+13] Mathias Bourel, José Díaz, Eduardo Lacués, Freddy Rabín, and Julio Sabattino. Algunas cuestiones para pensar sobre el ingreso de los estudiantes a las carreras de Ingeniería en Uruguay. pages 1–8, Montevideo, sep. 2013.

[Bla06] Jorge Blanco. *Introducción al Análisis Multivariado: Teoría y aplicaciones a la realidad latinoamericana*. Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República, Montevideo, 2006.

[BM03] Lercy Barros and Pablo Míguez. Aplicación de Procesos de Markov para el estudio del Rendimiento Académico Universitario. Master's thesis, Montevideo, Uruguay, 2003.

[Boa11] Marcelo Boado. *La deserción estudiantil universitaria en la Udelar y en Uruguay entre 1997 y 2006*. Departamento de Publicaciones, Unidad de Comunicación de la Universidad de la República, Montevideo, Uruguay, 1 edition, 2011.

[BOP+12] Jaroslav Bayer, Tomas Obsivac, Lubomir Popelinsky, Jan Geryk, and Hana Bydzovska. Predicting drop-out from social behaviour of students. volume 5, pages 103–109, Chania, Greece, June 2012. International Educational Data Mining Society. OCLC: 911593280.

[Bou12] Mathías Bourel. Métodos de agregación de modelos y aplicaciones. *Memoria de Trabajos de Difusión Científica y Técnica*, 10:19–32, 2012.

[BY09] Ryan S. J. d Baker and Kalina Yacef. The state of educational data mining in 2009: a review and future visions. *JEDM | Journal of Educational Data Mining*, 1(1):3–17, October 2009.

[CHP15] Chi-Yang Chu, Daniel Henderson, and Christopher Parmeter. Plug-in bandwidth selection for kernel density estimation with discrete data. *Econometrics*, 3(2):199–214, March 2015.

[CMP+15] Sergio Celis, Luis Moreno, Patricio Poblete, Javier Villanueva, and Richard Weber. Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería. *Revista Ingeniería de Sistemas*, 29(1):5–24, 2015.

[con67] Constitución de la República O. del Uruguay, 1967.

[Cus09] Lorena Custodio. Caracterización de los desertores de la UdelaR (año 2006): desde la inversión y el consumo hacia la exclusión académica y la deserción voluntaria, 2009.

[DGP16] DGPlan. Sistema de indicadores para la evaluación universitaria: Indicadores de enseñanza de grado. Technical report, Montevideo, Uruguay, 2016.

[DGP17] DGPlan. Estadísticas Básicas 2016. Technical report, Montevideo, Uruguay, 2017.

[Don17] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[Dor93] Sherman Dorn. Origins of the "dropout problem". volume 33, pages 353–373. ACM Press, 1993.

[Duo18] Tarn Duong. *ks: Kernel Smoothing*, 2018. R package version 1.11.3.

[DW11] H Deng and Hadley Wickham. Density estimation in R. Technical report, 2011.

[Ela13] Obbey Ahmed Elamin. *Nonparametric Kernel Estimation Methods for Discrete Conditional Functions in Econometrics*. PhD thesis, Manchester, United Kingdom, 2013.

[Enr14] Heber Enrich. Desempeño estudiantil en FIng. ¿Dónde estamos ubicados?, 2014.

[Enr15] Heber Enrich. Desempeño estudiantil en la Facultad de Ingeniería. *InterCambios*, 2(1):41–47, June 2015.

[Faw04] Tom Fawcett. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. *Pattern Recognition Letters*, 31(8):1–38, 2004.

[FR13] Nicolás Fiori and Raúl Ramírez. Análisis de las trayectorias y perfil de los estudiantes desafiliados en la Universidad de la República (período 2007-2012). México DF, November 2013.

[Fri97] Jerome H. Friedman. Data mining and statistics: What's the connection. In *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics*, 1997.

[GC10] M.R. Gupta and Y. Chen. EM Demystified: An Expectation-Maximization Tutorial. Technical Report UWEETR-2010-0002, Seattle, WA, USA, 2010.

[GII18] Global innovation index, 2018.

[GR18] Sara González Gómez and Guillermo Ramón Ruiz. El acceso irrestricto de estudiantes a las universidades argentinas a través de los discursos de la prensa diaria (1982-1983). *História da Educação*, 22(54):113–134, April 2018.

[GUZ01] Juan José Goyeneche, Inés Urrestarazu, and Guillermo Zoppolo. ¿Cuándo me voy a recibir? Una aproximación para el análisis de la duración de la carrera estudiantil. *Revista Quantum*, 12(1):101–110, 2001.

[HCR01] Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C. Weng. A Note on Platt's Probabilistic Outputs for Support Vector Machines. 2001.

[HJS16] François Husson, Julie Josse, and Gilbert Saporta. Jan de Leeuw and the French School of Data Analysis. *Journal of Statistical Software*, 73(6):1–18, sep. 2016.

[HZZA18] Mushtaq Hussain, Wenhao Zhu, Wu Zhang, and Syed Muhammad Raza Abidi. Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Computational Intelligence and Neuroscience*, 2018:1–21, October 2018.

[IM10] María del Carmen Ibarra and Juan Carlos Michalus. Análisis del rendimento academico mediante un modelo Logit. *Revista Ingeniería Industrial*, 9(2):47–55, 2010.

[Ins13] Instituto Nacional de Estadística. Encuesta Continua de Hogares, 2013.

[JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, editors. *An introduction to statistical learning: with applications in R*. Number 103 in Springer texts in Statistics. Springer, New York, 2013. OCLC: ocn828488009.

[KM17] Alboukadel Kassambara and Fabian Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2017. R package version 1.0.5.

[KMH06] Alexandros Karatzoglou, David Meyer, and Kurt Hornik. Support Vector Machines in R. *Journal of Statistical Software*, 15(9), 2006.

[LBA+17] Carlos Daniel Luna, Pablo Babino, Daniel Alessandrini, Ximena Otegui, Luciana Chiavone, and Andrea Viscarret. Orientación estudiantil y desempeño académico en Ingeniería. *InterCambios. Dilemas y transiciones de la Educación Superior*, 4(1):96–103, June 2017.

[LJH08] Sébastien Lê, Julie Josse, and François Husson. FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 2008.

[LMP95] Ludovid Lebart, Alain Morineau, and Marie Piron. *Statistique exploratoire multidimensionnelle*, volume 1. Dunod, Paris, France, 1995.

[lVCI+07] Walter Álvarez Villar, Ada Czerwonogora, Gabriela Isolabella, Eduardo Lacués, Julia Leymonié, and Magdalena Pagano. La matemática al ingreso en la universidad. un estudio comparativo de cuatro Facultades en el Uruguay. *Revista Iberoamericana de Educación*, 42(4):1–9, abr. 2007.

[LW02] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[Mí08] Marina Míguez. *Análisis de las relaciones entre proceso motivacional, estrategias de aprendizaje y rendimiento académico en estudiantes del Área Científico – Tecnológica de la Universidad de la República*. PhD thesis, Montevideo, Uruguay, 2008.

[MBAP15] Marina Míguez, Lucía Blasina, Daniel Alessandrini, and Mauro Picó. Colaborando en la transición enseñanza media-universidad. Medellín, October 2015.

[MCC+07] Marina Míguez, Carolina Crisci, Karina Curione, Silvia Loureiro, and Ximena Otegui. Herramienta Diagnóstica al Ingreso a Facultad de Ingeniería: motivación, estrategias de aprendizaje y competencias disciplinarias. volume 8, pages 29–37, Río Cuarto, Córdoba, Argentina, jul. 2007.

[MDH+18] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2018. R package version 1.7-0.

[MLO05] Marina Míguez, Silvia Loureiro, and Ximena Otegui. *Aprendizaje, Enseñanza y Desempeño Curricular en la Facultad de Ingeniería: análisis cuantitativos y cualitativos*. Serie Análisis de Datos. Unidad de Enseñanza, Facultad de Ingeniería, Montevideo, Uruguay, 2005.

[OB16] Avaro Agustín Oñate Bowen. Análisis de la deserción y permanencia académica en la educación superior aplicando minería de datos. Master's thesis, Universidad Nacional de Colombia - Sede Bogotá, September 2016.

[PA14a] Alejandro Peña-Ayala, editor. *Educational Data Mining*, volume 524 of *Studies in Computational Intelligence*. Springer International Publishing, Cham, 2014.

[PA14b] Alejandro Peña-Ayala. Educational Data Mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 41(4):1432–1462, March 2014.

[Pag11] María Magdalena Pagano. El rendimiento académico y su vinculación con aspectos motivacionales. *InterCambios*, 59:11–17, 2011.

[RA07] María Noel Rodríguez-Ayan. *Análisis multivariado del desempeño académico de estudiantes universitarios de Química*. PhD thesis, Madrid, 2007.

[RA09] María Noel Rodríguez Ayan. Pruebas evaluatorias de final y mitad de carrera, informe de resultados 2008. Technical report, Montevideo, Uruguay, 2009.

[RARD11] María Noel Rodríguez-Ayán and Miguel Ángel Ruíz Díaz. Indicadores de rendimiento de estudiantes universitarios: calificaciones versus créditos acumulados. *Revista de Educación*, 355:467–492, May 2011.

[RM17] Pilar Rodríguez Morales. Creación, desarrollo y resultados de la aplicación de pruebas de evaluación basadas en estándares para diagnosticar competencias en matemática y lectura al ingreso a la universidad. *Revista Iberoamericana de Evaluación Educativa*, 10.1, 2017.

[Rom10] Carlos Romero. Un sistema universitario sin limitaciones de acceso: el caso de Uruguay. *Revista Iberoamericana de Evaluación Educativa*, 3(2):77–89, 2010.

[RPI17] Sergi Rovira, Eloi Puertas, and Laura Igual. Data-driven system to predict academic grades and dropout. *PLOS ONE*, 12(2):e0171207, February 2017.

[RQ03] Jeff Racine and Li Qi. Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, 86(2):266–292, August 2003.

[RV07] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146, July 2007.

[SB12] George Siemens and Ryan S. J. d. Baker. Learning analytics and educational data mining: towards communication and collaboration. page 252, Vancouver, British Columbia, Canada, 2012. ACM Press.

[Seo15] Mariana Seoane. Desempeño estudiantil en el primer y segundo año de la carrera de Odontología de la Universidad de la República. Análisis de trayectorias académicas de la cohorte 2009. *InterCambios*, 2(1):111–120, June 2015.

[SK15] M Saarela and T Kärkkäinen. Analysing Student Performance using Sparse Data of Core Bachelor Courses. *Journal of Educational Data Mining*, 7(1):3–32, 2015.

[SMG+18] Martin Solis, Tania Moreira, Roberto Gonzalez, Tatiana Fernandez, and Maria Hernandez. Perspectives to predict dropout in university students with machine learning. pages 1–6, San Carlos, July 2018. IEEE.

[SMN+05] Miguel Serna, Alina Machado, Laura Nalbarte, Fabiana Espínola, and Abadie Panambí. *Rendimiento escolar en la Universidad de la República: una propuesta de indicadores de desempeño de los estudiantes*, volume 05/01 of *Documentos de Trabajo*. Comisión Sectorial de Enseñanza, Universidad de la República, 2005.

[SSBL05] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, October 2005.

[Str08] Carolin Strobl. *Statistical issues in machine learning: towards reliable split selection and variable importance measures*. Cuvillier, Göttingen, 1. aufl edition, 2008. OCLC: 436287292.

[TA18] Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2018. R package version 4.1-13.

[UCU07] UCUR. Sexta etapa. La intervención (1973-1985), 2007. http://www.universidad.edu.uy/renderPage/index/pageId/98.

[UEF09a] UEFI. Informe Herramienta Diagnóstica Media. Technical report, Montevideo, Uruguay, 2009.

[UEF09b] UEFI. Predicción de la Herramienta Diagnóstica al Ingreso (HDI). Technical report, Montevideo, Uruguay, 2009.

[UEF13] UEFI. Informe de Deserción 1997-2009. Technical report, Montevideo, Uruguay, 2013.

[UEF18] UEFI. Informe de Avance 2017. Technical report, Montevideo, Uruguay, 2018.

[VR02] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Statistics and Computing. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[Wan15] Matt Wand. *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995)*, 2015. R package version 2.23-15.

[Wic07] Hadley Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007.

[Wic16] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[WLLR05] Claus Weihs, Uwe Ligges, Karsten Luebke, and Nils Raabe. klar analyzing german business cycles. In D. Baier, R. Decker, and L. Schmidt-Thieme, editors, *Data Analysis and Decision Support*, pages 335–343, Berlin, 2005. Springer-Verlag.

[ZBHB15] J. Zimmermann, K Brodersen, H Heinimann, and J Buhmann. A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. *Journal of Educational Data Mining*, 7(3):151–176, 2015.

Glosario

El siguiente glosario está inspirado en el homónimo que presentan L. Lebart y A. Salem en su libro *Statistique Textuelle (1994)*, a modo de guía conceptual.

Notas

- Los asteriscos indican otra definición en el presente glosario.
- Los números en rojo indican la página donde se encuentra la primer cita de cada concepto.
- Las siguientes abreviaciones que aparecen entre paréntesis refieren al dominio en el cual se aplica la correspondiente definición:
  - **(aa)** Aprendizaje Automático
  - **(est)** Estadística
  - **(ps)** Psicometría
  - **(ot)** Otros

Aprendizaje Supervisado (aa) Tarea dentro del aprendizaje automático, consistente en entrenar (en base a una muestra de entrenamiento*) y validar uno o varios modelos, a partir de una muestra de prueba* disponiendo de los correspondientes a la variable de interés, Y. 25

Bondad de Ajuste (est) Medidas que indican la discrepancia entre los datos observados y los valores esperados bajo un determinado modelo 28

Clasificador Nulo (aa) Aquel que siempre asigna la clase mayoritaria; es un "competidor natural" de cualquier clasificador. 78

Consistencia Interna (ps) Sirve para medir si distintas preguntas o ítems en una prueba, que fueron creados para medir un mismo constructo no observable o latente, producen resultados o valores similares. Una medida popularmente utilizada es el estadístico *Alfa de Cronbach*, una media ponderada de las correlaciones entre los ítems que forman parte de la escala:

$$\alpha = \frac{k}{k-1}\left(1 - \sum_{i=1}^{k} \frac{S_i^2}{S_t^2}\right),$$

siendo S_i^2 la varianza del ítem o pregunta i, S_t^2 la varianza de los valores totales observados y k la cantidad de preguntas 12

Estudiante Activo (ot) (para toda la UdelaR): aquel que registra al menos una actividad académica (p.ej. inscripción a curso o examen) en los dos años anteriores a la fecha de referencia. 18

Indice de Dificultad (ps) Proporción de personas que responden correctamente a un ítem de una prueba; a mayor proporción *menor* será la dificultad. 12

Indice de Discriminacion (ps) Permite distinguir, para los ítems de una prueba, entre estudiantes de altos y bajos rendimientos; generalmente se divide a la población evaluada en tres grupos de tamaño similar. A mayor valor, mejor diferenciación o discriminación entre estudiantes con altas o bajas calificaciones. 12

Meta-Algoritmo (aa) Métodos diseñados para solucionar problemas en forma general, bajo condiciones particulares; también denominado "metaheurística" 35

Validacion Cruzada (aa) Validación de modelos utilizando subconjuntos aleatorios, mediante p.ej. bootstrap (usa muestreo aleatorio simple con reposición) o jackknife (usa el método *leave-one-out*: extrae una observación, calcula el estadístico, agrega nuevamente esa observación y prosigue con los cálculos). Se obtienen dos grupos de datos:

- *Muestra de Entrenamiento (ME,MEnt)*: parte de los datos con los cuales se entrena al modelo en cuestión
- *Muestra de Prueba o Testing (MT)*: parte complementaria de los datos con la cual se cuantifica la precisión del modelo puesto a prueba

22

APÉNDICES

Apéndice 1

Información adicional: FIng, pruebas diagnósticas

Generalidades en Facultad de Ingeniería

En este apartado se busca -a grandes rasgos- conocer quienes llegan, desarrollan sus carreras de grado y egresan de las mismas dentro de la Facultad de Ingeniería de la Universidad de la República (FIng) en función de distintos resultados obtenidos a lo largo de los 20 últimos años de datos disponibles (1997-2016).

Ingreso a la vida universitaria

A la FIng pueden ingresar aquellos alumnos que hayan culminado la educación media superior, fundamentalmente aquellos que optan por un bachillerato científico, en los cuales la enseñanza de matemáticas, física y química ocupan un lugar mayoritario en la currícula[1]. De la información al ingreso solo quedan disponibles algunas variables "de base", tales como la fecha de nacimiento, donde y en qué subsistema culminó su educación media, además de las concernientes a la inscripción a la(s) carrera(s) escogida(s). No se conserva en modo alguno información sobre el desempeño de cada alumno antes de ingresar a la vida universitaria, con lo cual es necesario utilizar la información disponible y herramientas propias para obtener más datos, como es el caso de la Herramienta Diagnostica al Ingreso (HDI), ya descrita en el Capítulo 2.

Respecto a las características de los alumnos ingresantes, estos presentan rasgos similares a lo largo del tiempo: es una población fuertemente masculina, que proviene mayoritariamente de liceos públicos del interior y de públicos o privados de Montevideo -aunque en los últimos años con un aporte creciente de los estudiantes de la UTU o extranjeros-, con edades comprendidas mayoritariamente entre los 18 y 19 años de edad.

[1] Hay varias opciones válidas, que tienen distintos nombres pero comparten el principio mencionado (ver https://www.fing.edu.uy/bedelia). Además, otros estudios preuniversitarios con perfil más técnico son aceptados, aunque teniendo en cuenta la restricción anterior.

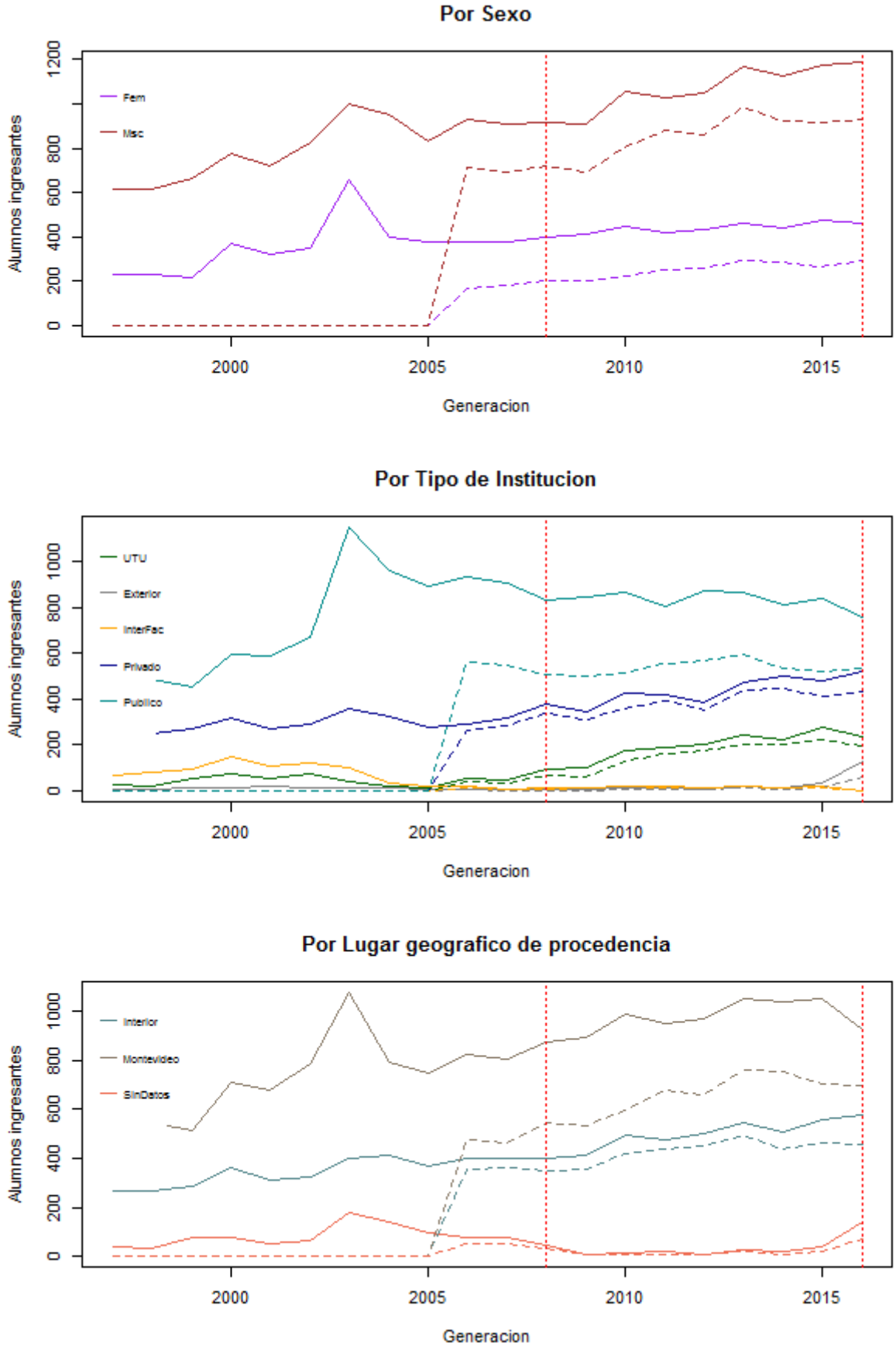

Figura A.1: Evolución de las características de los alumnos ingresantes a Facultad de Ingeniería según variables sociodemográficas utilizadas, entre 1997 y 2016. Los alumnos que realizan la HDI figuran en líneas punteadas horizontales; las punteadas verticales indican las generaciones consideradas en este trabajo. Fuente: elaboración propia, en base a datos SGB

Desde 2008 se puede ingresar en el primer o segundo semestre del año, contando para ello con cursos "a contra semestre" para estos alumnos, que suelen ser entre un 10-15 % del total de alumnos del primer semestre, con un promedio de edad y con una presencia de estudiantes del exterior levemente mayor que lo observado en los primeros semestres de cada año. La evolución anual[2] de algunas variables de interés se muestra en la Figura A.1.

En el período citado, las carreras más "tradicionales" (Ingenierías en Computación, Química, Civil, Eléctrica, Mecánica y en Alimentos, y finalmente Agrimensura) se llevan más del 85 % de las preferencias anualmente, dejando el resto de las inscripciones a carreras[3] para tecnólogos o carreras creadas más recientemente. Anualmente, Ingeniería en Computación se lleva más de un tercio de todas las inscripciones[4], seguidas de las Ingenierías Química, Eléctrica y Civil (intercalándose el orden entre ellas en algunos años, cada una entre el 10 % y el 15 % del total de inscriptos), y finalmente Ingeniería Industrial Mecánica (entre un 6 % y 8 %), en Alimentos (ésta más oscilante[5]) y Agrimensura (1 % promedio).

Tránsito de estudiantes por FIng

Según el último informe de avance por generaciones realizado por la Unidad ([UEF18]), unos 8593 estudiantes que ingresaron a carreras del Plan 1997 son activos a marzo de 2017. Más de un tercio de ellos (38 %) se encuentra en la franja de créditos correspondiente al primer semestre (entre 0 y 44 créditos), y trepa a 49 % si se agregan a los anteriores a los alumnos que no superan el segundo semestre. Esto es una característica recurrente en Ingeniería: se la denomina como "facultad de primer año", dado que la mitad de su población activa no supera el primer año en relación a los créditos obtenidos desde su ingreso.

Otro dato interesante es sobre la movilidad de los estudiantes: según datos de las generaciones 1997 a 2016, los que deciden por una sola carrera durante toda su vida activa oscilan entre un 65 % y un 85 % de cada generación, disminuyendo dicho guarismo a medida que los alumnos tienen más tiempo de actividad. Los restantes realizan actividades en más de una carrera, ya sea desde el inicio de su vida universitaria como incluso pasados muchos años después.

[2] La evolución anual incluye los ingresos de ambos semestres, si corresponde.

[3] Se cuentan inscripciones a *primer carrera*; es decir que si el estudiante se anota a más de una se seleccionará la primera que aparece en el registro; este orden -en caso de anotarse en la *misma fecha*- es aleatorio.

[4] Los inscriptos a Computación alcanzaron a más del 50 % para 1997; luego este valor se estabilizó entorno al 40 % ya comenzado el presente milenio.

[5] El caso de Ing. en Alimentos es extremo: pasó de menos del 1 % en 1997 al 24 % en 2003; la mediana 1997-2016 está entorno al 5 % de toda la matrícula.

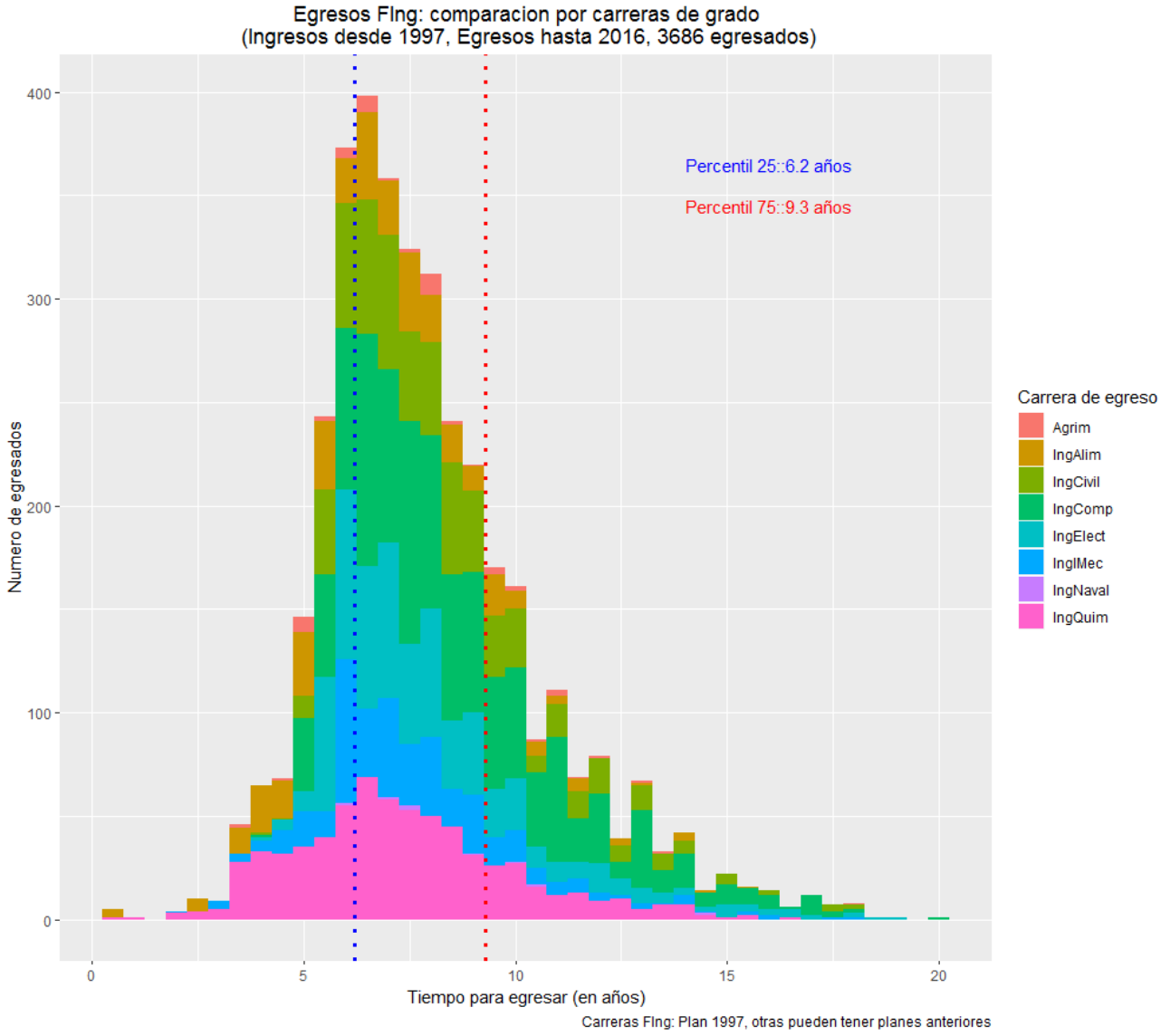

Figura A.2: Histograma del tiempo de egreso en carreras de grado FIng. Fuente: elaboración propia, en base a datos SGB

Egreso en FIng

Considerando datos de los 20 años citados anteriormente, los estudiantes de las carreras tradicionales de grado -sean o no compartidas con otras facultades- tienen una mediana de 7,5 años para egresar, con un 90 % central de los egresos entre 5 y 13 años, y con una gran concentración de casos entre los 6 y los 9 años desde el ingreso, tal como se aprecia en el histograma de la Figura A.2.

Indicadores generales: ingreso, permanencia y egreso

La siguiente es una tabla incluida en todos los informes de avance anuales que realiza la Unidad, actualizada para el último informe disponible ([UEF18]), donde se muestra -por generación de ingreso- qué ocurre con las subpoblaciones de activos, egresados y alumnos desvinculados.

Se definen para ello *Desvinculación Neta*, como $DN_t = \frac{Ingresos_t - Activos_t - Egresos_t}{Ingresos_t}$, tasa de *Activos* y *Eficiencia de Titulación Real* como $ETR_t = \frac{Egresos_t}{Egresos_t + Activos_t}$ para una generación t dada[6], todas ellas observadas a marzo de 2017. Esta tabla a continuación muestra la evolución de dichos guarismos:

Generación	Tasa DN	Activos	ETR
1997	0,64	0,04	0,89
1998	0,62	0,06	0,83
1999	0,63	0,07	0,81
2000	0,62	0,08	0,79
2001	0,51	0,13	0,75
2002	0,55	0,12	0,72
2003	0,5	0,15	0,7
2004	0,55	0,16	0,65
2005	0,6	0,19	0,52
2006	0,48	0,25	0,52
2007	0,49	0,31	0,4
2008	0,42	0,39	0,33
2009	0,39	0,46	0,25
2010	0,49	0,41	0,19
2011	0,52	0,44	0,07
2012	0,34	0,65	0,01

Sintetizando lo que postula la tabla anterior, se puede afirmar que:

- La deserción a las carreras de grado varía entre un 50 y un 60 % del total de ingresantes por generación (incluso supera ese porcentaje para algunas de ellas)
- De forma complementaria, la proporción de egresados tiene un "techo" que es el 40 % de los ingresantes por generación (considerando a los desertores en el total de casos), mientras que si se mira extrayendo a los desertores, la titulación neta al cabo de 15 años supera el 70 % (es decir, de cada 10 activos por lo menos 7 de ellos egresará en un período de 15 años o más)
- La absoluta mayoría de estudiantes que egresa lo hace de una sola carrera, ya sea de grado o técnica. De todos modos, un pequeño porcentaje de egresados de carreras de grado obtiene un título más *del mismo nivel* (0,4 %), en su mayoría siendo egresados primero de Ingeniería Química y luego de Ingeniería en Alimentos, o viceversa.

[6]Se consideran solamente las generaciones que, al momento de efectuar la medición, cuentan con 5 años en la institución, que es el tiempo teórico de egreso para las carreras del Plan 1997

Pruebas diagnósticas en UdelaR más allá del inicio

Experiencia en FIng: Herramienta Diagnóstica Media Durante los años 2008 y 2009 la Facultad de Ingeniería instrumentó una prueba denominada "Herramienta Diagnóstica Media" (HDM), para evaluar conocimientos y competencias generales de los estudiantes activos en distintas carreras de la Facultad, durante su "tránsito medio". Tenía algunas similitudes con la HDI[7], aunque las preguntas relacionadas con saberes eran diferentes, pues muchas hacían referencia a temas teóricamente vistos por estos alumnos dentro de Facultad (preguntas específicas por carrera, preguntas generales); además se consultaban otras cosas diferentes (preguntas sobre dominio en inglés, situación laboral, uso de becas económicas) en pos de obtener información objetiva que permita mejorar la coordinación curricular.

Los resultados obtenidos muestran que, si bien entre ambas pruebas los resultados fueron algo diferentes, respecto a los componentes generales (comunes a todas las carreras) el nivel de suficiencia estuvo entorno a un 25 % de los casos. La suficiencia global presentó diferencias importantes (48 % en 2008, 32.5 % en 2009) al igual que la suficiencia por componentes específicos por carrera. Para algunas de estas preguntas, la influencia de alguna asignatura en particular, la formación específica escogida o no haber cursado alguna UC puede haber tenido que ver en las elecciones de algunos distractores. [UEF09a]

Otras experiencias: Facultad de Química La Facultad de Química de la UdelaR fue otros de los servicios que implementó pruebas fuera del inicio de las carreras. La Unidad Académica de Educación Química (UNADEQ) se encargó de realizar dos pruebas diagnósticas fuera del comienzo: una a la mitad de la carrera y otra al final, ambas en 2008. Para la segunda ocurrieron algunos problemas operativos con una consecuente baja participación de estudiantes en general, mientras que para la primera de las pruebas se logró una participación mayor. En general se concluye que los estudiantes dominan satisfactoriamente especificidades importantes (terminología, nomenclatura, convenciones, unidades) al tiempo que usan e interpretan correctamente la información en particular con preguntas no completamente abiertas, en donde una mayor autonomía genera caídas importantes en los niveles de suficiencia. [RA09]

Pruebas de egreso preuniversitario y/o ingreso universitario en algunos países de América Latina

La siguiente tabla muestra a grandes rasgos que requisitos de salida de la educación preuniversitaria y qué requisitos de entrada a la universidad son pedidos por país

[7] Se podía realizar 1 sola vez, era obligatoria (para estudiantes con entre 150 y 200 créditos y cursando una UC específica por carrera), otorgaba puntos adicionales para asignaturas específicas en caso de superar los mínimos establecidos, contaba con el mismo Cuestionario CEAM

	EXAMEN(ES)		
	Preuniversitario Salida	**Universidad Entrada**	Comentarios adicionales
Bolivia		PSA*	
Brasil	ENEM	Vestibular*	Algunas exigen exámenes adicionales
Chile		PSU***	Universidades que forman el Consejo de Rectores de las Universidades Chilenas (CRUCH)
Colombia	Saber 11	Pruebas de Admisión*	
Cuba		Prueba de Ingreso a la Universidad	
Ecuador	ENES	PAA**	
México	(EXANI-I) (al ingreso del bachillerato)	EXANI-II	Mayoría de universidades lo exige, otras tienen pruebas propias y algunas no exigen prueba alguna para entrar
Paraguay		Examen de Ingreso	En algunas facultades de la UNA (cursos propedéuticos posteriores)
Perú		Examen de Ingreso**	Las pruebas exigidas son diferentes para cada universidad
Venezuela		(no hay ingreso por examen)	Cupos según OPSU o c/universidad

Notas: (*) solo lo piden (algunas) universidades públicas, (**) solo lo exigen algunas universidades (públicas o privadas), (***) muchas de las universidades elaboran rankings adicionales usando las calificaciones de secundaria y la posición de esas calificaciones en la generación de ingreso del alumno (pueden incluso pedir pruebas adicionales)

Fuentes consultadas

- Generales: https://www.nosequeestudiar.net/orientacion
- Específicas por país:
  - Bolivia: http://www.umsa.bo/web/guest/pregrado
  - Brasil:
    https://www.educamaisbrasil.com.br/enem
    https://www.educamaisbrasil.com.br/enem/para-que-serve
    https://enem.inep.gov.br/
  - Chile: https://psu.demre.cl/
  - Colombia:
    http://www2.icfes.gov.co/estudiantes-y-padres/saber-11-estudiantes/informacion-general-del-examen

- Cuba: http://www.mes.gob.cu/es/como-acceder-la-educacion-superior
- Ecuador: https://www.puce.edu.ec/admisiones_GP.php
- México: http://www.ceneval.edu.mx/exani-ii
- Paraguay: http://www.ing.una.py/?page_id=30
- Venezuela: https://snibarinas.wordpress.com/

Apéndice 2

Aspectos Metodológicos

Algoritmo EM

Este algoritmo es ampliamente usado en distintas ramas del saber, para estimar parámetros en modelos con datos faltantes, para ajustar modelos mezcla de normales (GMM) o cadenas de Markov ocultas (HMM), entre otros.

Generalidades

En el proceso de estimación hay tres cosas diferentes:

- variables directamente observables que asumen una distribución de probabilidad
- *parámetros*, cantidades fijas no observables, generalmente asociadas a valores de momentos (ejs: esperanza, varianza) de las distribuciones asumidas para las variables (p.ej. la media μ o el desvió estándar σ^2 para la distribución normal)
- las *variables latentes*, que son cantidades no directamente observables y como tales no poseen una distribución de probabilidad (como p.ej. las probabilidades a priori, α_j)

Se introducen dos conceptos cruciales:

- *Función de Verosimilitud*[1]: $L(\theta|x)$, o simplemente $L(\theta)$.
- *Función Q*: es la esperanza del logaritmo de la verosimilitud. Como este problema -clase Y desconocida para X- se enmarca en el de *datos faltantes*, los datos completos Z se definen como $Z = (X, Y)$, siendo Y una variable latente (ej: pertenencia a una determinada clase). Así, la función Q cambia su zona de integración (será sobre Z, que es la única parte aleatoria de la función)

[1] Ya definida al introducir los modelos lineales generalizados, en la página 28. Como se dejó en claro en su momento, θ puede ser un *vector* de parámetros

Algoritmo 4 Seudocódigo algoritmo EM: general

Datos: parámetros $\theta^{(m=0)}$, datos observados x, umbral de convergencia δ
Resultado: estimador final de los parámetros, θ^{FINAL}
Inicio: $\theta^{(m)} = \theta^{(0)}$ **mientras** *no convergencia* $|L^{(m+1)} - L^{(m)}| > \delta$ *para cada* m **hacer**

1. **Paso-E** (estimación de *variables latentes*): dada la estimación para la m-ésima iteracion, $\theta^{(m)}$, formar la función $Q(\theta|\theta^{(m)})$
2. **Paso-M** (estimación de *parámetros* mediante máxima verosimilitud): calcular la $(m+1)$-ésima estimación de θ así:

$$\theta^{(m+1)} = \arg\max_{\theta \in \Theta} Q(\theta|\theta^{(m)})$$

fin

$$\begin{aligned} Q(\theta|\theta^{(m)}) &= E_{Z|x,\theta^{(m)}}[\log(p_Z(z|\theta))] \\ &= E_{Y|x,\theta^{(m)}}[\log(p_Z(x,y|\theta))] \\ &= \int_{\mathcal{Y}} \log(p_Z(x,y|\theta)) \cdot p_{Y|X}(y|x,\theta^{(m)}) dy \end{aligned} \tag{2.1}$$

Seudocódigo

Una versión generalmente utilizada -y simplificada- de los pasos dados en el algoritmo EM es la que figura en el algoritmo 4.

Propiedades

- El algoritmo 4 converge al valor máximo verosímil buscado debido a la propiedad de *monotonicidad*[2] -bajo ciertas condiciones de regularidad- que cumple la secuencia "esperanza-maximización" ([GC10, pp.6-8])
- La convergencia del EM es lenta, en algunos casos a máximos locales
- Es sensible a los valores de arranque $\theta^{(m=0)}$ (mientras más cerca de los valores originales, más rápida será su convergencia)

Desarrollo para casos particulares

A continuación se detallan dos casos particulares de interés del método EM: el de mezcla de normales (conocido comúnmente como Gaussian Mixture Models, GMM) y el modelo *ad-hoc* utilizando funciones de densidad por núcleos, en base al artículo de Gupta y Chen ([GC10]).

[2]Esta propiedad postula que la verosimilitud de la estimación $\theta^{(m+1)}$ nunca sera menor que la estimación anterior, $\theta^{(m)}$ (es decir, "*nunca empeorará* respecto al valor hallado en la iteración anterior").

Caso Particular 1: mezcla de normales

Los GMM son vectores de observaciones de los cuales solamente se sabe que provienen de una distribución que surge de mezclar distribuciones normales de parámetros desconocidos. Los parámetros son en este caso tanto los valores de las medias y las varianzas/covarianzas de cada distribución individual, como así también el ponderador que viene con cada una, que indica la probabilidad de pertenencia de cada observación a una de las densidades normales desconocidas, $\alpha_j \phi_j(x_i|\mu_j, \Sigma_j)$, para cada una de las $j = 1, ..., k$ clases definidas y de las n observaciones en la muestra iid $i = 1, ..., n$, de una población con distribución desconocida $p(x|\theta)$

En el caso de los modelos de mezcla normales (GMM), se definen previamente:

- Conjunto paramétrico a ser estimado: $\theta = \{(\alpha_j, \mu_j, \Sigma_j)\}_{j=1}^k$, siendo μ_j los vectores de medias, Σ_j la matriz de varianzas y covarianzas de las distribuciones normales que se desconocen y α_j los ponderadores de cada una de las normales por cada clase
- Funciones de densidad normales (simbolizada por $\phi j(\cdot)$), para cada observación $i = 1, ..., n$:

$$\phi_j(x_i|\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{p/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j)\right) \tag{2.2}$$

- Log-verosimilitud en m-ésima iteracion:

$$L^{(m)}(\theta|x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} \alpha_j^{(m)} \phi_j(x_i|\mu_j^{(m)}, \Sigma_j^{(m)})\right),$$

siendo $p(x|\theta) = \alpha_1\phi_{j1} + \alpha_2\phi_{j2} + ... + \alpha_k\phi_{jk}$ la mezcla de densidades normales resultantes.

Sea entonces $\phi_j(x_i|\mu_j, \Sigma_j)$ definida por (2.2), se busca conocer el conjunto de parámetros $\theta = \{(\alpha_j, \mu_j, \Sigma_j)\}_{j=1}^k$. Se define $\gamma_{ij}^{(m)}$ como la estimación (en la m-ésima iteración) de la probabilidad (a posteriori) de pertenencia de la i-ésima observación a la j-ésima distribución normal de la mezcla:

$$\gamma_{ij}^{(m)} = \frac{\alpha_j^{(m)} \phi_j(x_i|\mu_j^{(m)}, \Sigma_j^{(m)})}{\sum_{j=1}^k \alpha_j^{(m)} \phi_j(x_i|\mu_j^{(m)}, \Sigma_j^{(m)})},\ i = 1, ..., n;\ j = 1, ..., k \qquad (\text{satisface } \sum_{j=1}^{k} \gamma_{ij}^{(m)} = 1)$$

Al tratarse de una muestra iid, es sencillo demostrar -como en ([GC10, p.10])- que la función Q en (2.1) para este caso es igual a la suma de las funciones Q_i para cada observación, es decir

$$\begin{aligned} Q(\theta|\theta^{(m)}) &= \sum_{i=1}^{n} Q_i(\theta|\theta^{(m)}) \\ &= \sum_{i=1}^{n} E_{Z_i|x_i,\theta^{(m)}}[\log(p(Z_i|\theta))]. \end{aligned} \tag{2.3}$$

De este modo y luego de hacer algunas operaciones, se obtiene la expresión para la función Q del "Paso-E":

$$Q(\theta|\theta^{(m)}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}^{(m)}\left(\log\alpha_j - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}\left(x_i-\mu_j\right)^T\Sigma_j^{-1}\left(x_i-\mu_j\right)\right)$$

Si se define $n_j^{(m)} = \sum_{i=1}^{n}\gamma_{ij}^{(m)}$, se puede reescribir la expresión anterior como

$$Q(\theta|\theta^{(m)}) = \sum_{j=1}^{k} n_j^{(m)}\left(\log\alpha_j - \frac{1}{2}\log|\Sigma_j|\right) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}^{(m)}\left(x_i-\mu_j\right)^T\Sigma_j^{-1}\left(x_i-\mu_j\right).$$

De este modo, el "Paso-M" surge de resolver el siguiente problema de optimización[3]:

$$\begin{cases} \text{máx}_\theta & Q(\theta|\theta^{(m)}) \\ \text{s.a.} & \alpha_j \geq 0,\ \sum_{j=1}^{k}\alpha_j = 1,\ j=1,...,k\ , \\ & \Sigma_j \succ 0,\ j=1,...,k \end{cases} \tag{2.4}$$

siendo $\Sigma_j \succ 0$ la condición de que cada Σ_j sea definida positiva.

Al ser un problema con restricciones impuestas, para resolverlo éste debe ser relajado a través del planteo de los multiplicadores de Lagrange,

$$Q_\lambda(\alpha_j,\lambda) = \sum_{j=1}^{k} n_j^{(m)}\log\alpha_j + \lambda\left(\sum_{j=1}^{k}\alpha_j - 1\right)\ j=1,...,k.$$

Para encontrar extremos, se obtiene la derivada primera del Lagrangiano Q_λ respecto a α_j y se iguala a cero:

$$\frac{\partial Q_\lambda}{\partial\alpha_j} = 0 \implies \frac{n_j^{(m)}}{\alpha_j} + \lambda = 0 \iff \lambda = -\frac{n_j^{(m)}}{\alpha_j}\ j=1,...,k.$$

Como además se debe cumplir $\sum_{j=1}^{k}\alpha_j = 1$, los ponderadores α_j satisfacen

$$\alpha_j^{(m+1)} = \frac{n_j^{(m)}}{\sum_{j=1}^{k} n_j^{(m)}} = \frac{n_j^{(m)}}{n}\ j=1,...,k.$$

[3]Porque se buscan estimadores máximo verosímiles, que maximizan la función de verosimilitud ($Q(\theta|\theta^{(m)})$ en este caso)

Para el caso de las medias, se debe resolver $\frac{\partial Q}{\partial \mu_j} = 0$

$$\frac{\partial Q}{\partial \mu_j} = 0 \Longrightarrow \Sigma_j^{-1} \left(\sum_{i=1}^{n} \gamma_{ij}^{(m)} x_i - n_j^{(m)} \mu_j \right) = 0,$$

lo cual implica (dado que la matriz de covarianzas es definida positiva por restricción)

$$\mu_j^{(m+1)} = \frac{1}{n_j^{(m)}} \cdot \sum_{i=1}^{n} \gamma_{ij}^{(m)} x_i.$$

Para el caso de la matriz de covarianzas:

$$\frac{\partial Q}{\partial \Sigma_j} = 0 \Longrightarrow -\frac{1}{2} n_j^{(m)} \frac{\partial}{\partial \Sigma_j} \log|\Sigma_j| - \frac{1}{2} \sum_{i=1}^{n} \gamma_{ij}^{(m)} \frac{\partial}{\partial \Sigma_j} \left(x_i - \mu_j\right)^T \Sigma_j^{-1} \left(x_i - \mu_j\right) = 0 \; j = 1, ..., k$$

Operando se arriba a la solución para Σ_j:

$$\Sigma_j^{(m+1)} = \frac{1}{n_j^{(m)}} \sum_{i=1}^{n} \gamma_{ij}^{(m)} \cdot \left(x_i - \mu_j^{(m+1)}\right) \cdot \left(x_i - \mu_j^{(m+1)}\right)^T, \; j = 1, ..., k$$

Utilizando el algoritmo 4 como referencia, se detallan los pasos del algoritmo EM para este caso particular

- **Inicio**:
  - *Paso-1*: elegir un valor inicial de θ , $\theta^{(m=0)}$ y computar

$$L^{(0)} = \frac{1}{n} \cdot \sum_{i=1}^{n} \log \left(\alpha_j^{(0)} \phi_j(x_i | \mu_j^{(0)}, \Sigma_j^{(0)}) \right)$$

- **Paso-E**:
  - *Paso-2*: calcular $\gamma_{ij}^{(m)}$ y $n_j^{(m)}$, para $i = 1, ..., n; j = 1, ..., k$:

$$\gamma_{ij}^{(m)} = \frac{\alpha_l^{(m)} \phi(x_{obs,j} | \mu_l^{(m)}, \Sigma_l^{(m)})}{\sum_{l=1}^{k} \alpha_l^{(m)} \phi(x_{obs,j} | \mu_l^{(m)}, \Sigma_l^{(m)})}, \; i = 1, ..., n; \; j = 1, ..., k \tag{2.5}$$

$$n_j^{(m)} = \sum_{i=1}^{n} \gamma_{ij}^{(m)}, \; j = 1, ..., k \tag{2.6}$$

- **Paso-M**:
  - *Paso-3*: calcular las nuevas estimaciones

$$\alpha_j^{(m+1)} = \frac{n_j^{(m)}}{n}, \; j = 1, ..., k$$

$$\mu_j^{(m+1)} = \frac{1}{n_j^{(m)}} \sum_{i=1}^{n} \gamma_{ij}^{(m)} \cdot x_i,\ j = 1, ..., k$$

$$\Sigma_j^{(m+1)} = \frac{1}{n_j^{(m)}} \sum_{i=1}^{n} \gamma_{ij}^{(m)} \cdot \left(x_i - \mu_j^{(m+1)}\right) \cdot \left(x_i - \mu_j^{(m+1)}\right)^T,\ j = 1, ..., k,$$

encontrando de esta manera

$$L^{(m+1)} = \frac{1}{n} \cdot \sum_{i=1}^{n} \log\left(\alpha_j^{(m+1)} \phi_j(x_i | \mu_j^{(m+1)}, \Sigma_j^{(m+1)})\right)$$

.

- Se realizan nuevamente pasos 2 y 3 si $|L^{(m+1)} - L^{(m)}| \leq \delta$; sino terminar.

La salida se compone de los distintos parámetros máximo verosímiles hallados en la última iteración:

$$\hat{\alpha}_1^{(m+1)}, \ldots, \hat{\alpha}_k^{(m+1)}; \hat{\mu}_1^{(m+1)}, \ldots, \hat{\mu}_k^{(m+1)}; \hat{\Sigma}_1^{(m+1)}, \ldots, \hat{\Sigma}_k^{(m+1)}$$

Caso Particular 2: mezcla de densidades estimadas por nucleos (KDE)

De la misma manera que para la mezcla de normales vista en el *Caso Particular 1*, EM puede ser adaptado a la búsqueda de coeficientes en una mezcla de densidades que es estimada mediante estimadores de densidades por núcleo. Es decir, la función objetivo cambia a $p(x|\theta) = \alpha_1 f_1 + \alpha_2 f_2 + ... + \alpha_k f_k$, donde $f_j,\ j = 1, ..., k$ son estimadores de densidad por núcleo como los que se introdujeron en la Sección 3.1.5. Ahora, lo único que se busca estimar son los pesos desconocidos, α_j. Todo esto se resume en el algoritmo 5.

Algoritmo 5 Seudocódigo algoritmo EM: caso particular para hallar α_j en (3.24)

Datos: parámetros $\alpha_j^{(m=0)}$, datos observados x, densidades kernel $\hat{f}_j^{KDE}$, umbral δ
Resultado: estimador final de los parámetros, α^{FINAL}
Inicio: $\alpha_j^{(m)} = \alpha_j^{(0)}$ **mientras** *no convergencia* $|L^{(m+1)} - L^{(m)}| > \delta$ *para cada* m **hacer**

1. **Paso-E** (estimación de *variables latentes*): dada la estimación para la m-ésima iteracion, $\alpha_j^{(m)}$

   a) computar $L^{(m)}(\alpha_j|x) = \frac{1}{n} \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} \alpha_j^{(m)} \hat{f}_j^{KDE}(x_i)\right)$

   b) calcular $\gamma_{ij}^{(m)}$ y $n_j^{(m)}$ para $i = 1, ..., n; j = 1, ..., k$ sustituyendo $\phi(x|\alpha_j)$ por $\hat{f}_j^{KDE}(x_i)$ en (2.5)

2. **Paso-M**: calcular la $(m+1)$-ésima estimación de α_j:

$$\alpha_j^{(m+1)} = \frac{n_j^{(m)}}{n},\ j = 1, ..., k$$

y realizar nuevamente pasos 1-2 si $|L^{(m+1)} - L^{(m)}| \leq \delta$
fin

EM: *Ejemplo de aplicación*

En base a un ejemplo real[4], se simulan 10 datos provenientes de dos distribuciones normales, $j = 1, 2$; de las cuales se conoce solamente su varianza, que por simplicidad se asume común a ambas: $\sigma_j^2 = 1,\ j = 1, 2$. De estos 10 datos observados, 7 de $X_1 \sim N(13; 1)$ y 3 provienen de $X_2 \sim N(7; 1)$. El vector de datos univariados obtenido es:

$$x = \{6.1; 7.18; 8.59; 11.87; 12.92; 13.13; 13.71; 12.76; 14.98; 12.86\}$$

Lo que el investigador realmente conoce es el valor de las observaciones anteriores (fija en cada iteración) y que existen dos subpoblaciones disjuntas de donde provienen esos datos. Para continuar se debe hallar:

1. la probabilidad a posteriori de pertenencia a cada una de las subpoblaciones para cada x, llamadas α_j
2. las medias de cada subpoblación, μ_j.

A continuación, se desglosan los pasos establecidos en el algoritmo 4 y en el denominado "Caso Particular 1" de la sección anterior.

[4]Este ejemplo se basa en resultados reales obtenidos en la HDI para dos subpoblaciones particulares.

Valores iniciales: al ser las incógnitas las probabilidades a posteriori de pertenencia a cada subpoblación y las medias de cada una, se inician: (valores arbitrarios)

$\theta^{(m=0)} = \{\alpha_1 = 0.1; \alpha_2 = 0.9; \mu_1 = 1; \mu_2 = 0\}$. Se supone además un umbral de tolerancia $\delta = 0.01$

Iteración 1

- **Inicio**:
  - *Paso-1*: dado $\theta^{(m=0)} = \{\alpha_1 = 0.1; \alpha_2 = 0.9; \mu_1 = 1; \mu_2 = 0\}$ computar:

$$L^{(1)} = \frac{1}{10} \cdot (\log(0.1 \cdot \phi_1(x|1;1)) + \log(0.9 \cdot \phi_2(x|0;1))) = -61.48407$$

- **Paso-E**:
  - *Paso-2*: calcular $\gamma_{ij}^{(1)}$ y $n_j^{(1)}$, para $i = 1, ..., 10; j = 1, 2$:

$$\gamma_{i1}^{(1)} = \frac{0.1 \cdot \phi_1(x_i|1;1)}{\sum_{j=1}^{2} \alpha_j^{(1)} \phi_j(x_i|\mu_j^{(1)}, \Sigma_j^{(1)})} \Rightarrow \boldsymbol{\gamma}_1^{(1)} = \{0.968; 0.989; 0.997; 1.000; 1.000; ...; 1.000\}$$

$$\gamma_{i2}^{(1)} = \frac{0.9 \cdot \phi_2(x_i|0;1)}{\sum_{j=1}^{2} \alpha_j^{(1)} \phi_j(x_i|\mu_j^{(1)}, \Sigma_j^{(1)})} \Rightarrow \boldsymbol{\gamma}_2^{(1)} = \{0.032; 0.011; 0.003; 0.000; 0.000; ...; 0.000\}$$

$$n_1^{(1)} = \sum_{i=1}^{10} \gamma_{i1}^{(1)} \Rightarrow n_1^{(1)} = 9.954$$

$$n_2^{(1)} = \sum_{i=1}^{10} \gamma_{i2}^{(1)} \Rightarrow n_2^{(1)} = 0.046$$

- **Paso-M**:
  - *Paso-3*: calcular las nuevas estimaciones (con estimadores máximo verosímiles de los parámetros desconocidos)

$$\alpha_1^{(2)} = \frac{n_1^{(1)}}{n} = \frac{9.954}{10} = 0.9954$$

$$\alpha_2^{(2)} = \frac{n_2^{(1)}}{n} = \frac{0.046}{10} = 0.0046$$

$$\mu_1^{(2)} = \frac{1}{9.954} \sum_{i=1}^{10} \gamma_{i1}^{(1)} \cdot x_i = \frac{113.8085}{9.954} = 11.4338$$

$$\mu_2^{(2)} = \frac{1}{0.0046} \sum_{i=1}^{10} \gamma_{i2}^{(1)} \cdot x_i = \frac{0.303}{0.0462} = 6.5494$$

  - *Paso-4*: para chequear convergencia, calcular

$$\begin{aligned} |L^{(2)} - L^{(1)}| &= |\tfrac{1}{10} \cdot (\log(0.9954 \cdot \phi_1(x|11.4338;1)) + \log(0.0046 \cdot \phi_2(x|6.5494;1))) - (-61.48407)| \\ &= |-3.769 + 61.48407| = 57.7145 > 0.01 \end{aligned}$$

por esto debe continuar con la siguiente iteración.

Iteración 2

- **Inicio**:
  - *Paso-1*: dado $\theta^{(2)} = \{\alpha_1 = 0.9954; \alpha_2 = 0.0046; \mu_1 = 11.4338; \mu_2 = 6.5494\}$ computar:

$$L^{(2)} = \frac{1}{10} \cdot (\log(0.9954 \cdot \phi_1(x|11.4338; 1)) + \log(0.0046 \cdot \phi_2(x|6.5494; 1))) = -3.769$$

- **Paso-E**:
  - *Paso-2*: calcular $\gamma_{ij}^{(2)}$ y $n_j^{(2)}$, para $i = 1, ..., 10; j = 1, 2$:

$$\gamma_{i1}^{(2)} = \frac{0.9954 \cdot \phi_1(x_i|11.4338; 1)}{\sum_{j=1}^{2} \alpha_j^{(2)} \phi_j(x_i|\mu_j^{(2)}, \Sigma_j^{(2)})} \Rightarrow \boldsymbol{\gamma}_1^{(2)} = \{0.000; 0.031; 0.968; 1.000; 1.000; ...; 1.000\}$$

$$\gamma_{i2}^{(2)} = \frac{0.0046 \cdot \phi_2(x_i|6.5494; 1)}{\sum_{j=1}^{2} \alpha_j^{(2)} \phi_j(x_i|\mu_j^{(2)}, \Sigma_j^{(2)})} \Rightarrow \boldsymbol{\gamma}_2^{(2)} = \{1.000; 0.969; 0.032; 0.000; 0.000; ...; 0.000\}$$

$$n_1^{(2)} = \sum_{i=1}^{10} \gamma_{i1}^{(2)} \Rightarrow n_1^{(2)} = 7.9985$$

$$n_2^{(2)} = \sum_{i=1}^{10} \gamma_{i2}^{(2)} \Rightarrow n_2^{(2)} = 2.0015$$

- **Paso-M**:
  - *Paso-3*: calcular las nuevas estimaciones (con estimadores máximo verosímiles de los parámetros desconocidos)

$$\alpha_1^{(3)} = \frac{n_1^{(2)}}{n} = \frac{7.9985}{10} = 0.79985$$

$$\alpha_2^{(3)} = \frac{n_2^{(2)}}{n} = \frac{2.0015}{10} = 0.20015$$

$$\mu_1^{(3)} = \frac{1}{7.9985} \sum_{i=1}^{10} \gamma_{i1}^{(2)} \cdot x_i = \frac{100.7674}{7.9985} = 12.59826$$

$$\mu_2^{(3)} = \frac{1}{2.0015} \sum_{i=1}^{10} \gamma_{i2}^{(2)} \cdot x_i = \frac{13.3441}{2.0015} = 6.667106$$

  - *Paso-4*: para chequear convergencia, calcular

$$\begin{aligned} |L^{(3)} - L^{(2)}| &= |\tfrac{1}{10} \cdot (\log(0.79985 \cdot \phi_1(x|12.5983; 1)) + \log(0.20015 \cdot \phi_2(x|6.6671; 1))) - (-3.769)| \\ &= |-2.168 + 3.769| = 1.6017 > 0.01 \end{aligned}$$

por esto debe continuar con la siguiente iteración.

Iteración 3

- **Inicio**:
  - *Paso-1*: dado $\theta^{(3)} = \{\alpha_1 = 0.79985; \alpha_2 = 0.20015; \mu_1 = 12.5983; \mu_2 = 6.6671\}$ computar:

$$L^{(3)} = \frac{1}{10} \cdot (\log(0.7998 \cdot \phi_1(x|12.5983; 1)) + \log(0.2002 \cdot \phi_2(x|6.6671; 1))) = -2.171$$

- **Paso-E**:
  - *Paso-2*: calcular $\gamma_{ij}^{(3)}$ y $n_j^{(3)}$, para $i = 1, ..., 10; j = 1, 2$:

$$\gamma_{i1}^{(3)} = \frac{0.7998 \cdot \phi_1(x_i|12.5983; 1)}{\sum_{j=1}^{2} \alpha_j^{(3)} \phi_j(x_i|\mu_j^{(3)}, \Sigma_j^{(3)})} \Rightarrow \boldsymbol{\gamma}_1^{(3)} = \{0.000; 0.000; 0.009; 0.968; 1.000; ...; 1.000\}$$

$$\gamma_{i2}^{(3)} = \frac{0.2002 \cdot \phi_2(x_i|6.6671; 1)}{\sum_{j=1}^{2} \alpha_j^{(3)} \phi_j(x_i|\mu_j^{(3)}, \Sigma_j^{(3)})} \Rightarrow \boldsymbol{\gamma}_2^{(3)} = \{1.000; 1.000; 0.992; 0.000; 0.000; ...; 0.000\}$$

$$n_1^{(3)} = \sum_{i=1}^{10} \gamma_{i1}^{(3)} \Rightarrow n_1^{(3)} = 7.0083$$

$$n_2^{(3)} = \sum_{i=1}^{10} \gamma_{i2}^{(3)} \Rightarrow n_2^{(3)} = 2.9917$$

- **Paso-M**:
  - *Paso-3*: calcular las nuevas estimaciones (con estimadores máximo verosímiles de los parámetros desconocidos)

$$\alpha_1^{(4)} = \frac{n_1^{(3)}}{n} = \frac{7.0083}{10} = 0.70083$$

$$\alpha_2^{(4)} = \frac{n_2^{(3)}}{n} = \frac{2.9917}{10} = 0.29917$$

$$\mu_1^{(4)} = \frac{1}{7.0083} \sum_{i=1}^{10} \gamma_{i1}^{(3)} \cdot x_i = \frac{92.30697}{7.0083} = 13.1711$$

$$\mu_2^{(4)} = \frac{1}{2.9917} \sum_{i=1}^{10} \gamma_{i2}^{(3)} \cdot x_i = \frac{21.80454}{2.9917} = 7.2883$$

  - *Paso-4*: para chequear convergencia, calcular

$$\begin{aligned} |L^{(4)} - L^{(3)}| &= |\tfrac{1}{10} \cdot (\log(0.70083 \cdot \phi_1(x|13.1711; 1)) + \log(0.29917 \cdot \phi_2(x|7.2883; 1))) - (-2.171)| \\ &= |-1.965 + 2.171| = 0.2057 > 0.01 \end{aligned}$$

por esto debe continuar con la siguiente iteración.

Iteración 4

- **Inicio**:
  - *Paso-1*: dado $\theta^{(4)} = \{\alpha_1 = 0.70083; \alpha_2 = 0.29917; \mu_1 = 13.1711; \mu_2 = 7.2883\}$ computar:

$$L^{(4)} = \frac{1}{10} \cdot (\log(0.7008 \cdot \phi_1(x|13.1711; 1)) + \log(0.2992 \cdot \phi_2(x|7.2883; 1))) = -1.965$$

- **Paso-E**:
  - *Paso-2*: calcular $\gamma_{ij}^{(4)}$ y $n_j^{(4)}$, para $i = 1, ..., 10; j = 1, 2$:

$$\gamma_{i1}^{(4)} = \frac{0.7008 \cdot \phi_1(x_i|13.1711; 1)}{\sum_{j=1}^{2} \alpha_j^{(4)} \phi_j(x_i|\mu_j^{(4)}, \Sigma_j^{(4)})} \Rightarrow \boldsymbol{\gamma}_1^{(4)} = \{0.000; 0.000; 0.001; 1.000; 1.000; ...; 1.000\}$$

$$\gamma_{i2}^{(4)} = \frac{0.2992 \cdot \phi_1(x_i|7.2883; 1)}{\sum_{j=1}^{2} \alpha_j^{(4)} \phi_j(x_i|\mu_j^{(4)}, \Sigma_j^{(4)})} \Rightarrow \boldsymbol{\gamma}_2^{(4)} = \{1.000; 1.000; 0.999; 0.000; 0.000; ...; 0.000\}$$

$$n_1^{(4)} = \sum_{i=1}^{10} \gamma_{i1}^{(4)} \Rightarrow n_1^{(4)} = 7.0001$$

$$n_2^{(4)} = \sum_{i=1}^{10} \gamma_{i2}^{(4)} \Rightarrow n_2^{(4)} = 2.9999$$

- **Paso-M**:
  - *Paso-3*: calcular las nuevas estimaciones (con estimadores máximo verosímiles de los parámetros desconocidos)

$$\alpha_1^{(5)} = \frac{n_1^{(4)}}{n} = \frac{7.0001}{10} = 0.70001$$

$$\alpha_2^{(5)} = \frac{n_2^{(4)}}{n} = \frac{2.9999}{10} = 0.29999$$

$$\mu_1^{(5)} = \frac{1}{7.0001} \sum_{i=1}^{10} \gamma_{i1}^{(4)} \cdot x_i = \frac{92.23669}{7.0001} = 13.1764$$

$$\mu_2^{(5)} = \frac{1}{2.9999} \sum_{i=1}^{10} \gamma_{i2}^{(4)} \cdot x_i = \frac{21.87483}{2.9999} = 7.2919$$

  - *Paso-4*: para chequear convergencia, calcular

$$\begin{aligned} |L^{(5)} - L^{(4)}| &= |\tfrac{1}{10} \cdot (\log(0.70001 \cdot \phi_1(x|13.1764; 1)) + \log(0.29999 \cdot \phi_2(x|7.2919; 1))) - (-1.965)| \\ &= |-1.964979 + 1.964991| = 0.000012 < 0.01 \end{aligned}$$

y así el algoritmo converge.

De esta manera, los valores estimados de α_j y μ_j para $j = 1, 2$ son (con precisión a 4 dígitos):

- $\boldsymbol{\hat{\alpha}_1 = 0.7}$, $\boldsymbol{\hat{\mu}_1 = 13.1764}$
- $\boldsymbol{\hat{\alpha}_2 = 0.3}$, $\boldsymbol{\hat{\mu}_2 = 7.2919}$

Apéndice 3

Aspectos Informáticos

Estimación de densidades por núcleo en R

Lo que sigue es un compendio de información extraída del artículo de Deng y Wickham ([DW11]), comentando las principales características de las funciones más importantes para cada paquete estudiado.

- Función `hist()`, paquete `graphics`: permite generar histogramas para los datos considerados. El argumento `'breaks'` sirve para especificar la cantidad de barras (b) del histograma, o bien se puede especificar una función interna que calcule el b óptimo. Por defecto, esta función crea un gráfico; asignando `'plot=FALSE'` se obtienen valores de evaluación y resultados para el histograma.
- Función `density()`, paquete `stats`: permite implementar estimación univariada de densidades, con varios kernels disponibles (normal, rectangular, triangular y coseno). El ancho de banda se puede estimar mediante `'bw'`, utilizando reglas generales (`bw.nrd0`), o de validación cruzada (`bw.bcv, bw.ucv`); incluso se pueden proporcionar pesos a las observaciones.
- Función `kde()`, paquete `ks`: este paquete permite estimación multivariada de funciones de densidad, de hasta 6 dimensiones. El ancho de banda puede estimarse mediante varios métodos de las dos clases vistas (aproximación teórica de b: `Hpi`; basados en datos: `Hamise.mixt, Hmise.mixt, Hbcv, Hlscv, Hscv`). Con esta función se permite estimar valores para una densidad obtenida previamente, con el argumento `'eval.points'`.
- Función `npudens()`, paquete `np`: esta función utiliza tres etapas: primero calcula el ancho de banda, luego las densidades y finalmente permite estimar una nueva densidad, todo mediante el denominado "producto generalizado de funciones kernel" (Li y Racine, citados en [Ela13]). Permite trabajar con cualquier tipo de datos y sus rutinas de cálculo de b son por métodos basados en datos, muchas veces con tiempos de cálculo elevados.

La siguiente tabla resume lo anterior:

Función	**Paquete**	**Estimación** b		**Tipo de estimador** b		**Comentarios**
		Aprox. Teórica	Data-Driven	Fijo	Variable	
`hist`	`graphics`	√		√		Ancho de rectángulos, b'
`density`	`stats`	√	√	√		Version por defecto en `R`
`kde`	`ks`	√	√	√		Solo variables numéricas
`npudens`	`np`		√	√	√	Variables numéricas y categóricas

Tabla 3.1: `R`: algunos paquetes para estimación no paramétrica de funciones de densidad

Código de función asm2()

Código de función asm2(), donde se genera el loop de ajuste simultáneo de modelos predictivos y los principales resultados del trabajo[1].

```
asm2 <- function(frml,datFULL,columnaY,modelosML,parti=10,propme=0.75,predicts.out=FALSE,
semillas=NULL,metadatos=NULL,error.resum=TRUE,moc=c("vm","mp","waauc"),
cart.ctrl=rpart::rpart.control(cp=.01), #CART
lgt.stepaic=FALSE, #logit
costsvm=1,kernel="radial",svm.gamma=0.5, #SVM
ntree=100, mtry=2, #RF
b.mfinal=100, #Boost
uso.em=TRUE,em.alfa10=0.5,em.iter=100,em.tol=1e-5,cdb.dens="density",...){
###1.Modelos "usuales"
init <- Sys.time()
mimp <- c("CART","Logit","SVM","RF","Boosting","CDB")
mimp <- mimp[sort(match(x=modelosML,table=mimp))]
parti <- 1:parti
fo <- as.formula(frml)
largo <- length(attr(terms(fo),"term.labels"))
if(!is.null(datFULL))
tipoY <- class(datFULL[,columnaY])
else
tipoY <- class(datT[,columnaY])
nivY <- levels(as.factor(datFULL[,columnaY]))
dpg <- FALSE
if(sum(sapply(datFULL,function(x) sum(is.na(x)),simplify=TRUE))>0){  #if(missingdata)
datFULL <- model.frame(formula=fo,data=datFULL)
columnaY <- 1
dpg <- TRUE
}
##Funcion ml_dfs
ml_dfs <- function(dat,mod,tipo,i,rownm=FALSE){
pms <- c("pm.cart.df","pm.logit.df","pm.svm.df","pm.rf.df","pm.boost.df","pm.cdb.df","pmc.cart.
    df","pmc.logit.df","pmc.rf.df","pmc.boost.df","pmc.svm.df","pmc.cdb.df")
ml <- unique(gsub(pattern="([a-z]+\\.)([a-z]+)(\\.[a-z]+)",replacement="\\2",x=pms))
ml_dfs <- strtrim(tolower(mimp[sort(match(x=mod,table=ml))]),5) #asi evitamos problemas con mimp
    ::Boosting
modelo <- paste0("Yp_",ml_dfs,collapse="")
if(paste0(tipo,".",ml_dfs,collapse="") %in% c("pm.cart","pm.rf","pm.cdb")){
```

[1]Resto del código en GitHub: https://github.com/dalessandrini/uefi2

```
#1.pm: cart/rf, cdb OK
attr(names)=NULL
if(is.null(attr(dat[[i]],which="names")))
out <- try(data.frame(id = rownm, modelo = dat[[i]]),silent = TRUE)
else
out <- try(data.frame("id"=attr(dat[[i]],which="names"),modelo=dat[[i]]),silent=TRUE)
} else if(paste0(tipo,".",ml_dfs,collapse="") %in% c("pm.svm","pm.logit","pm.svm2","pm.logit2"))
    {
#2.pm: svm logit
out <- try(dat,silent=TRUE)
} else if(paste0(tipo,".",ml_dfs,collapse="") %in% "pm.boost"){
#3.pm boost
out <- try(data.frame("id"=rownm,modelo=dat[[i]]$class),silent=TRUE)
} else if(paste0(tipo,".",ml_dfs,collapse="") %in% c("pmc.cart","pmc.rf")){
#4.pmc: cart/rf OK
if(is.null(attr(dat[[i]],which="dimnames")[[1]]))
out <- try(data.frame(id=rownm, modelo=dat[[i]]),silent=TRUE)
else
out <- try(data.frame("id"=attr(dat[[i]],which="dimnames")[[1]],modelo=dat[[i]]),silent=TRUE)
} else if(paste0(tipo,".",ml_dfs,collapse="") %in% "pmc.logit") {
#5.pmc logit
out <- try(data.frame("id"=attr(dat,which="row.names"),modelo=dat),silent=TRUE)
} else if(paste0(tipo,".",ml_dfs,collapse="") %in% "pmc.svm") {
#6.pmc svm
out <- try(data.frame("id"=attr(dat,which="names"),attr(dat,which="probabilities")),silent=TRUE)
} else if(paste0(tipo,".",ml_dfs,collapse="") %in% "pmc.boost"){
#7.pmc boost
out <- try(data.frame("id"=rownm,modelo=dat[[i]]$prob),silent=TRUE)
names(out)[2:3] <- c("Boost_0","Boost_1")
} else {
out <- try(data.frame("id"=attr(dat[[i]],which="names"),modelo=dat[[i]]),silent=TRUE)
}
if(is.data.frame(out) && ncol(out)==2)
names(out)[2] <- modelo
#C:tryCatch a A/B
out <- tryCatch(out,
error=function(cond){
message("Error, va mensaje R")
message(cond)
return(NA)
},
warning=function(cond){
message("Advertencia, revisar")
message(cond)
return(NULL)
},
finally=message(paste0(" -- ",ml_dfs," iteracion :",i)))
return(out)
}
#a <- cart_dfs(pm.cart[[i]],i); aa <- cart_dfs(pmc.cart[[i]],i)
trf <- tbag <- t.mlg1 <- tsvm <- tnb <- tcb <- tcart <- papl <- perme <- permt <- em.alf <- m.
    logit.sa <-
tboost <- pm.cart <- pm.logit2 <- pm.svm2 <- pm.rf <- pm.boost <- pm.cdb <- comp.vs.Yobs <-
    dfpmlg <-
pmc.cart <- pmc.rf <- comp.vs.papc <- mg <- vml <- mpl <- waul <- vector(mode="list",length=max(
    parti))
for(i in parti){
if(!is.null(semillas)) set.seed(semillas[i])
perm_df <- uefi2:::mute(df=datFULL,nroperm=1,pme=propme)
```

```
datE <- perm_df$dat_me[[1]]
datT <- perm_df$dat_mt[[1]]
perme[[i]] <- perm_df$indME
permt[[i]] <- perm_df$rnMT
nf.dT <- row.names(datT)
if("CART" %in% modelosML){
#requireNamespace("rpart",quietly=TRUE)
m.cart <- rpart::rpart(fo,data=datE,method="class",control=cart.ctrl)
pm.cart[[i]] <- predict(m.cart,newdata=if(dpg) datT else datT[,-columnaY],type="class")
pmc.cart[[i]] <- predict(m.cart,newdata=if(dpg) datT else datT[,-columnaY],type="prob")
tcart[[i]] <- table(datT[,columnaY],factor(pm.cart[[i]],levels=nivY),dnn=c("YObs","YPred"))
t.cart <- mdcmed(tcart[[i]])
if(predicts.out){
pm.cart.df <- ml_dfs(dat=pm.cart,mod="cart",tipo="pm",i=i,rownm=nf.dT)
pmc.cart.df <- ml_dfs(dat=pmc.cart,mod="cart",tipo="pmc",i=i,rownm=nf.dT)
if(!is.null(dim(pmc.cart.df)))
names(pmc.cart.df)[2:(length(nivY)+1)] <- paste("CART_",nivY,sep="")
}
}
if("Logit" %in% modelosML){
#requireNamespace("stats",quietly=TRUE)
m.logit <- stats::glm(fo,family="binomial",data=datE,...)
if(lgt.stepaic)
m.logit.sa[[i]] <- MASS::stepAIC(m.logit,scope=list(upper=fo,lower=~1),direction="ba",trace=0)
pm.logit <- stats::predict(m.logit,newdata=if(dpg) datT else datT[,-columnaY],type="response")
#devuelve P(exito)=P(Y=1)
dfpmlg[[i]] <- data.frame(1-pm.logit,pm.logit)
names(dfpmlg[[i]]) <- nivY
pm.logit2[[i]] <- data.frame("id"=attr(dfpmlg[[i]],"row.names"),"Ypred"=factor(colnames(dfpmlg[[
    i]])[apply(dfpmlg[[i]],1,which.max)],levels=nivY))
t.mlg1[[i]] <- table(datT[,columnaY],pm.logit2[[i]]$Ypred,dnn=c("YObs","YPred"))
t.mlg <- mdcmed(t.mlg1[[i]])
if(predicts.out){
pm.lgt.df <- ml_dfs(dat=pm.logit2[[i]],mod="logit",tipo="pm",i=i)
pmc.lgt.df <- ml_dfs(dat=dfpmlg[[i]],mod="logit",tipo="pmc",i=i)
if(!is.null(dim(pmc.lgt.df)))
names(pmc.lgt.df)[2:(length(nivY)+1)] <- paste("Lgt_",nivY,sep="")
}
}
if("SVM" %in% modelosML){
#requireNamespace("e1071",quietly=TRUE)
m.svm <- svm(fo,data=datE,cost=costsvm,kernel=kernel,decision.values=TRUE,gamma=svm.gamma,
    probability=TRUE)
pm.svm <- predict(object=m.svm,newdata=if(dpg) datT else datT[,-columnaY],probability=TRUE,
    decision.values=TRUE)
pm.svm2[[i]] <- data.frame("id"=attr(pm.svm,"names"),"Ypred"=factor(colnames(attr(pm.svm,"
    probabilities"))[apply(attr(pm.svm,"probabilities"),1,which.max)],levels=nivY))       #dfpmlg
    $ypred
tsvm[[i]] <- table(datT[,columnaY],pm.svm2[[i]]$Ypred,dnn=c("YObs","YPred"))
t.svm <- mdcmed(tsvm[[i]])
if(predicts.out){
pm.svm.df <- ml_dfs(dat=pm.svm2[[i]],mod="svm",tipo="pm",i=i)
pmc.svm.df <- ml_dfs(dat=pm.svm,mod="svm",tipo="pmc",i=i)
if(!is.null(dim(pmc.svm.df)))
names(pmc.svm.df)[2:(length(nivY)+1)] <- paste("SVM_",nivY,sep="")
}
}
if("RF" %in% modelosML){
#requireNamespace("randomForest",quietly=TRUE)
```

```
m.rf <- randomForest(fo,data=datE,mtry=mtry,importance=TRUE,keep.forest=TRUE,ntree=ntree,
    proximity=TRUE,...)
pm.rf[[i]] <- predict(object=m.rf,newdata=if(dpg) datT else datT[,-columnaY])
pmc.rf[[i]] <- predict(object=m.rf,newdata=if(dpg) datT else datT[,-columnaY],type="prob")
trf[[i]] <- table(datT[,columnaY],factor(pm.rf[[i]],levels=nivY),dnn=c("YObs","YPred"))
t.mrf <- mdcmed(trf[[i]])
if(predicts.out){
pm.rf.df <- ml_dfs(dat=pm.rf,mod="rf",tipo="pm",i=i)
pmc.rf.df <- ml_dfs(dat=pmc.rf,mod="rf",tipo="pmc",i=i)
if(!is.null(dim(pmc.rf.df)))
names(pmc.rf.df)[2:(length(nivY)+1)] <- paste("RF_",nivY,sep="")
}
}
if("Boosting" %in% modelosML){
#requireNamespace("adabag",quietly=TRUE)
m.boost <- boosting(fo,data=datE,mfinal=b.mfinal)
pm.boost[[i]] <- predict(object=m.boost,newdata=if(dpg) datT else datT[,-columnaY],type="class")
tboost[[i]] <- table(datT[,columnaY],factor(pm.boost[[i]]$class,levels=nivY),dnn=c("YObs","YPred
    "))
t.boost <- mdcmed(tboost[[i]])
if(predicts.out){
pm.boost.df <- ml_dfs(dat=pm.boost,mod="boost",tipo="pm",i=i,rownm=nf.dT)
pmc.boost.df <- ml_dfs(dat=pm.boost,mod="boost",tipo="pmc",i=i,rownm=nf.dT)
}
}
if("CDB" %in% modelosML){
#A.cdb(...)
if(dpg){
datEm <- datE
datTm <- datT
Ym <- columnaY
} else {
datEm <- model.frame(fo,data=datE)
datTm <- model.frame(fo,data=datT)
Ym <- 1
}
efet <- dk(dfe=datEm,dft=datTm,vclase=Ym,dens=cdb.dens)
if(uso.em){ ##switch con "indicador de salida"
tsf <- efet$tsf
alf <- emalg2k(alfa1=em.alfa10,f1=switch(EXPR=tsf,snum=efet$f_num$fy1_Xte$fnum_Xte_prod,
scat=efet$f_cat$fcat_1$fcat_prod,mixt=efet$f_mix$fmix_1$pf_ctcl),
f2=switch(EXPR=tsf,snum=efet$f_num$fy2_Xte$fnum_Xte_prod,
scat=efet$f_cat$fcat_2$fcat_prod,mixt=efet$f_mix$fmix_2$pf_ctcl),
tol=em.tol,plota=FALSE,listalfa=TRUE)
#B.predict.cdb(...)
papl[[i]] <- pape(f_dk=efet,yr=datTm[,Ym],x=datTm[,-Ym],datfalt=efet$datfalt,
alfas=c(alf$alfa1,alf$alfa2))
} else {
papl[[i]] <- pape(alfas=NULL,f_dk=efet,yr=datTm[,Ym],x=datTm[,-Ym],
datfalt=efet$datfalt,nomb.obs=attr(datTm,"row.names"))
}
pm.cdb[[i]] <- papl[[i]]$PaP$ypred
#C.MdConf
tcb[[i]] <- table(datTm[,Ym],factor(pm.cdb[[i]],levels=nivY),dnn=c("YObs","YPred"))
if(uso.em){
em.alf[[i]] <- c("alfa1"=alf$alfa1,"alfa2"=alf$alfa2)
}
if(predicts.out){
pm.cdb.df <- data.frame("id"=attr(datT,which="row.names"),"Yp_CDB"=pm.cdb[[i]])
```

```
pmc.cdb.df <- data.frame("id"=attr(datT,which="row.names"),papl[[i]]$PaP[,1:length(nivY)])
if(!is.null(dim(pmc.cdb.df)))
names(pmc.cdb.df)[2:(length(nivY)+1)] <- paste("CDB_",nivY,sep="")
}
}
###Predicts
if(predicts.out){
###2.Predicts para TODOS (menos pa MCs)
listdfs <- list("YObs"=data.frame("id"=attr(datT,which="row.names"),"Yobs"=datT[,columnaY]),
if(is.data.frame(pm.cart.df)) "cart"=pm.cart.df,
if(is.data.frame(pm.svm.df)) "svm"=pm.svm.df,
if(is.data.frame(pm.rf.df)) "rf"=pm.rf.df,
if(is.data.frame(pm.lgt.df)) "logit"=pm.lgt.df,
if(is.data.frame(pm.boost.df)) "boost"=pm.boost.df,
if(is.data.frame(pm.cdb.df)) "cdb"=pm.cdb.df)
if(any(unlist(lapply(listdfs,function(x) any(is.data.frame(x)))))){
#DF generado x Reduce(merge(...)), ordenado x nro 'correcto' obs
cvY_noOrd <- Reduce(f=function(...) merge(...,by="id",all=TRUE),listdfs)
#cvY_noOrd$t1s_pred <- round(rowSums(cvY_noOrd[,4:ncol(cvY_noOrd)])/length(mimp),2) comp.vs.Yobs
    [[i]] <- cbind("OrdxObs"=rownames(cvY_noOrd)[order(rownames(cvY_noOrd))],cvY_noOrd[order(
    rownames(cvY_noOrd)),])
}
##Prob a posteriori
listdfsp <- list("YObs"=data.frame("id"=attr(datT,which="row.names"),"Yobs"=datT[,columnaY]),
if(is.data.frame(pmc.cart.df)) "cart"=pmc.cart.df,
if(is.data.frame(pmc.svm.df)) "svm"=pmc.svm.df,
if(is.data.frame(pmc.rf.df)) "rf"=pmc.rf.df,
if(is.data.frame(pmc.lgt.df)) "logit"=pmc.lgt.df,
if(is.data.frame(pmc.boost.df)) "boost"=pmc.boost.df,
if(is.data.frame(pmc.cdb.df)) "cdb"=pmc.cdb.df)
if(any(unlist(lapply(listdfsp,function(x) any(is.data.frame(x)))))){

cvY_noOrd_p <- Reduce(f=function(...) merge(...,by="id",all=TRUE),listdfsp)
comp.vs.papc[[i]] <- cbind("OrdxObs"=rownames(cvY_noOrd_p)[order(rownames(cvY_noOrd_p))],cvY_
    noOrd_p[order(rownames(cvY_noOrd_p)),])
}
}
}
###3.ModelosConsenso (afuera de loop xq precisan YprVsYob,YprVsPaP)
mc <- FALSE
mocimp <- c("vm","mp","waauc")
mocimp <- mocimp[sort(match(x=moc,table=mocimp))]
if(length(mocimp)>0)
mc <- TRUE
if(mc){
lmi <- length(mimp)
part <- max(parti)
ypr <- comp.vs.Yobs
pap <- comp.vs.papc
mg <- vector(mode="list",length=part)
if("vm" %in% moc){
for(k in 1:part){
### MeCo1::VM (1 si >50% son 1, 0 resto; en caso de empate asignar aleatoriamente)
vm0 <- apply(X=ypr[[k]][,4:lmi+3],MARGIN=1,FUN=function(x) sum(x=="0"))
vm1 <- apply(X=ypr[[k]][,4:lmi+3],MARGIN=1,FUN=function(x) sum(x=="1"))
dfvm <- data.frame("VM_0"=vm0,"VM_1"=vm1)
ypr[[k]]$VM <- factor(levels(ypr[[k]]$Yobs)[max.col(dfvm)])
pap[[k]]$VM_0 <- dfvm$VM_0
pap[[k]]$VM_1 <- dfvm$VM_1
```

```
}
}
if("mp" %in% moc){
for(k in 1:part){
### MeCo2::MeanProb
mp0 <- apply(X=pap[[k]][,seq(from=4,to=(lmi*2)+2,by=2)],MARGIN=1,FUN=function(x) sum(x)/length(
    mimp))
mp1 <- apply(X=pap[[k]][,seq(from=5,to=(lmi*2)+3,by=2)],MARGIN=1,FUN=function(x) sum(x)/length(
    mimp))
dfmp <- data.frame("MP_0"=mp0,"MP_1"=mp1)
ypr[[k]]$MP <- factor(levels(pap[[k]]$Yobs)[max.col(dfmp)])
pap[[k]]$MP_0 <- dfmp$MP_0
pap[[k]]$MP_1 <- dfmp$MP_1
}
}
if("waauc" %in% moc){
### MeCo3::WA-AUC (promedio ponderado)
wa <- data.frame(matrix(nrow=part,ncol=lmi))
vec <- seq(from=5,to=(lmi*2)+3,by=2)
for(k in 1:part){
for(j in seq_along(vec)){
wa[k,j] <- furoc(pred=pap[[k]][,vec[j]],patroro=pap[[k]][,3],med="auc")
}
}
names(wa) <- paste("AUROC_",mimp,sep="")
for(k in 1:part){
m0 <- pap[[k]][,seq(from=4,to=(lmi*2)+2,by=2)]
wm0 <- as.matrix(m0)%*%t(as.matrix(wa))
pap[[k]]$wau0 <- rowSums(wm0)
m1 <- pap[[k]][,seq(from=5,to=(lmi*2)+3,by=2)]
wm1 <- as.matrix(m1)%*%t(as.matrix(wa))
pap[[k]]$wau1 <- rowSums(wm1)
pap[[k]]$WAU <- factor(levels(pap[[k]]$Yobs)[max.col(data.frame(pap[[k]]$wau0,pap[[k]]$wau1))])
}
}
#Juntamos todo en df
for(k in 1:part){
mg_df <- merge(ypr[[k]],pap[[k]],by.x="OrdxObs",by.y="OrdxObs")
mg[[k]] <- mg_df[,c(1:10,which(names(mg_df) %in% c("MP","WAU")))]
if("vm" %in% moc)
vml[[k]] <- table(mg[[k]]$Yobs.x,factor(mg[[k]]$VM,levels=nivY),dnn=c("YObs","YPred")) if("mp" %
    in% moc)
mpl[[k]] <- table(mg[[k]]$Yobs.x,factor(mg[[k]]$MP,levels=nivY),dnn=c("YObs","YPred"))
if("waauc" %in% moc)
waul[[k]] <- table(mg[[k]]$Yobs.x,factor(mg[[k]]$WAU,levels=nivY),dnn=c("YObs","YPred"))
}
}
# Listas de salidas: indicadores de matrices de confusion
le.trf <- if("RF" %in% modelosML) errp(trf)
le.tmlg <- if("Logit" %in% modelosML) errp(t.mlg1)
le.tsvm <- if("SVM" %in% modelosML) errp(tsvm)
le.tboost <- if("Boosting" %in% modelosML) errp(tboost)
le.tcb <- if("CDB" %in% modelosML) errp(tcb)
le.tcart <- if("CART" %in% modelosML) errp(tcart)
if(mc){
le.erpv <- if("vm" %in% moc) errp(vml)
le.erpm <- if("mp" %in% moc) errp(mpl)
le.erpw <- if("waauc" %in% moc) errp(waul)
```

```
lge.gr <- list(le.tcart$todos.erp,le.tmlg$todos.erp,le.tsvm$todos.erp,le.trf$todos.erp,le.tboost
    $todos.erp,le.tcb$todos.erp,
le.erpv$todos.erp,le.erpw$todos.erp,le.erpm$todos.erp)
} else {
lge.gr <- list(le.tcart$todos.erp,le.tmlg$todos.erp,le.tsvm$todos.erp,le.trf$todos.erp,le.tboost
    $todos.erp,le.tcb$todos.erp)
}
#Tiempo de corridas:
fint <- Sys.time()
tej <- fint-init
###4.SALIDA FINAL: 1)infoY, 2)metadat$, 3)asm$(mod1,mod2,...)$(MdConf,i...,pred_)
out <- list("Metadatos"=list("Y"=names(datE[columnaY]),tipoY=tipoY,"Formula"=frml,mimp=mimp,moc=
    moc,
parti=max(parti),propme=propme,permt=permt,perme=perme,"TiempoEjec"=round(tej,2)),
"ASM"=list("CART"=list("MdConf_CART"=tcart,"iCART"=le.tcart,"pred_CART"=pm.cart),
"Logit"=list("MdConf_Logit"=t.mlg1,"iLogit"=le.tmlg,"SA_Logit"=m.logit.sa,"pred_Logit"=pm.logit2
    ),
"SVM"=list("MdConf_SVM"=tsvm,"iSVM"=le.tsvm,"pred_SVM"=pm.svm),
"RF"=list("MdConf_RF"=trf,"iRF"=le.trf,"pred_RF"=pm.rf),
"Boost"=list("MdConf_Boost"=tboost,"iBoost"=le.tboost,"pred_Boost"=pm.boost),
"CDB"=list("MdConf_CDB"=tcb,"iCDB"=le.tcb,"pred_CDB"=pm.cdb,em.alf=em.alf),
"MC_VM"= if(predicts.out & "vm" %in% moc) list("MdConf_VM"=vml,"iVM"=le.erpv),
"MC_MP"= if(predicts.out & "mp" %in% moc) list("MdConf_MP"=mpl,"iMP"=le.erpm),
"MC_WA"= if(predicts.out & "waauc" %in% moc) list("MdConf_WA"=waul,"iWA"=le.erpw),
"MC_Todos"= if(predicts.out) list("Pred"=mg),
lge.gr=lge.gr,"YprVsYob"=comp.vs.Yobs,"YprVsPaP"=comp.vs.papc))
class(out) <- "asm"
return(invisible(out))
}
```