COMPARACIÓN ENTRE EL MÉTODO TRADICIONAL Y ALGUNOS BASADOS EN INTELIGENCIA ARTIFICIAL PARA EL ESTUDIO DEL RIESGO CREDITICIO EN INSTITUCIONES FINANCIERAS COLOMBIANAS

DIANA MARCELA ARANGO CORREA
LAURA JULIANA COLMENARES COLMENARES
ISABEL CRISTINA RAVE CONTRERAS

UNIVERSIDAD EAFIT

ESCUELA DE ADMINISTRACIÓN

MAESTRÍA EN ADMINISTRACIÓN DE RIESGOS

MEDELLÍN

2018

COMPARACIÓN ENTRE EL MÉTODO TRADICIONAL Y ALGUNOS BASADOS EN INTELIGENCIA ARTIFICIAL PARA EL ESTUDIO DEL RIESGO CREDITICIO EN INSTITUCIONES FINANCIERAS COLOMBIANAS

Trabajo presentado como requisito parcial para optar al título de magíster en Administración de Riesgos

DIANA MARCELA ARANGO CORREA[1]
LAURA JULIANA COLMENARES COLMENARES[2]
ISABEL CRISTINA RAVE CONTRERAS[3]

Asesor: Milton Alfonso Martínez Negrete, M. Sc.

UNIVERSIDAD EAFIT

ESCUELA DE ADMINISTRACIÓN

MAESTRÍA EN ADMINISTRACIÓN DE RIESGOS

MEDELLÍN

2018

[1] diarango27@hotmail.com
[2] lauracolmenares@outlook.com
[3] Isabel.rave@gmail.com

Resumen

Los modelos de inteligencia artificial son un problema abierto para aplicación en diversos campos de la ciencia y la identificación de relaciones entre variables, en especial cuando la distribución de los sucesos no depende de una función lineal. Al considerar lo anterior, esta investigación busca establecer una comparación entre un método tradicional utilizado para el seguimiento del comportamiento crediticio y modelos avanzados de inteligencia artificial.

Las guías que en la actualidad existen en el país y que están dadas por la Superintendencia Financiera de Colombia para el manejo del riesgo de crédito, así como los estándares internacionales, como Basilea II, Basilea III y Solvency, están basados en regresiones logísticas y en análisis discriminantes, modelos que usan las entidades financieras para medir el comportamiento crediticio, por lo que se propone explorar la utilidad de nuevas metodologías.

En este trabajo se aborda uno de los métodos tradicionales utilizados en las instituciones financieras, esto es, la regresión logística, y se compara con métodos alternativos, como las redes neuronales y los bosques aleatorios.

A partir de la revisión de literatura y mediante la utilización de una base de datos suministrada por una entidad bancaria, se seleccionaron las variables dependientes y la variable de respuesta, se calibraron los modelos de regresión logística, lo mismo que bosques aleatorios y redes neuronales en el aplicativo *Microsoft Azure Machine Learning* y se compararon entre sí con indicadores de precisión y exactitud, como la curva de tipo ROC (por las iniciales de la expresión en inglés *receiver operating characteristic*) y la matriz de confusión; se obtuvieron para los modelos de inteligencia artificial resultados tan buenos como el tradicional, por lo que pueden ser utilizados por el sector financiero como métodos alternos o complementarios en el análisis de riesgo de crédito.

Palabras clave: riesgo crediticio, puntaje de crédito, inteligencia artificial, redes neuronales, bosques aleatorios.

Abstract

Artificial intelligence models are an open problem for application in various fields of science and search of variable relationships especially when the distribution of events doesn't depend on a linear function; through this work we want to compare the traditional method most used for credit behavior monitoring with advanced models of artificial intelligence.

The guides that exist in Colombia for management of credit risk are given by the Financial Superintendence of Colombia, international standards such as Basel II, Basel III and Solvency are based on the logistic regression and the discriminant analysis, models used by financial institutions in Colombia to measure credit behavior, thus we carried out an investigation to explore the utility of new models.

This paper addresses one of the traditional methods used in financial institutions, that is, logistic regression, and compares it with alternative methods such as neural networks and random forests.

From the literature review and using a database provided by a banking entity, the dependent variables and the response variable are selected, the logistic regression models, random forests and neural networks are calibrated in the Microsoft Azure Machine Learning application and they are compared to each other with indicators of precision and accuracy such as ROC (from receiver operating characteristic) curve and confusion matrix, obtaining for the models of artificial intelligence, results as good as the traditional one; so they can be used by the financial sector as alternate and / or complementary methods in the analysis of credit risk.

Key words: *credit risk, credit scoring, artificial intelligence, neural networks, random forests.*

Introducción

La Superintendencia Financiera de Colombia (1995) definió que las entidades financieras deben velar por mantener una adecuada gestión del riesgo de crédito y para ello estableció, en el capítulo II de la circular básica contable y financiera, las normas mínimas que debe contener un sistema de administración de riesgo crediticio (SARC) con el fin de supervisar el ciclo completo de crédito, que se inicia con los procesos de otorgamiento y seguimiento y que termina con la regulación de la cobranza.

Asobancaria (2016) plantea que, como resultado de la desaceleración económica durante dicho año, generada sobre todo por el bajo crecimiento de la actividad productiva, el aumento en las tasas de interés y una menor demanda interna, junto con las expectativas ante una nueva reforma tributaria, que empezó a regir en el país el 1 de enero de 2017, la cartera crediticia del sector financiero se ha visto afectada y muestra un crecimiento en los indicadores de deterioro frente a los últimos años. Dado lo anterior, las entidades deben mejorar sus sistemas de seguimiento de crédito con el fin de mantener una adecuada gestión del riesgo y disminuir pérdidas futuras.

Por su parte, Samaniego Medina (2007, p. 16) define el riesgo de crédito como la probabilidad de cambio en la calidad crediticia de un emisor y los bancos e instituciones financieras son las entidades más vulnerables frente al riesgo de crédito.

Los métodos tradicionales, como regresiones logísticas y multivariada, entre otros, utilizan por lo general variables cuantitativas que modelan la variable de respuesta (binaria), en este caso el incumplimiento del cliente. Sin embargo, este tipo de modelos puede arrojar resultados con baja precisión, por lo que se hace necesario recurrir a métodos alternativos para mejorar el tratamiento de los datos.

Los modelos de redes neuronales y de bosques aleatorios son esquemas matemáticos que mediante algoritmos e iteraciones permiten relacionar variables

de entrada y, a través del aprendizaje, modelar la variable de respuesta, por lo que dichas metodologías toman un papel fundamental en la evaluación y el seguimiento del riesgo de crédito.

Desarrollar una adecuada herramienta para el otorgamiento, el control y el monitoreo del riesgo de crédito es una actividad muy importante para las entidades bancarias. En la actualidad, las instituciones financieras en Colombia utilizan el modelo de regresión logística y el análisis discriminante para predecir el comportamiento futuro de sus clientes.

En la primera parte de esta investigación se propone una revisión de literatura en la que se exploran los modelos estadísticos más usados para la clasificación crediticia y las técnicas de inteligencia artificial, como un enfoque alternativo de clasificación. Más tarde se identifican las variables de mayor correlación con la de respuesta, luego se construyen un bosque aleatorio y una red neuronal y se comparan con el método tradicional.

Objetivo general

Realizar un análisis comparativo entre el método tradicional utilizado para el seguimiento de crédito de personas naturales y algunos modelos de inteligencia artificial en carteras crediticias de vehículos de una entidad financiera colombiana.

Objetivos específicos

- Describir el método tradicional actual para el seguimiento del riesgo crediticio.
- Describir los modelos de inteligencia artificial como bosques aleatorios, redes neuronales y árboles de decisión.
- Identificar las variables cualitativas y cuantitativas que presentan alta correlación con el comportamiento de pago de las obligaciones crediticias.
- Comparar los resultados entre los modelos construidos, es decir, entre el método tradicional y los modelos de inteligencia artificial.

Marco de referencia

1. Referentes normativos

Este apartado pretende exponer el contexto actual de las diferentes normas que rigen en Colombia la gestión del riesgo de crédito para entidades financieras y se concentra en la etapa de seguimiento de cartera.

1.1. Superintendencia Financiera de Colombia

La Superintendencia Financiera de Colombia tiene la función de supervisar la gestión integral de los riesgos de crédito (RC) a través de la circular 100, capítulo II (Reglas relativas a la gestión del riesgo crediticio), que señala los principios, los criterios generales y los parámetros mínimos que se deben observar para el diseño, el desarrollo y la aplicación del Sistema de Administración del Riesgo Crediticio (SARC). Para el seguimiento de cartera de consumo cuenta con un modelo de referencia para calificar y provisionar a los clientes, que parte del método tradicional, esto es, la regresión logística, que se describe a continuación:

Modelo de referencia para cartera de consumo (MRCO)

Mediante el análisis del comportamiento del cliente, las compañías financieras buscan pronosticar el comportamiento futuro y, con base en ello, diseñar estrategias de cobranza y recuperación de cartera.

La Superintendencia Financiera de Colombia (1995) describió, en la circular 100, capítulo 2, anexo 5, un modelo de referencia para la cartera de consumo (MRCO). Allí precisó que dicho modelo permite clasificar y calificar según el riesgo a los sujetos de crédito, de acuerdo con los siguientes segmentos: automóviles, tarjeta de crédito y otros.

El modelo permite calcular un puntaje (*score*) a través de una regresión logística y está dado por la aplicación de la siguiente fórmula:

$$Puntaje = \frac{1}{1+e^{-z}}$$

Donde Z se modela como una función lineal de variables explicativas, lo que como resultado la función que se puede observar en el anexo 1 para el segmento de automóviles para personas naturales.

La descripción del modelo MRCO, la fórmula descrita en el anexo 1 y la presentación de las categorías de riesgo se tomaron de la circular 100, capítulo 2, anexo 5, de la Superintendencia Financiera de Colombia (2005).

A partir de la función descrita, el marco normativo establece una serie de categorías de riesgo por probabilidad de incumplimiento para el segmento de automóviles, que se limitan de la siguiente manera:

- **Categoría AA**: se incluyen aquellos créditos con una capacidad de pago óptima y un comportamiento excelente, en los que el recaudo se da en los términos acordados.
- **Categoría A:** reflejan una capacidad de pago y un comportamiento apropiado porque hay estabilidad en los recaudos.
- **Categoría BB**: los créditos en esta categoría muestran debilidades en la capacidad de pago y un comportamiento crediticio que de manera potencial puede afectar el recaudo de las obligaciones.
- **Categoría B:** crédito caracterizados por demostrar insuficiencia en la capacidad de pago y un comportamiento crediticio deficiente que afecta el recaudo.
- **Categoría CC:** en ella se incluyen aquellos créditos que presentan graves insuficiencias en la capacidad de pago y su comportamiento.

La tabla 1 indica los puntajes de referencia dados por la Superintendencia Financiera para cada una de las categorías antes mencionadas.

Tabla 1. Categorías de riesgo por probabilidad de incumplimiento

	Puntaje hasta	
Calificación	**General - automóviles**	**Compañías de financiamiento comercial - automóviles**
AA	0.2484	0.21
A	0.6842	0.6498
BB	0.81507	0.905
B	0.94941	0.9847
CC	1	1

Fuente: Superintendencia Financiera de Colombia (2005)

1.2. Normas internacionales de contabilidad financiera (IFRS o NIIF)

International Accounting Standards Board (2000) definió, en la NIIF 9 (Instrumentos financieros) los principios para el reconocimiento y la valoración de los activos financieros y para el efecto cuenta con un apartado de deterioro e incobrabilidad en el que se determina la importancia de evaluar los cambios en el riesgo crediticio a partir de los modelos de puntuación para estimar el deterioro.

La norma no es explícita en la forma de identificar el deterioro del cliente; no obstante, se deben considerar criterios como el comportamiento, el nivel de riesgo y la situación financiera del cliente, lo que les permite a las entidades financieras, con base en los mencionados modelos, construir los propios para monitorear el riesgo de crédito a través de modelos estadísticos que permitan reflejar las pérdidas reales en los estados financieros de las compañías, con el fin de buscar el mejor de ellos por medio de la comparación entre los diferentes que sea posible plantear.

1.3. Basilea III

Este marco de referencia presenta los requisitos mínimos que debe tener una entidad en los sistemas de calificaciones, entre los que se tienen: dimensión, estructura de los sistemas, horizonte de evaluación, documentación y aplicación de los modelos para calificar y mantenimiento de datos (BIS, 2017).

Basilea III recoge los estándares mínimos que deben cumplir las entidades con el fin de fortalecer la regulación, la supervisión y la gestión de riesgos del sector bancario. Surgió con el objetivo de estabilizar el sistema bancario internacional en un momento de crisis económica. En 1988 se creó Basilea I, en el que se definió un capital regulatorio mínimo para hacer frente a los riesgos y en este acuerdo se estableció como capital el 8% sobre los activos ajustados por riesgo. Las urgencias del sector y la necesidad de incluir nuevos riesgos condujeron a que en 2004 se promulgase Basilea II, que buscó mejorar la asignación del capital para cubrir el riesgo. Incluye el riesgo operacional y permite a las entidades bancarias el uso de modelos internos aprobados con anticipación por el supervisor. Luego, en 2010, nació Basilea III, que se fundamenta en el incremento del nivel del capital, la mejora de la cobertura del riesgo, las limitaciones en el apalancamiento bancario, la mejora de liquidez bancaria y las limitaciones de la prociclicidad. Este marco tuvo su última modificación en 2017 y constituye una mejora en el tratamiento del riesgo de crédito porque ofrece un método de ponderación de riesgo más detallado, pues reduce la utilización de calificaciones externas (Gutiérrez López, 2014).

2. Modelos de seguimiento de crédito

Entre los métodos tradicionales usados para establecer un puntaje de comportamiento crediticio se encuentran la regresión logística y el análisis discriminante. A continuación se describe cada uno:

- Regresión logística: es el método más popular aplicado en el puntaje crediticio. Describe la relación existente entre una variable dependiente y una o más variables independientes. La principal característica es que la variable de respuesta o dependiente es binaria o dicotómica (Hosmer y Lemeshow, 2000).
- Análisis discriminante: es una técnica estadística propuesta por primera vez por Fisher en 1936 y busca clasificar un nuevo elemento observado en grupos ya conocidos. Es un método sencillo y genera resultados fáciles de interpretar; no obstante, es estático y puede presentar problemas, porque no se adapta con facilidad al entorno, por lo que surgieron otros métodos, asociados con la inteligencia artificial, que pueden tener más exactitud en la predicción y en la disminución de errores de tipo II (Ince y Aktan, 2009).
- Inteligencia artificial: fue introducida por primera vez por John McCarthy, pero solo hasta la conferencia de Dartmouth en 1956 se formalizó. La inteligencia artificial es la ciencia y la ingeniería de hacer máquinas inteligentes, en especial programas informáticos inteligentes. Está relacionada con la tarea de usar computadores para entender la inteligencia humana (McCarthy, 2007).

  Entre las nuevas técnicas adaptadas, Salazar Villano (2013, p. 420) afirma que

  > Con el desarrollo de las herramientas informáticas aplicadas a la actividad bancaria, conduce a que nuevas técnicas se incorporen para el tratamiento del riesgo, entre ellas las redes neuronales y los algoritmos evolutivos (Frydman, Altman y Kao, 1985), matrices de transición (Altman y Kao, 1992), máquinas de vectores de soporte (Vapnik, 1995; Moreno y Melo, 2011), y el esquema del valor en riesgo o VaR (Romero Meza, 2005), el cual se

complementa con métodos auxiliares como son las simulaciones de tipo Montecarlo (De Lara, 2006) o la simulación histórica con crecimientos absolutos o logarítmicos (Jorion, 2007).

A continuación se explican las técnicas de interés de inteligencia artificial aplicadas en la evaluación del riesgo crediticio en la presente investigación:

- Redes neuronales: son unos modelos que simulan el funcionamiento del cerebro humano con el fin de resolver problemas de clasificación. Se componen de un conjunto de nodos o neuronas que procesan información mediante modelos matemáticos multivariantes que utilizan procedimientos iterativos. Están dotadas de las siguientes características: capacidad de aprendizaje, algoritmos de aprendizaje inductivo que permiten su entrenamiento a partir de ejemplos, alta velocidad de procesamiento y tolerancia a fallas (Cazorla et al., 1999).
- Bosques aleatorios: son clasificadores conformados por una colección de árboles de decisión, en el que cada uno depende de vectores aleatorios independientemente distribuidos en forma idéntica y en el que arrojan un voto unitario para cada variable X (Breiman, 2001).

  Los árboles aleatorios han demostrado tener una buena capacidad predictiva porque mejoran los resultados de métodos tradicionales (Ala'raj y Abbod, 2016).

Metodología

La presente investigación tuvo un enfoque cuantitativo con alcance correlacional, debido a que el propósito fue examinar la relación que existe entre algunas variables del comportamiento de pago del cliente con la probabilidad de que se materialice el riesgo de crédito.

1. Determinación de los datos

Para construir los modelos de puntuación crediticia con el propósito de medir el comportamiento se utilizaron datos secundarios, privados y pertenecientes un solo producto (créditos de vehículos), que fueron proporcionados por una de las principales entidades financieras en Colombia.

La información suministrada sirvió para construir variables dependientes cualitativas y cuantitativas y una variable dicotómica de respuesta, con el fin de modelar las diferentes metodologías investigadas.

La entidad cuenta con una arquitectura en la que a través de aplicativos (fuentes) recoge el comportamiento de pago de cada una de las obligaciones y alimentan un repositorio de información que almacena datos de manera histórica, a partir del cual se extrae la información para la construcción de modelos y para realizar la analítica de riesgo de crédito de la entidad.

2. Técnica utilizada

La información histórica del comportamiento de los clientes puede ser analizada mediante diversas técnicas que permiten disminuir el riesgo de crédito, con la finalidad de ayudar a las entidades financieras en el desarrollo de alertas tempranas.

El proceso de extracción del conocimiento a partir de datos denominado *Knowledge Discovery in Databases* (KDD), según Fayyad, Piatetsky-Shapiro y Smyth (1996, p. 37), es “un proceso no trivial de identificar patrones válidos, nuevos, potencialmente útiles y en última instancia comprensible en los datos” (p. 40). De acuerdo con los autores, este proceso consta de las siguientes etapas, que sirvieron de guía para procesar la información utilizada en la presente investigación:

2.1. Selección y exploración: en esta fase se recopila e integra la información que se requiere, se define la fuente de la que se tomará la información relevante en el desarrollo del problema y se seleccionan las variables.

2.2. Preprocesamiento: se compone de las siguientes fases:

2.2.1. Limpieza de los datos: en ella se busca eliminar datos atípicos, valores faltantes e información incompleta que pueda afectar el resultado final.

2.2.2. Transformación: cuando se requiera que los datos deban ser transformados se utilizan los siguientes tipos:

- Transformaciones lógicas: se unen categorías y se convierten variables de intervalos en ordinales o nominales.
- Transformaciones lineales: surgen de alguna operación sobre las observaciones originales que no cambien la forma de distribución ni la distancia ni el orden entre los valores.

2.3 Minería de datos: en ella se decide cuál técnica de procesamiento de datos se utilizará en la minería. En la presente investigación se utilizaron los bosques aleatorios, las redes neuronales y la regresión logística como método de referencia; la decisión se tomó con base en resultados de estudios previos que demuestran que estas técnicas son las que mejores resultados entregan en la predicción (Ince y Aktan, 2009).

2.4 Evaluación e interpretación de resultados: busca evaluar patrones y comparar modelos. En esta investigación se utilizaron el análisis de tipo ROC y la matriz de confusión como medidas de precisión, exactitud y error.

Instrumento

1. Construcción de la base de datos

Se extrajo la información correspondiente al total de créditos de vehículos y los datos se dividieron en una primera ventana de tiempo de observación, comprendida entre enero de 2014 y septiembre de 2015, que se usó para la construcción de los

modelos, y otra ventana de tiempo de análisis, que contiene el comportamiento de la obligación un año después y que se usó para construir la variable de respuesta. Las fechas fueron elegidas con el fin de incorporar el mayor número de datos posibles de 2016, dado que fue un año en el que se presentó alto deterioro del portafolio de vehículos.

Se cuenta con cortes mensuales, lo que condujo a nueve puntos de análisis. La base de datos extraída contiene la siguiente información, utilizada para la creación de modelos de comportamiento:

Tabla 2. Base de datos inicial

NOMBRE DEL CAMPO	NOMBRE DE LA VARIABLE	DESCRIPCIÓN DE LA VARIABLE	TIPO
FA	Fecha de Análisis	Fecha de referencia para el análisis en la cual se recopila la información con formato DDMMAAAA	Fecha
FO	Fecha de Otorgamiento	Fecha de desembolso del crédito (DDMMAAAA)	Fecha
CEFV21	Fecha de Vencimiento	Ultima fecha de pago del cliente, formato DDMMAAAA	Fecha
CLASE	Clase	Indica si es un crédito de vehículo otorgado para consumo privado o para trabajar (1: trabajo, 2: privado)	Número
ID	Identificación	Campo único que identifica al cliente, que se encuentra encriptado con el fin de proteger la confidencialidad de la información	Número

GH	Segmento de mercado	Clasificación del portafolio según la política de la compañía y está dado con base en los ingresos del cliente (preferencial, personal plus, personal y emprendedor)	Texto
NUEVO	Vehículo nuevo	Registra si el crédito desembolsado es para un vehículo nuevo o usado (0: usado, 1: nuevo)	Dicotómico
RESTFA_OBLIG	Vehículo reestructurado en fecha de análisis	Marca si el cliente hizo una restructuración del crédito de vehículo en la fecha de análisis	Dicotómico
REST12M_OBLIG	Vehículo reestructurado en el año siguiente a la fecha de análisis	Marca si el cliente hizo restructuración del crédito de vehículo un año después de la fecha de análisis	Dicotómico
CASTFA_OBLIG	Vehículo castigado en fecha de análisis	Marca si el cliente se castigó en la fecha de análisis	Dicotómico
CAST12M_OBLIG	Vehículo castigado en el año siguiente a la fecha de análisis	Marca si el cliente se castigó un año después de la fecha de análisis	Dicotómico
ATMFA_OBLIG	Altura de mora de la obligación en la fecha de análisis	Número de días de mora del crédito de vehículo en la fecha de análisis	Número
MAX12MD_OBLIG	Máxima altura de mora de la obligación un año después	Máximo número de días de mora del crédito de vehículo un año después de la fecha de análisis	Número

MAX12MA_OBLIG	Máxima altura de mora de la obligación un año antes	Máximo número de días de mora del crédito de vehículo un año antes de la fecha de análisis	Número
SALDOK_FA	Saldo de capital	Saldos del crédito de vehículo en la fecha de análisis	Decimal
CEAM21	Valor de los desembolsos	Valor desembolsado al momento del otorgamiento	Decimal
MADUREZ	Plazo remanente	Plazo que le falta al cliente para terminar de pagar el crédito de vehículo; está dado en número de meses	Número
PLAZO	Plazo del crédito	Plazo en el que fue otorgado el crédito de vehículo; está dado en número de meses	Número
VL_GARANTIA_FA	Valor de la garantía	Valor de mercado de la garantía en la fecha de análisis	Decimal
M01	Máxima altura de mora en el mes 1	Máximo número de días de mora de todas las obligaciones del cliente en el mes 12 antes de la fecha de análisis	Número
M02	Máxima altura de mora en el mes 2	Máximo número de días de mora de todas las obligaciones del cliente en el mes 11 antes de la fecha de análisis	Número
M03	Máxima altura de mora en el mes 3	Máximo número de días de mora de todas las obligaciones del cliente en el mes 10 antes de la fecha de análisis	Número
M04	Máxima altura de mora en el mes 4	Máximo número de días de mora de todas las obligaciones del	Número

		cliente en el mes 9 antes de la fecha de análisis	
M05	Máxima altura de mora en el mes 5	Máximo número de días de mora de todas las obligaciones del cliente en el mes 8 antes de la fecha de análisis	Número
M06	Máxima altura de mora en el mes 6	Máximo número de días de mora de todas las obligaciones del cliente en el mes 7 antes de la fecha de análisis	Número
M07	Máxima altura de mora en el mes 7	Máximo número de días de mora de todas las obligaciones del cliente en el mes 6 antes de la fecha de análisis	Número
M08	Máxima altura de mora en el mes 8	Máximo número de días de mora de todas las obligaciones del cliente en el mes 5 antes de la fecha de análisis	Número
M09	Máxima altura de mora en el mes 9	Máximo número de días de mora de todas las obligaciones del cliente en el mes 4 antes de la fecha de análisis	Número
M10	Máxima altura de mora en el mes 10	Máximo número de días de mora de todas las obligaciones del cliente en el mes 3 antes de la fecha de análisis	Número
M11	Máxima altura de mora en el mes 11	Máximo número de días de mora de todas las obligaciones del cliente en el mes 2 antes de la fecha de análisis	Número
M12	Máxima altura de mora en el mes 12	Máximo número de días de mora de todas las obligaciones del	Número

		cliente en el mes 1 antes de la fecha de análisis	
MAX12MD_CLIENTE	Máxima altura de mora del cliente un año después	Máximo número de días de mora del cliente en todos sus productos activos durante los 12 meses después a la fecha de análisis	Número
MAX12MA_CLIENTE	Máxima altura de mora del cliente un año antes	Máximo número de días de mora del cliente en todos sus productos activos durante los 12 meses antes a la fecha de análisis	Número
ATMFA_CLIENTE	Máxima altura de mora del cliente en la fecha de análisis	Máximo número de días de mora del cliente en todos sus productos activos en la fecha de análisis	Número
G	Calificación	Calificación interna de riesgo otorgada por la entidad a través de un puntaje de comportamiento; va desde G1 hasta G8, con G1 para el menor riesgo y G8 para el mayor riesgo G1 < 417 puntos G2 entre 417 y 519 puntos G3 entre 520 y 599 puntos G4 entre 600 y 649 puntos G5 entre 650 y 689 puntos G6 entre 690 y 729 puntos G7 entre 730 y 769 puntos G8 entre 770 y 830 puntos	Texto

Fuente: elaboración propia

2. Creación de nuevas variables a partir de la información de origen

Con el fin de caracterizar de mejor manera el comportamiento de los datos, se crearon nuevas variables de comportamiento extraídas del vector de mora de los 12 meses anteriores a la fecha de análisis, que se observan en la tabla 3; dichas variables se construyeron a partir de medidas estadísticas.

Tabla 3. Creación de nuevas variables

NOMBRE DEL CAMPO	NOMBRE DE LA VARIABLE	DESCRIPCIÓN DE LA VARIABLE	TIPO
MAX_6MA	Máxima mora en los últimos seis meses	Máximo número de días de mora del cliente en el crédito de vehículo entre el mes 1 y el mes 6 antes de la fecha de análisis	Número
MAX_3MA	Máxima mora en los últimos tres meses	Máximo número de días de mora del cliente en el crédito de vehículo entre el mes 1 y el mes 3 antes de la fecha de análisis	Número
CONT_0MA	Conteo de mora cero	Número de observaciones con mora cero en los últimos 12 meses del crédito de vehículo	Número
CONT_30MA	Conteo de mora 30+	Número de observaciones con mora entre 30 y 59 días en los últimos 12 meses del crédito de vehículo	Número
CONT_60MA	Conteo de mora 60+	Número de observaciones con mora entre 60 y 89 días en los últimos 12 meses del crédito de vehículo	Número
CONT_90MA	Conteo de mora 90+	Número de observaciones con mora mayor o igual que 90 días en los	Número

		últimos 12 meses del crédito de vehículo	
STD_12MA	Desviación estándar	Desviación estándar del número de días en mora del crédito de vehículo en los últimos 12 meses	Decimal

Fuente: elaboración propia

3. Transformación de variables categóricas en numéricas:

- Variable de segmento (GH)

Esta variable en su origen es un conjunto de categorías que establecen el grupo al que pertenecen los créditos otorgados a los clientes, por ejemplo: personal, personal plus, preferencial y emprendedor y se trasforma en numérica en función del riesgo del segmento; para el efecto se estima el porcentaje de clientes en incumplimiento respecto del total de clientes del segmento (clasificación interna del banco de acuerdo con los ingresos del tipo de cliente) y se numeran de 1 a 4, con 1 para la categoría de menor riesgo y 4 para la de mayor riesgo; para la asignación se utilizó la siguiente fórmula en cada segmento y se obtuvo como resultado la siguiente transformación, indicada en la tabla 4.

Nivel de riesgo del segmento = $\frac{Número\ de\ clientes\ en\ incumplimiento\ del\ segmento}{Número\ de\ clientes\ totales\ del\ segmento}$

Tabla 4. Transformación de la variable de segmento

GH	GHT
Preferencial	1
Personal Plus	2
Personal	3
Emprendedor	4

Fuente: elaboración propia

- Variable de calificación (G)

Es una variable que la entidad asigna según el nivel de riesgo y de la misma forma se realiza la transformación propuesta en la tabla 5.

Tabla 5. Transformación de la variable de calificación

G	GT
G1	1
G2	2
G3	3
G4	4
G5	5
G6	6
G7	7
G8	8

Fuente: elaboración propia

Como resultado se obtuvo la creación de nuevas variables indicadas en la tabla 6.

Tabla 6. Creación de variables transformadas

NOMBRE DEL CAMPO	NOMBRE DE LA VARIABLE	DESCRIPCIÓN DE LA VARIABLE	TIPO
GT	Calificación transformada	Transformación de la variable de calificación en numérica con base en el nivel de riesgo asignado por la compañía	Número
GHT	Segmento transformado	Transformación de la variable de segmento en numérica con base en	Número

	el nivel de riesgo asignado por la compañía

Fuente: elaboración propia

4. Construcción de la variable de respuesta

La variable de respuesta para el modelo fue el incumplimiento en los 12 meses después de la fecha de análisis; para ello se creó una variable llamada Default12M, que cumple las siguientes características:

- Variable dicotómica (dos categorías): 1 si el cliente entró en incumplimiento, 0 si el cliente no lo hizo.
- Clientes que tenga un número de días de mora mayor o igual a 90 días en los 12 meses de comportamiento después de la fecha de análisis.
- Clientes que en los 12 meses hubiesen sido reestructurados.
- Clientes que en los 12 meses hubiesen sido castigados.

5. Selección de variable explicativa

Las variables explicativas se seleccionaron por medio de una matriz de correlaciones, con el fin de evitar problemas de alta correlación entre ellas; cuando tal cosa pasa se busca eliminar la que tenga menor correlación con la variable de respuesta (incumplimiento).

6. Selección de variables correlacionadas con la variable de respuesta

En la tabla 7 se puede observar la matriz de correlación de las variables explicativas con la variable de respuesta (Default12M), para seleccionar las que en realidad se correlacionan y explican el incumplimiento; este análisis se realizó mediante el método de Spearman (Díaz et al., 2014).

Tabla 7. Resultados de coeficientes de correlación mediante el método de Spearman

VARIABLES	CLASE	NUEVO	RESTFA_OBLIG	RESTFA_CLIENTE	CASTFA_CLIENTE
DEFAULT 12M	-0,0053	-0,00342	.	0,04933	.
	<.0001	0,0002	.	<.0001	.
	MAX12MA_OBLIG	AX12MA_CLIENTE	ATMFA_OBLIG	ATMFA_CLIENTE	SALDOK_FA
	0,35642	0,31571	0,42396	0,39214	-0,08143
	<.0001	<.0001	<.0001	<.0001	<.0001
	MADUREZ	PLAZO	VL_GARANTIA_FA	M01	M02
	0,02177	-0,0515	-0,04568	0,39743	0,36588
	<.0001	<.0001	<.0001	<.0001	<.0001
	M03	M04	M05	M06	M07
	0,33918	0,33492	0,32308	0,31225	0,29963
	<.0001	<.0001	<.0001	<.0001	<.0001
	M08	M09	M10	M11	M12
	0,29037	0,28026	0,27209	0,26433	0,2573
	<.0001	<.0001	<.0001	<.0001	<.0001
	MAX_6MA	MAX_3MA	CONT_0MA	CONT_30MA	CONT_60MA
	0,37869	0,39435	-0,22523	0,4131	0,32372
	<.0001	<.0001	<.0001	<.0001	<.0001
	CONT_90MA	STD_12MA	GHT	GT	
	0,23598	0,36933	0,0841	0,34102	
	<.0001	<.0001	<.0001	<.0001	

Nota: nivel de confianza del 99%

Fuente: elaboración propia

Cuando se calcularon los coeficientes de correlación entre las variables explicativas y el incumplimiento, no se sabía de antemano cuáles valores tomarán los mismos (debido a que provienen de una cartera de datos reales); para determinar cuáles valores se tomarían como altos se utilizaron las medidas de posicionamiento enlistadas en la tabla 8 y se dividieron en tres grandes grupos los resultados observados en la matriz; de esta manera se usaron las variables clasificadas en los percentiles más altos (mayor al 66%).

Tabla 8. Categorías correlacionadas con la variable de respuesta

PERCENTIL	COEFICIENTE DE CORRELACIÓN	NIVEL DE CORRELACIÓN
0 a 33%	0 a 0.259	Bajo
33% a 66%	0.260 a 0.331	Medio
Mayor que 66%	Mayor que 0.337	Alto

Fuente: elaboración propia

Al calcular los coeficientes de correlación se encontraron valores desde 0.02 hasta 0.42 y se observó que las correlaciones fueron relativamente bajas para las variables; sin embargo, de estos valores se tomaron los más altos, para lo cual se dividió el rango de las correlaciones en tres grupos iguales, alto, medio y bajo, por medio de los percentiles, que se pueden observar en la tabla 8; de esta manera, la categoría baja fue aquella que iba desde el primer dato hasta el percentil 33%, la media fue la que iba desde el percentil 33% hasta el 66% y, por último, la categoría alta fueron los datos que estaban por encima del percentil 66% y cuyo nivel de correlación fue alto, como se puede observar en la tabla 9.

Tabla 9. Selección de variables superiores al percentil 66%

NOMBRE DEL CAMPO	NOMBRE DE LA VARIABLE	COEFICIENTE DE CORRELACIÓN
MAX12MA_OBLIG	Máxima altura de mora de la obligación un año antes	0.35642
ATMFA_OBLIG	Altura de mora de la obligación en la fecha de análisis	0.42396
ATMFA_CLIENTE	Máxima altura de mora del cliente en la fecha de análisis	0.39214
M01	Máxima altura de mora en el mes 1	0.39743
M02	Máxima altura de mora en el mes 2	0.36588
M03	Máxima altura de mora en el mes 3	0.33918
MAX_3MA	Máxima mora en los últimos tres meses	0.39435
MAX_6MA	Máxima mora en los últimos seis meses	0.37869
CONT_30MA	Conteo de mora 30+	0.4131
STD_12MA	Desviación estándar	0.36933
GT	Calificación transformada	0.34102

Fuente: elaboración propia

Además, para las variables numéricas se aplicó el método de correlación de Pearson con el mismo criterio utilizado para el de Spearman y se obtuvieron como resultado dos nuevas variables que entraron al modelo, que se indican en la tabla 10.

Tabla 10. Selección de variables superiores al percentil 66% por método de Pearson

NOMBRE DEL CAMPO	NOMBRE DE LA VARIABLE	COEFICIENTE DE CORRELACIÓN
MAX12MA_CLIENTE	Máxima altura de mora del cliente un año antes	0.31571
M04	Máxima altura de mora en el mes 4	0.33492

Fuente: elaboración propia

7. Eliminación de variables explicativas altamente correlacionadas entre sí

Después de seleccionar las variables que tuvieron alta correlación con el incumplimiento, en la tabla 11 se observa el análisis de aquellas que podrían contener la misma información entre sí con el fin de no obtener un modelo parsimonioso y dejar las variables que de mejor manera pueden predecir el incumplimiento del cliente.

En la tabla 11 se indican las variables excluidas con base en la prueba de correlación de Spearman entre las variables seleccionadas en el punto anterior, con una correlación mayor al percentil 80 y se obtuvieron como resultado las variables que se muestran en la tabla 12.

Tabla 11. Selección de variables superiores al percentil 80%

	MAX12MA_OBLIG	MAX12MA_CLIENTE	ATMFA_OBLIG	ATMFA_CLIENTE	M01	M02	M03	M04	MAX_3MA	MAX_6MA	CONT_30MA	STD_12MA	GT
MAX12MA_OBLIG		0.81761	0.53713	0.50072	0.624	0.6119	0.5918	0.5817	0.79709	0.89828	0.59028	0.99479	0.6064
MAX12MA_CLIENTE	0.81761		0.45701	0.5359	0.5214	0.5155	0.5011	0.4956	0.65843	0.73819	0.52551	0.81709	0.6603
ATMFA_OBLIG	0.53713	0.45701		0.84925	0.5895	0.5417	0.5024	0.4882	0.59535	0.5713	0.4825	0.54187	0.506
ATMFA_CLIENTE	0.50072	0.5359	0.84925		0.5281	0.4871	0.4538	0.4408	0.54266	0.52701	0.44029	0.50499	0.5731
M01	0.62395	0.52143	0.58949	0.52806		0.5703	0.5264	0.5038	0.77431	0.68948	0.52665	0.62353	0.4893
M02	0.61188	0.51545	0.54166	0.48708	0.5703		0.556	0.5294	0.75915	0.67578	0.55966	0.61126	0.4892
M03	0.59176	0.50106	0.50242	0.45376	0.5264	0.556		0.5607	0.73564	0.65365	0.54795	0.59088	0.4839
M04	0.58168	0.49562	0.48815	0.44076	0.5038	0.5294	0.5607		0.56946	0.64209	0.55952	0.58037	0.48
MAX_3MA	0.79709	0.65843	0.59535	0.54266	0.7743	0.7592	0.7356	0.5695		0.88565	0.56719	0.80077	0.5771
MAX_6MA	0.89828	0.73819	0.5713	0.52701	0.6895	0.6758	0.6537	0.6421	0.88565		0.58206	0.90044	0.6107
CONT_30MA	0.59028	0.52551	0.4825	0.44029	0.5267	0.5597	0.548	0.5595	0.56719	0.58206		0.58659	0.4465
STD_12MA	0.99479	0.81709	0.54187	0.50499	0.6235	0.6113	0.5909	0.5804	0.80077	0.90044	0.58659		0.634
GT	0.60637	0.66032	0.50602	0.5731	0.4893	0.4892	0.4839	0.48	0.57707	0.6107	0.44645	0.634	

Nota: nivel de confianza del 99%

Fuente: elaboración propia

Tabla 12. Variables explicativas seleccionadas para el modelo

NOMBRE DEL CAMPO	NOMBRE DE LA VARIABLE
MAX12MA_OBLIG	Máxima altura de mora de la obligación un año antes
ATMFA_OBLIG	Altura de mora de la obligación en la fecha de análisis
M01	Máxima altura de mora en el mes 1
M02	Máxima altura de mora en el mes 2
M03	Máxima altura de mora en el mes 3
M04	Máxima altura de mora en el mes 4
MAX_3MA	Máxima mora en los últimos tres meses
CONT_30MA	Conteo de mora 30+
GT	Calificación transformada

Fuente: elaboración propia

8. Selección de la muestra

La población obtenida tiene un número de observaciones total de 1,217,810 (base original entregada por la entidad financiera).

Martínez Bencardino (2012, p. 688) indica que para calcular el tamaño de la muestra se utiliza la fórmula de muestreo basada en la estimación de la proporción para población finita con la fórmula descrita a continuación:

- Tamaño de muestra para estimar la proporción de una población infinita:

$$n_0 = \frac{z_{\alpha/2}^2 pq}{\mathrm{e}^2}$$

Nivel de confianza	99%
Error admisible	0.01
Proporción esperada	0.032
Tamaño de muestra	**2,086**

- Tamaño de muestra ajustado para poblaciones finitas:

$$n = \frac{n_0}{1 + \frac{n_0}{\mathrm{N}}}$$

Tamaño de población	1,217,810
Tamaño muestral	2,086
Tamaño de muestra ajustado	**2,082**

Para estimar la proporción esperada se generó el indicador de cartera en mora mayor o igual a 90 días del portafolio de vehículos de la entidad financiera durante 2016 y se obtuvo un promedio de 3.2% y de esta forma se concluyó que la muestra apropiada tiene 2,082 observaciones como base para la construcción de los modelos.

Dado que la proporción esperada de clientes morosos fue baja, se debe balancear la muestra y para ello se eligió el 50% de los clientes que entraron en incumplimiento y 50% de los clientes que no lo hicieron, en ambos casos elegidos de manera aleatoria.

9. Modelo

Para la construcción de los modelos se utilizó la plataforma de Microsoft Corporation denominada *Azure Machine Learning Studio*, que es de la modalidad de aprendizaje de ca mencionada compañía que permite preparar datos, desarrollar experimentos e implementar modelos a escala en la nube.

En esta sección se pretende explicar el procedimiento realizado para la construcción y el entrenamiento de los modelos. Para este trabajo se utilizaron el bosque de decisión, la regresión logística y las redes neuronales como algoritmos de aprendizaje supervisado. Se entrenó el 70% de la base de datos obtenida después del análisis de correlación y el 30% restante se utilizó para probar los modelos. La anterior es una regla heurística, pues no existe un resultado claro en la literatura estadística que indique la mejor manera de partir una base de datos; todo depende del número de datos, del ruido que contienen y de su representatividad (Pérez López y Santín González, 2006).

Para parametrizar los modelos se estableció un rango en el que los parámetros podían variar y que se definió mediante pruebas de ensayo y error. Con posterioridad se utilizó el módulo de hiperparámetros, que permite optimizar el modelo y se seleccionó el indicador de exactitud para ser optimizado.

A continuación se describen los algoritmos y los parámetros utilizados para cada modelo:

9.1. Bosque de decisión

El algoritmo funciona mediante la construcción de múltiples árboles de decisión. La votación es una forma de agregación, en la que cada árbol en un bosque de decisión genera un histograma de frecuencias. El proceso de agregación suma estos histogramas y normaliza el resultado para obtener las "probabilidades" para cada etiqueta. Los árboles que tienen una alta confianza de predicción tendrán un mayor peso en la decisión final del conjunto (Microsoft Corporation, 2017).

Tabla 13. Parámetros del bosque aleatorio de decisión óptimo

PARÁMETRO	VALOR
Número mínimo de muestras por nodo	3
Número de divisiones aleatorias por nodo	4
Profundidad máxima de los árboles de decisión	4
Número de árboles de decisión	11

Fuente: elaboración propia

9.2. Regresión logística

El algoritmo busca predecir la probabilidad de que ocurra un evento por medio del ajuste un conjunto de datos a una función logística. Los parámetros con los que se optimiza el modelo y que se utilizaron para el entrenamiento se enumeran en la tabla 14.

Tabla 14. Parámetros de la regresión logística óptima

PARÁMETRO	VALOR
Tolerancia de la optimización	1.80535317E-05
Peso L1	0.577968538
Peso L2	0.731765747

Fuente: elaboración propia

9.3. Red neuronal

La mayoría de las tareas predictivas se pueden realizar con facilidad con solo una o algunas capas ocultas. Investigaciones recientes han demostrado que las redes neuronales profundas (DNN, por las iniciales de la expresión en inglés *deep neuronal network*) pueden ser muy efectivas en tareas complejas, como reconocimiento de imágenes o de voz, en las que se utilizan capas sucesivas para modelar niveles crecientes de profundidad semántica.

Para calcular la salida de la red para cualquier entrada dada se calcula un valor para cada nodo en las capas ocultas y en la de salida. Para cada nodo, el valor se establece mediante el cálculo de la suma ponderada de los valores de los nodos en la capa anterior y la aplicación de una función de activación a dicha suma ponderada (Microsoft Corporation, 2017).

La selección de los parámetros de la red se llevó a cabo en forma experimental por medio de pruebas con diferentes capas ocultas y al variar la cantidad de neuronas. Más tarde se utilizó el módulo de hiperparámetros para optimizar el rango seleccionado.

Tabla 15. Parámetros de la red neuronal óptima

PARÁMETRO	VALOR
Diámetro inicial de los pesos	0.1
Tasa de aprendizaje	0.0188296642
Función de pérdida	Error al cuadrado

Fuente: elaboración propia

En la tabla 16 se presentan los experimentos realizados, en los que se variaron la cantidad de capas ocultas y el número de neuronas por cada capa. Se seleccionó una red de tres capas y diez neuronas con una función de activación sigmoidal. Hubo un resultado similar con cuatro capas y 50 neuronas pero se seleccionó el primero porque optimizó el tiempo de procesamiento.

Tabla 16. Modelación de activación sigmoidal

Una capa					
Número de neuronas	**Área bajo la curva (AUC)**	**Exactitud**	**Precisión**	**Sensibilidad**	**Calificación**
10	0.904	0.832	0.895	0.773	0.829
50	0.904	0.834	0.899	0.773	0.831
100	0.904	0.832	0.895	0.773	0.829
200	0.904	0.832	0.895	0.773	0.829
300	0.904	0.834	0.899	0.773	0.773
400	0.904	0.832	0.895	0.773	0.829

Dos capas					
Número de neuronas	**Área bajo la curva (AUC)**	**Exactitud**	**Precisión**	**Sensibilidad**	**Calificación**
10	0.904	0.834	0.899	0.773	0.831
50	0.903	0.832	0.895	0.773	0.829
100	0.903	0.832	0.895	0.773	0.829
200	0.904	0.837	0.9	0.777	0.834
300	0.904	0.832	0.895	0.773	0.829
400	0.904	0.832	0.895	0.773	0.829
Tres capas					
Número de neuronas	**Área bajo la curva (AUC)**	**Exactitud**	**Precisión**	**Sensibilidad**	**Calificación**
10	**0.904**	**0.837**	**0.90**	**0.777**	**0.834**
50	0.903	0.832	0.895	0.773	0.829
100	0.094	0.837	0.9	0.777	0.834
200	0.904	0.832	0.895	0.773	0.829
300	0.904	0.834	0.899	0.773	0.831
400	0.904	0.834	0.899	0.773	0.831
Cuatro capas					
Número de neuronas	**Área bajo la curva (AUC)**	**Exactitud**	**Precisión**	**Sensibilidad**	**Calificación**
10	0.903	0.832	0.895	0.773	0.829
50	0.904	0.837	0.900	0.777	0.834

100	0.904	0.834	0.899	0.773	0.831
200	0.904	0.832	0.895	0.773	0.829
300	0.904	0.832	0.895	0.773	0.829
400	0.903	0.832	0.895	0.773	0.829

Fuente: elaboración propia

Resultados

Con la aplicación de la metodología se eligió el modelo con mayor capacidad de predicción del comportamiento del cliente y que fuese capaz de disminuir los niveles de error para de esta forma concluir si los modelos de inteligencia artificial aplicados en la cartera seleccionada tenían mejor capacidad predictiva que los métodos tradicionales.

Para la construcción de las pruebas se usaron los mismos datos y las mismas variables con el fin de permitir las comparaciones.

A continuación se presentan los resultados para cada uno de los modelos:

1. Regresión logística

Tabla 17. Pesos de las variables del modelo

PARÁMETRO	VALOR
GT	3.73193
Sesgo	-2.6956
ATMFA_OBLIG	2.65259
CONT_30MA	0.604678
MAX_3MA	0.480543
MAX12MA_OBLIG	0.449547
M02	0.411061
M01	0.343743
M03	0
M04	0

Fuente: elaboración propia

Según se muestra en la tabla 17, se concluye que la variable de calificación interna (GT) que otorga el banco a un cliente es la más influyente en el modelo, lo que demuestra que si un cliente fue calificado con riesgo alto es más probable que entre en incumplimiento.

Otra variable clave es la altura máxima de mora en la fecha de análisis. Cuando el banco hace la evaluación y detecta clientes que a la fecha presentan varios días de mora, es una señal de alerta porque es probable que los mismos dejen de pagar sus obligaciones financieras.

Las variables que resultaron poco relevantes fueron la mora en el mes 3 (M03) y la mora en el mes 4 (M04).

2. Bosque aleatorio

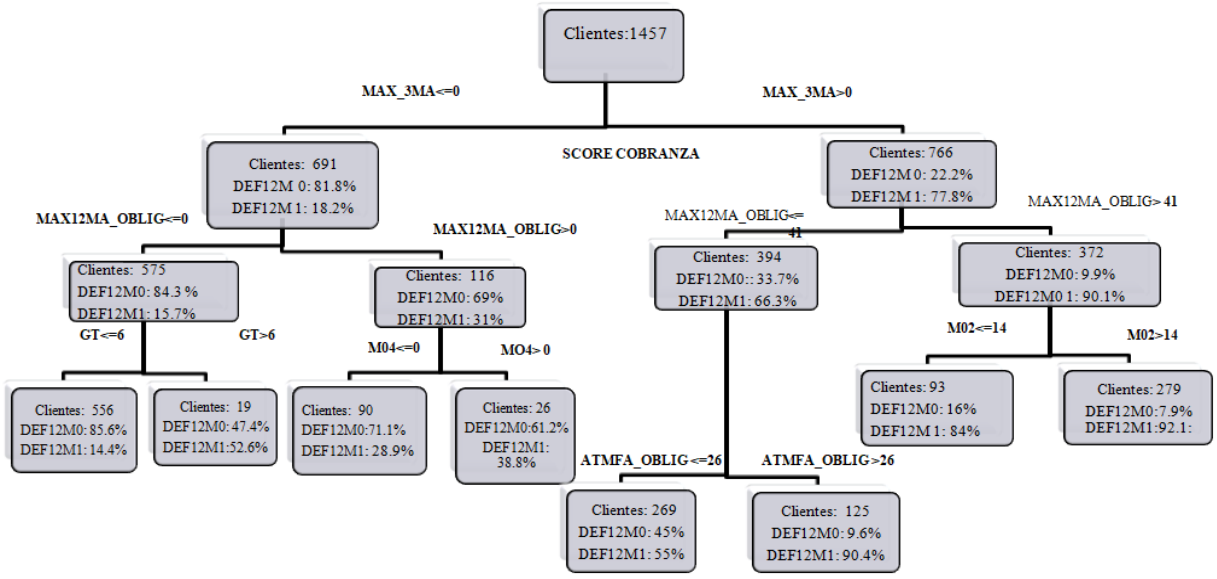
En la tabla 18 se muestran las reglas obtenidas para uno de los árboles construidos en el bosque aleatorio:

Tabla 18. Reglas obtenidas en un árbol construido en el bosque aleatorio

REGLA	PROBABILIDAD DE INCUMPLIMIENTO
MAX_3MA <= 0 MAX12MA_OBLIG <=0 GT > 6	52.6%
MAX_3MA <= 0 MAX12MA_OBLIG > 0 M04 > 0	38.8%
MAX_3MA > 0 MAX12MA_OBLIG <= 41 ATMFA_OBLIG > 26	90.4%
MAX_3MA > 0 MAX12MA_OBLIG > 41 M02 > 14	92.1%

Fuente: elaboración propia

Figura 1. Ejemplo de árbol de clasificación tomado del bosque aleatorio



Fuente: elaboración propia

A manera de ejemplo se toma un cliente cuya altura de mora en los últimos tres meses fue superior a 0: tiene un 77% de probabilidad de entrar en incumplimiento. Si, además, en los últimos 12 meses ha tenido retrasos en el pago de sus obligaciones superiores a 41 días, la probabilidad se incrementa al 90.1% y puede llegar a un 92.1% si el cliente ha tenido una mora mayor a 14 días en los últimos dos meses.

En el anexo 2 se muestran los árboles obtenidos con el programa *Azure Machine Learning* para el modelo optimizado.

3. Red neuronal

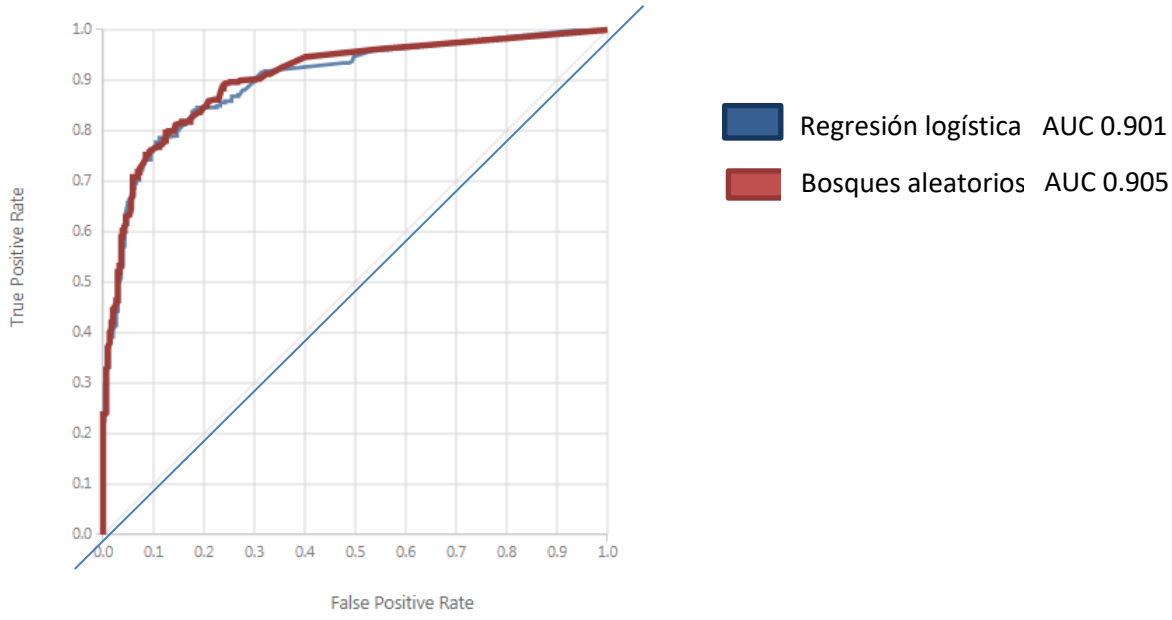
Las redes neuronales son modelos matemáticos que mediante un conjunto de iteraciones minimizan el error de salida por medio de la disminución de la distancia entre el resultado real y el de aprendizaje y se le asigna a cada variable un peso en cada nodo. Los resultados obtenidos en la modelación se probaron con datos reales y se observó un alto grado de precisión. En la red obtenida se evidenció que no era

necesario un número de neuronas mayor que diez ni un número de capas mayor que tres puesto que se obtuvieron resultados muy similares con menor tiempo de procesamiento.

4. Análisis de tipo ROC

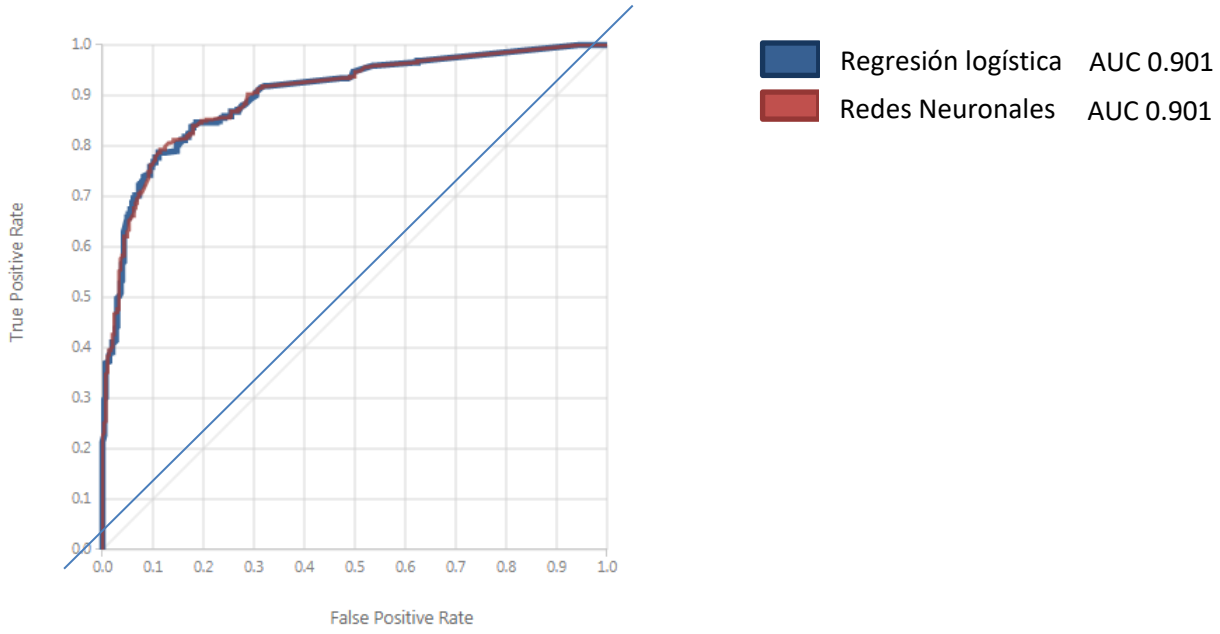
Para demostrar la capacidad de los modelos y su desempeño se generó la curva ROC en la que se compararon los modelos de inteligencia artificial con la regresión logística y se obtuvieron los siguientes resultados:

Figura 2. Análisis de tipo ROC entre regresión logística y bosque aleatorio



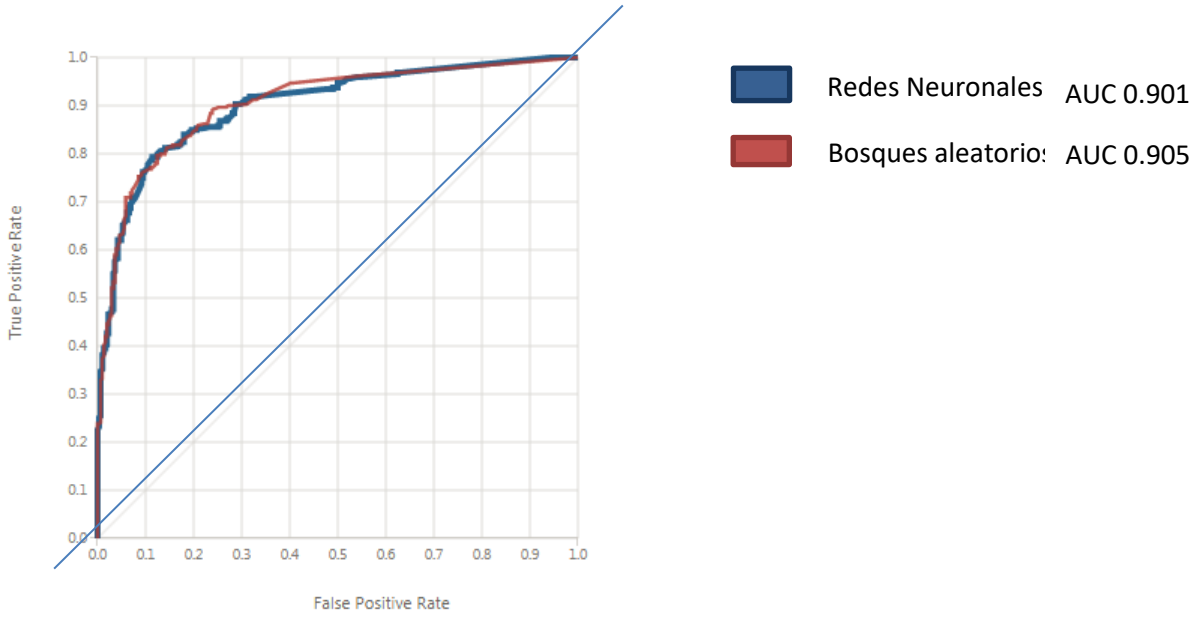
Fuente: elaboración propia

Figura 3. Análisis de tipo ROC entre regresión logística y redes neuronales



Fuente: elaboración propia

Figura 4. Análisis de tipo ROC entre bosques aleatorios y redes neuronales



Fuente: elaboración propia

En la curva ROC se comparó la sensibilidad, definida como la probabilidad de clasificar de manera correcta a un individuo cuyo estado real es positivo, con la especificidad, que es la probabilidad de clasificar en forma correcta a una persona cuyo estado real es negativo. Ambas características son necesarias para obtener un modelo óptimo.

Si la prueba fuera perfecta, la curva solo tendría el punto (0, 1), lo que significa que la sensibilidad y la especificidad son el 100%; por el contrario, si la prueba fuera inútil, la curva sería la diagonal de (0, 0) a (1, 1), línea conocida como no discriminante. Todos los puntos bajo esta diagonal representan resultados pobres (Hanley y McNeil, 1982).

En los modelos analizados se observó que todos los puntos estuvieron por encima de dicha diagonal, lo que permite afirmar que la capacidad de los modelos analizados fue buena. Por otro lado, al comparar el área bajo la curva de los tres

modelos no se observa una variación significativa, lo que indica que con cada uno se tiene una capacidad similar y cualquiera de ellos podría satisfacer las necesidades de las instituciones financieras.

5. Matriz de confusión

Los resultados de las matrices de confusión de los tres modelos fueron similares. Según la tabla 19, se puede observar que en la red existió mayor riesgo de clasificar mal a clientes que son buenos y, a su vez, clasifica como buenos a los clientes que son malos.

Tabla 19. Matriz de confusión

			Predicción	
			Positivo	**Falso**
Real	**Positivo**	Regresión logística	251	68
		Bosque aleatorio	253	66
		Red neuronal	243	76
	Falso	Regresión logística	34	272
		Bosque aleatorio	38	268
		Red neuronal	30	276

Fuente: elaboración propia

6. Métrica para medir el rendimiento para la clasificación

En la mayoría de indicadores el modelo de regresión logística arrojó mejores resultados. excepto en la precisión, que fue superada por la red neuronal.

Al comparar el bosque aleatorio con las redes neuronales se observó que las últimas tuvieron mejores resultados.

Tabla 20. Indicadores de rendimiento

INDICADOR	RED NEURONAL	BOSQUE ALEATORIO	REGRESIÓN LOGÍSTICA
Exactitud	0.830	0.814	0.837
Precisión	0.890	0.846	0.881
Sensibilidad	0.762	0.777	0.787
Calificación	0.821	0.810	0.831

Fuente: elaboración propia

Conclusiones e implicaciones

Las entidades financieras en Colombia, en general, utilizan métodos tradicionales de regresión logística para determinar el nivel, la calificación y la probabilidad de incumplimiento de un cliente. Estos modelos son promovidos por la Superintendencia Financiera de Colombia, que estableció guías para su aplicación; sin embargo, en esta investigación se puso en evidencia que existen otras alternativas, como las redes neuronales y los bosques aleatorios, que son dinámicas y adaptativas y que generan resultados o indicadores de rendimiento tan buenos como los métodos por lo común usados para predecir el riesgo de crédito, por lo que podrían dar un aporte interesante en la toma de decisiones de las entidades financieras.

Con base en los resultados obtenidos se encontró que tanto la regresión logística como los bosques aleatorios y la red neuronal contribuyen a una satisfactoria toma de decisiones. La primera es un modelo más simple y fácil de interpretar; sin embargo, para una entidad financiera la función de costo-beneficio es muy importante, por lo que otorgar créditos a clientes potencialmente morosos puede salir muy costoso. En este sentido, la matriz de confusión ayudó a elegir la red neuronal como el modelo ideal, dado que es el que mejores resultados ofreció frente a falsos positivos, aspecto que es esto muy valioso porque ayudaría a la salud de la cartera.

En el bosque aleatorio las variables de comportamiento seleccionadas para determinar la probabilidad de que un cliente entre en incumplimiento permitieron la construcción de un conjunto de reglas con las cuales fue posible determinar las características de un cliente que tiende a deteriorar su capacidad de pago. Estas reglas permitirían la implementación de alertas tempranas, que serían de gran ayuda para las entidades financieras porque podrían planear una gestión preventiva de cobro.

Con base en lo anterior, fue posible elegir un modelo de acuerdo con el uso o la necesidad que se le pueda dar. Por ejemplo: la red neuronal se podría utilizar para

tomar decisiones en procesos de preaprobados porque tiene menor error de clasificar como bueno a un cliente malo, el bosque aleatorio sería apropiado para disminuir el riesgo de crédito debido a que se podrían implementar sistemas de monitoreo temprano y la regresión logística es sencilla por su facilidad para explicar la calificación del cliente.

Entre los resultados obtenidos se destaca que la calificación interna que otorga el banco al cliente y el comportamiento del cliente en los meses cercanos a la fecha de análisis fueron las variables más relevantes para determinar si un cliente puede entrar en incumplimiento o no.

En la investigación y con las variables usadas se comprobó, además, que los modelos de seguimiento no requieren una estructura de redes de alta complejidad; con una red simple se pueden alcanzar buenos resultados, lo que es una ventaja si se quiere disminuir el tiempo y el costo en el procesamiento de los datos.

Para estudios futuros se recomienda aplicar los modelos antes referidos a las diferentes líneas de productos que manejan las entidades financieras y probar los modelos en las diferentes etapas del crédito, tales como otorgamiento y seguimiento.

Sería importante revisar la relevancia que tiene la opción de incluir nuevas variables de tipo demográfico, tales como edad, el estado civil o el tamaño de la familia, así como otras de índole macroeconómica de acuerdo con varios escenarios probables puesto que podrían influir en la capacidad de endeudamiento del cliente.

Se recomienda probar estas metodologías con información de otros países, al tener en cuenta que la mayoría de las instituciones financieras son bancas multilatinas o internacionales y estarían interesadas en tener modelos comunes para sus subsidiarias.

Referencias

Ala'raj, M., & Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems, 104,* 89-105. doi: 10.1016/j.knosys.2016.04.013

Asociación Bancaria y de Entidades Financiaras de Colombia, Asobancaria (2016). *Balance del sector bancario del 2016 y perspectivas crediticias para el 2017*. Bogotá: Asobancaria. Recuperado de http://www.asobancaria.com/2017/02/06/edicion-1076-balance-del-sector-bancario-en-2016-y-perspectiva-creditica-2017/

Bank of International Settlements, BIS (2017). *Basilea III: marco regulador internacional para los bancos*. Basilea: BIS. Recuperado de https://www.bis.org/bcbs/basel3_es.htm

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1) 5-32. doi: 10.1023/A:1010933404324

Cazorla Quevedo, M. Á., Colomina Pardo, O., Escolano Ruiz, F., Gallardo López, D., Rizo Aldeguer, R., y Satorre Cuerda, R. (1999). *Técnicas de inteligencia artificial.* Alicante: Digitalia.

Díaz I, García C, León M., Ruiz, F., Torres, F., Lizama, P., y Boccardo, G. (2014, noviembre). *Guía de asociación entre variables (Pearson y Spearman en SPSS). Ayudantía Estadística I 2014*. Santiago de Chile: Universidad de Chile, Facultad de Ciencias Sociales (FLACSO), Departamento de Sociología. Recuperado de https://www.u-cursos.cl/facso/2014/2/SO01007/1/material_docente/bajar?id_material=994690

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-54. Recuperado de https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131

Gutiérrez López, C. (2014). Evolución e impacto de la regulación bancaria internacional hasta Basilea III: el caso de América Latina. *Pecvnia*, *16/17*, 147-173. Recuperado de http://revpubli.unileon.es/index.php/Pecvnia/article/view/1339

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29-36. Recuperado de https://pubs.rsna.org/doi/pdf/10.1148/radiology.143.1.7063747

Hernández Sampieri, R., Fernández Collado, C., y Baptista Lucio, P. (2014). *Metodología de la investigación*, 6ª ed. Ciudad de México: McGraw-Hill.

Hosmer, Jr., D. W., & Lemeshow, S. (2000). *Applied logistic regression*. Nueva York, NY: John Wiley & Sons.

Ince, H., & Aktan, B. (2009). A comparison of data mining techniques for credit scoring in banking: a managerial perspective. *Journal of Business Economics & Management*, *10*(3), 233-240. doi:10.3846/1611-1699.2009.10.233-240

International Accounting Standards Board (2000). *Norma internacional de contabilidad Nº 9 (NIIF 9).* Recuperado el 10 de marzo 2017 de http://www.ifrs.org/search/Pages/Results.aspx?k=nic%2039

Martínez Bencardino, C. (2012). *Estadística y muestreo*, 13ª ed. Bogotá: Ecoe Ediciones.

McCarthy, J. (2007). What is artificial intelligence? Stanford, CA: Stanford University, Computer Science Department. Recuperado de http://www-formal.stanford.edu/jmc/

Microsoft Corporation (2017). *Microsoft Azure Machine Learning Studio*. Redmont, WA: Microsoft Corporation. Recuperado de https://msdn.microsoft.com/en-us/library/azure/dn905994.aspx

Pérez Ramírez, F. O., y Fernández Castaño, H. (2007). Las redes neuronales y la evaluación del riesgo de crédito. *Revista Ingenierías Universidad de Medellín*, *6*, 77-91. Recuperado de http://www.scielo.org.co/pdf/rium/v6n10/v6n10a07.pdf

Pérez López, C., y Santín González, D. (2006). *Minería de datos. Técnicas y herramientas*. Madrid: Thomson.

Salazar Villano, F. E. (2013). Cuantificación del riesgo de incumplimiento en créditos de libre inversión: un ejercicio econométrico para una entidad bancaria del municipio de Popayán, Colombia. *Estudios Gerenciales*, *29*(129), 416-427. doi:10.1016/j.estger.2013.11.007

Samaniego Medina, R. (2007). *El riesgo de crédito, en el marco del acuerdo Basilea II*. Madrid: Delta Publicaciones.

Sheaffer, R., y Mendenhall, W. (2007). *Muestreo de elementos.* México: Grupo Editorial Iberoamericana.

Superintendencia Financiera de Colombia (1995). *Circular básica contable y financiera (circular externa 100 de 1995), capítulo II, reglas relativas a la gestión del riesgo crediticio*. Bogotá: Superintendencia Financiera de Colombia. Recuperado el 10 de marzo 2017 de https://www.superfinanciera.gov.co/publicacion/circular-basica-contable-y-financiera-circular-externa-100-de-1995--15466

Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques*, 2ª ed. San Francisco, CA: Morgan Kaufmann.

Anexos

Anexo 1. Función lineal del segmento de automóviles (Superintendencia Financiera de Colombia, 1995, capítulo II, anexo 5):

$$Z = -2.779 + AM_B * 1.855 + AM_C * 3.0205 + MM_B * 1.668 + MM_C * 1.7234 + MM_D * 5.4605 + GI * 0.4960 + CA_R * 0.683 + CA_M * 1.5784 + CRB * 0.2505$$

La Superintendencia Financiera de Colombia (1995) describió las variables de la función de la siguiente manera:

AMB (altura de mora actual entre 31 y 60 días): toma el valor 1 si la altura de mora del cliente en el momento de la calificación para este tipo de crédito en la entidad es mayor o igual a 31 días e inferior o igual a 60 días y vale 0 si no.

AMC (altura de mora actual entre 61 y 90 días): toma valor el 1 si la altura de mora actual del cliente en el momento de la calificación para este tipo de crédito en la entidad es mayor o igual a 61 días e inferior o igual a 90 días y vale 0 si no.

MMB (máxima atura de mora entre 31 y 60 días): toma el valor 1 si la máxima altura de mora del cliente en los últimos tres años en la entidad y para este tipo de crédito es mayor o igual a 31 días y menor o igual a 60 días y vale 0 si no.

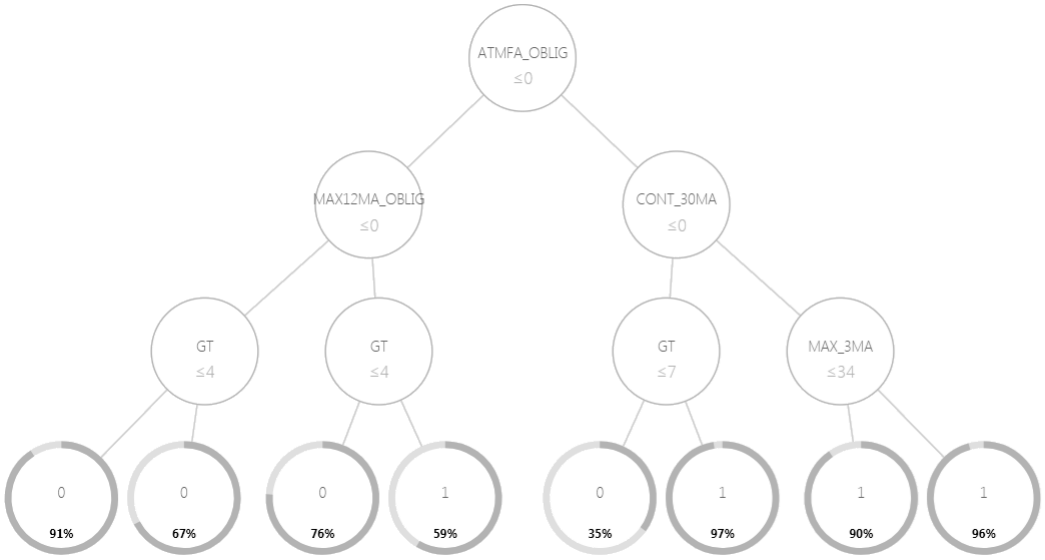
MMC (máxima altura de mora entre 61 y 90 días): toma el valor 1 si la máxima altura de mora del cliente en los últimos tres años en la entidad y para este tipo de crédito es mayor o igual a 61 días y menor o igual a 90 días y vale 0 si no.

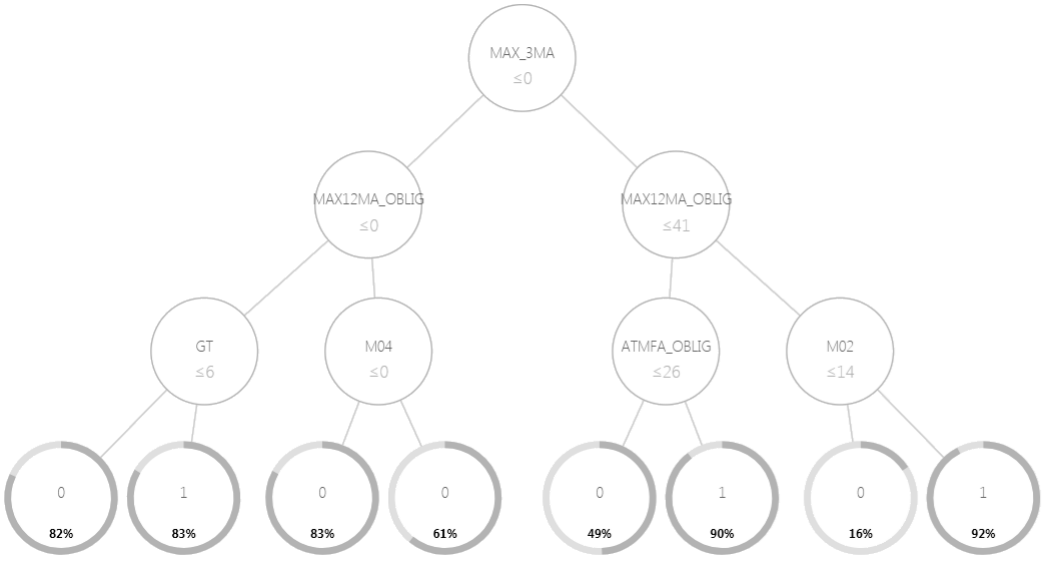
MMD (máxima altura de mora mayor a 90 días): toma el valor 1 si la máxima altura de mora del cliente en los últimos tres años en la entidad en este tipo de crédito es mayor a 90 días y vale 0 si no.

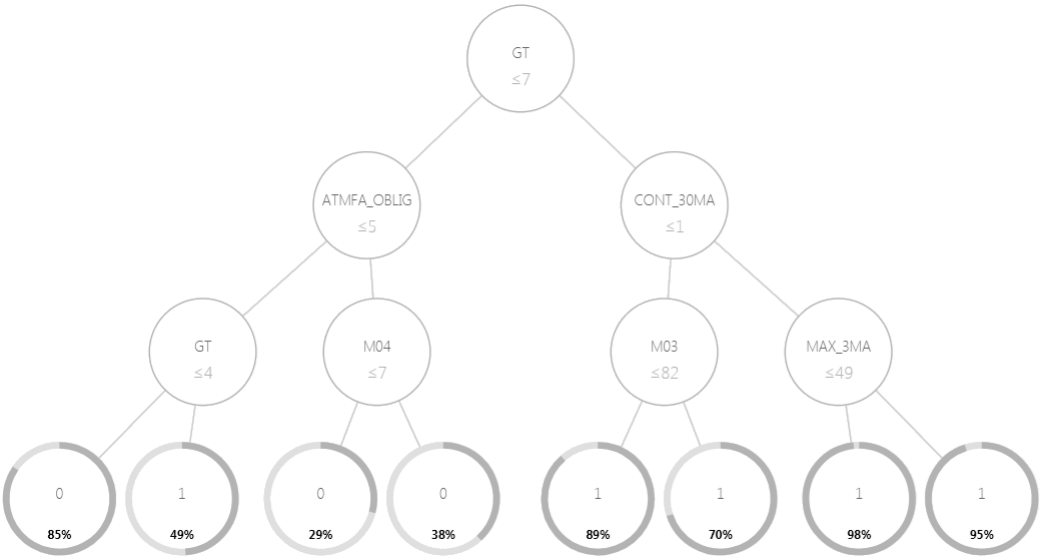
CRB (créditos activos): toma el valor 1 si el cliente en el momento de la calificación tiene activos con la entidad otros créditos de consumo diferentes al del segmento y vale 0 en caso contrario.

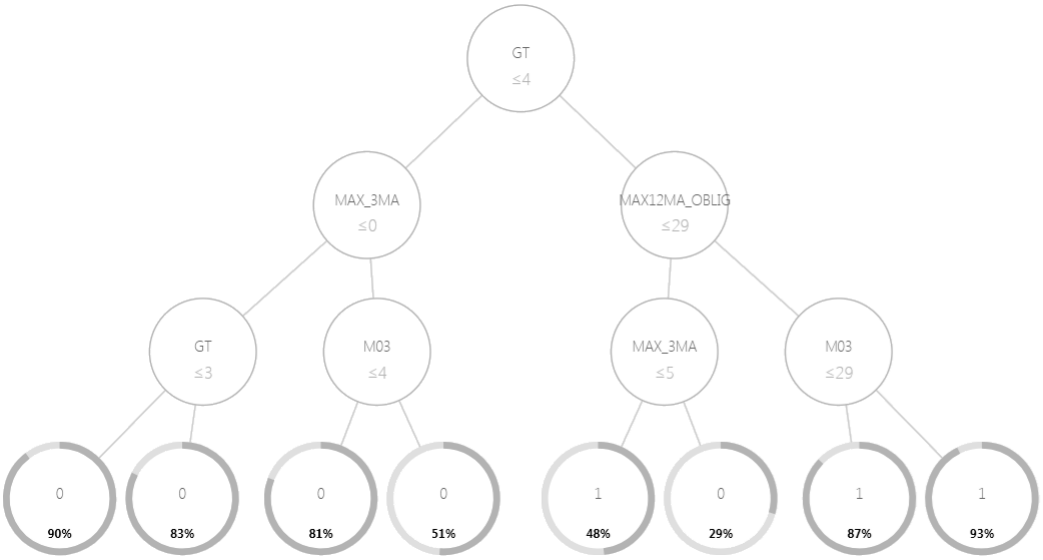
GI (garantía idónea): toma valor 1 si el cliente no tiene asociada a su crédito una garantía idónea de acuerdo con la definición del literal d del numeral 1.3.2.3.1 del capítulo y vale 0 en caso contrario.

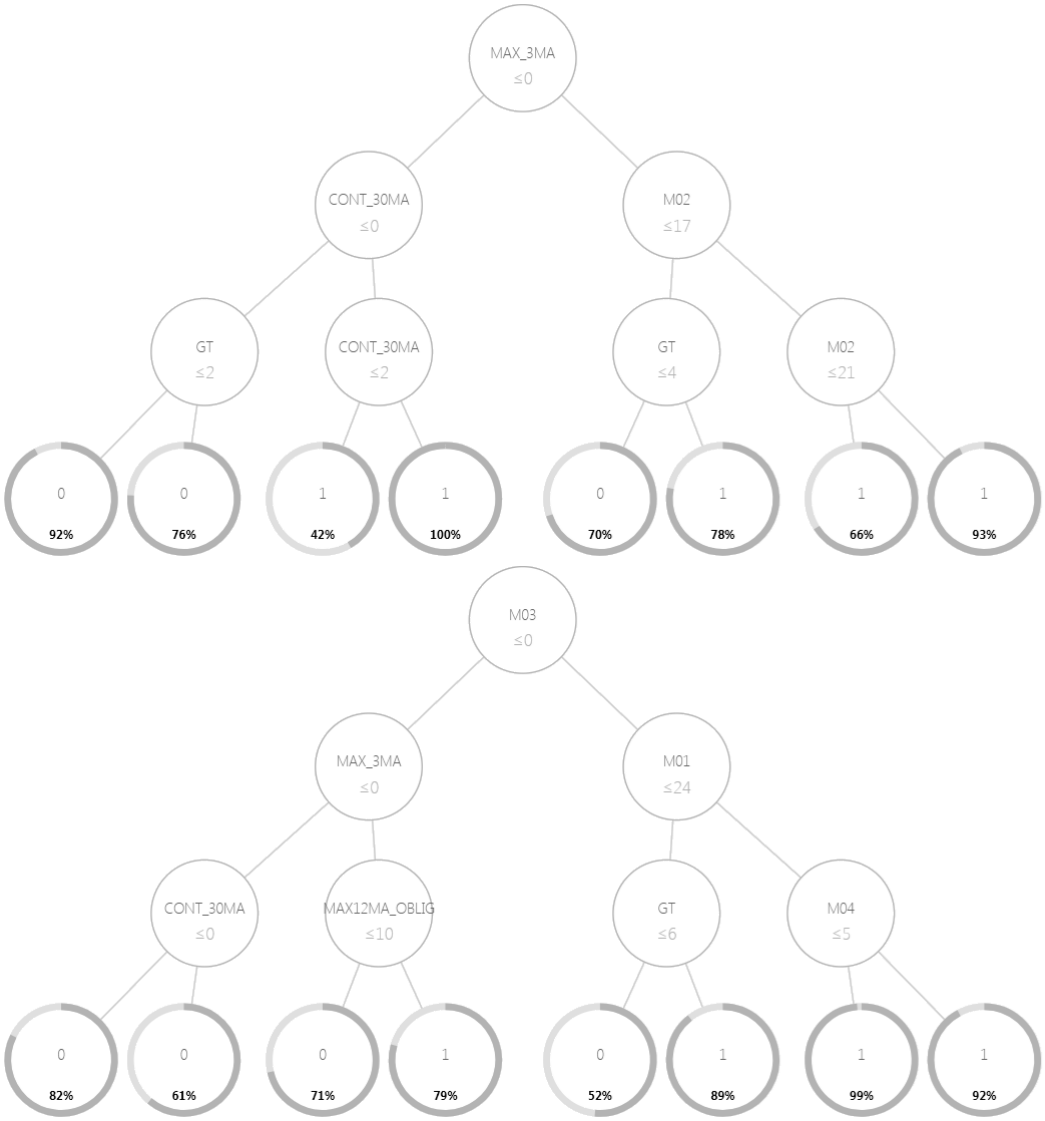
Anexo 2. Árboles óptimos obtenidos con *Microsoft Azure Machine*

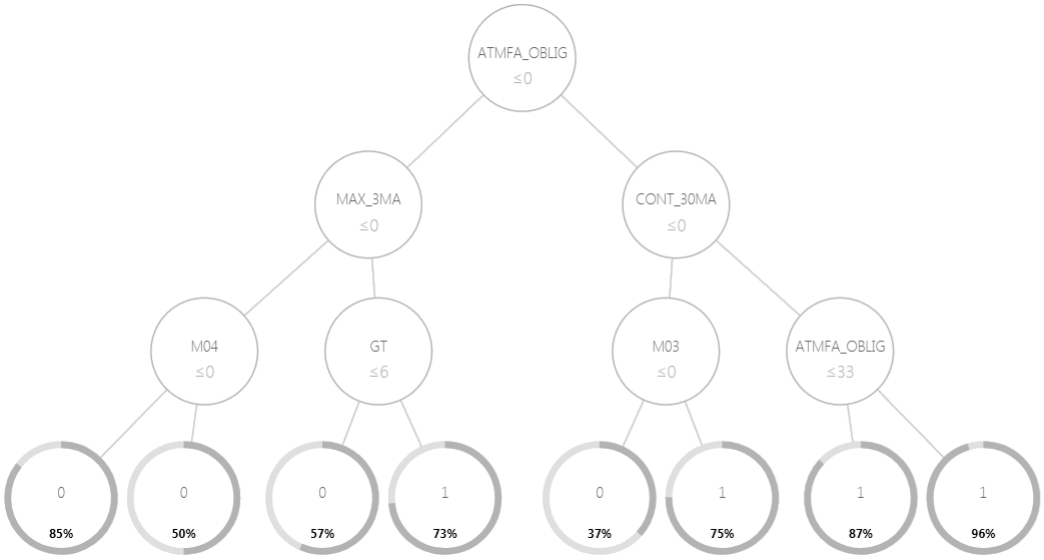


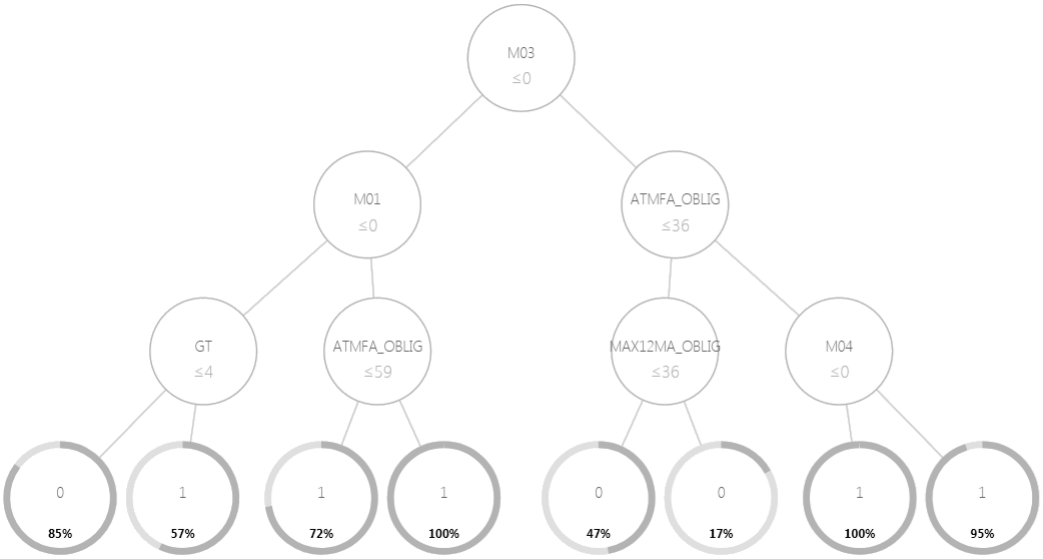


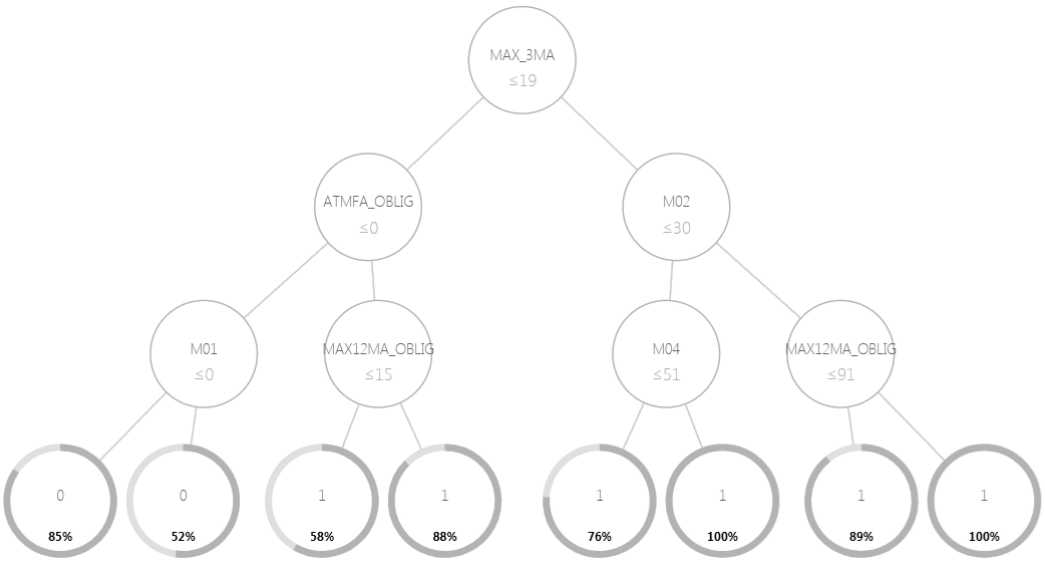

GT
≤7
ATMFA_OBLIG
≤5
CONT_30MA
≤1
GT
≤4
M04
≤7
M03
≤82
MAX_3MA
≤49
0
85%
1
49%
0
29%
0
38%
1
89%
1
70%
1
98%
1
95%

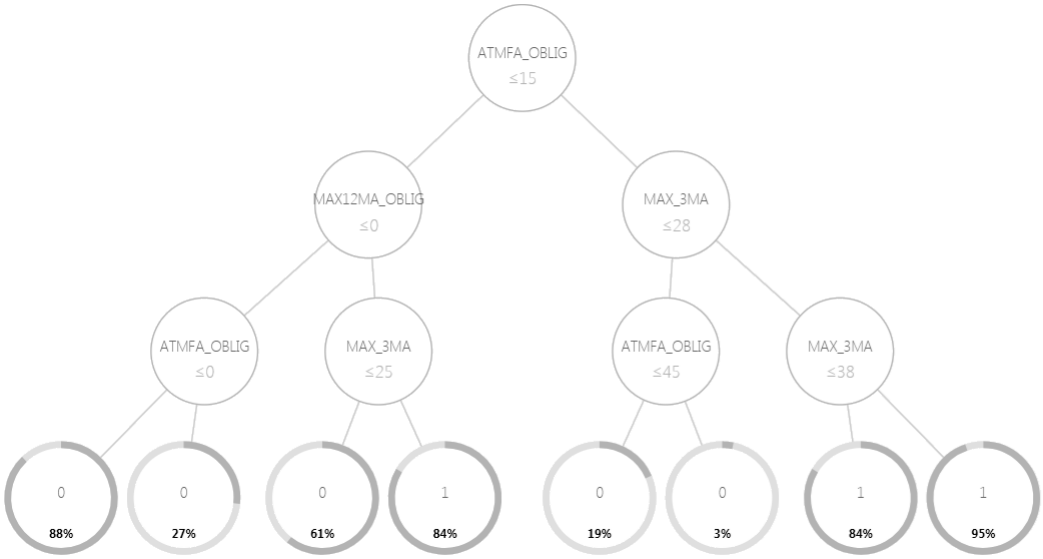

GT
≤4
MAX_3MA
≤0
MAX12MA_OBLIG
≤29
GT
≤3
M03
≤4
MAX_3MA
≤5
M03
≤29
0
90%
0
83%
0
81%
0
51%
1
48%
0
29%
1
87%
1
93%


MAX_3MA
≤0
CONT_30MA
≤0
M02
≤17
GT
≤2
CONT_30MA
≤2
GT
≤4
M02
≤21
0
92%
0
76%
1
42%
1
100%
0
70%
1
78%
1
66%
1
93%
M03
≤0
MAX_3MA
≤0
M01
≤24
CONT_30MA
≤0
MAX12MA_OBLIG
≤10
GT
≤6
M04
≤5
0
82%
0
61%
0
71%
1
79%
0
52%
1
89%
1
99%
1
92%


ATMFA_OBLIG
≤0
MAX_3MA
≤0
CONT_30MA
≤0
M04
≤0
GT
≤6
M03
≤0
ATMFA_OBLIG
≤33
0
85%
0
50%
0
57%
1
73%
0
37%
1
75%
1
87%
1
96%


M03
≤0
M01
≤0
ATMFA_OBLIG
≤36
GT
≤4
ATMFA_OBLIG
≤59
MAX12MA_OBLIG
≤36
M04
≤0
0
85%
1
57%
1
72%
1
100%
0
47%
0
17%
1
100%
1
95%


MAX_3MA
≤19
ATMFA_OBLIG
≤0
M02
≤30
M01
≤0
MAX12MA_OBLIG
≤15
M04
≤51
MAX12MA_OBLIG
≤91
0
85%
0
52%
1
58%
1
88%
1
76%
1
100%
1
89%
1
100%


ATMFA_OBLIG
≤15
MAX12MA_OBLIG
≤0
MAX_3MA
≤28
ATMFA_OBLIG
≤0
MAX_3MA
≤25
ATMFA_OBLIG
≤45
MAX_3MA
≤38
0
88%
0
27%
0
61%
1
84%
0
19%
0
3%
1
84%
1
95%