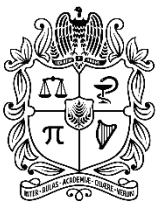


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Analítica de datos en la gestión de recuperación de la cartera financiera

Jonathan Steven Herrera Román

Universidad Nacional de Colombia
Facultad de Minas, Maestría en Ingeniería de Sistemas
Medellín, Colombia
2020

Analítica de datos en la gestión de recuperación de la cartera financiera

Jonathan Steven Herrera Román

Tesis o trabajo de investigación presentado como requisito parcial para optar al título de:

Magister en Ingeniería de Sistemas

Director:

Ph.D. John Willian Branch Bedoya

Línea de Investigación:

Analítica de Datos

Grupo de Investigación:

GIDIA

Universidad Nacional de Colombia

Facultad de Minas, Maestría en Ingeniería de Sistemas

Medellín, Colombia

2020

Persevera, pues si te convences de que puedes lograrlo, no hay forma de que fracases

Jonathan Herrera

Resumen

La inclusión financiera es una necesidad social que toma cada vez más fuerza en países en desarrollo. Los microcréditos son una efectiva forma de permitir la inclusión financiera, pero representan un reto en la gestión de cartera. Este trabajo aplica técnicas de analítica de datos y aprendizaje de máquina para predecir el comportamiento de la cartera en una entidad no financiera. Luego de ejecutar varios modelos predictivos, los árboles de decisión han mostrado el mejor desempeño para predecir si un crédito será pagado o se convertirá en cartera irrecuperable.

Palabras clave: Entidades no financieras, microcréditos, cartera de crédito, analítica de datos

Data analytics in financial portfolio recovery management

Abstract

Financial inclusion is a social need that is gaining more and more strength in developing countries. Microcredit or social lending is an effective way to enable financial inclusion, but it represents a challenge in loan default management. This work applies data analytics and machine learning techniques to predict the behavior of the loan default in a non-financial entity. After running various predictive models, decision trees have shown the best performance in predicting whether a loan will be paid or will become irrecoverable.

Keywords: Default scoring, P2P lending, social lending, data analytics

Contenido

Pág.

Lista de figuras

Pág.

Lista de tablas

Pág.

1. Introducción

El sistema financiero de Colombia incluye cada vez a más población adulta, mientras que en 2014 el indicador de inclusión financiera era de 73,9%, para finales de 2019 ya se encontraba en 81,4%. Esto deja el reto de incluir financieramente a 6,3 millones de adultos colombianos (Banca de las Oportunidades & Superintendencia Financiera de Colombia, 2018).

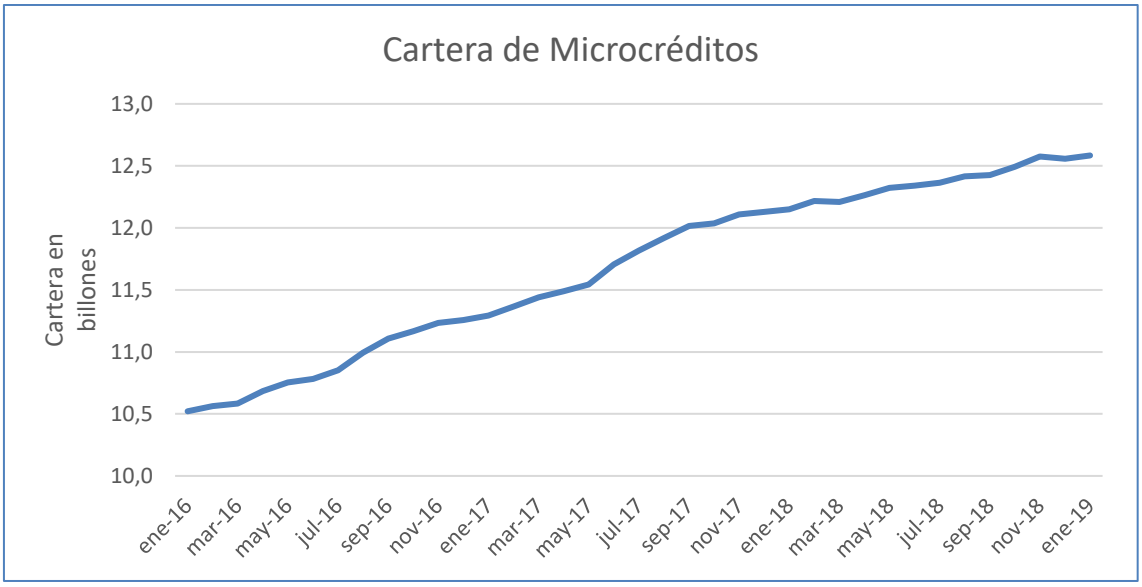
Buscando ser una alternativa a los bancos, existen entidades no financieras que pretenden ser un puente entre los usuarios sin historia crediticia y el sistema bancario, confiando en las personas y permitiendo la inclusión financiera. Estas entidades se especializan en los llamados microcréditos, mientras que los bancos tienen productos como los créditos de consumo y tarjetas de crédito.
La cobertura de microcréditos en población adulta sólo aumentó de 3 a 3.3 millones de adultos entre 2014 y 2017, lo cual significa un incremento del 10%, sin embargo, para 2018 se redujo a 3,1 millones de adultos (Banca de las Oportunidades & Superintendencia Financiera de Colombia, 2017) (Banca de las Oportunidades & Superintendencia Financiera de Colombia, 2018).

Incentivar estos microcréditos como forma de inclusión financiera se convierte en algo esencial para los clientes sin historial crediticio, pero enfrenta a las organizaciones a una exposición cada vez más alta al riesgo.

La cartera financiera, en general, ha crecido. (Superintendencia Financiera de Colombia, 2019) Pero las entidades no financieras, quienes ofrecen en su gran mayoría microcréditos, se enfrentan a un nivel de riesgo mayor que el sistema bancario, entre otras razones, porque cobijan un gran número de población sin ingresos fijos como estudiantes, amas de casa e independientes no formales. Si bien el aumento en la cartera de estas entidades puede verse como algo proporcional, debido al aumento mismo de los microcréditos (Ilustración 1-1), el aumento porcentual de la cartera vencida es mayor para

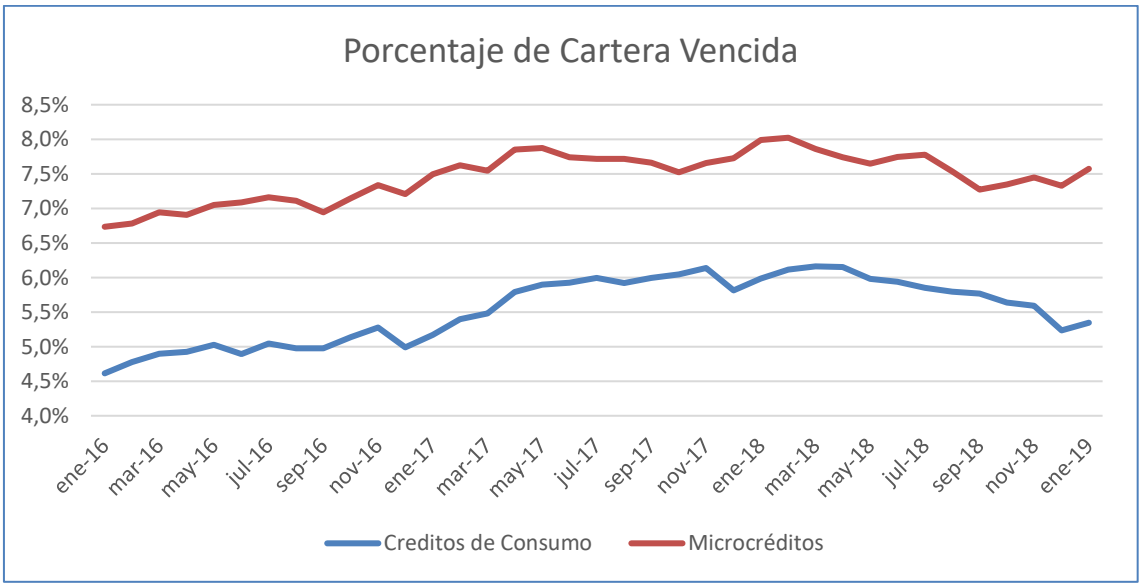
los microcréditos, manejados por entidades no financieras, que, para los créditos de consumo, manejados por los bancos (Ilustración 1-2). Lo anterior representa un riesgo alto, pues amenaza con hacer más estrictos los requisitos a los que se enfrentan las personas que desean un microcrédito, con la consecuente exclusión financiera de los mismos, contrario a lo que se desea.

Ilustración 1-1 Aumento de la cartera en microcréditos desde enero 2016.



(Superintendencia Financiera de Colombia, 2019).

Ilustración 1-2 Cartera vencida en microcréditos vs Créditos de consumo.



(Superintendencia Financiera de Colombia, 2019).

A pesar de que el manejo de cartera es mejor en los bancos que en las entidades no financieras, éstas últimas no pueden beneficiarse de la experiencia de los bancos debido a varios factores como que el conocimiento de los bancos rara vez es compartido con entidades fuera de la banca, a que su manejo de cartera requiere de departamentos enteros con muchos empleados dedicados a labores muy especializadas, y finalmente, a que los bancos se exponen a un menor riesgo negando créditos que las entidades no financieras si están dispuestas a otorgar.

1.1 Definición del problema

En las entidades no financieras el área de cartera puede limitarse a un grupo de personas que ubican telefónicamente a los clientes con impagos, en otros casos pueden tener sistemas automáticos que envían correos electrónicos o mensajes de texto a los clientes, pero sin métricas que permitan indicar si dichos canales son efectivos o no. Si a lo anterior se le adiciona que el proceso interno de cartera es totalmente reactivo, sin ningún tipo de análisis de los datos, vemos que un primer acercamiento al problema de cartera de las entidades no financieras consiste en permitirles entender sus propios datos y realizar análisis sobre los mismos.

Desde hace varios años se ha comenzado a utilizar las tecnologías de información y comunicaciones, para ayudar en la gestión de la cartera, usando técnicas como un sistema de cobranza integral, marcación predictiva, gestiones automatizadas, segmentación de cartera automática y sistemas de reporte. (Deloitte, 2012)

En 2014 se presenta como un servicio corporativo el uso de *machine learning* en el cobro de cartera, prediciendo aquellos créditos que tienen más probabilidad de ser pagados, para concentrar la gestión de la cobranza en ellos. Dicha técnica supera el 85% de efectividad en la predicción y ahora es usada por importantes *call centers* y entidades de gestión de cobranza. (CNN Periodismo Digital, 2016)

Otras alternativas empresariales que existen en el mercado permiten a un deudor realizar el pago de su deuda sin necesidad de la intervención de una entidad de cobro. Tal es el

caso de la plataforma *eResolve* que lanzó Experian en 2017. Esta técnica se basa en una disponibilidad las 24 horas del día y sin necesidad de una llamada u otras tácticas más agresivas de cobro. (Experian, 2017)

El uso de técnicas de inteligencia artificial para la gestión de cartera ha afianzado la necesidad de manejar grandes volúmenes de datos, por lo que la analítica de datos también ha tomado su parte en el proceso. Algunas empresas se han especializado en el uso de la analítica para ofrecer soluciones de administración del riesgo financiero y de la optimización del crédito y la cobranza. (Infórmese, 2018)

La presente propuesta busca aplicar analítica de datos en una entidad no financiera, específicamente en su proceso de cartera, para encontrar las relaciones entre sus datos que le permitan realizar una gestión de cartera efectiva y eficiente.

1.2 Objetivos

Objetivo general

Diseñar un modelo para el cobro de cartera en una entidad no financiera, empleando técnicas de Analítica de Datos.

Objetivos específicos

- Definir los datos correspondientes a un proceso de cartera en una entidad no financiera.
- Preprocesar los datos para ser utilizados en el modelo de analítica de datos.
- Diseñar un modelo para el cobro de cartera en una entidad no financiera, empleando técnicas de Analítica de Datos.
- Evaluar el desempeño del modelo creado, entrenado con la información real de cobro de cartera en una entidad no financiera.

1.3 Trabajos previos

Los microcréditos a los que hacemos referencia en la presente propuesta se refieren a aquellos realizados por entidades no bancarias, conocidos en inglés como *peer to peer lending* o *social lending*. Aclarado esto, presentamos investigaciones de inteligencia artificial aplicada a los microcréditos, la predicción de su comportamiento, el uso de técnicas que aprovechan tanto los datos estructurados como aquellos que no lo son para definir quién será un buen o mal cliente y otros trabajos relacionados al ámbito del microcrédito, el riesgo crediticio y la cartera.
Se resaltan trabajos realizados en los últimos cuatro años, variando en la aplicación de métodos como las redes neuronales, los modelos bayesianos, el análisis de supervivencia e inclusive, analítica de datos

Concentrándonos en las investigaciones de los últimos cuatro años, vemos que el microcrédito aún es joven en el sector financiero, debido a la larga curva de aprendizaje de las entidades no financieras por su desconocimiento del sector que ha sido siempre manejado por los bancos y entidades similares.

Un acercamiento desde las redes neuronales permite evidenciar que esta aplicación sirve para tomar mejores decisiones sobre a quién adjudicarle un microcrédito. Comparadas con modelos de regresión reflejan un comportamiento más preciso, permitiendo reducir el riesgo crediticio. (Byanjankar et al., 2015)

Evaluando desde un punto de vista estadístico, se han implementado modelos bayesianos con regresiones no lineales, haciendo uso tanto de datos estructurados como aquellos que no lo son. Esto con el objetivo de predecir el comportamiento de los microcréditos, logrando identificar las variables en los movimientos del mercado de microcréditos. (Bitvai & Cohn, 2015)

Una investigación interesante, fue la llevada a cabo por Ruyi Ge et al. Donde usaron datos de redes sociales para predecir el comportamiento de clientes morosos de microcréditos. Los autores usaron dos *dataset*, uno con la información de microcréditos y otro con datos de redes sociales. Curiosamente lograron una reducción de la cartera, pero, además, los

clientes estaban más dispuestos a pagar nuevamente luego de ser contactados por medio de redes sociales. (Ge et al., 2017)

Un nuevo trabajo de Byanjankar para predecir riesgos crediticios en microcréditos permitió algo más que clasificar los clientes en buenos o malos. Usando análisis de supervivencia consiguió predecir la probabilidad de supervivencia del crédito en período específicos. Sin embargo, sólo usaron un *dataset* para el análisis, lo que no permite generalizar el resultado para los microcréditos. (Byanjankar, 2018)

Retomando la clasificación habitual de cliente bueno y malo, tenemos el trabajo de Archana Gahlaut, sólo que esta vez utilizando un modelo de minería de datos para la clasificación. La predicción realizada se generó con un *dataset* orientado a los bancos y permite tomar de manera temprana la decisión de aceptar o negar un crédito, para continuar con otros aplicantes, cabe resaltar que las variables más representativas encontradas por el autor para el análisis del crédito fueron edad, duración y monto del crédito. (Gahlaut et al., 2017)

Teniendo en cuenta que los métodos hasta ahora emplean datos denominados ‘duros’ por ser verificados y dejan de lado los datos ‘suaves’ como texto no estructurado. Cuiqing Jiang nos muestra un trabajo donde precisamente estos datos suaves demuestran tener tanta o más información que los utilizados normalmente por los modelos de predicción. El autor usa el método LDA (*Latent Dirichlet Allocation*) para extraer características importantes de los textos descriptivos de los créditos, luego diseña un método de selección de características de dos estados para generar variables efectivas en el modelado, finalmente, usa un modelo de predicción de cartera de crédito que utiliza cuatro métodos de clasificación, todo esto con datos reales de microcréditos de plataformas en bancarrota de China. El resultado demuestra que un modelo tradicional que incluya también datos suaves podría mejorar su rendimiento basado en el mismo análisis. (Jiang et al., 2018)

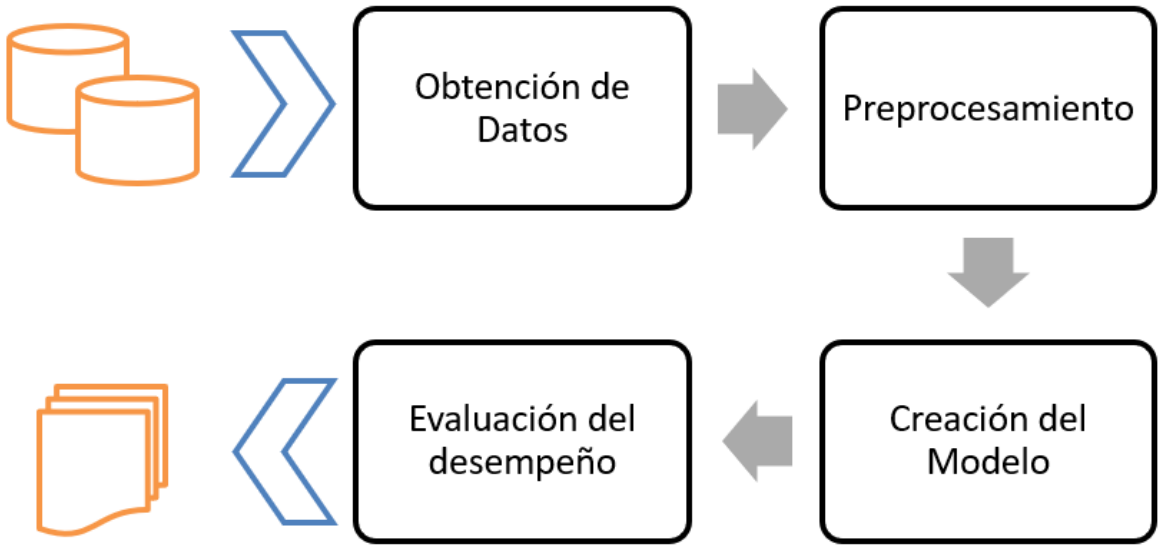
En un trabajo diferente y enfocado a la prevención de cartera en créditos de hipoteca en bancos, los autores nos presentan un modelo de tres pasos para la selección que incluye un marco de trabajo probabilístico para la cartera. El resultado demuestra tener mejor desempeño que los modelos de regresión por mínimos cuadrados ordinarios. El novedoso modelo representa un impacto en la eficiencia del uso de capital por parte de los bancos. (Do et al., 2018)

Finalmente, en el novísimo trabajo de Kim y Cho, relacionan un modelo que combina el algoritmo de propagación de etiquetas con una máquina transductiva de soporte vectorial (TSVM) y la teoría de Dempster–Shafer para ajustar una exacta predicción de microcréditos con datos no etiquetados. Los autores comparan los resultados de su método con varios de los anteriores trabajos mencionados, obteniendo hasta un 10% más de exactitud en la predicción. (Kim & Cho, 2019)

1.4 Estructura

El método por desarrollar se compone de cuatro fases; la obtención de los datos, el preprocesamiento, la creación del modelo y la evaluación del desempeño.

Ilustración 1-3 Pasos para el desarrollo del trabajo de grado.



Obtención de datos: El primer paso a realizar es la extracción de los datos que serán utilizados, para esto se revisará la estructura de la información que almacena la compañía sobre su cartera, se definirá el método a emplear para obtener los datos (base de datos, tablas, archivos planos…) y se procederá a extraer los mismos, previa firma de acuerdo de confidencialidad

Preprocesamiento: Una vez obtenidos los datos, se procederá a realizar una depuración de estos, descartando datos duplicados, incompletos o registros nulos. A su vez, los datos serán clasificados según algunos parámetros importantes para el estudio, como la época del año de su origen, la región del país al que pertenecen u otro particular a considerar.

Creación del modelo: Después de haber sido comprendida la información que representan los datos y las relaciones entre los mismos, se comenzará a crear un modelo de analítica de datos que permita extraer características representativas de los datos que apoyen la toma de decisiones basadas en los lineamientos estratégicos de la compañía.

Evaluación del desempeño: Teniendo los datos clasificados, sea por fecha, región… Se verificará el nivel de predicción del modelo desarrollado, afinándolo para que su comportamiento sea óptimo y aceptable para la compañía y se analizarán los resultados encontrados. Además, se tendrá en cuenta el trabajo posterior que podría realizarse sobre una investigación similar.

2. Marco teórico

Aunque la cartera crediticia ha sido objeto de estudio desde hace décadas, el enfoque del presente trabajo de grado está centrado en la cartera de crédito de entidades no financieras, las cuales también aplican como *Fintech*. Podría también hablarse de microcréditos, pero enmarcados a los que son otorgados por las entidades no financieras diferentes a la banca tradicional.

En este aspecto ya no existen tantos estudios al respecto, sin embargo, el trabajo de las entidades no financieras y las clasificadas como *Fintech* es una vía importante para lograr la inclusión financiera de la población que no puede acceder directamente a servicios de la banca.

Es en este asunto de la inclusión financiera donde países como Colombia le están apostando fuertemente al tema, con la creación de Colombia Fintech (Colombia Fintech, 2019), asociación que busca crear un ambiente de trabajo colaborativo y dinámico para los negocios *Fintech* del país. Por otro lado, la Banca de las Oportunidades (Banca de las Oportunidades, 2019) hace seguimiento a los avances en temas de inclusión financiera.

Teniendo claro el tipo de entidad de estudio, el análisis llevado a cabo usa técnicas de analítica de datos orientadas a poder realizar mejoras tempranas y reales dentro de las actividades de cartera de la entidad no financiera.

2.1 Cartera de crédito

La cartera, el crédito y el comportamiento crediticio de los clientes de las entidades no financieras difieren a lo conocido en la banca, principalmente porque el mercado objetivo de estas entidades es el grupo de personas que no ha podido acceder a los productos de la banca tradicional, teniendo que buscar alternativas que permitan la inclusión financiera de estas personas.

2.1.1 Créditos

Los créditos de los que trata el presente trabajo de grado tienen algunos factores diferenciadores con respecto a los créditos de la banca tradicional.

Microcréditos: La cartera procede de los créditos denominados microcréditos, aquellos préstamos realizados por montos inferiores a los 25 salarios mínimos legales mensuales vigentes (cerca de 22 millones de pesos colombianos a 2020) y no tienen una finalidad definida y delimitada previamente como los créditos típicos de la banca, donde existe el crédito de vivienda, comercial, entre otros (Superintendencia Financiera de Colombia, 2013).

Pago temprano: Estos créditos tienen un pago temprano de sus montos y un cobro igualmente temprano en su cartera, principalmente porque son realizados a plazos muy cortos en comparación con la banca. Estos créditos tienen un tiempo estipulado de pago entre 1 y 8 meses, mientras que la banca tiene plazos más extensos, lo cual también se refleja en montos mayores en sus créditos.

Canales: Los créditos del estudio son otorgados por diferentes canales al cliente, tanto presenciales como virtuales. En algunos casos, inclusive, son realizados a través de un tercero, una entidad diferente a aquella a la que se le adeudará el crédito, esto es una diferencia importante con respecto a la banca, quien realiza personalmente el crédito.

2.1.2 Cobro de cartera

Un crédito se considera en cartera o en mora cuando ha pasado al menos un día luego de la fecha de pago oportuna pactada entre el cliente y la entidad que otorga dicho crédito. Sin embargo, en Colombia se considera un crédito como normal si la mora es menor a un mes, un crédito aceptable si la mora es entre 1 y 2 meses, deficiente con mora de hasta 3 meses, de difícil cobro para moras entre 3 y 6 meses, e irrecuperable para moras superiores a los 6 meses (Superintendencia Financiera de Colombia, 2001). Pero lo anterior aplica para créditos de consumo, no existe aún una calificación oficial para microcréditos, tampoco tiene en cuenta a las entidades no financieras sino a las entidades bancarias.

Luego de que un crédito supere los 30 días de mora, el cliente puede ser reportado negativamente a las centrales de riesgo, dichas centrales sirven como referente para la

banca y las entidades no financieras, para conocer el estado de las deudas de una persona, previo al otorgamiento de un crédito, en su fase de estudio.

Definir cuándo un crédito se considera en cartera irrecuperable es crucial para el análisis, y aunque existe la categoría E definida por la Superintendencia Financiera de Colombia, esta cartera irrecuperable es para una mora superior a los 6 meses, demasiado permisiva para los microcréditos objetos del estudio. Por ello se hace necesario encontrar alternativas para la definición del riesgo crediticio en donde la cartera irrecuperable pueda ser determinada para períodos más tempranos de mora.

Las matrices de transición son útiles para este fin y pueden ayudar a definir unos parámetros propios para cada entidad en cuanto a su cartera irrecuperable.

Las matrices de transición fueron definidas en 1997 por Morgan como la probabilidad de que una obligación o crédito migre de un estado aceptable a un estado irrecuperable en un período de tiempo definido (Morgan & Morgan Grenfell, 1997).

Usualmente se define una matriz con las filas ‘i’ como el estado inicial de un crédito y las columnas ‘j’ como el estado final, la intersección define la cantidad de créditos que cambiaron entre dichos estados.

La probabilidad de que un crédito en un estado aceptable ‘i’ cambie a un estado irrecuperable ‘k’ se define como la cantidad de créditos que pasaron de un estado ‘i’ al estado ‘k’, sobre el total de créditos que iniciaron en el estado ‘i’. El punto donde la probabilidad de que un crédito entre en cartera irrecuperable desde un estado aceptable sea mayor a 50% se considera el punto de *default*, es decir, donde debe evitarse que la cartera empeore, pues será mayoritariamente irrecuperable.

Múltiples estudios se han realizado al respecto en Colombia, (Zapata, 2003) por ejemplo, concluye que la probabilidad de que un crédito sea irrecuperable depende del estado de los ciclos económicos, por lo tanto, se hace necesario establecer las transiciones de dichos ciclos para combinarlas con las transiciones de los créditos y con ello, anticipar pérdidas anticíclicas.

2.2 Analítica de datos

La analítica de datos es una parte de la estadística que busca limpiar, transformar y modelar datos con el objetivo de encontrar información útil que apoye la toma de decisiones.

Básicamente la analítica puede entenderse desde tres enfoques:

Analítica descriptiva: Se analizan de manera objetiva los datos históricos y se busca extraer información estadística sobre ellos. Es útil para entender cómo han evolucionado los datos y de qué manera funciona la relación entre los mismos.

Analítica predictiva: Usa técnicas matemáticas avanzadas para predecir datos faltantes o clasificar los datos en grupos definidos o implícitos. Es útil para la toma de decisiones al adelantarse al comportamiento específico de un conjunto de datos, según una variable de salida o resultado.

Analítica prescriptiva: Utiliza algoritmos y reglas propias de un negocio para automatizar la toma de decisiones. Es útil para optimizar el uso de recursos y para aumentar la eficiencia en los procesos críticos de la empresa.

Por la naturaleza de los datos y la necesidad de predecir o clasificar si un crédito entrará en default o no, la analítica predictiva es la herramienta que se usará en el presente trabajo. En el apartado 4.3 se hablará en más detalle de los tipos de análisis posibles.

3. Revisión de la literatura

Los siguientes son trabajos aplicados a entidades colombianas:

En el trabajo de (Velásquez, 2013) el objetivo es categorizar a un cliente en un nivel de riesgo cualitativo. El método aplicado usa un conjunto de indicadores obtenidos a raíz de variables del cliente, de sus créditos y del mercado, a cada uno de estos indicadores se les otorga una escala numérica según los valores en los que puedan variar
La manera de relacionar las variables fue a través de una función lineal, donde cada indicador tiene un peso, la sumatoria de sus pesos es la variable de salida definida como el riesgo del cliente. Aunque el modelo predictivo empleado no usa técnicas de analítica de datos, se ajusta a la entidad para la cual fue diseñado, en últimas, la ventaja de un modelo es justo su utilidad práctica y su potencial para el apoyo en la toma de decisiones.

Otro trabajo importante, que también busca predecir el comportamiento en cartera de los clientes es el de (Daza Sandoval, 2015). En este caso, la autora usó un proceso sistemático que incluía el juicio de expertos para la selección de variables explicativas con respecto a la cartera de un cliente. Interesante el hecho de que cada experto tenía una relevancia en su aporte al proceso y no pertenecían a la misma jerarquía organizacional. Los cuestionarios realizados por dichos expertos tenían ponderaciones de peso por variable para tener en cuenta en el estudio, el resultado final del peso de cada variable fue definido de acuerdo con la relevancia del experto y a las ponderaciones de las variables. Esta metodología logró una reducción de la dimensionalidad del problema de 66 variables a sólo 19.
Posteriormente la autora usó árboles de decisión como modelo predictivo para los datos, obtuvo un árbol de profundidad 8 y de 6 hojas, pequeño pero eficiente, con un porcentaje de precisión superior al 92%.
Lo aportante del trabajo no son sólo sus resultados, sino también la generación de políticas de directo impacto en el manejo de la cartera de la entidad bancaria objeto del estudio.

También a través del uso de técnicas de aprendizaje de máquina, pero esta vez redes neuronales y clasificación Bayesiana, tenemos el trabajo de (León Sánchez, 2015), que evalúa el riesgo de liquidez de cartera colectiva colombiana.
El análisis es importante debido a que usa variables macroeconómicas del país principalmente y tiene en cuenta su comportamiento en el riesgo de liquidez.

Por otra parte, en la cartera del sector salud, el artículo de (Castaño & Ayala, 2016) también buscó predecir el comportamiento de la misma usando técnicas de aprendizaje de máquina. El autor utilizó inicialmente regresiones logísticas, pero el nivel de predicción no mostró ser aceptable, mientras que el uso de árboles de decisión demostró ser el modelo adecuado para el análisis. A diferencia de (León Sánchez, 2015) el conjunto de datos que usó el autor carecía de información macroeconómica y es justamente una recomendación resultante el uso de información exógena relativa al mercado para ajustar la capacidad predictiva del modelo.

***

Teniendo en cuenta los trabajo realizados en predicción de la cartera o riesgo crediticio en el mercado colombiano y otros trabajos relacionados fuera de Colombia encontramos que hay varios factores comunes que la literatura evidencia.
El uso inicial de conjuntos de datos con muchas variables es necesario, (Zhu et al., 2019) tenían un *dataset* de más de 100 variables, por su parte, (Zhou et al., 2019) usaron más de mil variables iniciales.
Sin embargo, iniciar con un conjunto de datos grande exige un preprocesado de datos y el uso de técnicas para reducir la dimensionalidad; el análisis de correlación es una de ellas (Zhu et al., 2019), el juicio de expertos no debe ser descartado también como técnica útil, más aún si tenemos datos específicos de una entidad (Daza Sandoval, 2015), inclusive el uso de técnicas de aprendizaje pueden ser útiles para este fin (Zhou et al., 2019).
La aplicación de varias técnicas es indispensable para poder tener un análisis comparativo entre las mismas, la sensibilidad, especificidad y exactitud de los modelos son factores importantes a la hora de decidir cuál de ellos predice mejor el comportamiento que esperamos.

4. Analítica de datos en la gestión de recuperación de la cartera financiera

Teniendo en cuenta la naturaleza de entidad no financiera de la empresa en donde se llevó a cabo la investigación, se realizó un trabajo inicial de comprensión de la necesidad de la empresa, conocimiento de su información, estructura de sus bases de datos y en general, entendimiento del negocio, para asegurar un trabajo enfocado hacia el cumplimiento de los objetivos estratégicos de la entidad.

Como se mencionó anteriormente, la entidad no financiera del estudio clasifica en el neologismo de las *Fintech*, categoría que cobija a las empresas que prestan servicios financieros, usualmente como intermediarios en la inclusión financiera, apoyándose en las tecnologías de la información y la comunicación.

4.1 Obtención de los datos

A pesar de tener una base tecnológica como principio, el mercado cambia más rápido que el ritmo al que pueden responder algunas empresas, y las *Fintech* no son la excepción. La entidad del estudio no contaba con un área que realice analítica de datos, de hecho, su información ha crecido al ritmo del día a día, sin poder ser evaluado si algo de esta información es redundante, no tiene uso en la empresa o es inaportante.

El trabajo de obtener los datos se llevó a cabo durante varios meses, desarrollándose en 3 fases; definición de la información a extraer, creación la matriz de transición y creación del *dataset*.

4.1.1 Definición de la información a extraer

Una vez comprendida la actividad de negocio de la empresa del estudio, se definió conjuntamente que se consideraría un registro a la información relacionada con un crédito y no con una persona, esto debido principalmente a la complejidad de realizar el estudio por persona y al hecho de que muchos clientes tienen pocos créditos o incluso uno sólo, lo cual hace que no se pueda tener un comportamiento previo de la persona para estudiarla. La siguiente explicación permitirá entender la naturaleza de los datos.

Una persona puede tener varios créditos a la vez, estos pueden tener una cantidad de cuotas diferentes, ya que son elegidas por el cliente al momento del crédito. El cupo total del cliente limita la cantidad de créditos activos que puede tener en un momento dado, pero depende totalmente del monto de dichos créditos, pues un crédito puede consumir cualquier porcentaje del cupo del cliente.

Teniendo en cuenta la cobranza, ésta también se realiza por crédito, es decir, para cada crédito que entra en mora se realiza una cobranza enfocada en dicho crédito y no en la persona. Por ejemplo, si un cliente tiene dos créditos, uno en mora y el otro a pocos días de entrar en mora, al ser contactado sólo se le hará el cobro del crédito en mora, pudiendo ser contactado días más tarde para realizarse el cobro del crédito que apenas acaba de entrar en mora, obviando al anterior que ya se le hizo cobro.

Esta dinámica hace muy complejo identificar estrategias efectivas para la cobranza si nos enfocamos en la persona y no en el crédito.

Una vez aclarado que se analizará la información por crédito falta definir el período de análisis, pues podemos considerar por período un mes, un trimestre, semestre o un año entero.

En la empresa del estudio los créditos se pueden diferir a un plazo definido por el cliente. En general los créditos pueden ser pactados a pagar en plazos entre 1 y 18 meses, pero más del 80% de los créditos son pactados a 4 o menos meses. Teniendo en cuenta esto, se acordó definir como unidad de período, un mes.

4.1.2 Creación de la matriz de transición

Ya que se definió un período mensual para el análisis de los créditos y un plazo máximo de 4 meses para la mayoría de estos, se seleccionaron 1 millón de créditos y se analizaron durante los períodos diciembre 2018 hasta abril 2019, tomando diciembre como el período 0, así tendremos otros 4 períodos para que el crédito pueda estar pago o en cartera irrecuperable.

En cuanto al estado de la cartera, para cada período se definió la salud del crédito según la siguiente tabla:

Tabla 4-1 Categoría de salud de los créditos según estado de mora

Categoría	Estado de crédito
0	Crédito al día, sin mora
1	Mora menor a 30 días
2	Mora mayor o igual a 30 y menor a 60 días
3	Mora mayor o igual a 60 y menor a 90 días
4	Mora mayor o igual a 90 días
-1	Crédito cancelado en su totalidad

Ya que la categoría -1 indica que un crédito ya fue cancelado en su totalidad, debería hacer que este permanezca en la misma categoría para cualquier período futuro, se revisaron todos los créditos y se eliminaron de los análisis aquellos en los que no se cumplía dicha regla, pues suponen un error en la base de datos, ya sea por manipulación directa o por problemas de software. En total se eliminaron 751 créditos, dejando para el análisis un total de 999.249 créditos.

Para todos los créditos analizados se definió la matriz de transición, teniendo en cuenta la probabilidad de que un crédito que comienza en un estado ‘i’, pase a un estado ‘j’ dentro del rango de períodos analizados $P(i,j)$ (diciembre 2018 a abril 2019)

Tabla 4-2 Matriz de transición entre diciembre 2018 y abril 2019

Categoría inicial	Categoría final					
	0	1	2	3	4	-1
0	0,96%	0,06%	0,20%	0,43%	**3,83%**	94,53%
1	18,06%	6,58%	3,50%	4,19%	**7,83%**	59,83%
2	6,30%	0,12%	0,19%	1,10%	**59,39%**	32,89%
3	6,79%	0,01%	0,04%	0,05%	**74,49%**	18,62%
4	3,85%	0,00%	0,00%	0,00%	**92,36%**	3,78%

Si tomamos para cada categoría el $P(i,4)$, tendremos la probabilidad de que una categoría entre en *default*, es decir, sea irrecuperable. Así para un crédito que comience estando al día (0) es poco probable que termine en *default* (3,83%), pero para un crédito que ya tenga una mora entre 30 y 60 días se hace más probable que termine en irrecuperable (59,39%).

Identificar la categoría en donde los créditos empeoran en la mayoría de los casos es vital para definir la variable dependiente a usar en el *dataset*. Es así como para todos los créditos a analizar se definió como variable dependiente el estado 0 si para el siguiente período (enero 2019) estaban en una categoría 2 o peor y 1 en otro caso.

4.1.3 Creación del conjunto de datos

Hemos definido los datos a usar; 1 millón de créditos cada uno con un identificador único. Así mismo se ha definido la variable dependiente; un valor binario que indica 0 si un crédito entrará en *default* y 1 en otro caso.

Ahora, para definir las variables independientes se tomaron en cuenta tres categorías para las mismas; datos del cliente, datos del crédito y datos calculados.

Datos del cliente: Estas variables incluyen información asociada al cliente dueño del crédito, aun cuando el análisis se realizará por crédito es importante tomar en cuenta variables propias de las personas. Entre los datos obtenidos relacionados con el cliente se obtuvo información relacionada con su ocupación, género, estado civil, información de residencia y de lugar de trabajo y puntaje en centrales de riesgo.

Datos del crédito: Estas son variables relacionadas directamente con la obligación crediticia, entre la información obtenida se tiene la fecha de creación del crédito, valor de la cuota, monto total del crédito, saldo adeudado, valor pagado y edad de mora.

Datos calculados: Adicional a los datos obtenidos directamente de las bases de datos se definieron otros que podrían ser útiles para el análisis, estos campos son resultado de la operación de varios campos preexistentes, dando lugar a información que no existe como campo en ninguna base de datos de la entidad a analizar. Algunos de los datos calculados tienen en cuenta la cantidad de crédito activos por el mismo cliente, los días transcurridos desde la creación del crédito hasta la primera gestión de cobranza o la mora máxima tenida por el cliente del crédito teniendo en cuenta el total de crédito del último año.

Finalmente, el *dataset* se constituye de 999.249 registros o créditos y cada uno de ellos se compone de 67 campos, como sigue:

- **Índice:** 1 campo
- **Datos del cliente:** 21 campos
- **Datos del crédito:** 31 campos
- **Datos calculados:** 13 campos
- **Variable dependiente:** 1 campo

4.2 Preprocesamiento

En la fase de preparación de los datos se deben considerar varios factores que pueden afectar al resultado general del análisis, como las variables tipo fecha, las variables categóricas o los datos nulos, además ha de tenerse en cuenta las variables con varianza cero y una normalización de los datos antes de la creación del modelo de predicción.

4.2.1 Variables tipo fecha

Las variables tipo fecha deben ser identificadas y llevadas a un formato único para todas, de forma que puedan ser tenidas en cuenta en la normalización de los datos.

Para el *dataset* se tienen 9 variables tipo fecha y se observa que algunas no comparten un formato común, algunas inclusive no son detectadas como fecha sino como tipo *string*:

Ilustración 4-1 Variables tipo fecha antes del preprocesamiento

FechaNacimiento	LaboralFechaIngreso	FechaApertura
1995-08-14 09:33:01.000	15/07/2013	20150715
1990-03-12 15:41:28.000	18/08/2015	20150715
1996-06-11 17:37:23.000	14/07/2013	20150715
1995-08-14 09:33:01.000	15/07/2013	20150715
1994-03-06 14:34:08.000	NaN	20150715

Para cada variable se define entonces el formato *'%Y%m%d-%H%m%s*', el cual es equivalente al formato en que se encuentra la variable FechaNacimiento. Después de unificar el formato de cada variable y para permitir su integración con las demás variables numéricas, se cambia el valor de la fecha por su equivalente numérico, un número de control que permite identificar cada fecha de manera única, además convierte a las variables tipo fecha en variables continuas.

Ilustración 4-2 Variables tipo fecha como valor numérico

FechaNacimiento	LaboralFechaIngreso	FechaApertura
808392781	1373846400	1436918400
637256488	1439856000	1436918400
834514643	1373760000	1436918400
808392781	1373846400	1436918400
762964448	-2208988800	1436918400

4.2.2 Variables categóricas

Hay otro tipo de variables discretas que además son expresadas como texto, estas variables deben ser analizadas en cada caso para determinar qué hacer con cada una de

ellas. El *dataset* contiene 5 variables categóricas; Genero, DireccionResidencia, LaboralDireccion, Calificacion y DiaCreaCredito.

Para cada variable se evalúan la cantidad de categorías diferentes que poseen, en busca de poder codificarlas.

Genero: Esta variable sólo tiene dos categorías; Femenino y Masculino. Por esto se codifica en 1 o 2 como sigue: {"Femenino": 1, "Masculino": 2}

Calificacion: Esta variable tiene 5 categorías, cada una expresada como una letra; A, B, C, D, E. Se codifica como un número entre 1 y 5: {"A": 1, "B": 2, "C": 3, "D": 4, "E": 5}

DiaCreaCredito: Representa el día de la semana en que se creó el crédito, se codifica según un número entre 1 y 7: {"Monday": 1, "Tuesday": 2, "Wednesday": 3, "Thursday": 4, "Friday": 5, "Saturday": 6, "Sunday": 7}

Aunque para la variable *DiaCreaCredito* se podrían haber usado 7 variables *dummy*, cada una representando un día de la semana, se opta por asignarle un número a cada día de la semana, para no aumentar la cantidad de variables del conjunto de datos, sin embargo, de encontrarse que esta variable es determinante en la decisión del modelo se podrían plantear otras alternativas en su codificación.

Para las variables DireccionResidencia y LaboralDireccion se encuentra que tienen demasiadas categorías para ser codificadas, además son valores escritos libremente por lo que da lugar a errores tipográficos y a la interpretación de quien ingresa el dato, éstos son campos elegibles para ser eliminados del *dataset*.

4.2.3 Datos nulos

Los campos con datos nulos pueden ser un inconveniente a la hora de aplicar un modelo de predicción, sin embargo, eliminar los registros que contienen datos nulos podría significar la pérdida de información vital para el análisis, por lo que cada campo que contiene datos nulos debe ser analizado individualmente.

Primero se listan los campos que tienen datos nulos:

Tabla 4-3 Campos con datos nulos

Campo	Datos Nulos (%)
BarrioCodigo	0,000100
Estudiante	0,012910
Genero	0,107781
DireccionResidencia	0,263198
LaboralCargoCodigo	7,878617
LaboralSalarioCodigo	8,473464
PersonasACargo	8,618673
PaisCodigo	8,798107
OcupacionCodigo	8,913694
ScoreCentrales	10,195957
LaboralMunicipioCodigo	11,815874
LaboralFechaIngreso	12,582149
LaboralDireccion	13,177496
FechaUltimoPago	44,601095
FechaPago	82,487348

Para determinar qué hacer para cada variable con datos nulos, se acude al juicio de expertos, en este caso, al conocimiento que la entidad no financiera tiene de su información y de lo que ésta significa, por ejemplo, para la variable *BarrioCodigo* el valor 0 significa que no se tiene el dato del barrio en ese crédito específico, en ese caso, los nulos se reemplazan por 0 para significar que el dato no se tiene. Existe otro caso, como el de la variable *LaboralFechaIngreso*, en donde no existe un valor 0, pero si existe un valor por defecto que toma la variable en los casos donde no se tiene el valor para un crédito específico o se ingresó de manera errónea, para este caso, el experto define que los nulos pueden ser reemplazados por el valor por defecto para esta variable.

Para otros casos, como una práctica conocida (Skiena, 2017), los valores nulos son reemplazados por el valor mayoritario en la variable, esto se hizo para algunas variables como *Genero*.

Se tienen un total de 15 campos con datos nulos, para cada caso se explica su proceder:

1. BarrioCodigo, se reemplazan los nulos por 0 que significa sin datos
2. Estudiante, los valores nulos se reemplazan por 0, el valor predominante del campo
3. Genero, se reemplazan los nulos por 1, valor mayoritario
4. DireccionResidencia, se elimina el campo como se definió previamente
5. LaboralCargoCodigo, se reemplazan los nulos por 0 que significa sin datos
6. LaboralSalarioCodigo, se reemplazan los nulos por 2, valor mayoritario
7. PersonasACargo, se reemplazan los nulos por 0, valor mayoritario
8. PaisCodigo, se elimina el campo, pues se sabe que todos los clientes son del país código 57
9. OcupacionCodigo, se reemplazan los nulos por 4, valor mayoritario
10. ScoreCentrales, se reemplazan los nulos por 0 que significa sin datos
11. LaboralMunicipioCodigo, se reemplazan los nulos por 0 que significa sin datos
12. LaboralFechaIngreso, se reemplazan los nulos por una fecha por defecto (1900-01-01), tomada como dato inválido
13. LaboralDireccion, se elimina el campo como se definió previamente
14. FechaUltimoPago, se elimina el campo, pues tiene más de 30% de datos nulos (45%)
15. FechaPago, se elimina el campo, pues tiene más de 30% de datos nulos (82%)

4.2.4 Varianza de los datos

Aunque la varianza de un campo puede ser motivo de análisis para determinar qué tan variados o repetidos son sus datos, en este caso sólo se tendrán en cuenta aquellos campos cuya varianza es 0, pues significa que son datos que no varían para ninguno de los 999.249 registros que se tienen, siendo un campo para eliminar.

Tabla 4-4 Varianza de los campos

Campo	Varianza
PorcentajeAVAL	0,00E+00
PeriodoReporteCodigo	0,00E+00
ValorTotalCuotaCapital	0,00E+00
ValorTotalDeuda	0,00E+00
FechaCrea	0,00E+00
Estudiante	1,27E+04
ReportadoPagado	1,44E+05
…	
NroPagoRealizados	2,45E+06
CantidadCuotas	2,66E+06
CantCuotas	2,66E+06
DiaCreaCredito	4,07E+06
…	
FechaCreacionPersona	4,48E+22
FechaNacimiento	6,45E+22
LaboralFechaIngreso	1,47E+24

La tabla anterior muestra la varianza de algunos de los campos, ordenados de menor a mayor varianza. Observamos que 5 campos pueden ser eliminados, pues todos los registros tienen el mismo valor para esos campos, no siendo útil para el análisis del *dataset*.

Aunque también vemos datos con varianza muy grande no se tomarán acciones con dichos campos, pues podrían resultar útiles en el análisis de los modelos de predicción.

4.2.5 Normalización

Como fase final del preprocesamiento de los datos se proceden a normalizarlos, esto para evitar comparar datos con diferentes escalas entre sí.

Para cada campo, se realiza la normalización restándole al dato la media y dividiendo por la desviación estándar del campo:

$$X = \frac{X - \mu}{\sigma}$$

Se observa que ahora el *dataset* está normalizado y listo para ser usado por los modelos de predicción.

Ilustración 4-3 *Dataset* normalizado

LaboralFechaIngreso	LaboralMunicipioCodigo	TipoViviendaCodigo	ScoreCentrales	AlmacenCodigo
0.432542	0.236179	-1.106025	-0.910849	1.398727
-2.625676	-0.793946	-0.058824	0.845174	0.112100
0.450087	-0.178733	-0.058824	0.288386	0.344685

4.3 Creación del modelo

Luego de la preparación de los datos el *dataset* a trabajar tiene la siguiente estructura:

- **Índice:** 1 campo
- **Datos del cliente:** 17 campos
- **Datos del crédito:** 24 campos
- **Datos calculados:** 13 campos
- **Variable dependiente:** 1 campo

Los registros siguen siendo los mismos 999.249 tomados como válidos por la entidad.

Antes de tener un modelo a usar para el trabajo, ha de definirse qué tipos de algoritmos podrían aplicar, teniendo en cuenta que hay múltiples modelos, entre los de clasificación, regresión o de clusterización.

4.3.1 Clusterización

La clusterización es una técnica de aprendizaje no supervisado que identifica las similitudes entre los datos y los segmenta en grupos denominados clústeres. Luego del entrenamiento, al recibir un dato nuevo y no clasificado, la técnica define a qué clúster pertenece el dato.

Debido a que se tiene una variable dependiente que se pretende predecir no se usará clusterización, pues los datos sólo deben ser clasificados en las categorías permitidas por la variable dependiente, en este caso, el pago de un crédito o su ingreso a cartera irrecuperable.

4.3.2 Regresión

Los métodos de aprendizaje supervisado que son usados para problemas de regresión pretenden encontrar la relación entre la variable dependiente con las variables independientes. En el caso de la regresión lineal se busca una función lineal que aproxime el conjunto de datos, minimizando el error de dicha aproximación.

En nuestro problema la necesidad clara es de predecir si un crédito entrará en cartera irrecuperable o no, lo cual no es una regresión sino una clasificación, el próximo método a estudiar.

4.3.3 Clasificación

Los algoritmos de aprendizaje supervisado denominados de clasificación buscan encontrar una etiqueta de datos o variable de salida para un conjunto de datos definido.

En nuestro problema actual estos algoritmos se ajustan a la necesidad actual, pues basados en valores asociados a un crédito se pretende predecir o clasificar a cada uno de ellos en si será pagado o entrará en el denominado *default* de cartera, un claro caso de clasificación.

Siendo así se usarán varios métodos de clasificación con el conjunto de datos, se compararán los resultados de cada uno y se elegirá aquel que minimice el error en la predicción. Los algoritmos para evaluar serán la regresión logística, los árboles de decisión, el método de los K vecinos, las redes neuronales y las máquinas de soporte vectorial.

4.4 Evaluación del desempeño

Una vez definido el modelo a emplear, se ejecutará cada uno basado en el *dataset*, luego de esto se evaluará el desempeño de cada uno, teniendo en cuenta la sensibilidad, o capacidad de predecir correctamente los créditos que serán pagados, y la especificidad, o la capacidad de predecir correctamente los créditos que serán irrecuperables. Para la entidad es más la especificidad que la sensibilidad, esto es, predecir correctamente los créditos que no serán pagados y, por ende, que requieren una correcta gestión de cobranza. Antes de comparar cada modelo de clasificación se divide el *dataset* en tres conjuntos diferentes según su uso; entrenamiento, pruebas y validación, cada uno con el

70%, 10% y 20% de los datos respectivamente. Esta partición se realiza a través de validación cruzada aleatoria.

Una vez tenemos los tres conjuntos se evalúan diferentes métodos para la clasificación.

El tamaño de cada conjunto de datos es el siguiente:

- Datos de entrenamiento: 699.474 registros
- Datos de pruebas: 99.925 registros
- Datos de validación: 199.850 registros

A continuación, se muestran los resultados de los cinco algoritmos de aprendizaje supervisado para clasificación que se usaron.

4.4.1 Regresión logística

Este tipo de análisis es útil para predecir el comportamiento de una variable dependiente categórica, según un conjunto de variables independientes. Es ampliamente usado este método en implementaciones donde la variable dependiente es una variable cualitativa binaria, en cuyo caso la regresión logística aproxima la probabilidad de que la variable dependiente ocurra (sea 1) o no ocurra (sea 0).

Para nuestro caso usamos una regresión logística con tolerancia de $1E^{-5}$ y con 1000 iteraciones. Se observan los siguientes resultados para los datos de pruebas:

Tabla 4-5 Regresión logística: resultados de pruebas

--Regresión Logística-- Desempeño con datos de prueba			
	Predicción		
Observación	**0**	**1**	**Precisión**
0	2331	3234	Especificidad - 41,9%
1	983	93377	Sensibilidad - 99,0%
Exactitud			**95,780%**

Se observa una sensibilidad cercana al 100% para los créditos que no entrarán en *default*, pero una especificidad inferior al 50% para aquellos que sí lo harán.

Aun cuando no es tan buena la regresión logística para determinar si un crédito entrará en default se predice el comportamiento para los datos de validación:

Tabla 4-6 Regresión logística: resultados de validación

--Regresión Logística-- Desempeño con datos de validación			
	Predicción		
Observación	0	1	Precisión
0	4739	6359	Especificidad - 42,7%
1	1822	186930	Sensibilidad - 99,0%
Exactitud			95,906%

4.4.2 Árboles de decisión

Los árboles de decisión son modelos de clasificación que consiste en la comparación del valor a clasificar con un conjunto de características definidas, definiendo reglas del tipo “si, entonces”. Funcionan con variables numéricas y categóricas.

Tabla 4-7 Árboles de decisión: resultados de pruebas

--Árboles de decisión-- Desempeño con datos de prueba			
	Predicción		
Observación	0	1	Precisión
0	3618	1937	Especificidad - 65,1%
1	1709	92661	Sensibilidad - 98,2%
Exactitud			96,351%

Se observa que la sensibilidad sigue siendo alta, y en este caso se supera el 50% especificidad, funcionando mejor que la regresión logística.

Tabla 4-8 Árboles de decisión: resultados de validación

--Árboles de decisión-- Desempeño con datos de validación			
	Predicción		
Observación	0	1	Precisión
0	7149	3934	Especificidad - 64,5%
1	3491	185276	Sensibilidad - 98,2%
Exactitud			96,285%

Ya que los datos de entrenamiento están desbalanceados, pues hay muchos más créditos con clasificación ‘1’ que con clasificación ‘0’, se entrenó un árbol de decisión con el parámetro *class_weight="balanced"*, esto para que iguale o balancee las clases, teniendo en cuenta el desbalanceo original del *dataset*.

4.4.3 Método de los K Vecinos

En este método de clasificación, los datos se ubican en un espacio n dimensional y para cada dato a clasificar se tienen en cuenta los ‘k’ vecinos más cercanos espacialmente, el dato será asignado a la categoría a la que pertenezcan la mayoría de los vecinos evaluados.

Este método es muy lento en su clasificación cuando se tienen muchas variables, debido a lo complejo del cálculo de sus vecindades a medida que crecen las dimensiones del problema.

Para el problema se evaluó el método con 3 vecinos

Tabla 4-9 Método K Vecinos: resultados de pruebas

--Método de K Vecinos-- Desempeño con datos de prueba			
	Predicción		
Observación	0	1	Precisión
0	2474	3091	Especificidad - 44,5%
1	1130	93230	Sensibilidad - 98,8%
Exactitud			95,776%

Tabla 4-10 Método K Vecinos: resultados de validación

--Método de K Vecinos-- Desempeño con datos de validación			
	Predicción		
Observación	0	1	Precisión
0	4855	6243	Especificidad - 43,7%
1	2129	186623	Sensibilidad - 98,9%
Exactitud			95,811%

Se aprecia que se obtiene un resultado inferior a los algoritmos anteriores, la especificidad es inferior al 50%, al igual que para la regresión logística.

4.4.4 Redes Neuronales

Las redes neuronales están inspiradas en el funcionamiento de las neuronas cerebrales.

> "Las redes neuronales se componen de unidades llamadas neuronas, donde varias de ellas se encuentran conectadas entre sí por medio de una interconexión a la cual se le asigna un peso, cada neurona posee varias entradas, normalmente procedentes de otras neuronas, y una o varias salidas que serán las entradas de las demás neuronas o la salida del sistema." (Arroyave et al., 2008)

La red neuronal tiene pesos asociados a las interconexiones y un error definido por la comparación de la salida del sistema y el resultado esperado.

> "El sistema de aprendizaje de las redes neuronales consiste en cambiar los pesos de las interconexiones de acuerdo con el error producido entre el resultado de la red comparado con el resultado que se esperaba. Dependiendo de dicho error se aplican unas ecuaciones a cada interconexión modificando su respectivo peso con lo cual se genera un nuevo resultado y con él, un nuevo error generando así un proceso iterativo hasta que dicho error sea aceptable o nulo."(Arroyave et al., 2008)

Para el análisis se usó una red neuronal tipo perceptrón multicapa con 10 neuronas en la capa oculta y 2 neuronas en la capa de salida, dicha red se configuró con una función de activación sigmoidea logística.

Tabla 4-11 Redes neuronales: resultados de pruebas

--Redes neuronales-- Desempeño con datos de prueba			
	Predicción		
Observación	**0**	**1**	**Precisión**
0	3604	1961	Especificidad - 64,8%
1	1260	93100	Sensibilidad - 98,7%
Exactitud			**96,777%**

Tabla 4-12 Redes neuronales: resultados de validación

--Redes neuronales-- Desempeño con datos de validación			
	Predicción		
Observación	0	1	Precisión
0	7277	3821	Especificidad - 65,6%
1	2360	186392	Sensibilidad - 98,7%
Exactitud			96,907%

Los resultados de la red neuronal son similares a los logrados por los árboles de decisión, teniendo una exactitud ligeramente mejor en la red neuronal.

4.4.5 Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial son clasificadores lineales que ubican espacialmente los datos de entrenamiento y buscan encontrar un hiperplano que los separe en las categorías a clasificar. Los puntos que conforman dicho hiperplano se denominan vector de soporte, y aunque existen infinitos hiperplanos solución, se busca aquel que maximice el margen entre las categorías a clasificar.

Tabla 4-13 Máquinas de Soporte Vectorial: resultados de pruebas

--Máquinas de Soporte Vectorial-- Desempeño con datos de prueba			
	Predicción		
Observación	0	1	Precisión
0	1987	3366	Especificidad - 37,1%
1	691	93881	Sensibilidad - 99,3%
Exactitud			95,940%

Tabla 4-14 Máquinas de Soporte Vectorial: resultados de validación

--Máquinas de Soporte Vectorial-- Desempeño con datos de validación			
	Predicción		
Observación	**0**	**1**	**Precisión**
0	4207	6855	Especificidad - 38,0%
1	1293	187495	Sensibilidad - 99,3%
Exactitud			**95,923%**

Para el análisis se usó una máquina de soporte vectorial lineal. Se observa una buena sensibilidad, sin embargo, es el peor desempeño en cuanto a especificidad.

5. Conclusiones

Luego de aplicados seis algoritmos diferentes para el análisis de los datos, se consolidan sus resultados como lo muestra la siguiente tabla:

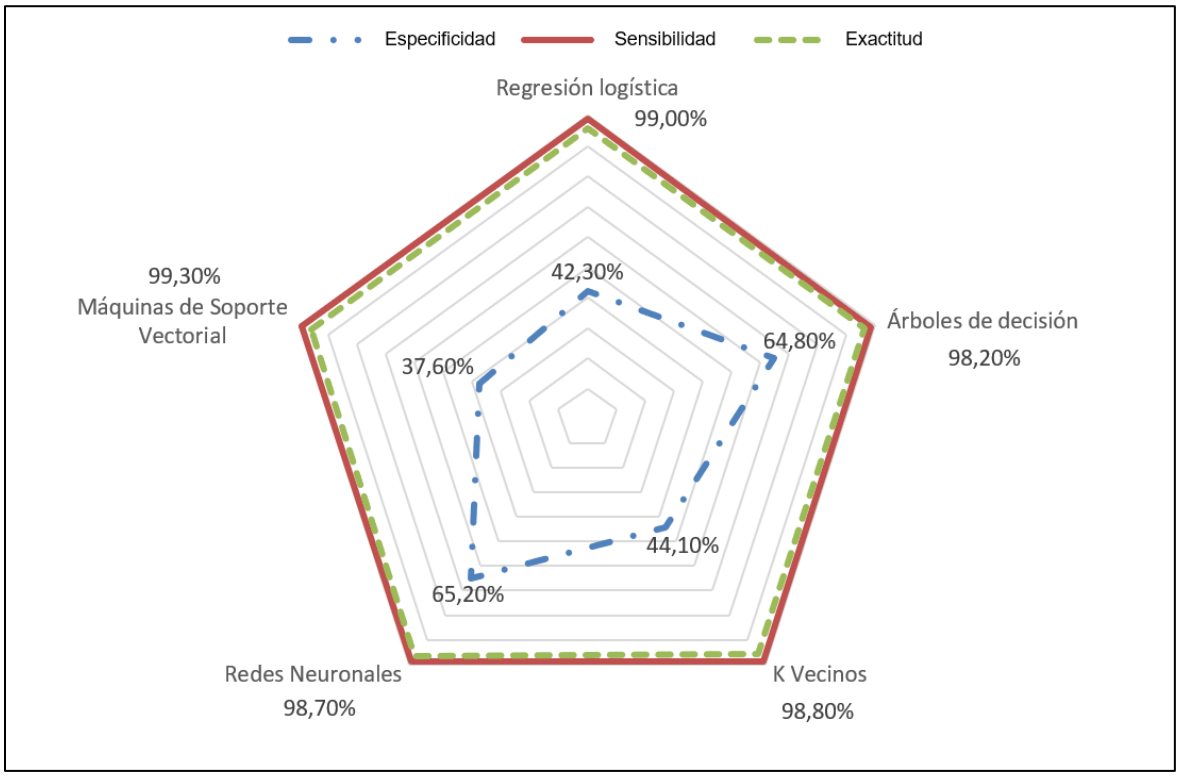
Tabla 5-1 Consolidado de resultados de clasificación

Precisión	Regresión logística	Árboles de decisión	K Vecinos	Redes Neuronales	SVM
Especificidad	42,3%	64,8%	44,1%	65,2%	37,6%
Sensibilidad	99,0%	98,2%	98,8%	98,7%	99,3%
Exactitud	**95,843%**	**96,323%**	**95,793%**	**96,842%**	**95,932%**

A simple vista parece fácil elegir a las redes neuronales por tener una exactitud mayor que los demás, sin embargo, también podría pensarse en las máquinas de soporte vectorial por su alta sensibilidad.

Si vemos la misma información en un gráfico podríamos decidir fácilmente los algoritmos que mejor se comportaron.

Ilustración 5-1 Resultados Clasificación



Vemos en el gráfico cómo la exactitud y sensibilidad es similar en todos los algoritmos, es en la especificidad donde se diferencian.
Claramente los árboles de decisión y las redes neuronales son los que obtuvieron mejor desempeño en este aspecto.

5.1 Comparativo: árboles de decisión vs redes neuronales

Teniendo en cuenta la similitud en los resultados para los árboles de decisión y para las redes neuronales, se hace necesario entender el funcionamiento de cada uno y conocer los beneficios que podrían traer para la clasificación de la cartera de crédito.

Los **árboles de decisión** reciben como entrada un conjunto de atributos que describen una situación y entregan como salida una decisión. Cuando se usan para una salida discreta se consideran de clasificación y cuando son para una salida continua se

consideran de regresión (Rusell & Norvig, 2004). En nuestro caso la salida será booleana por lo que haremos una clasificación.

Sabiendo que la clasificación se hará para dos estados; crédito recuperable o crédito irrecuperable, el árbol de decisión comienza encontrando la variable más representativa como primera prueba, es decir, qué variable de entrada sirve como discriminante para separar los registros o créditos en una categoría u otra. Esto se logra separando los datos en dos grupos según el valor de la variable elegida, por ejemplo, para la variable *Edad de Mora* se podrían separar los créditos en aquellos con una edad de mora superior a 30 días y los demás, para cada uno de estos grupos debe encontrarse otra variable discriminante, y el proceso se repite una y otra vez hasta poder llegar a la decisión de salida en todas las rutas posibles del árbol.

Una vez entrenado, un árbol de decisión puede ser visto gráficamente, además, permite determinar la importancia de cada variable en la decisión final de la clasificación, también puede conocerse el camino tomado por un registro específico y todas las variables que intervinieron en la decisión de clasificación.

Las **redes neuronales** perceptrón, como la usada en el presente trabajo, son usadas para la clasificación, teniendo como salida una variable booleana. Una red perceptrón simple servirá para clasificar datos que sean separables linealmente, es decir, que puedan ser separados en dos categorías por una línea recta o por un hiperplano. Para datos que no son separables linealmente, se deben agregar capas ocultas a la red perceptrón, lo cual genera regiones de decisión complejas, que responden a superficies de funciones polinomiales (Martín del Brio & Sanz Molina, 2005). En nuestro caso se usó una red perceptrón multicapa.

La red usada se compone de tres capas; la capa de entrada, con 54 neuronas, correspondientes a las variables de entrada, la capa oculta, con 10 neuronas y la capa de salida con 2 neuronas, correspondientes a los dos valores de la variable de salida. Cada neurona de una capa tiene un peso hacia cada neurona de la capa siguiente, así entre la capa de entrada y la oculta se tienen 540 pesos y entre la capa oculta y la de salida se tienen 20 pesos.

Para cada registro o crédito de entrenamiento que recibe la red neuronal, realiza operaciones matemáticas usando los pesos, al final, según una función sigmoidea, genera

una salida binaria, la cual indica si el crédito es marcado como irrecuperable o no, si dicha salida coincide con la clasificación del crédito, entonces se recibe el siguiente registro, si la salida no coincide entonces se realizan ajustes a los pesos, de manera que el error de predicción vaya siendo cada vez más bajo.

Al finalizar la etapa de entrenamiento, se tendrán unos pesos establecidos que aseguran el menor error posible para la clasificación de la red neuronal. No es posible encontrar una relación entre las variables de entrada y los pesos de las neuronas, tampoco ordenar las variables de entrada de acuerdo con su importancia en la clasificación.

Considerando las diferencias entre los árboles de decisión y las redes neuronales, se elegirá como modelo de clasificación al primero, por su potencial explicativo sobre las variables de entrada, además de su capacidad de clasificación.

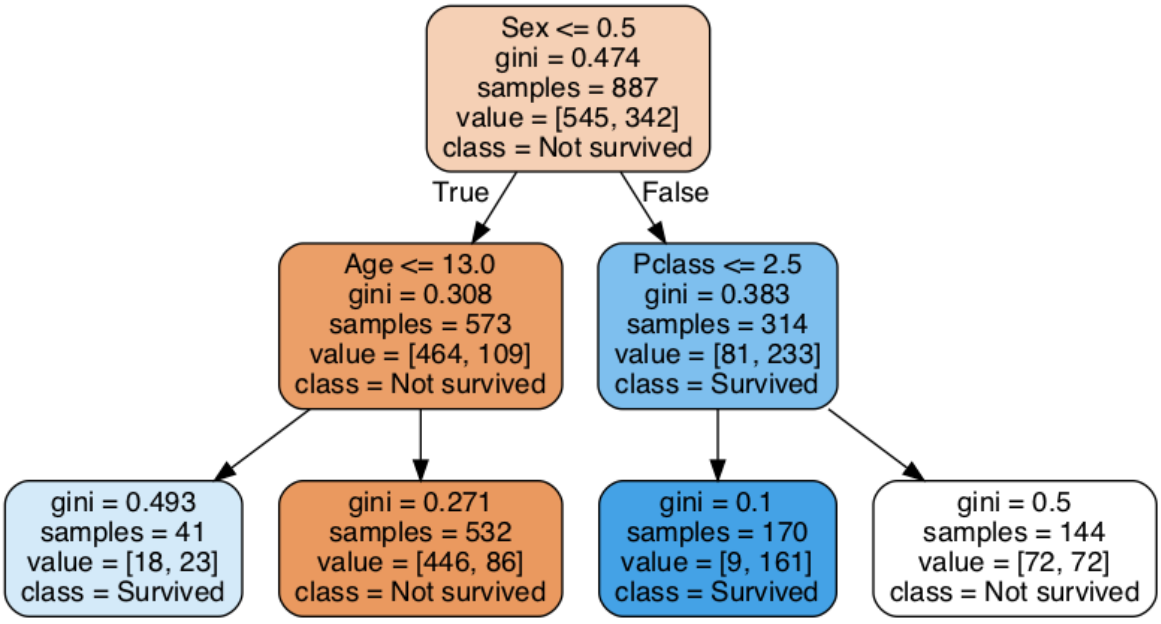
5.2 Análisis del árbol de decisión

Los árboles de decisión son muy útiles por su relativamente fácil forma de interpretarlos y por la parametrización que permiten para obtener la mejor relación entre complejidad y precisión al enfrentarse a un problema específico.

En la representación gráfica de un árbol de decisión existe un nodo raíz, ubicado en la parte superior del árbol, es la entrada y primera variable a evaluar.

La figura 4-2 representa un árbol de decisión para clasificar a las víctimas del Titanic según su posibilidad de sobrevivir teniendo en cuenta tres variables; género, edad y clase en que viajaban.

Ilustración 5-2 Representación gráfica de un árbol de decisión



Tomada de: https://towardsdatascience.com/an-introduction-to-decision-trees-with-python-and-scikit-learn-1a5ba6fc204f

Luego de evaluar el nodo raíz, se continúa descendiendo en el árbol evaluando otras variables a través de nodos interiores, hasta finalmente llegar a una clasificación en lo que se denomina una hoja. La distancia entre el nodo raíz y la hoja más lejana del árbol se considera la profundidad de este.

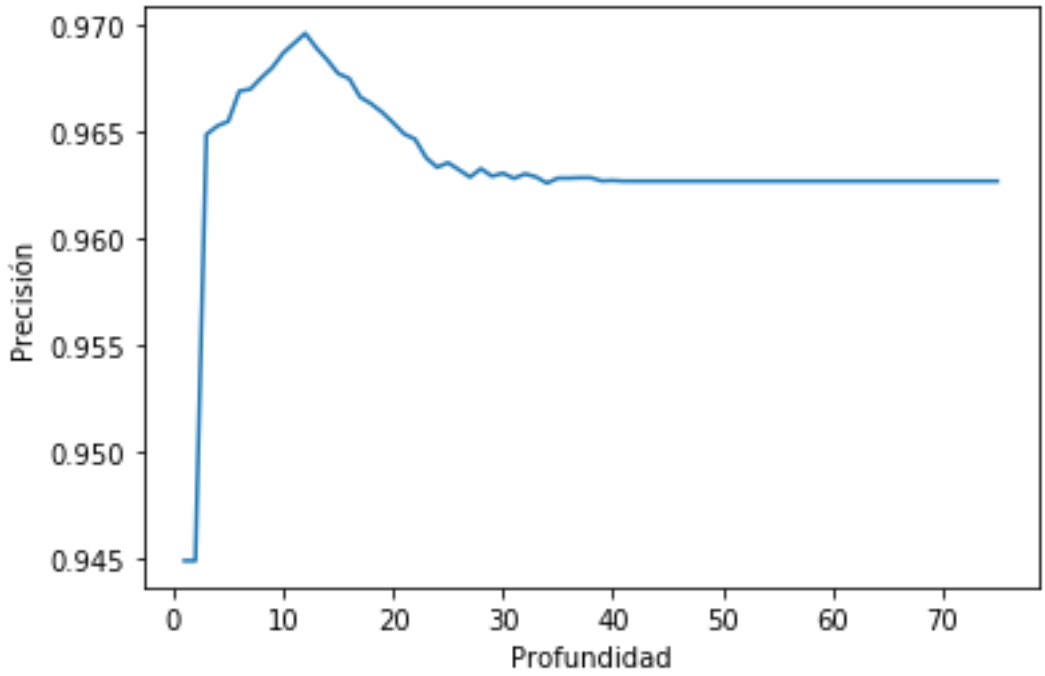
El árbol de la figura 4-2 tiene una profundidad de 2 y 4 hojas.

Para optimizar el árbol de decisión usado en el análisis, evaluamos su profundidad y cantidad de hojas y obtenemos los siguientes datos:

- **Profundidad**: 76
- **Cantidad de hojas**: 17.595

Estos datos dan muestra de un árbol sumamente complejo e imposible de mostrar gráficamente. Una mayor profundidad de un árbol de decisión no necesariamente se ve reflejada en una mayor precisión, por esto se ejecutan, para el mismo conjunto de datos de entrenamiento, árboles de precisión con profundidades entre 1 y 76, el resultado de dichas ejecuciones se refleja en la figura 4-3.

Ilustración 5-3 Precisión según profundidad del árbol de decisión



Sorprendentemente, no sólo se puede obtener la misma exactitud con una profundidad menor a 76, sino que inclusive hay mejores resultados para profundidades mucho menores a 76.

En este caso, la máxima exactitud se alcanzó con una profundidad de 12. Teniendo esta información se entrena un nuevo árbol de decisión, el cual queda con una profundidad de 12 y con 423 hojas. Esta nueva profundidad, 16% de la original y cantidad de hojas, menor al 3% de las hojas originales no hace necesario una poda minuciosa del árbol, debido a que sus resultados de especificidad y sensibilidad son más que aceptables para la entidad del trabajo.

Para este nuevo árbol se evalúan los datos de prueba y de validación.

Tabla 5-2 Árbol de decisión, profundidad 12: resultados de pruebas

--Árbol de decisión, profundidad 12-- Desempeño con datos de prueba			
	Predicción		
Observación	0	1	Precisión
0	5402	102	Especificidad - 98,1%
1	6452	87969	Sensibilidad - 93,2%
Exactitud			93,441%

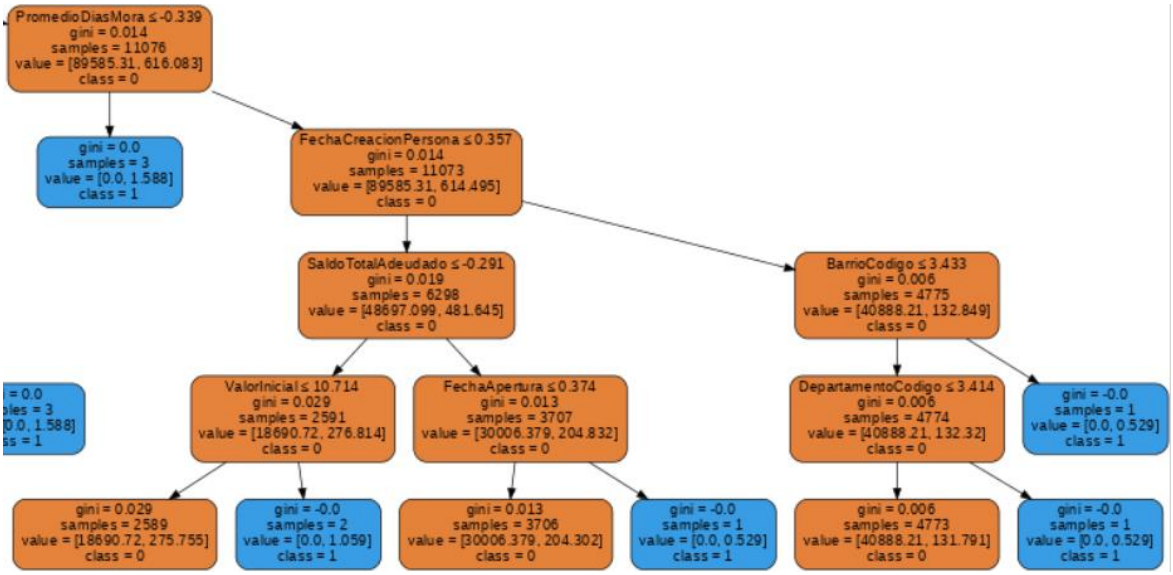
Tabla 5-3 Árbol de decisión, profundidad 12: resultados de validación

--Árbol de decisión, profundidad 12-- Desempeño con datos de validación			
	Predicción		
Observación	0	1	Precisión
0	10741	215	Especificidad - 98,0%
1	12874	176020	Sensibilidad - 93,2%
Exactitud			93,451%

Vemos como la especificidad se ha disparado hasta un resultado del 98%, cuando antes no se llegaba ni al 70% con ningún algoritmo. Esto reduce un poco la sensibilidad del modelo, pero es algo aceptable para la entidad del ejercicio. Además, una sensibilidad reducida en 5 puntos porcentuales a cambio de un aumento en la especificidad de más de 30 puntos porcentuales es un muy buen resultado desde el punto de vista de la aplicación del modelo.

Aunque el árbol ahora tiene mucho menos hojas, sigue siendo muy grande para ser mostrado visualmente, a modo de ejemplo, se muestra un subárbol.

Ilustración 5-4 Subárbol del árbol de decisión optimizado



Podemos evaluar la importancia que representa cada variable para la decisión llevada a cabo, esto es, qué tan preponderante es una variable para determinar la clasificación final de un dato.

Para nuestro caso se muestra el top diez de las variables más importantes para el árbol de decisión.

Tabla 5-4 Variables más importantes para clasificación

Índice	Variable	Importancia
49	TotalGestionesCarteraSMS	36,46%
18	FechaCreacionCredito	17,03%
40	EdadMora	12,03%
31	Saldo	11,67%
34	ValorTotalPagadoCapital	6,67%
35	FechaVenceUltCuota	3,49%
37	CuotasCanceladas	3,19%
50	TotalGestionesCarteraCorreo	2,16%
53	DiaCreaPrimerGestion	2,12%
46	NroPagoRealizados	1,64%

Este resultado es muy importante y explicativo para la entidad del análisis, pues le permite entender qué datos son útiles para definir si un crédito entrará en cartera irrecuperable o

no, además, permite centrar los esfuerzos en las variables más importantes, ya que algunas están relacionadas directamente con actividades en la entidad.

5.3 Trabajo futuro

En un análisis futuro de datos similares en la misma o en otra entidad es importante que se puedan obtener datos para varios ciclos de tiempo, según la manera de trabajar de cada entidad. De esta forma se tendrían matrices de transición para períodos de temporada de la entidad, para períodos normales y para períodos valle.

Realizar el análisis para estos períodos diferenciados permitiría entender el comportamiento de la cartera en cada período y podría usarse el modelo ajustado a cada período en el caso específico.

Aun cuando la cartera, típicamente se evalúa en períodos de 30 días, sería muy aportante poder realizar en una entidad un piloto con ciertos clientes o créditos y tomar acciones para esta cartera para períodos más cortos, como cada 10 días. Una vez hecho esto y teniendo datos del comportamiento de esos créditos, se podría realizar todo el análisis del presente trabajo de grado y seguramente se podrían observar edades tempranas de mora en las cuales es efectivo el cobro, quizá también pueda encontrarse el rango más preciso de días en donde el crédito entra en mora irrecuperable, pudiendo ser más oportuno el trabajo de la entidad en el cobro de la cartera antes de que llegue a estos días de mora.

Una reducción más drástica de la dimensionalidad de los datos podría realizarse buscando trabajar con menos variables, ya que al revisar la literatura encontramos que los modelos predictivos de cartera, y específicamente los árboles de decisión, funcionan adecuadamente con cerca de 20 variables. Esta reducción podría realizarse usando técnicas estadísticas y juicio de expertos.

El uso de diferentes técnicas garantiza un resultado más coherente y preciso en el modelo predictivo. Teniendo en cuenta las técnicas ya probadas en el presente estudio podrían

incluirse otras para verificar si la precisión del modelo puede mejorarse, específicamente hablando de árboles de decisión, se podría usar un *random forest*, el cual, por definición, sería más preciso que cualquier árbol de decisión del conjunto que lo conforman.

También se ve recomendable la generación de una base de datos de variables macroeconómicas actualizadas, que sirvan de insumo a la par de las variables definidas en el presente estudio, esto para asegurar un cubrimiento global del comportamiento de un crédito, desde la perspectiva del cliente y del mercado.

Referencias

Arroyave, J., Herrera, J., & Vásquez, E. (2008). *Sistema inteligente para detección de intrusos en redes informáticas SIDIRI*.

Banca de las Oportunidades. (2019). *Banca de las Oportunidades*. http://bancadelasoportunidades.gov.co/index.php/es

Banca de las Oportunidades, & Superintendencia Financiera de Colombia. (2017). *Reporte de Inclusión Financiera 2017*. https://bancadelasoportunidades.gov.co/sites/default/files/2018-07/RIF 2017 LIBRO FINAL_WEB 02_1.pdf

Banca de las Oportunidades, & Superintendencia Financiera de Colombia. (2018). *Reporte de Inclusión Financiera 2018*. http://bancadelasoportunidades.gov.co/sites/default/files/2019-06/RIF FINAL.pdf

Bitvai, Z., & Cohn, T. (2015). Predicting peer-to-peer loan rates using Bayesian non-linear regression. *Proceedings of the National Conference on Artificial Intelligence*, *3*, 2203–2209.

Byanjankar, A. (2018). Predicting credit risk in Peer-to-Peer lending with survival analysis. *2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings*, *2018-Janua*, 1–8. https://doi.org/10.1109/SSCI.2017.8280927

Byanjankar, A., Heikkila, M., & Mezei, J. (2015). Predicting credit risk in peer-to-peer lending: A neural network approach. *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*, 719–725. https://doi.org/10.1109/SSCI.2015.109

Castaño, L., & Ayala, C. (2016). *Diagnóstico de la Gestión de Cartera en una*

Empresa Proveedora del Sector Salud en Colombia. https://repository.eafit.edu.co/handle/10784/11500

CNN Periodismo Digital. (2016). *Ricard Bonastre: Cómo Generar más ingresos por campaña con algoritmos predictivos*. http://www.callcenternews.com.ar/entrevistas/320-cgmi

Colombia Fintech. (2019). *Asociación Colombiana de Empresas de Tecnología e Innovación Financiera*. https://www.colombiafintech.co/

Daza Sandoval, L. C. (2015). *Estrategias basadas en el modelo de análisis predictivo árbol de decisión para la mejora del proceso de recaudo de cartera de la línea de vehículo particular del banco Davivienda S.A.* Pontificia Universidad Javeriana.

Deloitte. (2012). *Tendencias de cobranza y recuperación de cartera en el sector financiero a partir de la crisis*. https://www2.deloitte.com/content/dam/Deloitte/pa/Documents/financial-services/2015-01-Pa-FinancialServices-CobranzaCartera.pdf

Do, H. X., Rösch, D., & Scheule, H. (2018). Predicting loss severities for residential mortgage loans: A three-step selection approach. *European Journal of Operational Research*, *270*(1), 246–259. https://doi.org/10.1016/j.ejor.2018.02.057

Experian. (2017). *Move debt collection practices into the digital age eResolve*. https://www.experian.com/consumer-information/virtual-debt-resolution-negotiation-eResolve.html

Gahlaut, A., Tushar, & Singh, P. K. (2017). Prediction analysis of risky credit using Data mining classification models. *8th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2017*. https://doi.org/10.1109/ICCCNT.2017.8203982

Ge, R., Feng, J., Gu, B., & Zhang, P. (2017). Predicting and Deterring Default with Social Media Information in Peer-to-Peer Lending. *Journal of Management Information Systems*, *34*(2), 401–424. https://doi.org/10.1080/07421222.2017.1334472

Infórmese. (2018). *Gestión Analítica de Crédito y Cobranza*. https://www.informese.co/gestion-analitica-credito-cobranza/

Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, *266*(1–2), 511–529. https://doi.org/10.1007/s10479-017-2668-z

Kim, A., & Cho, S. B. (2019). An ensemble semi-supervised learning method for predicting defaults in social lending. *Engineering Applications of Artificial Intelligence*, *81*(December 2017), 193–199. https://doi.org/10.1016/j.engappai.2019.02.014

León Sánchez, D. P. (2015). *Modelo predictivo para riesgo de liquidez de una entidad fiduciaria usando minería de datos* [Universidad Nacional de Colombia]. http://www.bdigital.unal.edu.co/51173/

Martín del Brio, B., & Sanz Molina, A. (2005). *Redes neuronales y sistemas difusos* (Alfaomega (ed.); 2nd ed.).

Morgan, J., & Morgan Grenfell, D. (1997). *Introduction to CreditMetrics*. http://www.jpmorgan.com

Rusell, S., & Norvig, P. (2004). *Inteligencia Artificial, un enfoque moderno* (Pearson Education (ed.); 2nd ed.).

Skiena, S. S. (2017). The data science design manual. In *Springer*. https://doi.org/10.1007/978-3-319-55444-0

Superintendencia Financiera de Colombia. (2001). *Superintendencia Financiera de Colombia - Cartera de Crédito*. https://www.superfinanciera.gov.co/publicacion/18575

Superintendencia Financiera de Colombia. (2013). *Superintendencia Financiera de Colombia - Conceptos*. https://www.superfinanciera.gov.co/jsp/loader.jsf?lServicio=Publicaciones&lTipo=publicaciones&lFuncion=loadContenidoPublicacion&id=60956&reAncha=1

Superintendencia Financiera de Colombia. (2019). *Evolución cartera de créditos*.

https://www.superfinanciera.gov.co/inicio/evolucion-cartera-de-creditos-60950

Velásquez, A. B. (2013). *Diseño de un modelo predictivo de seguimiento de riesgo de crédito para la cartera comercial, para una entidad financiera del Valle de Aburrá*.

Zapata, A. (2003). Modelando el riesgo de crédito en colombia: matrices de transición para la cartera comercial. *Apuntes de Banca y Finanzas*, *6*.

Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. *Physica A: Statistical Mechanics and Its Applications*, *534*, 122370. https://doi.org/10.1016/j.physa.2019.122370

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, *162*(Itqm 2019), 503–513. https://doi.org/10.1016/j.procs.2019.12.017