Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Caracterización y predicción de la demanda de crédito a partir de datos abiertos

Trabajo Fin de Máster
Máster Universitario en Gestión de la Información

Autor: Lady Katherine Riveros Perilla
Tutor: César Ferri Ramírez
Cotutor externo: Fabiola del Toro Osorio

2020/2021

Agradecimientos

Gracias a las personas que me acompañaron a lo largo de este proceso de mi maestría. A mi familia y amigos, gracias por su comprensión, apoyo y afecto. A mis tutores: César Ferri, Fabiola del Toro y Victoria Ospina muchas gracias por el aprendizaje construido, por guiarme, acompañarme y aconsejarme de la mejor manera.

Agradezco especialmente a la Escuela Colombiana de Ingeniería Julio Garavito por su programa de Becas de Posgrado del cual fui beneficiaria para cursar mis estudios de maestría y culminar en feliz término la doble titulación en convenio con la Universidad Politécnica de Valencia en España.

También un agradecimiento fraternal a la Escuela de Informática ETSINF de la Universidad Politécnica de Valencia porque me ha permitido expandir mis fronteras académicas y profesionales, mediante las prácticas empresariales en Segarra Abogados y Economistas y en la Conselleria de Política territorial, obras públicas y movilidad. Gracias a los profesionales de los equipos de trabajo en los que participe, sus comentarios me ayudaron a estructurar y desarrollar con profundidad mis ideas.

Resumen

En la industria financiera tradicional el acceso al crédito depende en gran medida del comportamiento de pagos del cliente en el pasado, el nivel de riqueza y las garantías. Esto implica barreras de acceso al crédito, por un lado, la exclusión de grupos de población con condiciones óptimas que no tienen experiencia previa con productos de crédito y carecen de información histórica sobre sus hábitos de pago en la banca. Y por otro lado, los reportes de historia de crédito pueden contener errores y rezagos en la información que pueden llevar a decisiones sesgadas.

En este sentido, se propone utilizar fuentes de datos abiertos para evaluar el acceso al crédito como una estrategia de inclusión financiera y mitigar las barreras de acceso al crédito. Se realiza un análisis de caracterización y predicción de acceso al crédito, se desarrollan tres técnicas supervisadas de clasificación: árbol de clasificación, bosque aleatorio y regresión logística, para evaluar las variables significativas y el modelo con mejor desempeño, con el objetivo de identificar los perfiles de bajo riesgo que acceden al crédito en las tres principales ciudades de Colombia: Bogotá, Medellín y Cali, utilizando datos abiertos de la Encuesta anual de carga financiera y educación financiera de los hogares (Iefic) del DANE. Finalmente, los resultados se presentan en un tablero de visualización analítica de datos.

Palabras clave: demanda de crédito, datos abiertos y técnicas supervisadas de clasificación.

Abstract

The traditional financial industry, the credit access depends on the customer's payment behavior in the past, the level of wealth and guarantees. This implies barriers to access to credit, on the one hand, the exclusion of population groups with optimal conditions that have no previous experience with credit products and lack historical information on their payment habits in banking. And on the other hand, credit history reports can contain errors and lags in information that can lead to biased decisions.

In this sense, it is proposed to use open data sources to evaluate access to credit as a financial inclusion strategy and to mitigate barriers to access to credit. This paper presents a characterization and prediction analysis of access to credit, using three supervised classification techniques: classification tree, random forest and logistic regression, to evaluate the significant variables and the model with the best performance, with the objective of identify the low-risk profiles that access credit in the three main cities of Colombia: Bogotá, Medellín and Cali, using open data from DANE's Annual Survey of Financial Burden and Financial Education of Households (Iefic). Finally, the results are presented on an analytical data visualization dashboard.

Key words: Credit demand, open data and supervised classification techniques.

Tabla de contenido

Tabla de gráficas

Lista de Tablas

1 Introducción

1.1 Problemática (Justificación)

El 82,5% de la población en Colombia cuenta con por lo menos un producto de depósito en el sector financiero, esta cifra se reduce para los productos de crédito, el 36,6% de los colombianos tienen al menos un producto con el sector financiero formal y el 20% de la población prefieren endeudarse de manera informal, ya sea con un prestamista, un familiar o amigo (Asobancaria, 2020:67). El análisis de demanda realizado por Banca de las oportunidades (2017:45) muestra las principales barreras de entrada para tomar un crédito: autoexclusión, trámites y requisitos, costos, ingresos insuficientes y reporte en centrales de información de crédito.

En la gestión de colocación de crédito se tiene gran interés en mantener el riesgo tan bajo como sea posible, y para esto, se utilizan métodos estadísticos que segmentan el perfil del cliente y la valoración del riesgo sobre el crédito solicitado. Así, las entidades financieras definen su apetito al riesgo e implementan métricas para captar y diferenciar clientes buenos o con menor riesgo de no pagar de aquellos clientes con probabilidad de ser morosos, usualmente estas métricas se construyen a partir de la información de experiencia crediticia del cliente con el banco y con el mercado financiero.

La información de la experiencia crediticia del cliente en el mercado financiero es consultada en las centrales de información de crédito, en el caso de Colombia Transunion y

Experian son las dos entidades principales. A partir de la historia de crédito, las áreas de analítica y de riesgo de crédito diseñan un perfil de riesgo del cliente y las evalúan frente a los requisitos de colocación de crédito del banco, si el cliente cumple las condiciones, su crédito será desembolsado.

Los principales requisitos para que una persona pueda acceder al crédito formal, tradicionalmente son: el historial de crédito, que incluye registro de la mora incurrida y montos de pago de los clientes, la antigüedad en la historia de crédito, los tipos de crédito que ha adquirido el usuario previamente, entre otros. (Hurley y Adebayo, 2017). Murcia (2007) afirma que cuando un cliente quiere acceder a crédito formal, el banco solicita información sobre el nivel de ingresos, revisa la historia de comportamiento crediticio y el valor de las garantías, con el fin de anticipar el nivel de endeudamiento del cliente y predecir la probabilidad de pago del crédito.

Consultar la historia de crédito a las centrales de información financiera, implica dos barreras de acceso al crédito tratadas en Hurley y Adebayo (2017), por un lado, los criterios para otorgar un crédito dependen en gran medida del historial crediticio del pasado, esto redunda en la exclusión de grupos de población con condiciones óptimas para el acceso al crédito que no tienen experiencia previa y por ende carecen de información relacionada en centrales de información crediticia.

Por otro lado, los datos del reporte crediticio pueden contener errores y rezagos en la información, la comisión de comercio federal de Estados Unidos en 2013 afirmó que está

imprecisión estuvo cercana al 26% y esto fue causal del 38% de rechazos de créditos para ese año. En el 2020 la Oficina de protección financiera del consumidor reveló que durante los meses de enero a julio las reclamaciones por información incorrecta en el reporte de crédito fueron del 86% (CFPB, 2020). En Colombia en el año 2020 las quejas de los usuarios relacionadas con el reporte de crédito represento el 2% del total de reclamaciones ante la Superintendencia financiera. (Superintendencia financiera de Colombia, 2020)

Las barreras de acceso al crédito y las grandes inversiones que realizan los bancos para consultar la experiencia de crédito de sus clientes en el mercado, permiten plantear la posibilidad de los bancos de utilizar fuentes de datos abiertos para realizar perfilamiento de la demanda de crédito formal, Este es un factor financiero de interés para los bancos, para replantear su modelo de negocio frente a las fintech como competidor en el mercado de crédito, que sustentan su operación en el uso de plataformas digitales de pagos, utilizan fuentes de datos alternativas y logran capturar información de los clientes a un menor costo.

En la competencia del mercado de crédito las entidades buscan adoptar capacidades digitales, desde el punto de vista de un área encargada de la explotación de datos, es relevante comprender la analítica de datos como un proceso iterativo soporte de la toma de decisiones. Por esto, en la sección teórica se abordan las fases de modelos de analítica de datos: integración de datos, procesamiento, construcción de modelos estadísticos, almacenamiento de datos, reutilización, análisis y visualización de datos.

Esta investigación académica utiliza una fuente de datos abiertos para aplicar las fases de un modelo de analítica de datos. Los microdatos utilizados están anonimizados, recogen información sobre el comportamiento financiero de los hogares colombianos y con el presente documento se promueve el uso, reutilización y visualización de datos abiertos en el país.

Este documento aporta evidencia empírica sobre la valoración de acceso al crédito utilizando información alternativa desde un estudio de demanda. Con lo cual se busca identificar y caracterizar nichos de población que podrían acceder al mercado de crédito formal. La pregunta que motiva esta investigación es ¿Cuál es el perfil de la población y la probabilidad de acceder a crédito formal utilizando técnicas supervisadas de clasificación sobre una fuente de datos abiertos?

1.2 Objetivos y pregunta de investigación

El propósito de este trabajo es caracterizar el perfil de la población y predecir la probabilidad de acceder a crédito formal utilizando técnicas supervisadas de clasificación sobre un conjunto de datos abiertos. Para lograrlo se proponen cuatro objetivos específicos:

1. Identificar y describir las fases de un modelo analítico de datos.
2. Recopilar fuentes de datos relevantes para la caracterización y predicción de la demanda de crédito.
3. Diseñar la integración y explotación de las fuentes de datos.
4. Estimar el modelo de caracterización y predicción de la demanda de crédito.

Este trabajo responde a la pregunta de investigación: ¿Cuál es el perfil de la población y la probabilidad de acceder a crédito en el sector financiero formal, utilizando técnicas supervisadas de clasificación sobre una fuente de datos abiertos?

1.3 Alcance y Limitaciones

El alcance de este estudio es realizar la caracterización y predicción de la demanda de crédito en las tres principales ciudades de Colombia: Bogotá, Medellín y Cali, a partir de datos abiertos. Para realizar el alcance se revisa el marco teórico sobre modelos analíticos de datos utilizando técnicas supervisadas de clasificación y predicción de la demanda de crédito financiero formal en Colombia.

Se selecciona una fuente de datos abiertos a partir de los criterios de reutilización de datos sugeridos por el modelo MELODA (*Metric for releasing open data*) y se seleccionó la *Encuesta anual de carga financiera y educación financiera de los hogares* (*Iefic*) realizada por el DANE (2019) y el Banco de la República (2019). Para asegurar la calidad de los datos seleccionados, se evaluaron criterios inherentes de la ISO/ IEC 25012.

Luego, se realiza la fase descriptiva mediante el análisis de estadística descriptiva de la fuente de datos y un análisis de correlación entre las variables preseleccionadas. Posteriormente, se desarrolla la fase del modelo predictivo, utilizando tres técnicas supervisadas de clasificación, árbol de decisión, bosque aleatorio y regresión logística.

En último lugar, se realiza una matriz de priorización con las variables más significativas que permiten describir los mejores perfiles de riesgo de los clientes y la probabilidad de acceder al crédito financiero formal, los resultados se presentan en un tablero de visualización analítica.

La *Encuesta anual de carga financiera y educación financiera de los hogares (Iefic)* es la principal fuente de microdatos y fue diseñada para obtener información de calidad sobre la situación financiera de los hogares según el DANE, (2009). Una limitación de este estudio es que la fuente de datos seleccionada solo contempla las tres principales ciudades del país: Bogotá, Medellín y Cali.

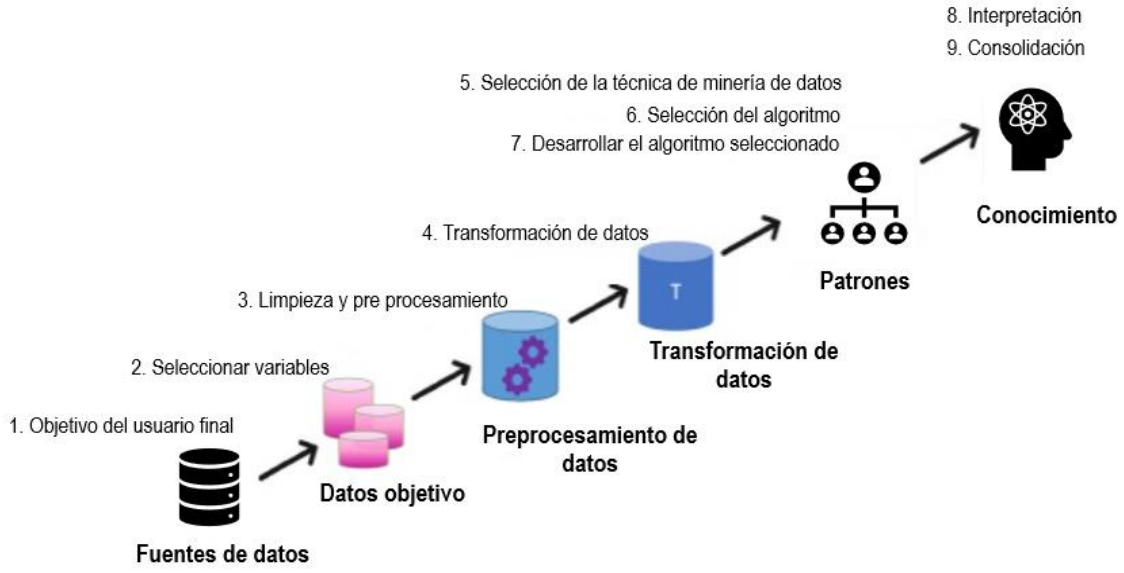
La *Iefic* se realiza sobre una muestra poblacional de la *Gran encuesta integrada de hogares* (GEIH) a partir de la tenencia de productos financieros de algún miembro del hogar encuestado. Y se encuentra disponible desde el año 2010 hasta 2016 sólo para muestras poblacionales de Bogotá, y para 2017 y 2018 se incluyeron Cali y Medellín. En este estudio, para el desarrollo del modelo se analizan las dos encuestas de 2010 hasta 2016 para Bogotá y 2017-2018 para Bogotá, Cali y Medellín, se realizan modelos de entrenamiento independientes.

La *Iefic* hace parte del sitio web de datos abiertos del *Ministerio de tecnologías y comunicaciones de Colombia* MINTIC (2019), los microdatos están anonimizados, por tratarse de información de identificación de personas en cumplimiento de la Ley Habeas Data, esto implica que el único tipo de asociación de bases de datos relacionales que puede realizarse sobre esta fuente de datos es con la GEIH y se excluyen otras fuentes.

1.4 Metodología

La metodología se desarrolla siguiendo los pasos propuestos por el proceso de descubrimiento de conocimiento en bases de datos KDD de Fayyad U. (1996) en la gráfica 1 se nombran los pasos y a continuación, se hace una breve descripción de lo que debe contener cada uno y cómo se desarrolló a lo largo de este trabajo.

Gráfica 1 Proceso de descubrimiento de conocimiento a partir de bases de datos KDD



Fuente: Fayyad, U. (1996)

Objetivos del usuario final: Se define el objetivo prioritario para el usuario final, se seleccionas las fuentes de datos a utilizar y se plantea si el producto final del proceso será utilizado para clasificación, visualización, exploración, sumarización.

Para cumplir con el primer paso, en este trabajo, se realiza en el estado del arte, una revisión de trabajos previos sobre el acceso al crédito y las fuentes de datos que se han

utilizado en las investigaciones preliminares. Y a partir de los objetivos definidos en este estudio, se expone la contribución de esta investigación.

En el marco conceptual, se explican las fases del modelo de descubrimiento en bases de datos y las técnicas de minería de datos utilizadas. Y la selección de la fuente de datos abiertos se desarrolla en el capítulo tres de fuentes de datos, se explican los criterios dados por el modelo MELODA 4.0 (*Metric for releasing open data*) para seleccionar la fuente de datos abiertos y la ISO 25012 para evaluar la calidad de la fuente de datos abiertos seleccionada.

Selección de variables: Se realiza la preselección de variables que son de interés para el negocio y que cumplen con los criterios de homogeneidad y completitud de los datos, definidos en la ISO 25012.

Limpieza y procesamiento: Se analiza la calidad de los datos y se aplican operaciones de estadística básica, para el manejo de datos desconocidos, nulos, duplicados y atípicos, para corregir los datos atípicos se utilizan histogramas y la técnica del rango del valor esperado IQR. Para corregir los datos nulos se usan métricas de tipo estadístico como media, moda, mínimo y máximo para reemplazarlos.

Transformación de variables: Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos (Fayyad et al., 1996). Se utilizan métodos de reducción de dimensión de bases de datos horizontal y vertical. La reducción horizontal,

consiste en la discretización de valores continuos basada en indicadores de entropía como el valor de la información y el peso de la evidencia. Para las variables continuas de precios, ingresos y gastos se construye un indicador para expresarlas en precios constantes de 2018. La reducción vertical, implica eliminar atributos que son insignificantes o redundantes en el contexto del problema. (Han y Kamber, 2001)

Selección de técnicas de minería de datos: De acuerdo con el objetivo definido, en este caso se aplican técnicas de sumarización, exploración y correlación, para el análisis descriptivo de caracterización y técnicas supervisadas de clasificación para la fase predictiva.

Selección de algoritmo: Para las técnicas de sumarización se utilizan estadísticos descriptivos básicos y un análisis de correlación, los algoritmos de las técnicas supervisadas que se seleccionan son árbol de decisión, bosque aleatorio y regresión logística.

Interpretación y consolidación de resultados: Se interpretan los resultados de los algoritmos, se extrae una lista de variables significativas de los modelos a la luz de los objetivos del negocio, y se crea un tablero de visualización para mostrar los patrones de comportamiento, el perfil de la población con mayor probabilidad de acceder al crédito.

1.5 Estructura del documento

En el capítulo dos, del estado del arte, se realiza una revisión de los trabajos previos de acceso al crédito, los objetivos perseguidos por la investigación, las fuentes de datos y el modelo de predicción utilizados en las investigaciones preliminares. Se explican los de modelos de descubrimiento a partir de bases de datos y sus respectivas fases.

En el capítulo tres, se selecciona la fuente de datos abiertos, evaluando los criterios dados por el modelo MELODA 4.0 (*Metric for releasing open data*) para medir el nivel de reutilización de la fuente de datos abiertos y la ISO 25012 para evaluar la calidad de la fuente de datos abiertos seleccionada y se describen las variables de la fuente de datos utilizada.

En el capítulo cuatro, se explican las tareas realizadas de limpieza, procesamiento y transformación de la base de datos seleccionada, para identificar las variables que están completas, corregir datos vacíos y atípicos, y transformar variables, a partir de la discretización de variables continuas con base en criterios de entropía como: el valor de la información y el peso de la evidencia. También, las variables continuas de precios, ingresos y gastos se expresan en un indicador a precios constantes de 2018 para que sean comparables en el tiempo.

En el capítulo cinco, se realiza el análisis de caracterización, en el cual se estiman las estadísticas descriptivas básicas sobre las variables, se realiza un análisis exploratorio gráfico y se calcula el índice de correlación entre las variables explicativas y la variable respuesta. Y se presentan las visualizaciones de la caracterización.

Posteriormente, en el capítulo seis, se desarrolla la fase del modelo predictivo, utilizando tres técnicas supervisadas de clasificación, árbol de decisión, bosque aleatorio y regresión logística, para cada una de las técnicas se discuten los supuestos, las ventajas, desventajas del modelo, y la selección de variables significativas de cada modelo, se compara el desempeño de los tres modelos en términos de disminución de la tasa de error, mejor desempeño en la clasificación y mayor generalidad de las reglas de decisión.

En la sección seis cuatro, sobre los resultados, se obtiene una comparación con las variables más significativas que permiten describir los mejores perfiles de riesgo de los clientes y la probabilidad de acceder al crédito financiero formal, los resultados se presentan en un tablero de visualización analítica y el análisis se documenta en el presente trabajo de fin de máster.

Finalmente, en el capítulo siete, se describen las conclusiones, el cumplimiento de los objetivos planteados en el presente estudio, los trabajos futuros y la relación del trabajo con los estudios cursados.

2 Estado del arte

La demanda de crédito en Colombia es objeto de investigación en tres perspectivas, según (Murcia, 2007) la primera, el mercado de crédito agregado y sus restricciones. El segundo enfoque es desde la oferta, el cual se desarrolla desde la perspectiva de los bancos, se han utilizado encuestas como la Encuesta trimestral sobre la situación del crédito en Colombia del Banco de la República (2021) para medir el desempeño de la demanda según la percepción de los desembolsos de crédito de los bancos. Finalmente, el enfoque de la demanda a partir de microdatos sobre el comportamiento de los usuarios de productos financieros.

2.1 Trabajos previos sobre la demanda de crédito

En las investigaciones previas sobre la demanda de crédito se encuentran trabajos empíricos desarrollados sobre la gestión de riesgo de los clientes como el de Butaru et al. (2016) quienes proponen una metodología para predecir el nivel de riesgo de los clientes existentes en seis bancos comerciales de Estados Unidos, combinando variables del comportamiento en el uso de las tarjetas de crédito, información del buró de crédito y variables macroeconómicas. Y realizan un modelo mediante herramientas de árboles de decisión para medir la capacidad de discriminación entre clientes con probabilidad de pago y de no pago, para cada una de las entidades de estudio de acuerdo con las políticas de otorgamiento de crédito internas de cada entidad.

En trabajos previos de investigación como el de FICO (2017) se demuestra que las fuentes de datos alternativas agregan valor predictivo a los modelos de riesgo de crédito basados en datos tradicionales. El análisis se realizó para una cartera de solicitud de préstamos personales, los atributos de datos de crédito tradicionales capturaron más valor que las características de datos alternativas, cuando se combinan las fuentes de datos tradicionales y alternativas, los datos alternativos capturaron aproximadamente el 60% del poder predictivo, y se obtuvo un alto grado de superposición entre los dos, se generó un modelo con mayor capacidad predictiva.

Murcia (2007) realiza un análisis sobre los determinantes del crédito en Colombia, a partir de los microdatos de la *Encuesta de calidad de vida de los hogares* elaborada por el DANE (2003), las variables más importantes fueron el ingreso, la riqueza, la posición geográfica, el acceso a la seguridad social, el nivel de educación y la edad, estos factores influyen en la posibilidad de adquirir tarjeta de crédito y crédito hipotecario en Colombia.

Murcia (2007) utiliza la metodología de componentes principales para construir índices de riqueza y de no pago, y se incluyen como variables explicativas del modelo junto con variables sociodemográficas como la ubicación demográfica y el sexo. Utiliza un modelo probit[1] para evaluar la probabilidad de adquirir un crédito hipotecario y tarjeta de crédito, su conclusión más relevante es que las variables más significativas son las características

[1] En Gujarati y Porter (2010) se explica que el modelo probit permite estimar la probabilidad de que un evento suceda, cuando la variable respuesta es dicotómica o presenta dos clases.

socioeconómicas y la riqueza, y que existe autoexclusión de los hogares para acceder al crédito.

Iregui, Melo, Ramírez y Tribín (2018) utilizan la *Encuesta longitudinal colombiana* (*elca*) elaborada por la Universidad de los Andes (2013) concluyen que la probabilidad de que un hogar tenga crédito está relacionada positivamente con el estado civil del jefe del hogar (casado), su educación, el ingreso, el tamaño del hogar, la propiedad de la vivienda y la participación laboral.

El estudio de Gutiérrez et al. (2011) utiliza la *Encuesta de carga y educación financiera de los hogares* de 2010 para Bogotá, y analiza las condiciones de endeudamiento y los determinantes de la probabilidad de incumplimiento y de sobreendeudamiento de los hogares. Su principal conclusión es que cuando aumenta el ingreso, el empleo y la edad del jefe del hogar disminuye la probabilidad del incumplimiento. (Iregui, Melo, Ramírez y Tribín, 2018).

Céspedes (2017) estima la elasticidad de la demanda de crédito respecto a la tasa de interés, mediante un modelo de regresión de mínimos cuadrados ordinarios calcula que ante un incremento de 1 % en la tasa de interés de mercado, la demanda de crédito disminuye en 0,29%. El autor crea una base de datos única resultado de la relación entre la base de Registro Consolidado de Créditos (RCC) y la Encuesta Nacional de Hogares de Perú (ENAHO), para los años 2008-2014. Las variables significativas son la heterogeneidad de

acuerdo con el tipo de crédito, la moneda en la cual se otorgó el crédito, el nivel de ingreso y el nivel educativo de las personas que acceden al crédito.

Mariño, Pacheco y Segovia (2018) utilizan la *Encuesta de carga y educación financiera de los hogares* elaborada por el DANE (2012-2018) y realizan un análisis de indicadores para Bogotá, Cali y Medellín, plantean un indicador de carga financiera, medido como la razón entre la amortización, y el pago mensual de intereses, y el ingreso mensual del hogar. Otro indicador de deuda total del hogar sobre el ingreso anual. Entre los principales hallazgos encontraron que Bogotá presentó doble digito de los indicadores comparado con las otras ciudades.

En la *tabla 1* se resumen los estudios previos que han utilizado microdatos sobre el comportamiento de crédito de los individuos, tales como Murcia (2007) utilizó la *Encuesta de calidad de vida de los hogares* elaborada por el DANE (2003), Gutiérrez et al. (2011) utilizó la *Encuesta de carga y educación financiera de los hogares* elaborada por el DANE (2010).

Iregui, Melo, Ramírez y Tribín (2018) utilizaron la *Encuesta longitudinal colombiana* (*elca*) de la Universidad de los Andes (2013). Y Céspedes (2017) en Perú utilizó la Encuesta Nacional de Hogares (2008-2014). Todos los autores resaltan la escasez de la literatura sobre estudios que hayan utilizado datos desagregados a nivel de cada uno de los créditos para caracterizar y predecir la demanda de crédito.

Tabla 1 Investigaciones previas sobre la demanda de crédito

Autor	Objetivo	Fuentes de datos	Modelo	Variables	Hallazgos
Murcia (2007)	Determinantes del acceso al crédito hipotecario y tarjeta de crédito en Colombia	Dane. Encuesta de calidad de vida (2003)	Regresión probabilística	Riqueza Sociodemográficas	Probabilidad observada: **4%(Hip) 10% (Tdc)** Probabilidad modelada: **2% (Hip) 4% (Tdc)**
Céspedes (2017)	Estimar la elasticidad de la demanda de crédito a nivel de personas	Encuesta Nacional de Hogares (2008-2014) Reporte consolidado de créditos	Regresión mínimos cuadrados ordinarios.	Tasa de interés Edad Ingreso Nivel educativo Región Actividad laboral Conocimiento financiero.	Elasticidad: **-0,29%** Concentración ingreso alto, Lima, Edad media, Más educada.
Iregui et al. (2018)	Determinantes del crédito formal e informal en zonas urbanas y rurales en Colombia	Universidad de los Andes Encuesta longitudinal colombiana (2016)	Regresión logística	Estado civil Educación Ingreso Tamaño hogar Propiedad vivienda Participación laboral	Aumento del ingreso aumenta la probabilidad del crédito formal, tanto en la zona urbana como en la rural, y reduce la probabilidad de tener un crédito informal.
Mariño, Pacheco, y Segovia (2018)	Indicadores de endeudamiento tenencia de productos de crédito	Dane, Encuesta de carga financiera y educación financiera de los hogares (2012-2018)	Análisis de sensibilidad Indicadores sintéticos	Indicador de carga financiera (CFI) Deuda (DSI) Promedio de cuota	Los indicadores CFI y DSI disminuyeron respecto al año anterior, esto se explica por un incremento del ingreso.

Fuente: Elaboración propia

2.2 Propuesta y contribución de la investigación

La contribución de esta propuesta de investigación es aportar evidencia sobre la predicción del acceso al crédito en Colombia, utilizando datos abiertos como una estrategia de inclusión financiera para mitigar dos barreras de acceso al crédito, definir un perfil de bajo riesgo para grupos de población con condiciones óptimas, que no tienen experiencia previa y carecen de información histórica sobre sus hábitos de pago. También la propuesta puede aplicarse cuando los reportes de historia de crédito contienen errores o rezagos de información.

El uso de datos abiertos, en particular, con microdatos, esto es un aporte en sí mismo, que permite desarrollar el mercado de la economía digital (Iglesias, 2020). Por las características de ser reutilizable, accesible, fácil de compartir y de calidad, tiene como impactó el menor costo de recabar información. Además, autores como Murcia, (2007), Gutiérrez et al. (2011) Iregui, Melo, Ramírez, & Tribín, (2018) y Céspedes (2017) resaltan la escasez de la literatura sobre estudios que hayan utilizado datos desagregados a nivel de cada uno de los créditos para la predicción de acceso al crédito.

En la práctica, los bancos diseñan sus modelos de predicción y riesgo de crédito a partir de datos desagregados del comportamiento de sus clientes, estos datos por regulación legislativa y protección de la privacidad del cliente no son públicos, pero son el insumo de los bancos para diseñar métricas de su oferta en el mercado. Esta investigación propone y desarrolla el análisis utilizando una fuente de datos abiertos, específicamente la *Encuesta*

anual de carga financiera y educación financiera de los hogares (*Iefic*) realizada por el DANE (2019) y el Banco de la República (2019).

Este estudio se aborda desde el ámbito de la informática y utiliza la metodología de modelos de descubrimiento de conocimiento a partir de bases de datos (KDD), el cual considera el componente de modelación estadística, un paso dentro del proceso automático, como explican Timarán, Hernández, Caicedo, Hidalgo, y Alvarado (2016) "El proceso KDD, combina descubrimiento y análisis, para extraer patrones de comportamiento en forma reglas generalizadas a partir de los datos, para que el usuario pueda comprender los datos de forma sistemática"

En este trabajo, las técnicas desarrolladas para extraer conocimiento son la sumarización y la clasificación, en la primera se realiza el análisis descriptivo y un análisis de correlación de variables y para la fase predictiva se desarrollan técnicas supervisadas de clasificación: árbol de decisión, bosque aleatorio y regresión logística. Para el análisis de los resultados se incorpora el componente de visualización analítica.

A continuación, se definen los conceptos utilizados: modelos de descubrimiento en bases de datos, las fases de un modelo descriptivo y predictivo, datos abiertos y visualización de datos. Para abordar la pregunta de investigación ¿Cuál es el perfil de la población y la probabilidad de acceder a crédito en el sector financiero formal, utilizando una fuente de datos abiertos?

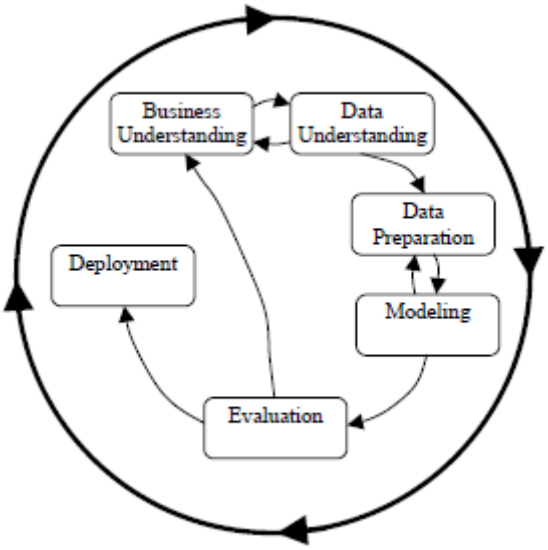
2.3 Modelos de descubrimiento de conocimiento de bases de datos (KDD)

"*Un modelo de descubrimiento de conocimiento KDD es el proceso de identificar y extraer patrones de comportamiento de los datos o conocimiento comprensible aplicando técnicas computacionales, típicamente presentan una fase descriptiva y otra predictiva*"

(Fayyad, 1996)

En Chapman (2000) se describe un modelo de descubrimiento de conocimiento como un proceso cíclico de seis pasos: comprensión del negocio, comprensión de los datos, preparación de los datos, modelación analítica, evaluación y despliegue. En el paso de modelación analítica concierne aplicar técnicas estadísticas apropiadas de la minería de datos: modelos descriptivos y predictivos (Alfred, 2005).

Gráfica 2 Modelo de descubrimiento en bases de datos

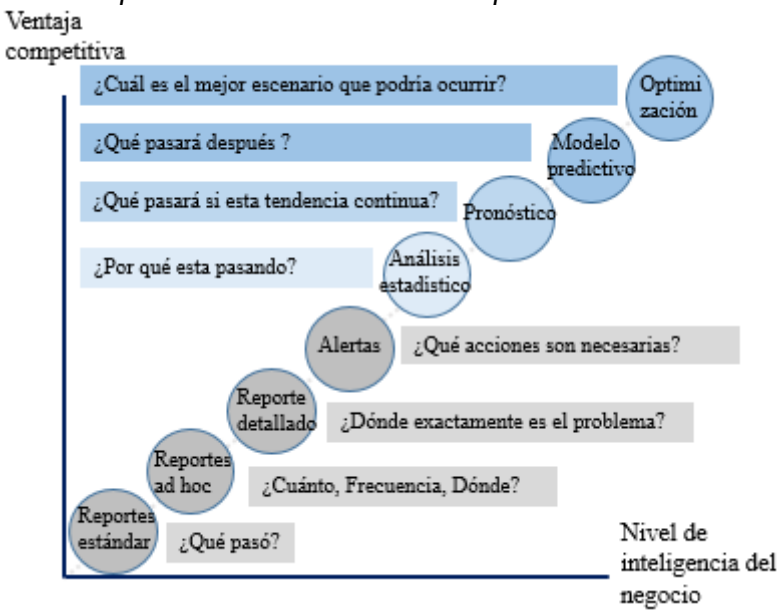


Fuente: Chapman (2000)

Hardoon y Schmueli (2016) muestran que las organizaciones dominan la inteligencia de negocios cuando tienen la capacidad e infraestructura de la compañía para generar reportes de datos automatizados, mientras que la analítica se entiende como la incorporación de métodos estadísticos y algoritmos sobre grandes volúmenes de datos para generar reportes y acciones que pueden implementarse dentro de la organización para alcanzar sus objetivos corporativos.

Hardoon y Schmueli (2016) afirman que los modelos de analítica comprenden las siguientes fases: Exploratorio y descriptivo: los datos hablan del pasado, predictivo, los datos hablan de un pronóstico futuro, prescriptivo, los datos brindan una recomendación optimizada sobre las decisiones futuras, y la visualización estratégica de los resultados.

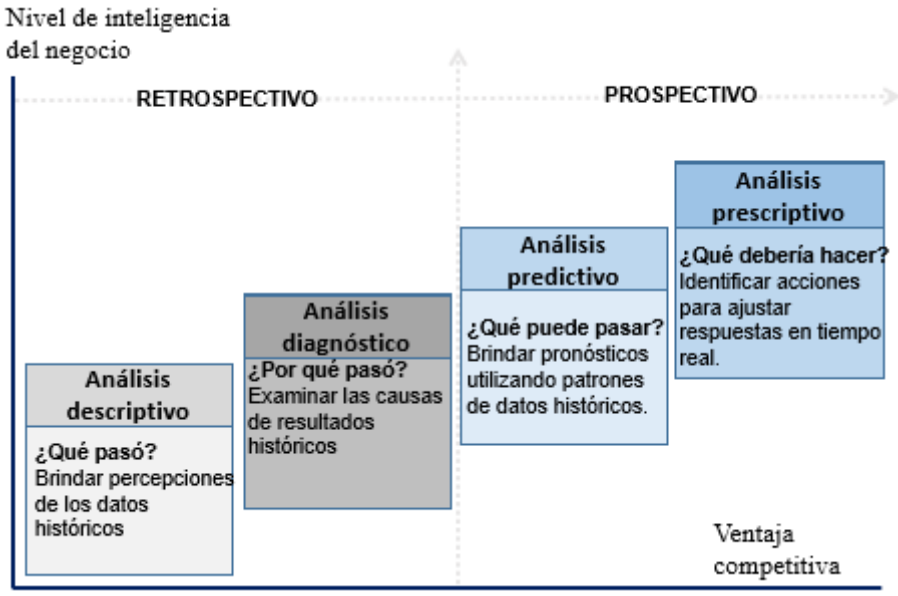
Gráfica 3 Modelo analítico de explotación de datos



Fuente: Hardoon y Schmueli (2016)

La Asociación bancaria del euro (2017) explica las fases de un ciclo analítico de datos, como se muestra en la gráfica 4, primero, un enfoque retrospectivo utilizando datos históricos, que incluye las fases descriptiva y diagnóstico. Y el enfoque de pronóstico de los datos, incluyendo las fases predictivas y prescriptivas, para obtener valor de cada una de las fases, se deberán aplicar las técnicas estadísticas apropiadas, según la disponibilidad de los datos y crear herramientas de visualización que puedan ser implementadas para la toma de decisiones de un banco, según el nivel de las preguntas e indicadores claves planteados.

Gráfica 4 Fases del modelo analítico de datos

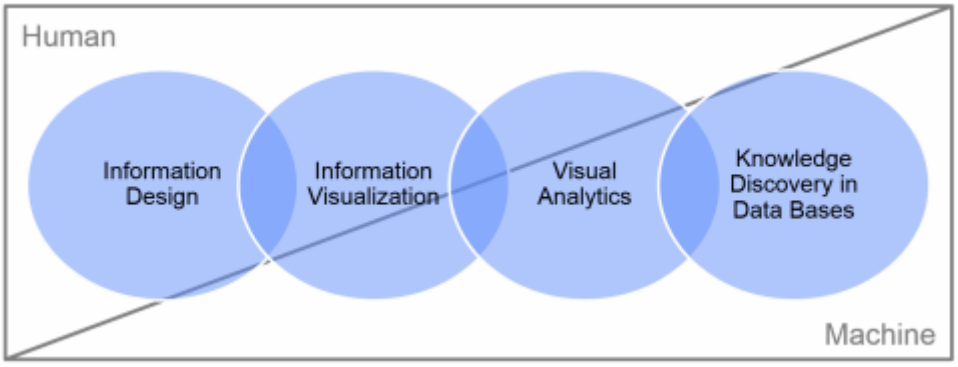


Fuente: Asociación bancaria del euro (2017)

La Asociación bancaria del euro (2017) explica que 1. El análisis descriptivo proporciona información basada en el análisis de datos históricos. Se espera que la visualización analítica de sus resultados sean la generación de informes estándar y tableros de control para monitorear los datos de la compañía. 2. La analítica de diagnóstico se utiliza para identificar

la causa de los resultados de eventos pasados con base en datos históricos, la visualización de sus resultados incorpora análisis de varianza y paneles interactivos. 3. La analítica predictiva ayuda a predecir eventos futuros mediante la búsqueda de patrones en datos históricos. 4. El análisis prescriptivo ayuda a la toma de decisiones, ya que permite identificar la decisión óptima para lograr el resultado deseado.

Gráfica 5 Metamodelo de descubrimiento de conocimiento KDM



Fuente: Ruppert (2018)

En la *tabla 2* se explican las categorías de visualización de datos y su correspondiente equivalencia con el *metamodelo de descubrimiento de conocimiento a partir de datos*, se resalta la diferencia entre inteligencia de negocios y la analítica de negocio, en la categoría de inteligencia de negocios los datos son estructurados, la compañía tiene la capacidad de la infraestructura tecnológica para producir datos estructurados.

Sin embargo, el esfuerzo de la explotación de los datos es totalmente realizada por el humano, y el objetivo alcanzado apenas permite explorar ideas clave <<insights>>, a partir de las visualizaciones obtenidas. Es todavía una fase retrospectiva del uso de los datos que

sugiere el uso de la estadística a nivel de diagnóstico y la generación de reportes automatizados para medir indicadores y alertas.

Tabla 2 Metamodelo de descubrimiento de conocimiento KDM

Categoría	Objetivo	Abstracción data	Interacción	Detección de patrones	Equivalencia DM
Diseño de información	Presentar "insights"	Agregada Data de reporte	Baja	Humano (guiado)	Reportes
Visualización información	Exploración "insights"	Datos estructurados	Media	Humano	Inteligencia de negocio
Analítica visual	Explorar y extraer "insights"	Datos no estructurados	Alta	Humano (+) Máquina	Analítica de negocio visual
KDD	Extraer "insights"	Datos no estructurados	Baja	Máquina	Analítica de negocio

Fuente: Ruppert (2018)

En la categoría de analítica de negocio, se incorporan datos externos y no estructurados dentro de los análisis y procesos de la compañía, el esfuerzo de la explotación de los datos es desarrollado por la combinación de los factores humano y máquina. El objetivo alcanzado permite explorar y extraer ideas clave <<insights>> de las visualizaciones obtenidas.

Desde las fases del ciclo analítico explicadas anteriormente, se considera el comienzo de la fase prospectiva del uso de los datos, que sugiere el uso de métodos estadísticos predictivos que dependen del esfuerzo de la máquina, y la generación de una herramienta automatizada reutilizable que permite visualizar los resultados según los cambios paramétricos dados por el humano.

Finalmente, la categoría de KDD se desarrolla completamente con datos no estructurados, la explotación de los datos depende del esfuerzo de la máquina, el objetivo alcanzado permite extraer “insights” de las visualizaciones obtenidas. Hace parte de la fase prospectiva de optimización del uso de los datos, sugiere el uso de métodos predictivos y aprendizaje automatizado <<*machine learning*>> que dependen del esfuerzo de la máquina, y la visualización de datos se muestra en formato de datos no estructurados y sistemas de recomendación.

Siguiendo a Ruppert (2018) una solución analítica de negocio visual pueda ser implementada dentro de una compañía, su viabilidad desde la perspectiva del diseño de sistemas y procesos se enmarca en los siguientes hitos con iteraciones e interacciones repetitivas:

a. Diseño exploratorio que responde a las preguntas ¿cómo se utilizará el sistema? y ¿cuáles serán las tareas para realizarlas?
b. La fase de implementación está compuesta por una fase predictiva ¿qué tan eficiente será el sistema? y otra formativa ¿cómo puede mejorarse el sistema?
c. Evaluación sumativa ¿qué tan bueno es el desempeño del sistema?

Tabla 3 Diseño de procesos de una solución analítica visual

Fuente: Ruppert (2018)

Desde el enfoque de diseño de procesos, siguiendo a Ruppert (2018) en la Tabla 3 se presentan los lineamientos para el desarrollo de herramientas de visualización de datos analítica, se tienen cuatro fases, exploratoria, predictiva, formativa y sumativa. En la fase exploratoria se realiza el diseño e identificación de los interesados <<stakeholders>>, los datos a utilizar y el modelo, esta fase se desarrolla a partir del requerimiento de la solución a las necesidades del grupo de interés, y la definición del plan de tareas a desarrollar.

En la fase predictiva, se evalúa cómo se mejora el rendimiento del usuario con la solución, se define cómo va a realizarse la operación del proceso y los tipos de datos a utilizar, y se desarrolla el prototipo de la solución. En la fase formativa, el interés se centra en cómo se mejora el sistema, se define la interacción entre el usuario y la máquina, Finalmente, en la fase sumativa se evalúa que tan bueno es el desempeño del sistema.

Siguiendo con la metodología propuesta y el objetivo del presente estudio que es determinar ¿cuál es el perfil de la población y la probabilidad de acceder al crédito?, se considera pertinente seguir el enfoque planteado por Hardoon y Schmueli (2016) y la Asociación bancaria del euro (2017). A continuación, se explican conceptualmente las fases del modelo descriptivo y predictivo, las fuentes de datos abiertos y visualización de datos

Específicamente, será necesario evaluar dos fases del ciclo analítico, la caracterización de la población para identificar las variables relevantes del conjunto de datos, y posteriormente la fase predictiva para evaluar la probabilidad de adquirir crédito o no.

2.4 Análisis descriptivo

Los modelos descriptivos tienen como objetivo resumir y representar la estructura del conjunto de datos, y mostrar tendencias históricas. Como se explica en Hernández, (2012) y en Molina y García, (2006) la analítica descriptiva proporciona técnicas de estadística sobre la muestra y las observaciones que se han realizado, tales como: promedios, sumas, conteos, agregaciones y correlación entre variables que ayudan a describir la relación entre los datos. Así, para las variables continuas se pueden observar algunas métricas unidimensionales, tales como:

- Medidas de posición: media, moda, mediana, posición no central
- Medidas de dispersión: rango, varianza, desviación estándar y percentiles,
- Medidas de forma: histograma
- Medidas de concentración: La curva de Lorenz e índice de concentración de Gini

Para las variables nominales:

- Frecuencias relativas
- Media y varianza de probabilidad estimada

2.4.1 Relaciones entre dos variables – Tablas de contingencias

Para analizar la dependencia entre los valores de variables nominales, se puede utilizar para dos variables una tabla de contingencia, para calcular la frecuencia de aparición o distribución, en ese sentido, se puede estimar la probabilidad marginal de un evento.

2.4.2 Relaciones entre múltiples variables

Estas técnicas se agrupan en tres categorías: Análisis de dependencia, interdependencia y otras. Para estudiar la relación entre múltiples variables, se busca: determinar la asociación entre las variables, la fuerza de la asociación, a través de la correlación, la forma de la relación, y posteriormente, la predicción de la variable objetivo, a partir de las variables independientes o explicativas.

2.5 Análisis predictivo

En el área del riesgo de crédito, la elegibilidad de los clientes que solicitan un crédito se define utilizando mediciones estadísticas, que parten de modelos de clasificación sobre muestras de población con la variable objetivo, se busca establecer reglas de decisión e incorporarlas para evaluar nuevas muestras (Tang, Cai, y Ouyang, 2019).

Los modelos predictivos tienen como principal objetivo representar la probabilidad de ocurrencia de un evento, expresado en una variable resultado, a partir de un conjunto de variables de entrada o variables explicativas. Depende del tipo de variable respuesta, el modelo predictivo puede ser de clasificación, si es una variable categórica o binaria, o de regresión si es una variable continua.

Existen varias técnicas de clasificación, las más adoptadas son: técnicas supervisadas y no supervisadas, en la primera, el conjunto de datos tiene una variable respuesta etiquetada y conocida, mientras que, para la segunda, no hay una única variable respuesta, el objetivo es conocer la estructura de los datos y las posibles relaciones entre las variables (Maglogiannis, Karpouzis y Wallace, 2007).

En las técnicas supervisadas el algoritmo especifica las clases de la clasificación y es necesario utilizar gran cantidad de datos de entrenamiento, para las técnicas no supervisadas, el usuario especifica el número de categorías en que se agrupan las variables, tiene mayor complejidad en la interpretación de los datos.

2.6 Técnicas supervisadas de clasificación

La característica principal del aprendizaje supervisado es que el conjunto de datos esta etiquetado, esto quiere decir que, para cada registro, se pueden observar características, atributos o variables explicativas y una variable etiqueta o variable respuesta conocida, la variable respuesta puede ser categórica, en este caso se utilizan modelos de clasificación, o

puede ser una variable continua, para estos se utilizan los modelos de regresión. (Doherty, Camiña, White, y Orenstein, 2016)

Es de carácter no paramétrico, esto implica que no se requiere evaluar la dependencia entre las variables, el objetivo es obtener una función que permita predecir la variable dependiente para los casos desconocidos.

Las técnicas supervisadas más populares para la valoración de riesgo y acceso al crédito, siguiendo la revisión empírica realizada por Tang, Cai, y Ouyang, (2019) son: la regresión logística, los árboles de decisión y redes neuronales. Los autores resaltan estudios relevantes como el de Press y Wilson (1978) quienes fueron los primeros en establecer la utilidad de las regresiones logísticas en este campo.

El trabajo de Sohn y Col. (2016) que utilizan información financiera y no financiera, y proponen un método de regresión logística difusa utilizando datos de la solicitud de crédito. También, el trabajo de Jena y Col (2017) quienes desarrollan una comparación entre la regresión logística, y algoritmos de clasificación como el árbol de decisión y el modelo de red neuronal, el último obtuvo el mejor desempeño.

De acuerdo con Maglogiannis, Karpouzis y Wallace, (2007) las técnicas supervisadas tienen modelos que mejor se ajustan según el tipo de variable respuesta, cuando se trata de una variable continua y multidimensional, será mejor utilizar, redes neuronales y máquinas de vectores de soporte, y cuando se trata de una variable respuesta categórica o discreta, se tiene mejor desempeño si se utilizan árboles de decisión o técnicas basadas en reglas.

Además, los árboles de decisión ofrecen alto criterio de precisión e interpretación que se traduce en reglas basadas en algoritmos que permiten la clasificación y ordenamiento de grupos del subconjunto, a partir del conjunto de datos de entrenamiento.

La regresión logística es una de las técnicas más tradicionales en el campo de la valoración de crédito y acceso al crédito, en el trabajo de Bartual, García, Giménez y Romero, (2012) se defiende su uso y se argumenta que esta técnica permite obtener un buen desempeño explicativo por el ajuste y precisión que puede alcanzar y obtener la probabilidad del evento clasificado.

Sin embargo, algunas desventajas presentes son el sesgo de selección de las variables explicativas, no indica los puntos de corte de las variables explicativas y presenta inestabilidad debido a la composición de la muestra utilizada para el modelo.

En esta línea se resalta el trabajo de Hargreaves, (2019) que utiliza la regresión logística para identificar la probabilidad de los nuevos clientes que acceden a crédito tendrán buen desempeño o entrarán en mora, utilizando un conjunto de datos del banco de Alemania obtiene un indicador del 74% sobre la precisión de su modelo.

La autora Hargreaves, (2019) afirma que la regresión logística es uno de los métodos más robustos cuando se tienen variables explicativas binarias. Además, es útil cuando la relación entre la variable objetivo y las variables explicativas no es lineal, requiere menor esfuerzo computacional y tiene mayor capacidad de interpretación que otros algoritmos.

2.6.1 Regresión logística

Este método se utiliza para explicar la relación entre una variable independiente binaria y las variables explicativas, la diferencia entre la regresión lineal y logística es que la primera da como respuesta una variable continua, mientras que la regresión logística da como resultado una variable binaria.

Este tipo de modelos se centra en minimizar el error de las observaciones de la muestra respecto a la recta promedio, que mejor se ajusta para predecir el comportamiento de los datos. La regresión logística permite predecir variables categóricas con variable respuesta 1 y 0. El modelo Logit básico está representado en la ecuación (1). (Allue et al. 2018)

$$Y = \frac{1}{X} = w^T X$$

Ecuación (1)

Tabla 4 Ventajas y desventajas de la regresión logística

Ventajas	Desventajas
El resultado son las variables significativas	Alta sensibilidad a valores atípicos
Facilidad de interpretación	Sólo puede modelar datos balanceados
Estabilidad en el resultado	Existencia de multicolinealidad puede reducir la precisión de la predicción del modelo.
No requiere una distribución específica para las variables	

Fuente: Allue et al. (2018)

2.6.2 Árboles de decisión

En los últimos años los problemas de clasificación se han desarrollado con base en técnicas estadísticas de partición recursiva también conocidos como árboles de decisión.

Existen diferentes algoritmos para ejecutar árboles de decisión, entre los más destacados y utilizados se encuentran: CHAID (Chi-square automatic interaction detection) desarrollado por Kass, (1980), CART (Classification and regression trees) desarrollado por Breiman, Friedman, Olshen, Stone (1984). ID3, C4.5 y C5 desarrollados por Quinlan (1993), Quest desarrollado por Loh y Shih (1997).

Y el método de inferencia condicional de Hothorn, Hornik, y Zeileis, (2006), los últimos autores, desarrollaron una propuesta basada en la distribución de la probabilidad condicional para superar los problemas de sobreajuste y el sesgo de selección de variables de los métodos tradicionales.

Hothorn et al., (2006) afirman que los métodos basados en partición binaria, tales como CART, ID3, C4.5 y C5 dividen el conjunto de observaciones, con el objetivo de aumentar la homogeneidad de las particiones, se basan en la validación cruzada y en la definición de criterios que buscan maximizar el indicador de ganancia de información, funcionan con base en hiper parámetros o criterios de poda y parada.

Pérez y Santín, (2007) resaltan que los hiper parámetros se establecen a priori por el investigador, estos son: la profundidad, la extensión máxima de los niveles permitidos después del nodo raíz, el tamaño del árbol, controla la cantidad de ramas del árbol, el número de observaciones mínimo y máximo en cada partición.

Hothorn et al., (2006) exponen que, en el enfoque tradicional, el sobreajuste se controla puntualmente definiendo el número de particiones permitidas, los criterios de poda permiten

definir el tamaño adecuado del árbol y la interpretación del árbol se ve afectada por el sesgo en la selección de variables.

Así que proponen utilizar el criterio de la distribución condicional entre la variable respuesta y las variables explicativas, en caso de que no exista una relación la partición se detiene. Y demuestran que sus resultados son una solución eficiente ya que se comporta como un árbol con criterios de poda óptimos.

En general, un modelo de clasificación busca definir las variables independientes que mejor explican la variable dependiente, separando grupos a partir de particiones de subconjuntos de datos, los cuales representan las categorías de las variables de entrada. Está compuesto por un nodo raíz, que describe las clases de la variable objetivo, nodos rama, que definen la partición del nodo, las hojas, representan los grupos clasificados y determinan los puntos de corte respectivos para la agrupación.

Además, permiten representar la probabilidad de que un evento ocurra, finalmente, se tiene el nodo terminal, que corresponde al último grupo clasificado de cada rama. El modelo observa los datos de entrada y divide los datos con mayores diferencias, se utilizan para evaluar las variables preliminares y da como respuesta una variable de clase. (SAS, 2019) A continuación, en la tabla 5 se resumen algunas ventajas y desventajas sobre los árboles de decisión.

Tabla 5 Ventajas y desventajas de los árboles de decisión

Ventajas	Desventajas
Variables de entrada y salida continuas o categóricas	Presenta sobreajuste
Prioriza las variables significativas	Presenta mayor tasa de error que otros modelos
Admite modelos no lineales	Inestabilidad: un cambio en los datos modifica la estructura del árbol
Facilidad de interpretación	Pérdida de información al categorizar variables continuas
No requiere escalamiento de variables	Baja capacidad de generalización

Fuente: Orellana (2018)

2.6.2.1 Algoritmos utilizados para estimar árboles de clasificación

A continuación, se describen los principales algoritmos utilizados para el desarrollo de árboles de clasificación: CHAID, CART, ID3, C4.5 y C5, Quest y el método de inferencia condicional. Los cuales han sido utilizados en los trabajos de Cardona, (2004), Tello, Eslava, y Tobías (2013) y Orellana (2018).

Algoritmo CHAID

Hothorn et al., (2006) resume que este método evalúa las variables explicativas con la menor significancia estadística chi-cuadrado, el objetivo es la separación de la selección de variables, la variable explicativa con la asociación más alta se selecciona para dividir. Y Orellana (2018) afirma que el cálculo se puede resumir como la suma de los cuadrados de la diferencia entre la frecuencia observada y la esperada, de la variable objetivo. A mayor valor del estadístico chi-cuadrado mayor valor de significancia.

$$chi - cuadrado = \frac{(X_{obs} - \widehat{X_{esp}})^2}{\widehat{X_{esp}}}$$

Ecuación (2)

Algoritmo CART

Desarrollado por Breiman et al. (1984) selecciona las variables predictoras en particiones distintas y no sobrepuestas. Utiliza el método de GINI para realizar la clasificación de las variables cuando son binarias. Para cada nodo de la división utiliza el score Gini ponderado. (Orellana, 2018). El cálculo del Gini es la suma de los cuadrados de probabilidad para cada una de las clases.

$$Gini = p^2 + q^2 \qquad \text{Ecuación (3)}$$

Algoritmo ID3

En el trabajo de Tello, Eslava, y Tobías (2013) utilizan este algoritmo, explican que tiene como criterio principal la selección de la variable más informativa, utilizando como base el concepto de entropía *ganancia de información* entre la variable explicativa y la variable objetivo. Los autores también afirman que este algoritmo tiene como desventaja privilegiar las variables con mayor volumen de valores, es decir, las variables más pobladas en el conjunto de datos.

Sancho, (2019) resume las principales desventajas sobre este algoritmo, menciona que está definido para variables categóricas y no funciona correctamente para variables continuas. Realiza las particiones por separado optimizando la homogeneidad local y no considera el orden de las particiones para alcanzar una solución de óptimo global. Puede ser inestable ante valores desconocidos o atípicos.

Algoritmo C4.5 y C5

Es una versión mejorada del ID3 corrige las desventajas del modelo predecesor ID3 y también utiliza como medida de entropía la ganancia de información. Mejora la estabilidad ante valores desconocidos, para el cálculo de la entropía, primero, segmenta los valores conocidos para cada atributo o variable explicativa y así mejora la homogeneidad. Mejora el óptimo global, aunque los autores afirman que no es una solución eficiente.

Y la diferencia considerable de este algoritmo frente al ID3 está en que se pueden incluir variables continuas, ordenando la variable en función de la ganancia de información y selecciona como punto de corte donde se genera el cambio de una clasificación a otra. (Sancho, 2019)

La entropía se calcula para cada nodo y la probabilidad de cada evento, se selecciona la variable con la entropía más baja respecto el nodo padre.

$$Entropía = p\ log_2(p) - q\ log_2(q)$$ Ecuación (4)

Algoritmo Quest

Loh y Vanichsetakul (1988) seleccionan las variables explicativas a partir de un análisis de varianza (ANOVA) de acuerdo con el criterio de significancia F. Loh y Shih (1997) demuestran que el enfoque anterior selecciona variables sesgadas que inducen la selección de variables nominales y abordan el problema seleccionando covariables en una escala de significancia del valor P, utilizando variables continuas, este enfoque reduce el sesgo de

selección de variables. Kim y Loh (2003) incluyen un modelo de análisis discriminante dentro de cada nodo de un árbol.

Algoritmo de inferencia condicional

Hothorn, Hornik, y Zeileis, (2006) compara los árboles de decisión basados en independencia condicional con los algoritmos CART, Quest y Guide y concluye que su propuesta presenta mejor desempeño ya que los árboles de inferencia condicional son insesgados, no presentan sobreajuste, y la exactitud de la predicción es óptima. La selección de las variables se realiza según el parámetro de significancia del valor p, esto determina el tamaño del árbol. Para evitar el sobreajuste y un tamaño excesivo del árbol, la estrategia utilizada es podar el árbol eliminando los nodos terminales hasta que se alcance un resultado igual al valor p del nivel de significancia determinado.

2.6.3 Bosque aleatorio

Es una técnica de modelos combinados que consiste en la iteración y ejecución de varios árboles de decisión a partir de múltiples subconjuntos o particiones de un conjunto de datos de entrenamiento, y se promedia el resultado de los distintos árboles, cada árbol se construye de manera independiente a los demás, comienza con un único nodo, llamado raíz del árbol y se evalúan las variables explicativas una única vez en cada partición, así, cada árbol será distinto de otro, y estarán menos correlacionados entre ellos, además, evalúa todos las posibles variables explicativas, y considera un número de variables similar entre cada árbol.

El objetivo del algoritmo es determinar la mejor separación entre las clases de la variable respuesta, el criterio para seleccionar la mejor separación, es determinado por la impureza

del índice de Gini, este índice mide la probabilidad de que un elemento seleccionado aleatoriamente sea clasificado en el grupo de clases elegido aleatoriamente, es decir, este índice permite medir la capacidad del modelo para discriminar la clasificación de quedar en un grupo u otro. La predicción de los árboles es altamente correlacionada. (Reis, Baron, y Shahaf, 2019)

Esta técnica tiene una ventaja principal sobre los árboles de decisión ya que mejora la exactitud del resultado, sin embargo, aumenta la complejidad de interpretación, el resultado obtenido se puede resumir con la importancia de cada variable explicativa del modelo y el respectivo índice de Gini asociado. (Duke University, 2014)

Para la ejecución de este modelo se deben especificar parámetros tales como: máximo número de árboles, número de variables seleccionadas para cada árbol, número de observaciones, tasa de observaciones usada para cada árbol.

Tabla 6 Ventajas y desventajas del bosque aleatorio

Ventajas	**Desventajas**
Reduce el sobreajuste	Interpretación de resultados compleja
Reduce la varianza	Selección de parámetros
Alta tasa de precisión de la predicción	Tiempo de procesamiento
Mayor estabilidad sobre los valores atípicos y faltantes.	
Admite varias dimensiones de los datos	
Su resultado son las variables significativas	

Fuente: Breiman (1999)

En el trabajo de Breiman (1999) se desarrolla el modelo de bosque aleatorio, en los últimos años se ha utilizado para evaluar el riesgo de crédito, en el trabajo de Tang, Cai, y

Ouyang, (2019) se afirma que la metodología de bosque aleatorio no se utiliza comúnmente en la evaluación del riesgo de crédito en China y su objetivo es construir evidencia empírica sobre el uso de este modelo.

Se resalta el trabajo de Haque, (2017) quien realizó una comparación entre la regresión logística y bosque aleatorio, para predecir el incumplimiento de préstamos, el bosque aleatorio presentó mejor desempeño en la precisión de la predicción, pero la regresión logística tuvo mayor volumen de casos clasificados correctamente

2.7 Validación de los modelos predictivos

Espino (2017) sugiere que la validación de los modelos consiste en dividir el conjunto de datos de acuerdo con un parámetro poblacional, una submuestra de datos será para el desarrollo del modelo, otra submuestra será para el entrenamiento y finalmente, el conjunto restante de datos para la prueba, sobre el cual se aplica la técnica de modelación para evaluar y compara los resultados con la submuestra del desarrollo. Además, como mencionan Öhman y Lundstrom, (2019) la matriz de confusión y la curva ROC son instrumentos útiles para evaluar la exactitud y el desempeño de los modelos.

2.7.1 Matriz de confusión

En la matriz de confusión se representa con una tabla de contingencias, para mostrar las clases de la variable respuesta comparado con las observaciones para cada clase entrenada en la predicción. Así, se pueden validar cuantas observaciones fueron ciertas en su resultado y cuantas fueron falsas. (Sarath, 2018)

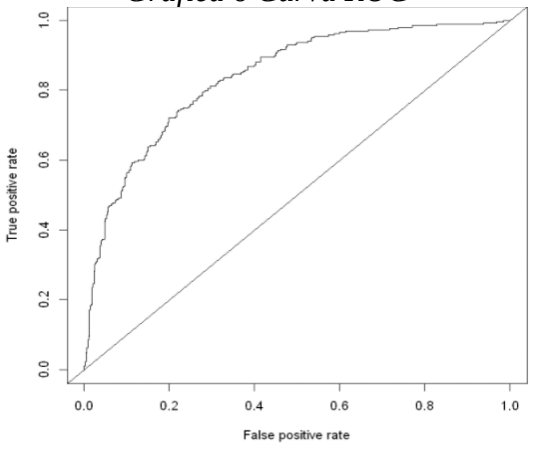
Tabla 7 Matriz de confusión

Predicción de la condición	**Condición verdadera**	
	Condición positiva	**Condición negativa**
Condición positiva	Verdadero – positivo	Falso - positivo
Condición negativa	Falso – negativo	Verdadero - negativo

2.7.2 Curva ROC

La curva ROC es una medida de rendimiento para clasificación, brinda la probabilidad y el grado de separación del modelo para diferenciar una clase de otra. Es la visualización de la clasificación binaria, presentada a partir de la sensibilidad de los casos positivos verdaderos, que no se contemplen como verdaderos (Sarath, 2018).

Gráfica 6 Curva ROC



Fuente: Sarath (2018)

3 Fuentes de datos

En la investigación de Hurley y Adebayo (2017) se afirma que los bancos tradicionales utilizan herramientas de medición de crédito que dependen principalmente del historial crediticio del pasado del cliente, esto hace que se excluyan grupos de población que pueden presentar condiciones óptimas para el acceso al crédito y no tengan experiencia previa o información en las centrales de información crediticia, con este hecho se refuerza la necesidad de incluir información alternativa para evaluar el acceso a la demanda de crédito. Además, los autores citan a la comisión de comercio federal de Estados Unidos en 2013 quien afirmó que los reportes de historial crediticio tradicionales presentan una tasa de error cercana al 26% y está fue la causa del 38% de negación de créditos ese año.

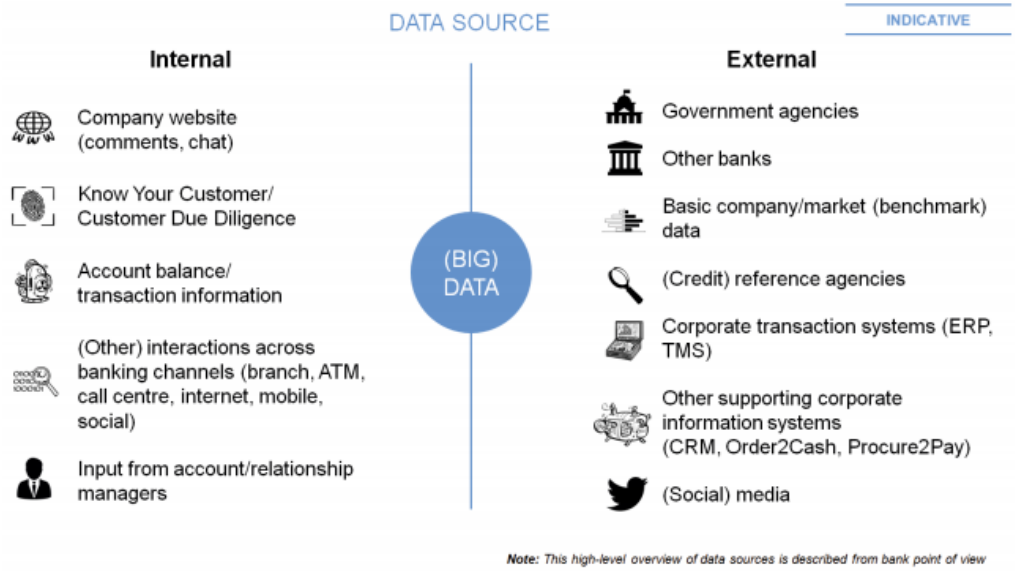
Para este trabajo es importante promover el uso de fuentes de datos abiertos y profundizar en el valor que pueden ofrecer en términos de su correcto tratamiento y explotación económica, ya sea para el perfilamiento de demanda de crédito formal en Colombia, como es el objetivo de este estudio, o cualquier tipo de segmentación y conocimiento de una población objetivo. Plantear propuestas analíticas para predecir las características de la población que accede al crédito financiero formal con fuentes de datos alternativas, ya que los bancos realizan grandes inversiones para consultar la experiencia de crédito de sus clientes en el mercado, y deben lograr capturar información de los clientes a un menor costo, como lo realizan sus competidores las Fintech.

3.1 Fuentes de datos internas y externas de los bancos

La Asociación bancaria del euro (2017) realiza una tipificación de las posibles fuentes de datos relevantes para un banco, en el rol de productor de información interna y consumidor de datos externos. Las fuentes de datos internos de un banco se generan a partir de procesos internos, productos, canales e interacciones humanas y se consolidan en tecnologías de información utilizando soluciones de inteligencia de negocios, que permiten unificar y estructurar los datos en un repositorio centralizado.

Las fuentes de datos externas relevantes para un banco pueden obtenerse de fuentes de terceros, estas fuentes pueden ser las empresas con las que tienen una relación contractual, entidades gubernamentales, otros bancos, compañías referentes en el mercado, redes sociales y otras agencias de referencia de crédito.

Gráfica 7 Tipificación de fuentes de datos de un banco



Fuente: Asociación bancaria del euro (2017)

FICO (2017) explica que las fuentes de datos externas relevantes y alternativas son todos los datos que no provengan de una agencia de crédito como por ejemplo: datos de empresas de telecomunicaciones, servicios públicos, alquiler, datos de transacciones de compras, utilización de productos, frecuencia de gasto y montos de transacciones en comercios minoristas, datos provenientes de redes sociales como Facebook, LinkedIn, Twitter, Instagram, Snapchat u otros sitios de redes sociales, pero advierte el riesgo regulatorio y los términos de privacidad serán un obstáculo para su uso aplicado en otorgamiento de crédito.

3.2 Identificación de fuentes de datos externas

Este trabajo promueve el uso de fuentes de datos abiertos, particularmente, microdatos, ya que, en investigaciones previas como el trabajo de Murcia, (2007) y en Iregui, et al. (2018) se ha mencionado la escasez de estudios sobre la demanda de crédito con microdatos. Una de las ventajas de los microdatos es que permiten conocer detalles sobre la modalidad del crédito adquirido, la fuente formal o informal, entre otras, estas características se han identificado como relevantes en las investigaciones previas.

A continuación, se recopilan fuentes de datos que pueden ser relevantes para los bancos siguiendo la tipología de la Asociación bancaria del euro (2017) principalmente se encontraron conjuntos de datos disponibles y relacionados con la demanda de crédito, en las dos primeras categorías, agencias de gobierno y otros bancos y se presentan en la Tabla 8.

Tabla 8 Fuentes externas de datos

Fuentes de datos externas	Nombre Dataset	Descripción	URL Origen
Entidades de gobierno	DANE Colombia- Microdatos Encuesta de carga financiera y educación financiera de los hogares 2017-2018	Microdatos nivel a hogares y personas sobre la situación patrimonial, nivel de endeudamiento, educación financiera y déficit de ingresos.	http://microdatos.dane.gov.co/index.php/catalog/626/data_dictionary
	DANE Colombia - Encuesta de Carga Financiera y Educación Financiera de los Hogares - IEFIC-2010 - 2016		http://microdatos.dane.gov.co/index.php/catalog/470/related_materials
	DANE Colombia – Encuesta nacional de calidad de vida		https://www.dane.gov.co/index.php/estadisticas-por-tema/salud/calidad-de-vida-ecv/encuesta-nacional-de-calidad-de-vida-ecv-2018#informacion-por-departamentos
	Superintendencia Financiera -Estudio de la demanda	Corte 2015 Corte 2017, 1.432 encuestas de muestreo aleatorio con representatividad nacional	https://www.superfinanciera.gov.co/publicacion/10098213
Otros bancos: BBVA	BBVA Data Open challenge on Kaggle	Microdatos de prueba anonimizados	https://www.kaggle.com/seussz/bbva-data-challenge/activity https://www.kaggle.com/c/bbvadatachallenge-recomendador/data
	BBVA API´S Pay Stats	Este archivo proporciona estadísticas (monto promedio de la transacción, la cantidad de	https://www.bbvaapimarket.com/documentation/bbva/paystats-download

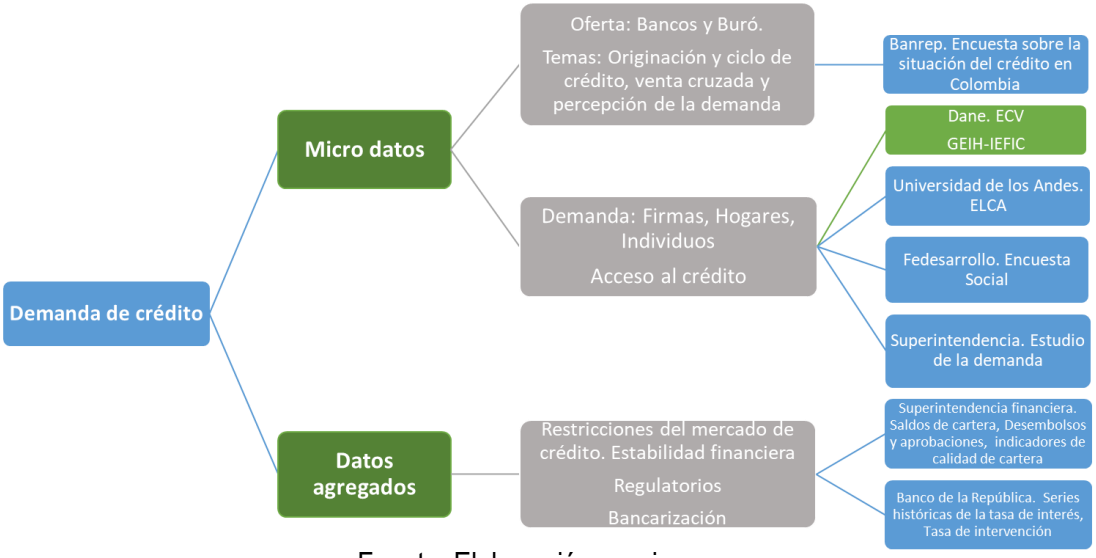
		transacciones, la cantidad de comerciantes y la cantidad de tarjetas, etc.) para un área particular y una categoría comercial.	
Otros bancos: Banco Santander	Banco Santander Open Challenge en Kaggle: Predicción de transacciones de clientes	Microdatos anonimizados que contienen variables para predecir las transacciones de los clientes	https://www.kaggle.com/c/santander-customer-transaction-prediction/data

Fuente: Elaboración propia

A partir, de la revisión teórica sobre investigaciones previas relacionadas con la literatura sobre los factores determinantes del acceso al crédito. En la *Gráfica 8* se resumen las fuentes de datos utilizadas en las investigaciones previas revisadas, siguiendo la clasificación que hace Murcia (2007) sobre la literatura de la estimación de la demanda de crédito en tres enfoques temáticos, el primero, el mercado de crédito y sus restricciones, para estos estudios se han utilizado datos macroeconómicos y agregados, para el enfoque de la oferta, estudios desde la perspectiva de los bancos, se han utilizado encuestas sobre la percepción de la demanda de crédito por parte de los bancos.

Finalmente, el enfoque de la demanda, los estudios previos han utilizado microdatos sobre el comportamiento de crédito de los individuos, tales como la *Encuesta de calidad de vida de los hogares* elaborada por el DANE (2003), la *Encuesta de carga y educación financiera de los hogares* elaborada por el DANE (2010) y la *Encuesta longitudinal colombiana* (*elca*) elaborada por la Universidad de los Andes (2013).

Gráfica 8 Fuentes de datos utilizados en investigaciones previas



Fuente: Elaboración propia

3.3 Definición de datos abiertos

Según MINTIC (2019) los datos abiertos son información pública puesta a disposición del público general, en formatos que permiten su uso y reutilización bajo licencia abierta y sin restricciones legales para su aprovechamiento. En Colombia, la Ley 1712 de 2014 sobre Transparencia y Acceso a la Información Pública Nacional, define los datos abiertos en el numeral sexto como "todos aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos". La Ley establece la obligatoriedad de las entidades públicas de "divulgar datos abiertos", teniendo en

cuenta las excepciones de acceso a la información, asociadas a información clasificada y reservada.

De acuerdo con el manual práctico para mejorar la calidad de los datos abiertos del Gobierno de España (2017) los datos abiertos deben cumplir con tres condiciones mínimas para ser clasificados como tal: accesibilidad, apertura y reutilización. El acceso debería poder realizarse a través de internet y de forma gratuita, para la apertura, los datos deberían contar con una licencia abierta o pertenecer a un dominio público y su reutilización es posible en formatos abiertos y legibles por máquinas. A continuación, se aplica el modelo MELODA, que permite medir el nivel de reutilización de los datos abiertos y los componentes de acceso, apertura, entre otros. Para seleccionar la fuente de datos abiertos.

3.4 Modelo de reutilización de datos abiertos: MELODA

Siguiendo Abella, Ortiz y De Pablos (2014) quienes proponen el modelo Meloda 3.0 para evaluar los datos abiertos en cuatro dimensiones: estándares técnicos, acceso, aspecto jurídico y modelo de datos. La dimensión de estándares técnicos evalúa que los datos estén almacenados en un formato no privativo. El acceso, evalúa que la información sea legible en formato automatizado, el aspecto legal, se mide en términos de las restricciones o barreras legales que presente el acceso y uso de los datos, considerando la obligatoriedad de la atribución a la fuente. Y el modelo de los datos, se mide en términos de la capacidad para compartir la estructura de datos.

En Abella, Ortiz y De Pablos (2017) se incorporan dos componentes en el modelo Meloda 4.0, información geolocalizada e información en tiempo real. La geolocalización hace referencia a evaluar si una fuente de datos contiene campos que ayuden a identificar la ubicación de los datos. Y el aspecto de tiempo real evalúa el período de actualización de la información.

En la Tabla 9 se aplica la metodología Meloda 4.0 para evaluar la calificación de datos abiertos de las fuentes de datos identificadas. Para una descripción de las escalas de puntuación y la descripción ver Abella, Ortiz y De Pablos (2017).

Tabla 9 Matriz MELODA para datos abiertos

Nombre	Jurídico	Acceso	Técnico	Modelo de datos	Geolocalización	Tiempo real	Calificación	Nivel de reutilización
Series históricas de la tasa de interés	100	90	60	100	50	40	73,3	Óptima
Encuesta de carga financiera y educación financiera de los hogares	100	50	60	90	50	15	60,8	Buena
Encuesta de calidad de vida	100	50	60	90	50	15	60,8	Buena
Estudio de la demanda	100	50	60	90	50	15	60,8	Buena
Saldos de cartera, Desembolsos y aprobaciones	100	50	60	90	15	15	55,0	Buena
BBVA API´S Pay Stats	10	100	100	50	50	15	54,2	Buena
Encuesta longitudinal colombiana	100	50	60	50	30	15	50,8	Buena
Encuesta sobre la situación del crédito en Colombia	100	50	60	50	15	15	48,3	Básica
Banco Santander Open Challenge en Kaggle	100	50	60	35	15	15	45,8	Básica
BBVA Data Open challenge en Kaggle	100	50	35	35	30	15	44,2	Básica
Encuesta social	10	50	60	50	50	15	39,2	Básica

Fuente: Elaboración propia

La serie histórica de la tasa de interés elaborada por el Banco de la República, según la calificación Meloda 4.0 tiene un nivel óptimo de reutilización, principalmente, se destaca en las categorías de acceso y tiempo real, ya que el sistema de datos SERANKUA permite hacer consultas individuales parametrizadas sobre los datos y descargarlos sin necesidad de descargar el histórico total de la base, y el tiempo de actualización es diario. Sin embargo, esta serie no aporta información sobre el acceso al crédito, razón por la cual se descarta.

Comparten el segundo lugar, con mayor puntuación y en un nivel de buena reutilización, las bases de datos abiertos de la Encuesta de carga financiera y educación financiera de los hogares *(Iefic*) elaborada por el DANE y el Banco de la República, la Encuesta de calidad de vida (*ECV*) elaborada por el DANE y el estudio de la demanda elaborada por la Superintendencia financiera de Colombia. En el siguiente apartado se discuten los criterios para la selección final de la fuente de datos.

3.5 Selección de la fuente de datos abiertos

El modelo Meloda 4.0 muestra un nivel de buena reutilización para los siguientes conjuntos de datos: la Encuesta de carga financiera y educación financiera de los hogares *(Iefic*) elaborada por el DANE y el Banco de la República, la Encuesta de calidad de vida (*ECV*) y los Datos de estudio de la demanda de la Superintendencia Financiera de Colombia.

Sobre las fuentes de datos sugeridas por Meloda 4.0 se aplican los siguientes criterios del proyecto que permite alcanzar el objetivo del estudio, caracterizar y predecir la demanda de

crédito, los datos seleccionados deben permitir identificar individualmente si un cliente obtuvo un crédito o no, ya que está es la variable objetivo, que se quiere predecir. Se debe contar con un conjunto variables que permitan la caracterización de la población y reflejar patrones de comportamiento financiero. Y la disponibilidad de los datos debe ser periódica ya que permite hacer comparaciones de las variables en el tiempo. Y un volumen de datos significativo es deseable para desarrollar modelos de predicción.

En la tabla 10 se realiza una matriz de manera que se puedan reflejar el cumplimiento de los criterios que permiten lograr el objetivo del presente estudio. Tiene tres ítems evaluados, cada uno con un punto si el data set contiene esa propiedad se califica con 1 sino con 0, así el puntaje total estará expresado en una escala de 0 a 3, la base de datos con el puntaje superior será la seleccionada.

Tabla 10 Matriz de selección de datos

Nombre	Datos individualizados el cliente tiene crédito o no.	variables financieras	Actualización y disponibilidad de los datos anual	Calificación
Encuesta de carga financiera y educación financiera de los hogares	1	1	1	3
Estudio de la demanda	1	1	0	2
Encuesta de calidad de vida	0	0	1	1

Fuente: Elaboración propia

De acuerdo con estos criterios se selecciona la *Encuesta anual de carga financiera y educación financiera de los hogares* que permite identificar la variable respuesta a predecir,

si el cliente tiene crédito o no, según el tipo de modalidad de crédito, además cuenta con un conjunto de variables sobre la situación patrimonial, nivel de endeudamiento, educación financiera y déficit de ingresos. Y se encuentra disponible anualmente desde el año 2010 hasta 2016 sólo para muestras poblacionales sobre Bogotá y para 2017 y 2018 se incluyeron Cali y Medellín.

Se descartan los microdatos de estudio de la demanda ya que se encuentran disponibles únicamente para los años 2015 y 2017, con un volumen de registros de 1.418 y 1.432 respectivamente. También, se descarta la encuesta de calidad de vida que, aunque tiene una periodicidad de aplicación y actualización anual, presenta pocas variables de componente financiero, fue diseñada para medir las condiciones de vida de los hogares e indicadores relacionados con pobreza.

3.6 Gestión de calidad de datos abiertos: ISO/IEC 25012

Para asegurar la calidad de los datos sobre el conjunto de datos abiertos seleccionado, la *Encuesta anual de carga financiera y educación financiera de los hogares,* el manual práctico para mejorar la calidad de los datos abiertos del Gobierno de España, (2017) recomienda que la utilización de los datos abiertos debe ir acompañada de un aseguramiento de la calidad de los datos, de no ser así, los datos abiertos presentarían barreras de uso por parte de terceros, y sugieren los criterios inherentes de la norma ISO / IEC 25012.

La norma ISO / IEC 25012 contiene quince características para definir un modelo de calidad de datos óptimo, estás características se clasifican en dos grupos: calidad de datos

inherentes y calidad de datos dependiente del sistema. El primer grupo, hacer referencia a las cualidades de los datos asociados con el dominio de datos y sus restricciones, la relación entre valores y los metadatos. El segundo grupo, define la calidad de datos preservada mediante los componentes del sistema informático del hardware y el software. (Calabrese, Esponda, Pasini, Boracchia y Pesado, 2019).

En la tabla 11 se explican las características inherentes asociadas a la calidad de datos de la norma ISO / IEC 25012 y se identifican dichos criterios sobre el conjunto de datos seleccionado la *Encuesta anual de carga financiera y educación financiera de los hogares.* Las características dependientes del sistema se dejan en el Anexo 1.

Tabla 11 Características inherentes del modelo de calidad de datos ISO/ IEC 25012

Grupo	Característica	Descripción	Método de validación	Documento asociado a la validación	Concepto sobre el conjunto de datos
Inherente	Exactitud	La información representa el valor deseado en un contexto específico.	Especificar reglas semánticas	Cuestionario permite estandarizar las posibles respuestas Diccionario de variables	La totalidad de los datos corresponden con las reglas definidas
Inherente	Completitud	La información que proveen los datos es suficiente para satisfacer la necesidad del usuario desde un punto de vista cuantitativo.	Indicar si el atributo es obligatorio La norma sugiere X>85%[2] son datos completos en su totalidad	Diccionario de variables	Total variables: 285 86 variables que representan el 30% del total se encuentran completas >80% 10 variables (3%) – 45%<=X<80% 28 variables (10%) – 10%<=X<45% 103 variables (36%) – 1%<=X<10%
Inherente	Consistencia	Información libre de contradicción	Uniformidad del significado del atributo en	Documentación de la Encuesta. El DANE	De acuerdo con la documentación, el cuestionario

[2] La norma ISO/ IEC 25012 sugiere que los atributos obligatorios deben representar por lo menos el 85% de completitud para ser catalogados como óptimos. En la aplicación práctica se considera el 80% sobre la valoración de las variables ya que estos campos no son obligatorios.

			el mismo contexto de uso, aplicación de convenciones generalizadas	verifica con el DMC (Dispositivo móvil de captura)	y la base de datos se evidencia generalidad de las convenciones
Inherente	Credibilidad	Los datos se consideran ciertos y creíbles	Origen de los datos y atribución	Documentación de la Encuesta	DANE-Dirección de metodología y producción estadística
Inherente	Actualidad	Periodicidad de actualización del dato	Analizar fecha de actualización de los datos	Fecha de última actualización de los datos	Actualización anual Fecha inicio:2010 Fecha última actualización: 2018

Fuente: ISO / IEC 25012. Elaboración propia

3.7 Descripción de la fuente de datos abiertos seleccionada

Para el presente análisis, se utilizará como insumo principal los microdatos de la *Encuesta anual de carga financiera y educación financiera de los hogares*, estos se encuentran publicados en el sitio web de datos abiertos del gobierno de Colombia, están desagregados a nivel hogar y dan información sobre variables financieras de los hogares, se encuentra disponible desde el año 2010 hasta 2016 para Bogotá con la primera metodología desarrollada, la cual contenía 285 variables. En la segunda metodología desarrollada para los años 2017 y 2018 se incorporaron las ciudades Cali, Medellín y Bogotá, con 331 variables.

Esta encuesta se realiza sobre una subpoblación de la *Gran encuesta integrada de hogares* (GEIH) se selecciona la población que tiene uno o varios productos financieros de algún miembro del hogar encuestado. Los datos son de tipo corte transversal, lo cual significa que se obtiene la información del mismo tipo, para diferentes poblaciones, en distintos cortes de tiempo. En la Tabla 12 se observa el número de observaciones para cada año de la encuesta.

Tabla 12 Cantidad de observaciones IEFIC

Año	Nº Observaciones por ciudad			Total observaciones año
	Bogotá	Cali	Medellín	
2010	3.955	NA	NA	3.955
2011	11.639	NA	NA	11.639
2012	17.953	NA	NA	17.953
2013	18.319	NA	NA	18.319
2014	18.041	NA	NA	18.041
2015	18.624	NA	NA	18.624
2016	18.147	NA	NA	18.147
2017	17.664	14.223	15.460	47.347
2018	22.573	19.939	20.326	62.838
Total	146.915	34.162	35.786	216.863

Fuente: Elaboración propia

Al observar las características inherentes de la norma ISO / IEC 25012 se resalta que el conjunto de datos cumple en su totalidad con criterios de exactitud, consistencia y credibilidad. Y respecto a los criterios del proyecto se resalta que los datos permiten identificar si el usuario tiene crédito o no, está compuesto por variables financieras y sociodemográficas que permiten reflejar patrones de comportamiento financiero, la actualización de los datos es anual, tiene disponibilidad histórica de ocho años desde 2010

hasta 2018 y el volumen de registros es 216.863, se considera suficientemente grande para realizar un modelo predictivo.

De acuerdo con los aspectos evaluados por el modelo Meloda 4.0, las características inherentes de la Norma ISO / IEC 25012 y los criterios del proyecto sobre el conjunto de datos, se evidencian las ventajas de utilizar la *Encuesta anual de carga financiera y educación financiera de los hogares* el resumen de todos los criterios se muestra en la Tabla 13.

Principalmente la fuente de datos es considerada de datos abiertos, ya que presenta un nivel de reutilización alto, los aspectos con mejor desempeño según Meloda 4.0 son: el jurídico ya que los datos se pueden utilizar sin restricciones, y el modelo de datos, el cual cumple estándares normativos internacionales de calidad.

Tabla 13 Criterios de selección de la fuente de datos

Meloda 4.0	**Características inherentes ISO /IEC 25012**	**Criterios del proyecto**
Aspecto jurídico: Puntaje: (100)- Superior: Reutilización sin restricciones, sólo atribución a la fuente original.	Exactitud: La totalidad de los datos corresponden con las reglas definidas en la metodología.	Datos individualizados: Permite identificar si el usuario tiene crédito o no.
Estándares técnicos: Puntaje: (60)- Medio. La fuente de datos se publica en estándares abiertos, pero como ficheros individuales csv y xls.	Completitud: 86 variables que representan el 30% del total 285, se encuentran completas >80%[3]	Variables financieras: Permite identificar patrones de comportamiento, contiene aspectos de tipo a. Sociodemográfico b. Endeudamiento c. Activos financieros d. Morosidad e. Solicitudes de préstamos, entre otras.
Acceso: Puntaje: (50) – Medio Acceso a la información a través de la web, pero permite acceder a cada uno de los conjuntos de datos de forma individual o a través de una URL única	Consistencia: De acuerdo con la documentación, el cuestionario y la base de datos se evidencia generalidad de las convenciones	Actualización y disponibilidad: Anual Desde 2010 hasta 2018.
Modelo de datos: Puntaje: (100)-Superior Existe un modelo de datos estandarizado por una entidad global con amplia adopción.	Credibilidad: El modelo de datos es diseñado por el DANE- Dirección de metodología y producción estadística	Volumen de datos: 216.863 registros individualizados.
Actualización: Puntaje: (15) – Baja El período de actualización es superior a una semana.	Actualización: Anual Fecha inicio:2010 Fecha última actualización: 2018	
Geolocalización Puntaje: (50)-Medio: La información geográfica son varios campos con texto descripción y son jerárquicos.		

[3] La norma ISO/ IEC 25012 sugiere que los atributos obligatorios deben representar por lo menos el 85% de completitud para ser catalogados como óptimos. En la aplicación práctica se considera el 80% sobre la valoración de las variables ya que estos campos no son obligatorios.

3.8 Descripción de las variables

En la metodología de la IEFIC (DANE, 2018) y (DANE, 2016) se definen diez secciones para clasificar las variables de la encuesta, ver Tabla 14. Para realizar la caracterización del usuario y definir la probabilidad de acceder al crédito, se propone agrupar las variables de manera similar a como lo haría una entidad financiera. Según el tipo de información que aporta la variable para reflejar comportamiento de crédito. En este sentido, se reordenan las agrupaciones y se proponen once categorías de agrupación.

Tabla 14 Descripción de variables de la IEFIC

Categoría DANE	**Descripción DANE**
1. Información de la vivienda	Características de identificación de la vivienda
2.Activos reales y deuda hipotecaria	El objetivo de esta sección es conocer si el hogar al momento de adquirir vivienda utilizó servicios financieros, tales como crédito hipotecario
3.Activos reales	Cuantifica los activos en electrodomesticos que posee el hogar
4.Consumo	Establecer los gastos básicos del hogar, y el ingreso destinado a ahorro
5.Activos y deuda hipotecaria	Nivel de endeudamiento, y que ha hecho para disminuirlo
6.Seguros y pensiones	Detectar el numero de hogares que cuentan con un seguro para cubrir la pérdida total o parcial de sus bienes y propiedades
7.Educación financiera	Indagar sobre la educación financiera asociada al crédito y funcionamiento del mercado financiero.
8.Deuda no hipotecaria	Establece cuantas tarjetas de crédito tienen las personas del hogar, créditos con casas comerciales, préstamos de libre inversión, crédito con prestamistas, tiendas de barrio, cajas de compensación, y cómo estan pagando los créditos

9. Activos financieros	Determinar si el hogar tiene inversiones financieras, tales como acciones, fondos mutuos de inversión, cuentas de ahorro, pensiones, certificados de depósito.
10. Percepción de carga financiera y restricciones al crédito	Indaga a cerca de la percepción de los hogares frente a su carga financiera y su actitud frente al endeudamiento.

Fuente: DANE (2018)

Con base en las observaciones identificadas en el Anexo 2.1, se proponen once categorías de agrupación, para ordenar las variables según el tipo de información que aportan y el aspecto que se quiere evaluar. Ver Anexo 2.2.

1. Características del individuo: recoge aspectos sociodemográficos.

2. Características del hogar: además de contener información sociodemográfica, contiene tenencia de vivienda y de subsidios, lo cual permite evaluar el nivel de riqueza y garantía o colateral, este es un requisito importante que tienen en cuenta los bancos para otorgar crédito.

3. Crédito formal: se compone de variables sobre los créditos adquiridos en el sector financiero.

4. Deuda informal: contiene variables sobre deuda adquirida en el sector informal.

5. Gastos: permite conocer los rubros y conceptos de los gastos del hogar.

6. Activos financieros: aportan información sobre los activos financieros que tiene el individuo, el monto de inversión y la renta percibida anual.

7. Conectividad: Se incluyen variables sobre si el hogar tiene teléfono e internet las cuales se infieren a partir de las variables de gasto en servicios de telefonía e internet.

8. Endeudamiento: son variables sobre percepción financiera del individuo respecto a su endeudamiento.

9. Mora: contiene una única variable, que indica la cantidad de veces que el individuo ha caído en mora en los últimos 12 meses.

10. Solicitudes de crédito formal y rechazos: las variables permiten identificar los aspectos de restricción al crédito.

11. Otras de poco interés: Se agrupan variables de identificación de la encuesta, conocimiento financiero, electrodomésticos y otras que su nivel de registros vacíos es cercano al 70% como el detalle de saldo, cuota, plazo de los diferentes tipos de crédito.

4 Procesamiento de la base de datos

A partir de los criterios de gestión de calidad que se revisaron para la base de datos, en la sección 3.6, se observaron: la exactitud, completitud, la consistencia, credibilidad y actualización de la base de datos. Esta identificación se utiliza y se desarrolla como parte del proceso de limpieza, procesamiento y transformación de la base de datos.

Las tareas de limpieza y procesamiento incluyen revisar: homogeneidad de variables entre las encuestas IEFIC 2010-2016 e IEFIC 2017-2018, el manejo de datos, nulos, vacíos y atípicos. Para identificar los datos vacíos, se utiliza el criterio de completitud de la gestión de calidad de datos y se consideran aquellas variables con más del 80% de registros. Para corregir los datos atípicos se utilizan diagramas de caja y la técnica del rango del valor esperado IQR. Para corregir los datos nulos se usan métricas de tipo estadístico como media, moda, mínimo y máximo para reemplazarlos. En 4.1 se explica en detalle cada paso.

En la fase de transformación de variables**:** Se utilizan métodos de reducción de dimensiones de bases de datos horizontal y vertical. La reducción horizontal, consiste en la discretización de valores continuos basados en indicadores de entropía como: el valor de la información y el peso de la evidencia. Para las variables continuas de precios, ingresos y gastos se construye un indicador para expresarlas en precios constantes de 2018. La reducción vertical, implica eliminar atributos redundantes para el problema. (Han y Kamber, 2001)

4.1 Limpieza y preprocesamiento de la base de datos

A continuación, se explican los mecanismos que se utilizaron para transformar la base de datos para el desarrollo de la fase de modelación:

a) Homogeneidad de variables en las encuestas IEFIC 2010-2016 e IEFIC 2017-2018

La muestra IEFIC 2010-2016 tiene 285 variables y la IEFIC 2017-2018 contiene 331 variables, las 44 variables de la IEFIC 2017-2018 adicionales están relacionadas con los medios de pago que utiliza el hogar, con variables como: realizar pagos móviles, transferencias entre cuentas, tarjetas prepago multiuso, entre otras. Se prescinde de estas variables para presentar homogeneidad en la cantidad de variables entre las muestras de datos.

Tabla 15 Número de variables de la IEFIC

Muestra de la encuesta	Cantidad de variables
IEFIC 2010-2016	285
IEFIC 2017-2018	331

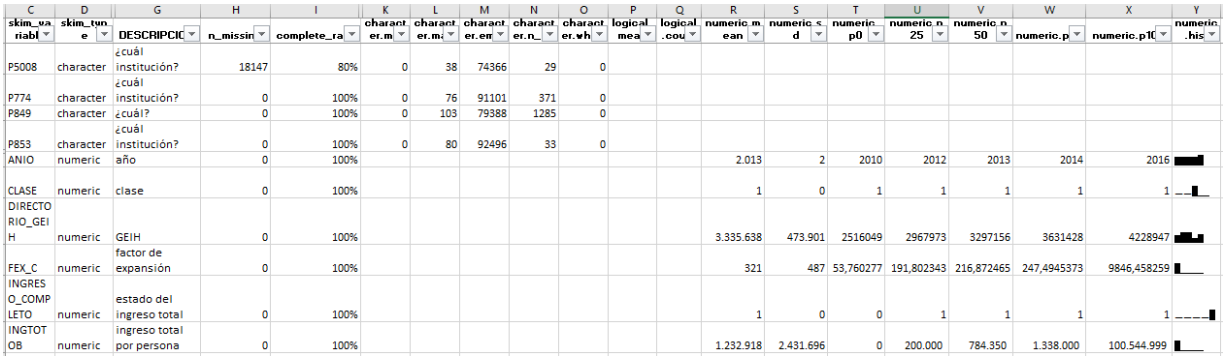
Fuente: DANE (2018).

b) Gestión de valores nulos o vacíos

Las variables con valores faltantes pueden causar sesgo en el modelo, utilizando el paquete “skimr” del software estadístico R sobre la base de datos, se detectan las variables con registros vacíos y se seleccionan las variables que tienen una tasa de completitud superior al 80%. Las variables que presentan una tasa de completitud mayor al 80% son 86

(30%) de 285 variables en total ver Tabla 16. Según sea necesario, para aquellas variables con registros vacíos se procede a aplicar técnicas de imputación, por la media, mediana o ceros.

Gráfica 9 Salida del paquete skimr - Software estadístico R Studio



C	D	G	H	I	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
skim_va riabl	skim_typ e	DESCRIPCIC	n_missin	complete_ra	charact er.m	charact er.ma	charact er.em	charact er.n_	charact er.wh	logical mea	logical .cou	numeric.m ean	numeric.s d	numeric. p0	numeric.p 25	numeric.p 50	numeric.p	numeric.p10	numeric .his
P5008	character	¿cuál institución?	18147	80%	0	38	74366	29	0										
P774	character	¿cuál institución?	0	100%	0	76	91101	371	0										
P849	character	¿cuál?	0	100%	0	103	79388	1285	0										
P853	character	¿cuál institución?	0	100%	0	80	92496	33	0										
ANIO	numeric	año	0	100%								2.013	2	2010	2012	2013	2014	2016	
CLASE	numeric	clase	0	100%								1	0	1	1	1	1	1	
DIRECTO RIO_GEI H	numeric	GEIH	0	100%								3.335.638	473.901	2516049	2967973	3297156	3631428	4228947	
FEX_C	numeric	factor de expansión	0	100%								321	487	53,760277	191,802343	216,872465	247,4945373	9846,458259	
INGRES O_COMP LETO	numeric	estado del ingreso total	0	100%								1	0	0	1	1	1	1	
INGTOT OB	numeric	ingreso total por persona	0	100%								1.232.918	2.431.696	0	200.000	784.350	1.338.000	100.544.999	

Fuente: Elaboración propia. Rstudio

Esta función indica el tipo de variable: numérica o categórica, el número de registros vacíos, la tasa de completitud y algunas estadísticas descriptivas básicas según el tipo de variable, la media, la desviación estándar, percentil 0, percentil 25, percentil 50, percentil 75, percentil 100. Y muestra un gráfico sobre la distribución de la variable.

Tabla 16 Tasa de completitud de las variables seleccionadas

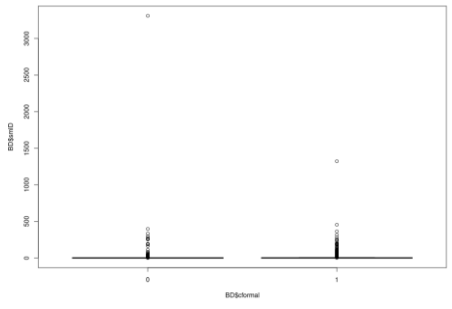
Tasa de completitud >= 80%	Cantidad de variables
80%	1
93%	1
96%	18
99%	44
100%	22
Total	86

Fuente: Elaboración propia. Datos obtenidos en Rstudio.

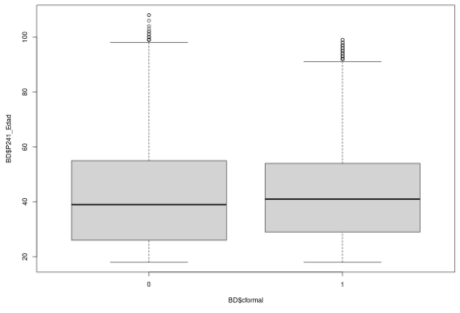
c) Gestión de datos atípicos

A partir de la descripción de valores, máximos, mínimos, percentiles y diagramas de caja, se ajustan según el rango IQR. En el gráfico 10 y 11 se representan los diagramas de caja y bigotes para evidenciar la desviación que tienen las variables de ingreso y edad respectivamente

Gráfica 10 Diagrama de caja y bigotes: Ingreso



Gráfica 11 Diagrama de caja y bigotes: Edad



4.2 Transformación de variables

Para la transformación de variables se utilizan métodos de reducción de dimensiones de tipo vertical y horizontal. A continuación, se explica la reducción vertical de eliminación de atributos redundantes para el problema. Y luego se explica la reducción horizontal que se realizó con dos técnicas: primero, se construye un indicador de precios constantes de 2018 para variables que se refieren a ingresos, gastos y precios. Y segundo, se discretizan variables continuas según los rangos sugeridos por los indicadores de entropía como: el valor de la información y el peso de la evidencia.

a) Reducción vertical de atributos

En la tabla 17 se observa la selección de 73 variables según las categorías de agrupación propuestas acorde con el tipo de información que aportan sobre el individuo.

Se seleccionan atributos que tienen la tasa de completitud mayor al 80% y presentan un buen indicador de ganancia de la información.

Tabla 17 Cantidad de variables seleccionadas

Categoría de variables	Total variables	Variables completas >=80%	Variables adicionales interés negocio <80%	Variables de interés	Aspectos
Características del individuo	7	7		5	Sociodemográficos
Características del hogar	8	3	3	6	Nivel de Riqueza
Crédito formal	87	9	2	11	Adquisición de créditos formales
Deuda informal	40	4		4	Adquisición de créditos informales
Gastos*	42	16	7	23	Nivel de gasto
Activos financieros*	23	9	3	12	Nivel de Riqueza y ahorros
Conectividad	2	2		2	Conectividad
Endeudamiento*	4	1	3	4	Nivel de endeudamiento
Mora	1	1		1	Morosidad
Solicitudes de crédito formal*	3	1	2	3	Restricción al crédito
Año	2	2		2	Tiempo
Subtotal	217	55	20	73	Variables interés
Otras de poco interés	68	31		0	
Total	285	86	13	73	

Fuente: Elaboración propia.

b) Agregación vertical de atributos

La técnica de agregación de atributos se aplica para resumir las variables: crédito formal ($cformal$), cantidad de créditos ($cant_c$), cantidad de deudas ($cant_{deuda}$) y gasto del mes. A continuación, se presentan las ecuaciones y definiciones utilizadas.

$cformal$ se construye a partir de $cant_c$.

$$cformal = 1; Si\ cant_c \geq 1$$
$$cformal = 0; Si\ cant_c < 1$$

Ecuación (5)

$cant_c$ es la agregación de las diferentes modalidades de crédito: crédito hipotecario, tarjeta de crédito, libre inversión, educativo y vehículo.

$$cant_c = hip + tdc + lib + edu + veh$$

Ecuación (6)

$Donde$,
hip : $crédito\ hipotecario$
tdc : $tarjeta\ de\ crédito$
lib : $crédito\ libre$
edu : $crédito\ educativo$
veh : $crédito\ vehículo$

$cant_{deuda}$ es la agregación de los diferentes tipos de deuda: deuda con una casa comercial, deuda gota a gota, deuda con la tienda de barrio, deuda con amigos y deuda con familia.

$$cant_{deuda} = dcomercial + dgota + dtiendabarrio + dfamilia + damigos$$

Ecuación (7)

$Donde,$
$dcomercial : deuda\ comercial$
$deudagota : deuda\ gota\ a\ gota$
$dtiendabarrio : deuda\ tienda\ de\ barrio$
$dfamilia : deuda\ familia$
$damigos : deuda\ amigos$

$Gastomes$ es la agregación de los diferentes tipos de gasto: gasto de educación, gasto de alimentación, gasto de internet, gasto de telefonía, gasto de transporte, gasto de manutención, gasto de vestuario, gasto de agua, gasto de luz, gasto de gas.

$$Gastomes = Gedu + Galim + Gint + Gtel + Gtrans + Gmanu + Gvest + Gagua + Gluz + Ggas$$

Ecuación (8)

$Donde,$
$Gedu : Gasto\ de\ educación$
$Galim : Gasto\ de\ alimentación$
$Gint : Gasto\ de\ internet$
$Gtel : Gasto\ de\ telefonía$
$Gtrans : Gasto\ de\ transporte$
$Gmanu : Gasto\ de\ manutención$
$Gvest : Gasto\ de\ vestuario$
$Gagua : Gasto\ de\ servicios\ público\ agua$
$Gluz : Gasto\ de\ servicios\ público\ luz$
$Ggas : Gasto\ de\ servicios\ público\ gas$

c) Transformación horizontal: precios constantes del año 2018

Se incluyen variables intermedias salario mínimo legal vigente y el índice de precios del consumidor (IPC) según el año correspondiente. Las variables de ingreso y gasto mensual se expresan en salario mínimo deflactado por el IPC a precios constantes del año 2018, para que su valor sea comparable en el tiempo. Las variables de precios de la vivienda, gasto telefonía

y gasto de internet se deflactan para expresarlas en precios constantes del año 2018. La ecuación para deflactar una serie de precios es:

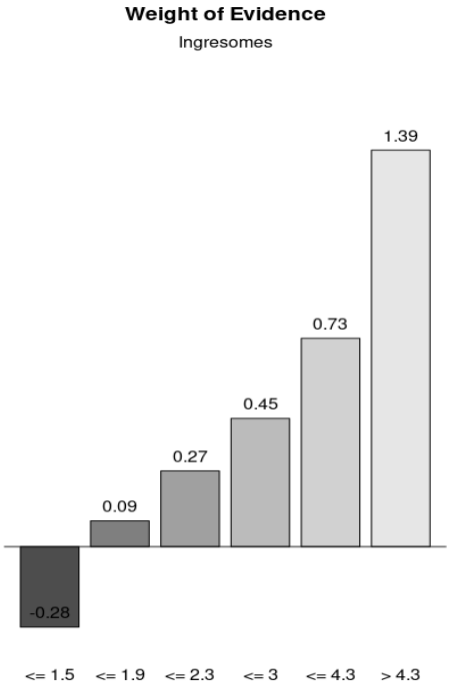
$$Índice\ deflactor\ de\ precios = \frac{IPC_{año\ x}}{IPC_{año\ 2018}}$$

Ecuación (9)

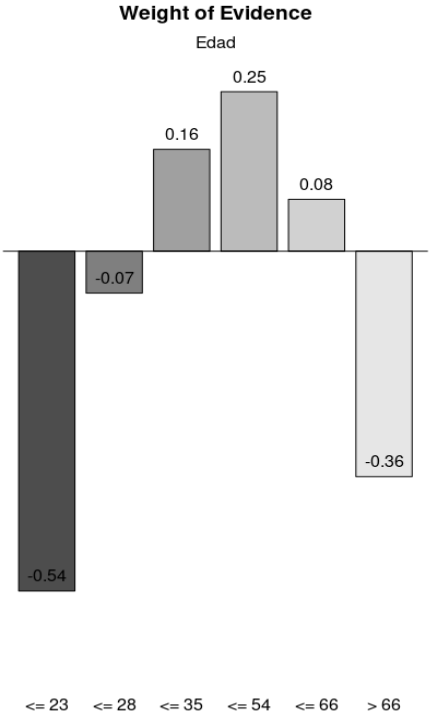
d) Reducción horizontal: Discretización de variables continuas según la entropía.

En el gráfico 12 y 13 se observan respectivamente las variables de ingreso y edad agrupadas en rangos definidos por los indicadores del valor de la información y el peso de la evidencia (WOE).

Gráfica 12 Indicador WOE: Ingreso



Gráfica 13 Indicador WOE: Edad

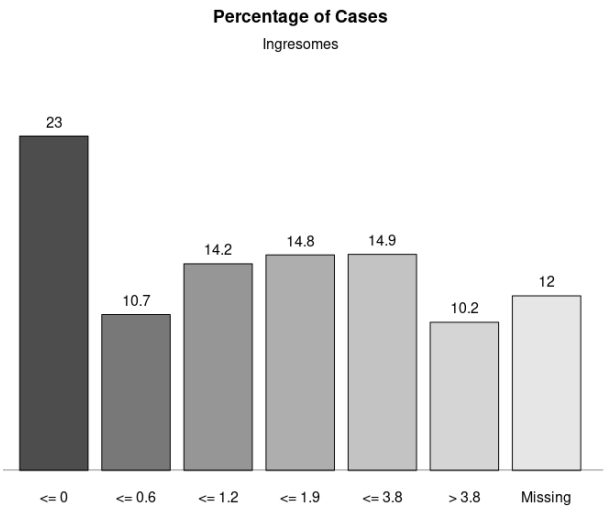


5 Desarrollo del análisis descriptivo

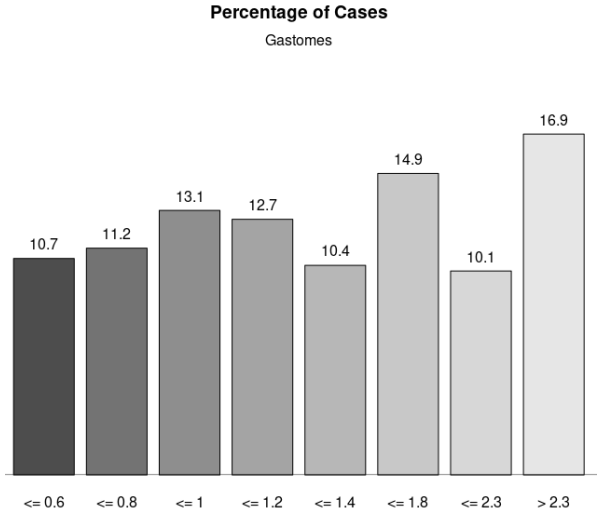
Para caracterizar las variables se desarrolla el análisis descriptivo según el tipo de variables, variables continuas o categóricas. Para las variables continuas, se observan las medidas de tendencia central y dispersión. Para las variables categóricas se observa la distribución de frecuencias y la forma de la distribución para observar en qué medida se separan o se aglomeran los datos.

A continuación, se realiza la descripción algunas variables representativas, en el Anexo 3: Análisis univariado se presentan las gráficas del análisis descriptivo univariado sobre el resto de las variables: características del individuo, características del hogar, crédito formal, gastos, activos financieros, conectividad, deuda informal, endeudamiento, mora y solicitudes.

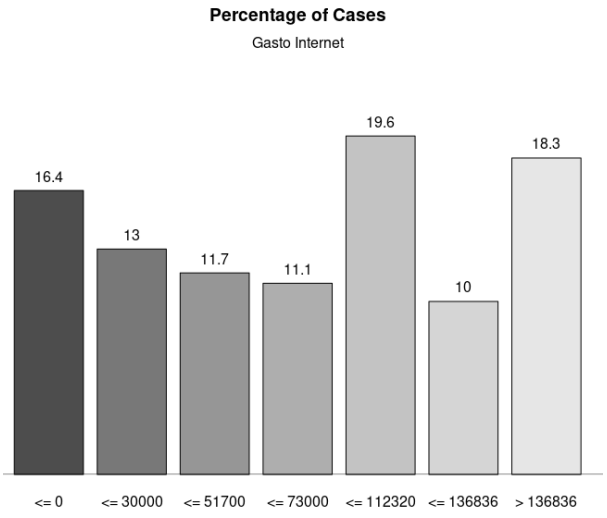
Gráfica 14 Distribución del Ingreso (smmlv)



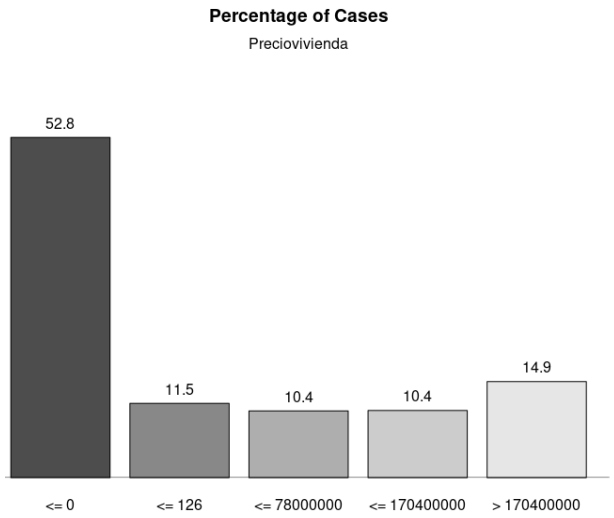
Gráfica 15 Distribución del Gasto (smmlv)



Gráfica 16 Distribución del gasto de internet



Gráfica 17 Distribución del precio de la vivienda



En el gráfico 14 se observa que el 62,7% de la población gana menos de 2 (smmlv) salarios mínimos legales vigentes y en el gráfico 15 se evidencia que el 73% de la población gasta menos de 2 smmlv. El 42,4 % del gasto en internet esta entre 51.700 pesos colombianos y 112.320 pesos. En el gráfico 17 se observa que el 35,7% de la población menciona que el precio de su vivienda es superior a 78 millones de pesos colombianos.

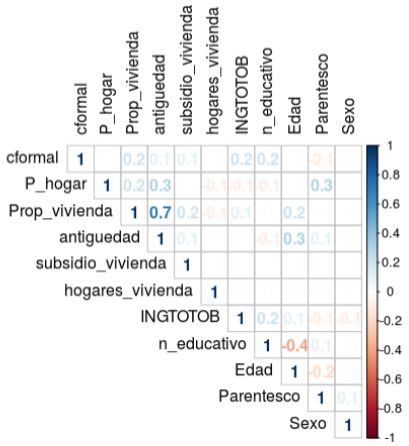
5.1 Análisis de correlación entre variables

Como parta del análisis descriptivo entre variables se incluye el análisis de correlación como se explica en Gujarati y Porter (2010) el objetivo principal es medir el grado de asociación lineal entre dos variables, para evaluar si existe una relación lineal entre las variables explicativas y la variable dependiente.

En la Tabla 18 se calculan las matrices de correlación para la muestra de datos IEFIC 2010-2016 y en la Tabla 19 respectivamente para la IEFIC 2017-2018. Las gráficas de las matrices se pueden ver en el Anexo 4. Matrices de correlación

Tabla 18 Variables significativas según matrices de correlación IEFIC 2010-2016

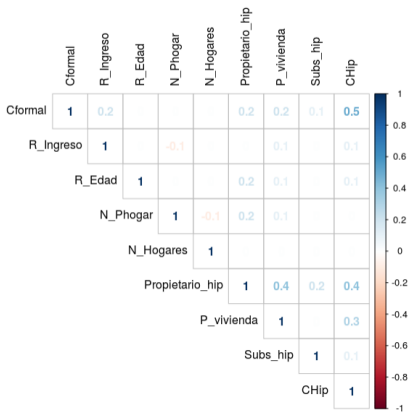
Variable	Índice de correlación
Ingreso	0,2
Nivel educativo	0,2
Propietario vivienda	0,2
Cuenta de ahorros	0,2
Computador	0,2
Televisor	0,2
Mora12m	0,2



Fuente: Elaboración propia.

Tabla 19 Variables significativas según matrices de correlación IEFIC 2017-2018

Variable	Índice de correlación
Gasto teléfono	0,3
Computador	0,3
Ingreso	0,2
Propietario vivienda	0,2
Precio vivienda	0,2
Cuenta de ahorros	0,2



Fuente: Elaboración propia

6 Desarrollo del análisis predictivo

En este documento se quiere caracterizar y predecir la demanda de crédito utilizando microdatos a nivel de hogar de las dos tomas de la Encuesta de Carga Financiera y Educación Financiera de los Hogares IEFIC 2010-2016 y 2017-2018. Por tanto, se realizarán modelos independientes para cada conjunto de datos.

Para la selección de los modelos de clasificación supervisada se deben satisfacer las siguientes condiciones:

- Capacidad de predicción
- Alto desempeño en la precisión
- Simplicidad en la interpretación
- Generalización de las reglas de decisión

Una técnica común en el desarrollo de análisis de crédito es medir la capacidad de predicción de las variables seleccionadas, utilizando los indicadores del valor de la información (IV) y el peso de la evidencia (WOE). Cada atributo o variable explicativa continua se agrupa mediante “bines” y se estima la capacidad de que ocurra el evento en cada “bin”. La agrupación óptima se caracteriza por presentar los puntos de corte donde se obtiene mayor discriminación o capacidad explicativa de que el evento objetivo ocurra. En la clasificación supervisada se optimiza el punto de corte de las variables para que ocurra el evento objetivo.

Validación cruzada

En los modelos predictivos de clasificación supervisada la precisión de la predicción se puede validar particionando el conjunto de datos, en datos de entrenamiento y datos de prueba, en una división aleatoria, en donde los conjuntos son mutuamente excluyentes. Típicamente se establecen en (2/3) 70% de los datos para entrenamiento y (1/3) 30% para prueba, este método es recomendable si el conjunto de entrenamiento queda con más de 1.000 observaciones, sino hay que utilizar algún tipo de submuestreo para balancear los conjuntos de datos. Con el conjunto de datos de entrenamiento se entrena el modelo y el modelo resultante se utiliza para predecir la clasificación en los datos de prueba. (García, Gámez y Alfaro, 2018)

Entrenamiento y prueba

IEFIC 2010-2016

```
> prop.table(table(train$cformal))

        0         1
0.4558318 0.5441682
> prop.table(table(test$cformal))

       0        1
0.457361 0.542639
```

IEFIC 2017-2018

```
        0         1
0.5632073 0.4367927
> prop.table(table(train$Cformal))

      0       1
0.56432 0.43568
> prop.table(table(test$Cformal))
```

Datos de entrenamiento: 70% del conjunto de datos

Datos de prueba: 30% del conjunto de datos

Para cada uno de los conjuntos de datos de entrenamiento y prueba, se mantiene la proporción de la variable objetivo, lo cual evidencia que el modelo está balanceado.

6.1 Desarrollo del modelo: árbol de decisión

Para la modelación del árbol se deben definir los siguientes parámetros con el fin de controlar el sobreajuste

- Definir el tamaño de cada nodo:
- Definir el mínimo de observaciones, para que un nodo sea considerado como una rama. Esto se define con la validación cruzada.
- Mínimo de observaciones para un nodo terminal: Valores bajos manifiestan clases desbalanceadas.
- Profundidad del árbol (Longitud vertical): Permite aprender relaciones específicas, se ajusta con validación cruzada.
- Máximo número de nodos hoja: Profundidad n= máximo 2^n hojas
- Máximo número de atributos para la ramificación: Selección aleatoria, como regla general se debería probar de un 30%-40% del total de atributos.

El diseño del árbol de decisión se realizó utilizando el software estadístico R Studio Cloud con el algoritmo CART implementado en el paquete estadístico RPART, este algoritmo identifica las variables independientes que presentan mayor discriminación para separar las clases de la variable objetivo.

Como se mencionó en la sección 2 los árboles de decisión presentan ventajas en la priorización de variables mediante el criterio de ganancia de información, no es necesario

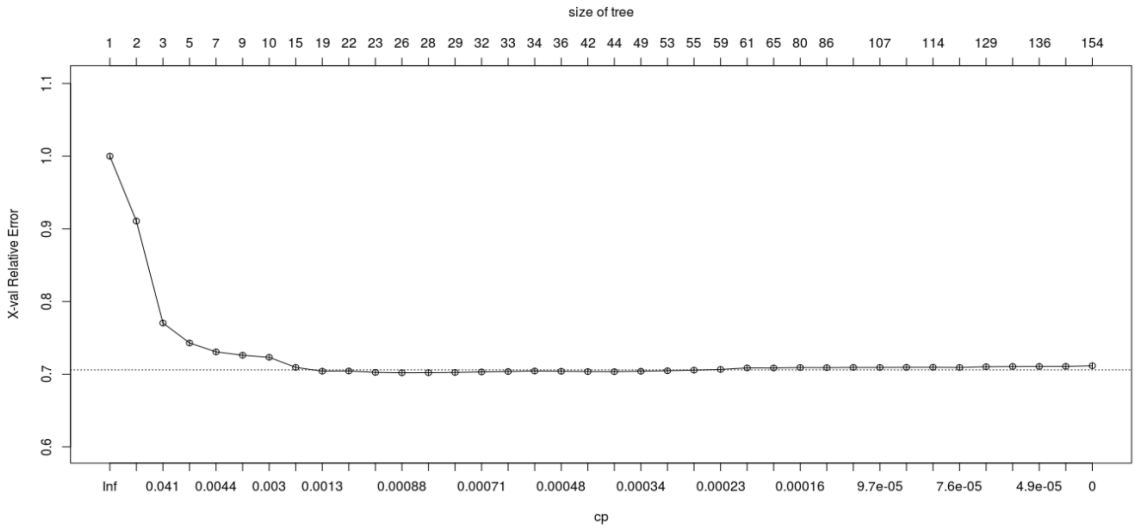
asumir supuestos de distribución normal para el desarrollo del modelo, admite diferentes tipos de variables y admite clases desbalanceadas.

Las desventajas que presenta este modelo son: el sobre ajuste, es decir que las reglas de decisión replica las características de entrenamiento, esto se debe controlar con los parámetros de tamaño del modelo y sus criterios de parada. En la Gráfica 9 se presenta el árbol de decisión del conjunto de datos de entrenamiento 70% de la IEFIC 2010-2016, sin determinar ningún hiper parámetro.

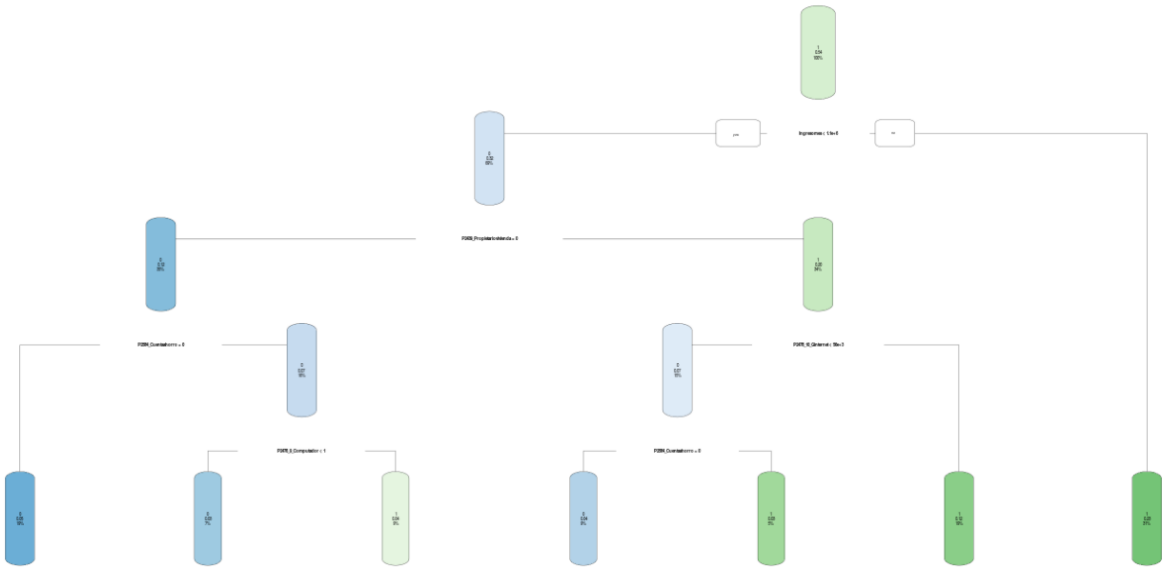
Definición de hiper parámetros para el desarrollo del árbol de decisión

Para encontrar el balance entre la profundidad, tamaño vertical dado por las relaciones de las variables, y la complejidad del árbol, el parámetro de costo de complejidad (CP) se estima cercano a cero y se crea el árbol con la mayor extensión computacional posible, esto permite determinar los hiper parámetros de control de poda del árbol para encontrar el tamaño óptimo del árbol.

Gráfica 18 Costo de complejidad del árbol de decisión IEFIC 2010-2016



Gráfica 19 Árbol de decisión IEFIC 2010-2016



Fuente: Elaboración propia.

Las variables explicativas con mayor significancia para el árbol de decisión son:

Tabla 20 Nodos del árbol de decisión IEFIC 2010-2016

Nodo	Variable	Índice de correlación	Punto de corte	Regla de decisión	% Tasa objetivo	% Población
1	Ingreso	0,3	<1140697	Si	20%	30%
2	Propietario vivienda	0,2	>=0,5	Si	13%	18%
3	Gasto internet	0,2	>=56400	Si	4%	7%
4	Cuenta ahorros	0,3	>=0,5	Si	20%	30%
5	Computador	0,2	>=0,5	No	5%	13%

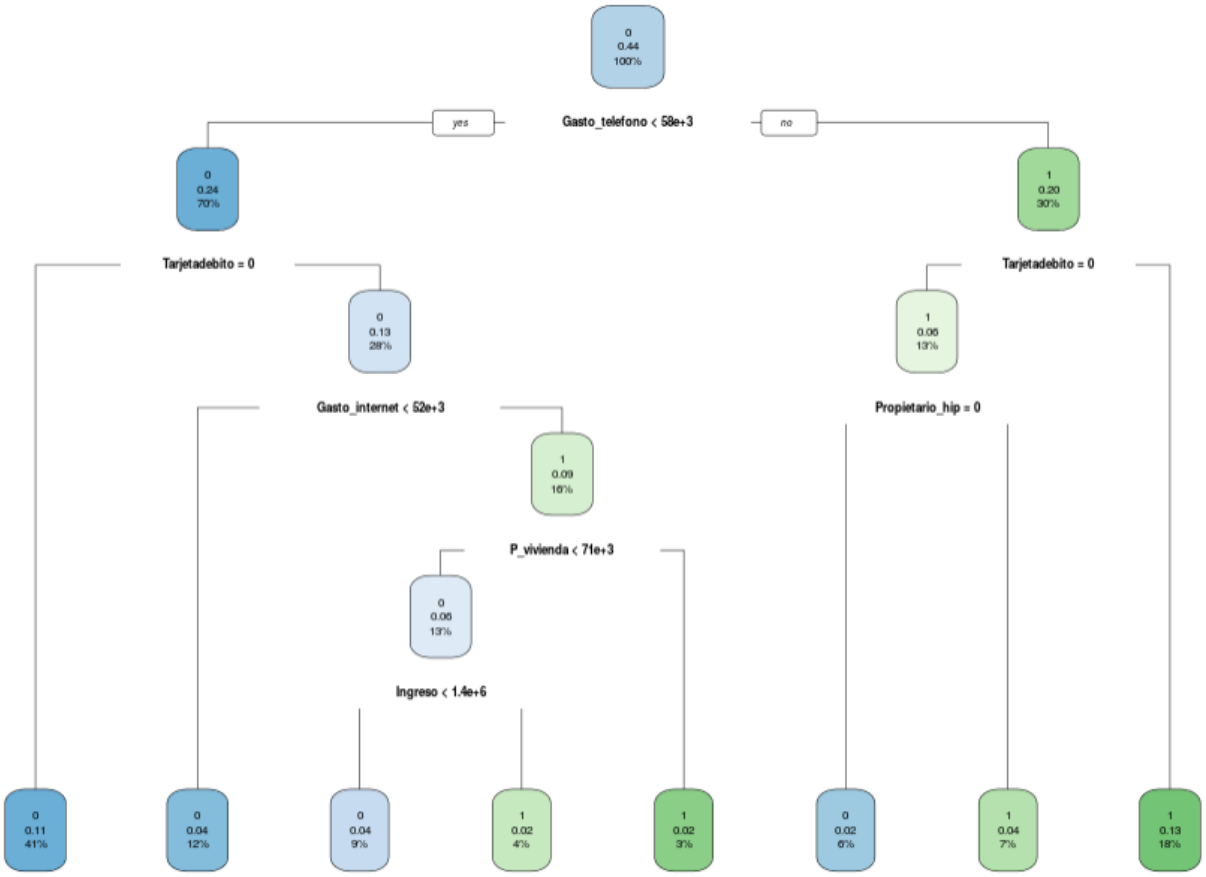
Matriz de confusión: Entrenamiento

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 15336 14206
         1  7648 27619

               Accuracy : 0.6628
                 95% CI : (0.6591, 0.6664)
    No Information Rate : 0.6454
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3078

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6672
            Specificity : 0.6603
         Pos Pred Value : 0.5191
         Neg Pred Value : 0.7831
             Prevalence : 0.3546
         Detection Rate : 0.2366
   Detection Prevalence : 0.4558
      Balanced Accuracy : 0.6638

       'Positive' Class : 0
```

Matriz de confusión: Prueba

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0  6631  6069
         1  3424 11644

               Accuracy : 0.6581
                 95% CI : (0.6525, 0.6637)
    No Information Rate : 0.6379
    P-Value [Acc > NIR] : 9.609e-13

                  Kappa : 0.2998

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6595
            Specificity : 0.6574
         Pos Pred Value : 0.5221
         Neg Pred Value : 0.7728
             Prevalence : 0.3621
         Detection Rate : 0.2388
   Detection Prevalence : 0.4574
      Balanced Accuracy : 0.6584

       'Positive' Class : 0
```

Gráfica 20 Árbol de decisión IEFIC 2017-2018



Las variables explicativas con mayor significancia para el árbol de decisión son:

Tabla 21 Nodos del árbol de decisión IEFIC 2017-2018

Nodo	Variable	Índice de correlación	Punto de corte	Regla de decisión	% Tasa objetivo	% Población
1	Gasto teléfono	0,3	>58.000	Si	20%	30%
1	Tarjetadebito	0,2	>=0,5	Si	13%	18%
2	Gasto teléfono	0,3	>58.000	Si	20%	30%
2	Tarjetadebito	0,2	>=0,5	No	5%	13%
2	Propietario Hip	0,2	>=0,5	Si	4%	7%
3	Gasto teléfono	0,3	>58.000	Si	20%	30%
3	Tarjetadebito	0,2	>=0,5	No	5%	13%
3	Propietario Hip	0,2	>=0,5	No	2%	6%
3	Gasto teléfono	0,3	>58.000	No	24%	70%
3	Tarjetadebito	0,2	>=0,5	No	5%	13%
3	Propietario Hip	0,2	>=0,5	No	2%	6%

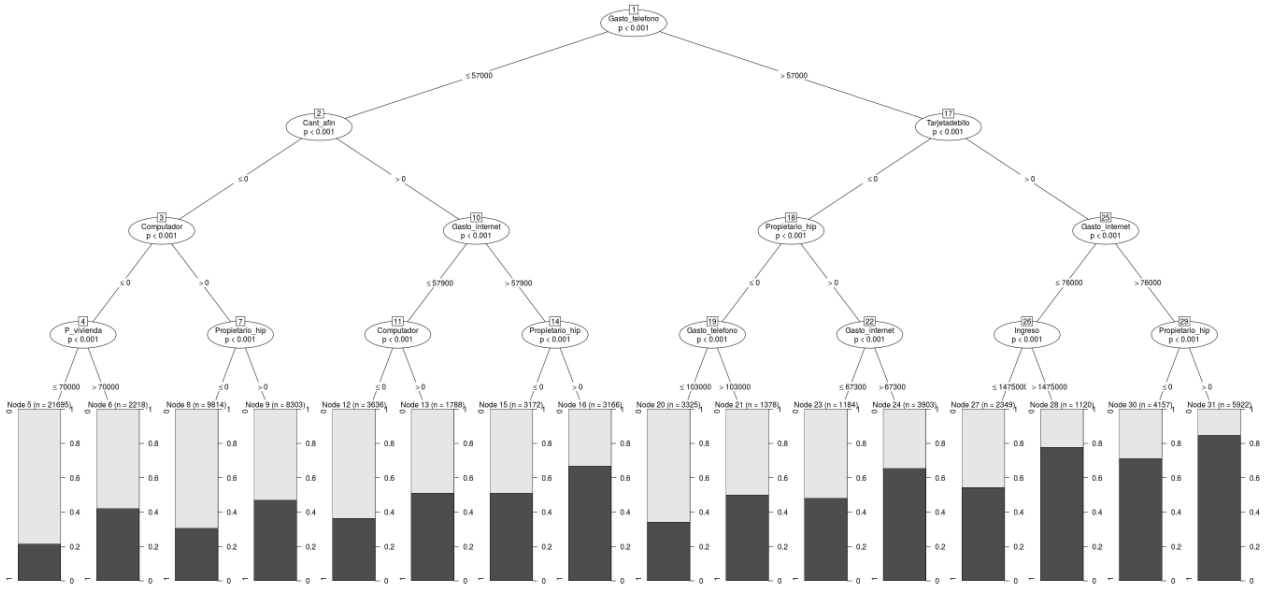
Fuente: Elaboración propia

Se observa que después de 8 nodos terminales, vemos rendimientos decrecientes en la reducción de errores a medida que el árbol se hace más profundo. Por lo tanto, podemos podar significativamente nuestro árbol y aun así lograr un error mínimo esperado.

Matriz de confusión: Prueba IEFIC 2017-2018

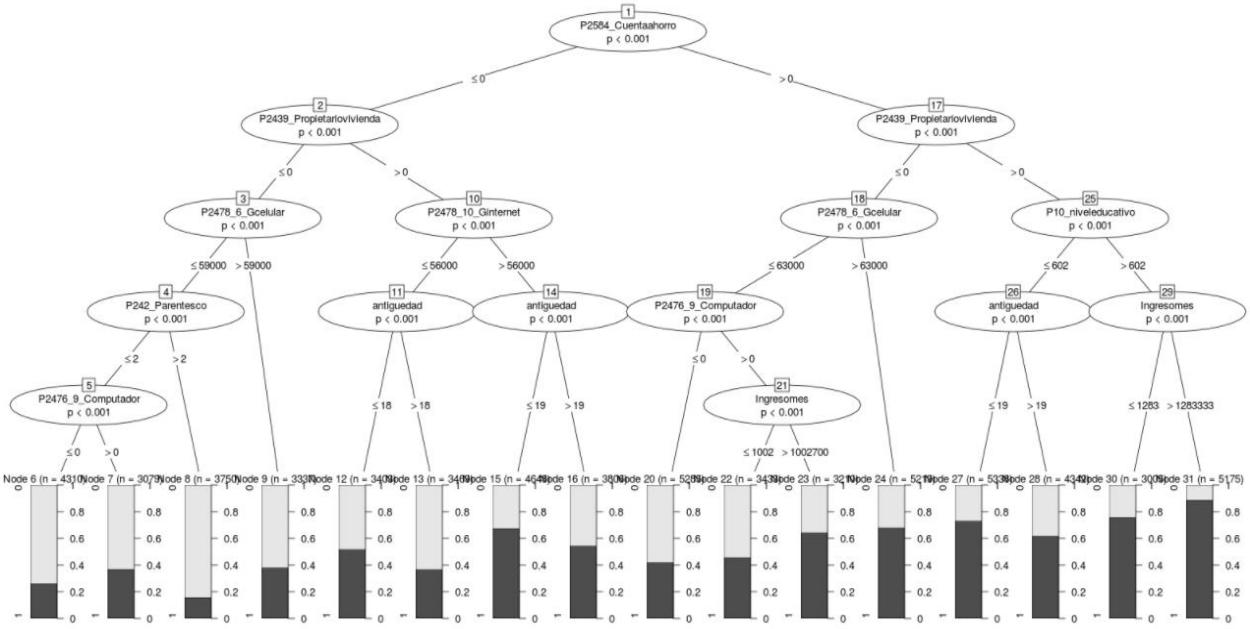
```
> confusionMatrix (test$Cformal, prediccion_1)
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 15461  3062
         1  7166  7350

               Accuracy : 0.6904
                 95% CI : (0.6854, 0.6954)
    No Information Rate : 0.6849
    P-Value [Acc > NIR] : 0.01479

                  Kappa : 0.3518

 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.6833
            Specificity : 0.7059
         Pos Pred Value : 0.8347
         Neg Pred Value : 0.5063
             Prevalence : 0.6849
         Detection Rate : 0.4680
   Detection Prevalence : 0.5606
      Balanced Accuracy : 0.6946

       'Positive' Class : 0
```

Gráfica 21 Árbol condicional IEFIC 2010-2016



Fuente: Elaboración propia.

Gráfica 22 Árbol condicional IEIFIC 2017-2018



Fuente: Elaboración propia.

Matriz de confusión: Entrenamiento

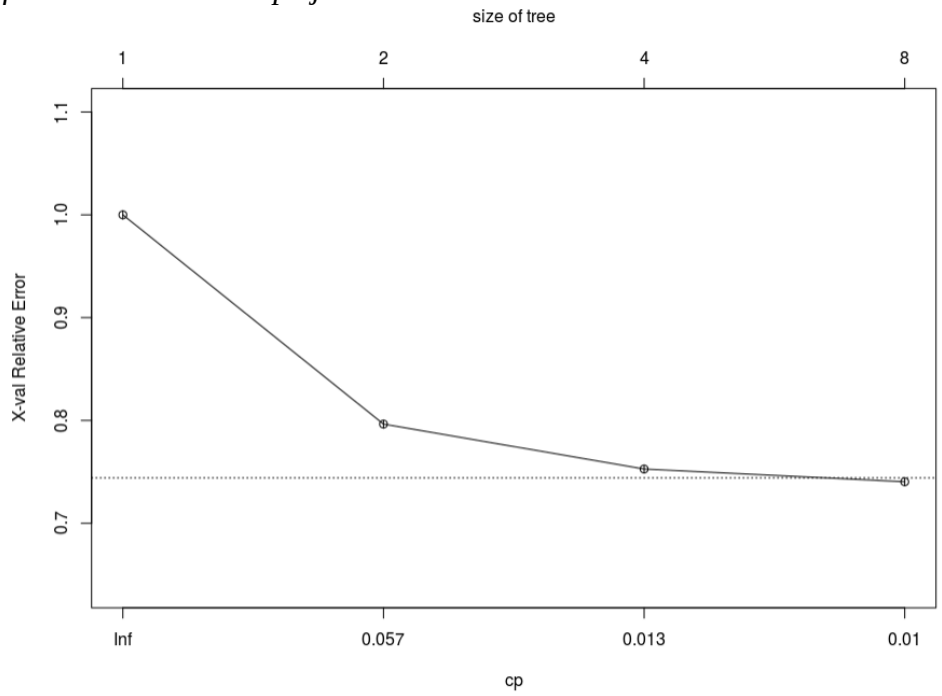
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 17475 12067
         1  9188 26079

               Accuracy : 0.672
                 95% CI : (0.6684, 0.6757)
    No Information Rate : 0.5886
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3336

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6554
            Specificity : 0.6837
         Pos Pred Value : 0.5915
         Neg Pred Value : 0.7395
             Prevalence : 0.4114
         Detection Rate : 0.2696
   Detection Prevalence : 0.4558
      Balanced Accuracy : 0.6695

       'Positive' Class : 0
```

Matriz de confusión: Prueba

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0  7550  5150
         1  4096 10972

               Accuracy : 0.667
                 95% CI : (0.6614, 0.6726)
    No Information Rate : 0.5806
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3248

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6483
            Specificity : 0.6806
         Pos Pred Value : 0.5945
         Neg Pred Value : 0.7282
             Prevalence : 0.4194
         Detection Rate : 0.2719
   Detection Prevalence : 0.4574
      Balanced Accuracy : 0.6644

       'Positive' Class : 0
```

6.1.1 Métricas de valoración del modelo

Gráfica 23 Costo de complejidad árbol de decisión condicional IEFIC 2017-2018



Fuente: Elaboración propia

Se observa que después de 8 nodos terminales, vemos rendimientos decrecientes en la reducción de errores a medida que el árbol se hace más profundo. Por lo tanto, podemos podar significativamente nuestro árbol y aun así lograr un error mínimo esperado.

6.2 Desarrollo del modelo: Bosque aleatorio

6.2.1 Cumplimiento de supuestos teóricos

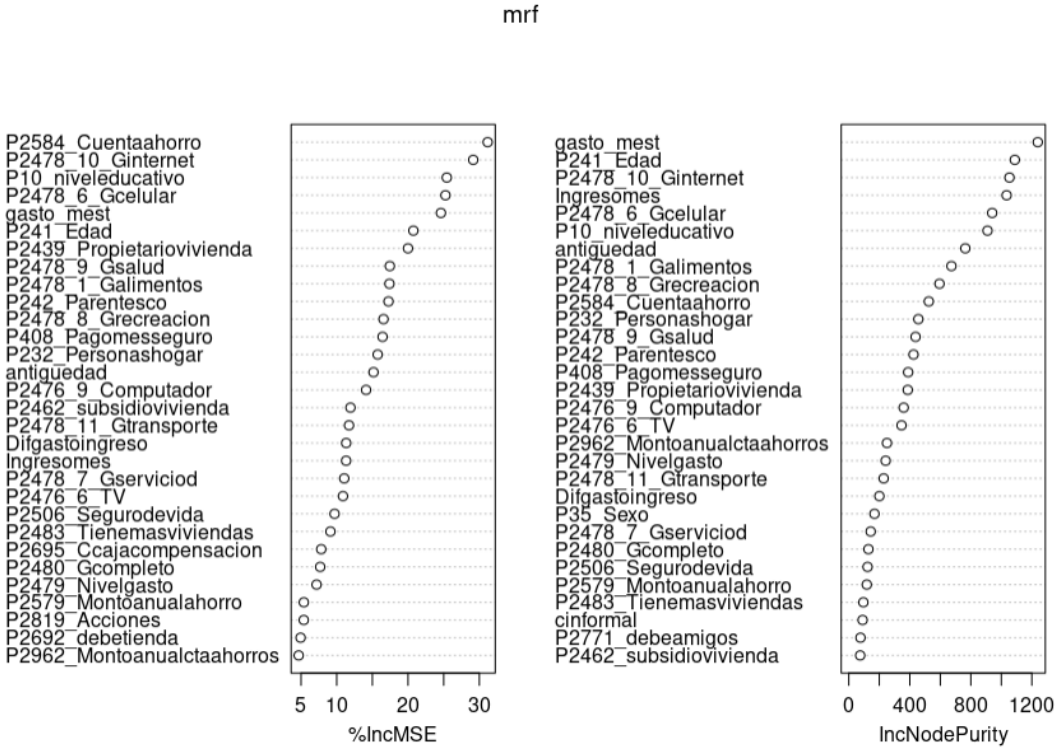
En el análisis empírico de Tang, Cai, & Ouyang, (2019) se explican algunas consideraciones relevantes para detectar las variables más importantes, el modelo de bosque aleatorio proporciona dos métodos, la precisión de la disminución media y la disminución del Gini promedio.

Además, se deben definir dos parámetros importantes que afectan la precisión del modelo: el número de árboles en el bosque aleatorio (ntree) y el número de variables explicativas seleccionadas al azar. Si el número de árboles es bajo, la clasificación puede ser deficiente, y un alto número de árboles puede ser ineficiente, hay un momento en el cual aumentar el tamaño del bosque ya no mejora la predicción del modelo.

Bernard et al (2007) mediante las pruebas empíricas realizadas logró establecer de manera general que, si el bosque es eficiente y supera un tamaño de 100 árboles, los cambios en la tasa de precisión son imperceptibles.

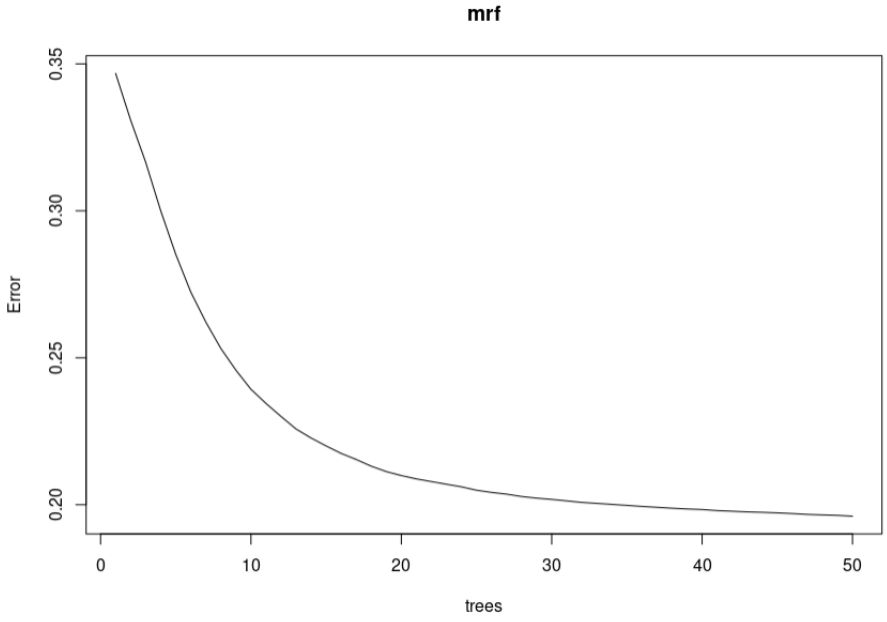
Para definir el número de variables explicativas y evaluar arboles independientes, teóricamente un número pequeño debería permitir mayores diferencias entre los árboles y mejores resultados de clasificación.

Gráfica 24 Bosque aleatorio IEFIC 2010-2016



Fuente: Elaboración propia

Gráfica 25 Costo de complejidad del Bosque Aleatorio IEFIC 2010-2016



Fuente: Elaboración propia

Matriz de confusión: Entrenamiento

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2
         1 29531    11
         2    85 35182

               Accuracy : 0.9985
                 95% CI : (0.9982, 0.9988)
    No Information Rate : 0.543
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.997

 Mcnemar's Test P-Value : 9.297e-14

            Sensitivity : 0.9971
            Specificity : 0.9997
         Pos Pred Value : 0.9996
         Neg Pred Value : 0.9976
             Prevalence : 0.4570
         Detection Rate : 0.4557
   Detection Prevalence : 0.4558
      Balanced Accuracy : 0.9984

       'Positive' Class : 1
```

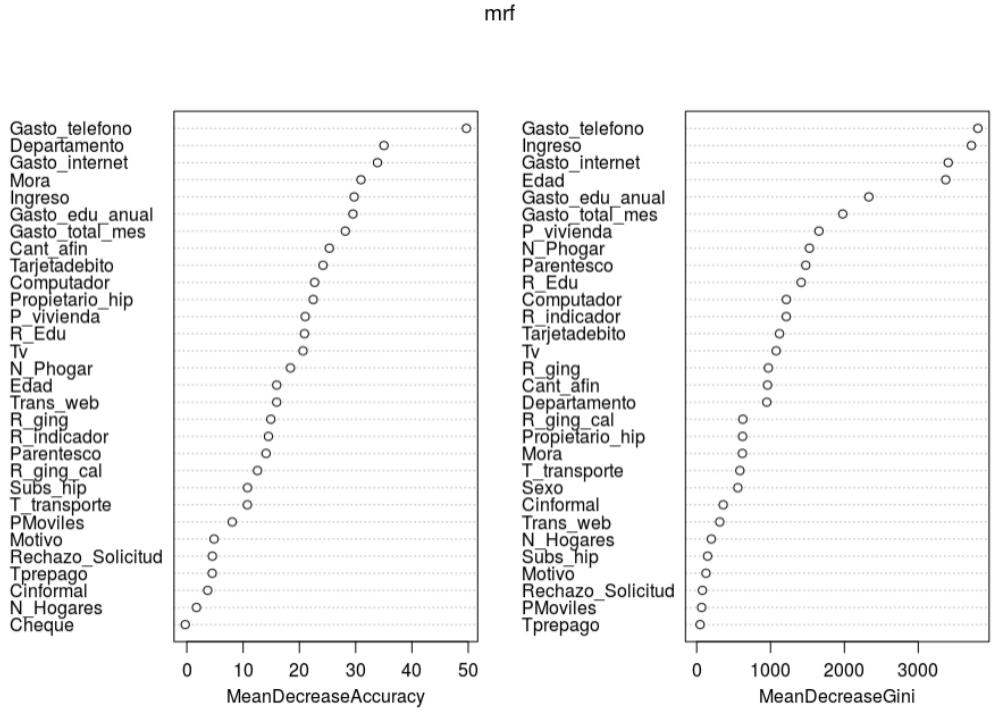
Matriz de confusión: Prueba

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2
         1  8454  4246
         2  3919 11149

               Accuracy : 0.706
                 95% CI : (0.7006, 0.7113)
    No Information Rate : 0.5544
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4064

 Mcnemar's Test P-Value : 0.0003088

            Sensitivity : 0.6833
            Specificity : 0.7242
         Pos Pred Value : 0.6657
         Neg Pred Value : 0.7399
             Prevalence : 0.4456
         Detection Rate : 0.3045
   Detection Prevalence : 0.4574
      Balanced Accuracy : 0.7037

       'Positive' Class : 1
```

Matriz de confusión: Prueba 2017-2018

```
> confusionMatrix (test$Cformal, preds)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  496  257
         1  645 2017

               Accuracy : 0.7359
                 95% CI : (0.7207, 0.7506)
    No Information Rate : 0.6659
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3515

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4347
            Specificity : 0.8870
         Pos Pred Value : 0.6587
         Neg Pred Value : 0.7577
             Prevalence : 0.3341
         Detection Rate : 0.1452
   Detection Prevalence : 0.2205
      Balanced Accuracy : 0.6608

       'Positive' Class : 0
```

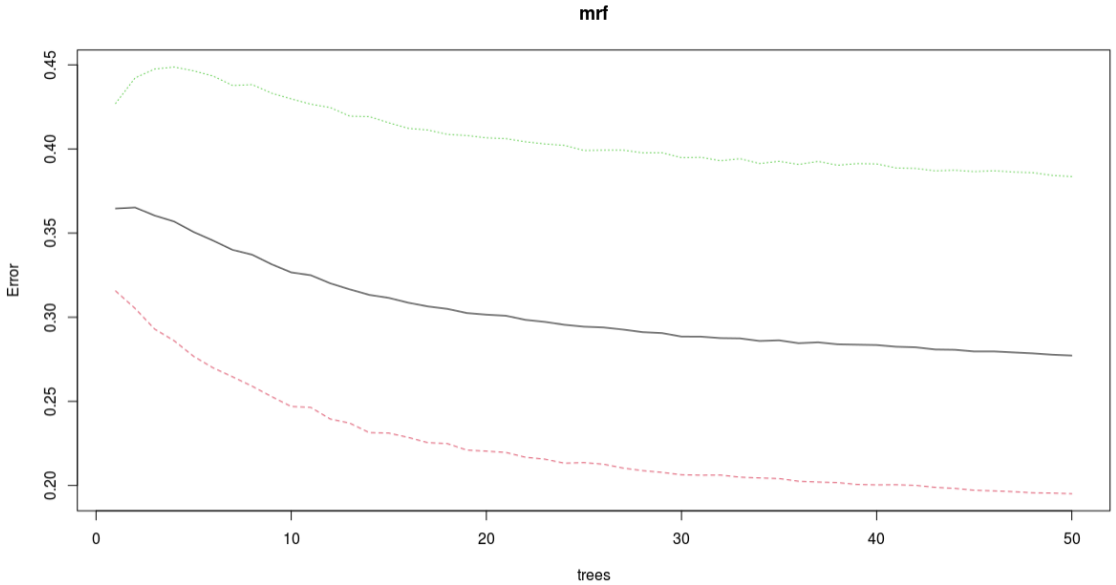
Gráfica 26 Bosque aleatorio IEFIC 2017-2018



Mrt=50

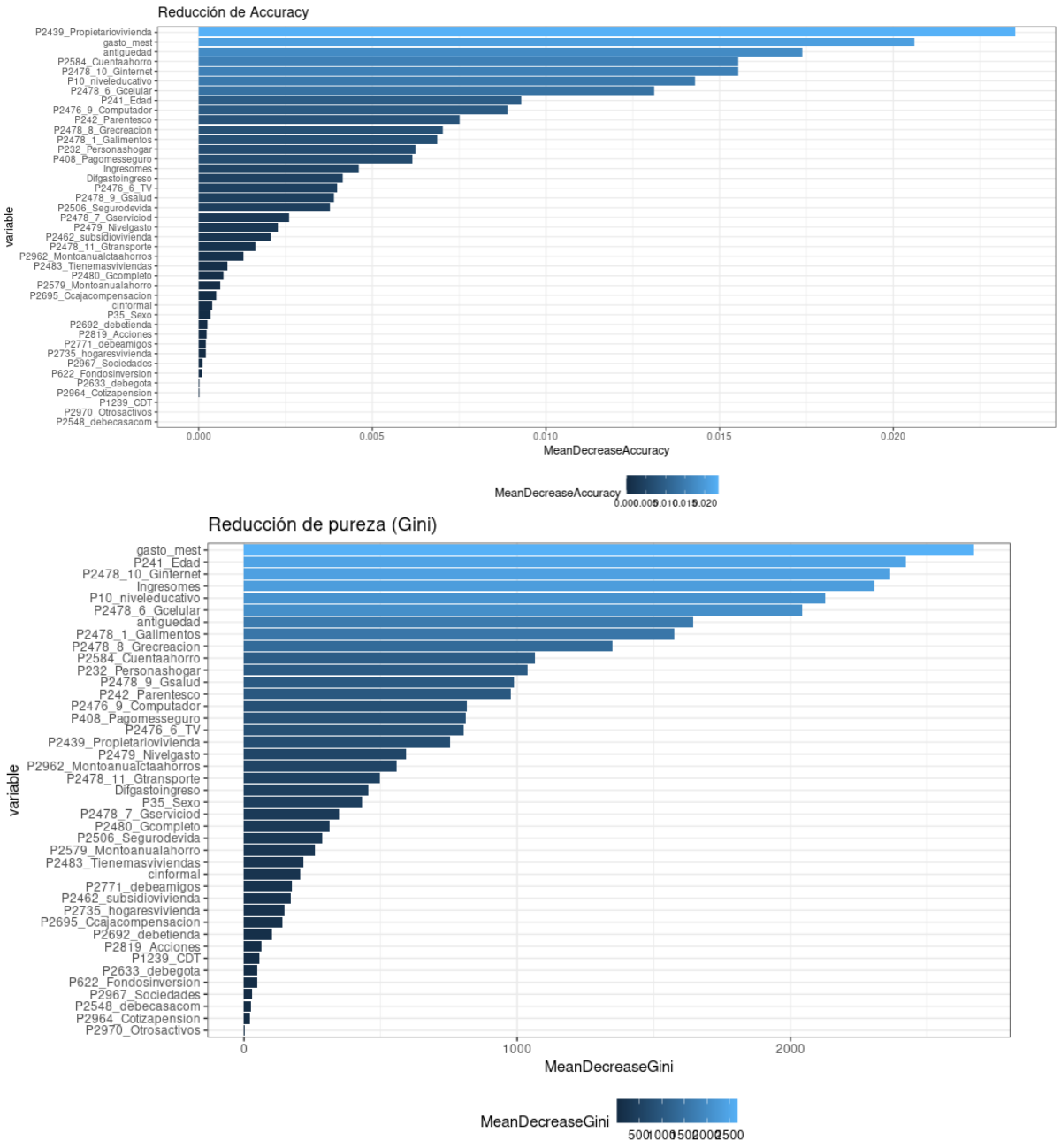
Fuente: Elaboración propia

Gráfica 27 Costo de complejidad del Bosque aleatorio IEFIC 2017-2018



Fuente: Elaboración propia

Gráfica 28 Indicador de precisión y GINI: Bosque aleatorio IEFIC 2010-2016



Fuente: Elaboración propia

6.3 Desarrollo del modelo: Regresión logística

6.3.1 Cumplimiento de supuestos teóricos

La función logística se representa con la estructura

$$p_i = \frac{1}{[1 + e^{(-\beta_0 - \beta_1 X_1 - \beta_2 X_2 + \cdots)}]}$$ Ecuación (10)

X_1 y X_2 son las variables explicativas, p_i representa la probabilidad de que la variable respuesta sea 1. (Pérez J. , 2017). En este caso 1 significa acceder al crédito y $1 - p_i$ representa la probabilidad de no acceder al crédito. Los aspectos positivos de este modelo son: es de fácil interpretación, requiere bajo esfuerzo computacional, admite múltiples variables explicativas, identifica el nivel de significancia o peso relativo de cada variable explicativa. Este modelo sólo es útil cuando la variable dependiente es dicotómica y su relación con las variables explicativas es lineal. (Moral, 2006)

Algunos supuestos que se deben tener en cuenta son:

- Variable respuesta dicotómica
- Se debe hacer un estudio previo de selección de variables previamente, puede ser mediante análisis de correlación bivariada u otro método de selección previo.
- El modelo admite variables explicativas de tipo continua, en el caso de ser categórica y tener más de dos categorías se deben transformar en variables dummy, es decir, se

construyen variables artificiales, que sólo pueden tomar valores de 0 y 1 para cada categoría de la variable original.

Selección de variables

IV	Decisión
>= 0,02	No predictivo
(0,02 – 0,1)	Predicción débil
(0,1 – 0,3)	Predicción media
(0,3 – 0,5)	Predicción fuerte
<= 0,5	Sobreajuste

	Variable	IV
3	gastoD	0.447464214
4	P2478_GcelularD	0.309810426
11	P2476_9_Computador	0.305687681
9	P2584_Cuentaahorro	0.295662443
5	P2461_PrecioviviendaD	0.290210779
2	smlD	0.273235914
6	P2478_GinternetD	0.250538938
8	P2439_Propietariovivienda	0.200722900
15	P2977_Mora12m	0.180604605
10	P2476_6_TV	0.157972917
7	P241_Edad	0.084822924
1	ANIO	0.082675847
14	P35	0.005184410
13	P2735_hogaresvivienda	0.004387205
12	P232_Personashogar	0.004148333

Se realiza la selección de variables incluyendo variables significativas que dieron como resultado en los modelos previos, y se analiza el criterio del valor de información (IV), de acuerdo con los parámetros se excluyen: Sexo (P35), hogaresvivienda y personas en el hogar.

Resultados de la estimación de regresión logística

```
> summary(logit)

Call:
glm(formula = cformal ~ smlD + gastoD + P2478_GcelularD + P2461_PrecioviviendaD +
    P2478_GinternetD + P241_Edad, family = "binomial", data = BD)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.9581  -0.7148   1.0929   1.9111

Coefficients:
                       Estimate       Std. Error z value            Pr(>|z|)
(Intercept)          -1.33704952362677  0.01498158924473  -89.25 <0.0000000000000002 ***
smlD                  0.22264114019367  0.00337893788513   65.89 <0.0000000000000002 ***
gastoD                0.43282195753346  0.00754007620864   57.40 <0.0000000000000002 ***
P2478_GcelularD       0.00000356839705  0.00000009904640   36.03 <0.0000000000000002 ***
P2461_PrecioviviendaD 0.00000000471588  0.00000000007117   66.26 <0.0000000000000002 ***
P2478_GinternetD      0.00000096782966  0.00000008199053   11.80 <0.0000000000000002 ***
P241_Edad            -0.00373437802820  0.00027805942209  -13.43 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 26727  6122
         1 16134 14845

               Accuracy : 0.6513
                 95% CI : (0.6476, 0.655)
    No Information Rate : 0.6715
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2955

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.6236
            Specificity : 0.7080
         Pos Pred Value : 0.8136
         Neg Pred Value : 0.4792
             Prevalence : 0.6715
```

6.4. Resultados

En la selección de modelos con técnicas supervisadas de clasificación se sugiere evaluar los siguientes indicadores: maximizar la precisión del modelo, minimizar la tasa de error u obtener el mayor AUC, también, es útil utilizar la técnica de validación cruzada.

Optimización de los parámetros usados en el algoritmo específico

La técnica de validación cruzada permite optimizar el modelo y asegurar que no hay sobreajuste sobre los datos, el modelo permite generalizar las reglas de decisión sobre diferentes conjuntos de datos y realizar predicciones adecuadas. Para llevarlo a cabo el conjunto de datos evaluados deberá ser dividido en dos subconjuntos, uno de entrenamiento y otro de validación, los datos de entrenamiento servirán para entrenar el modelo predictivo y se evaluara el nivel de predicción sobre el conjunto de datos de prueba o validación.

- Matriz de confusión: Es una tabla de contingencias que permite comparar la predicción de las clases de la variable objetivo con su valor real.
- Especificidad: Es la tasa de verdaderos negativos y se representa con la fórmula:

$$\frac{Tn}{(Fp + Tn)}$$ Ecuación (11)

- Sensibilidad: Es la tasa de positivos verdaderos, indica la proporción de valores positivos correctamente clasificados por la predicción sobre el total de positivos

$$\frac{Tp}{(Fp + Tn)}$$ Ecuación (12)

- Precisión: (Accuracy) Indica la capacidad de precisión del modelo

$$\frac{Tp + Tn}{(p + n)}$$ Ecuación (13)

Tabla 22 Estadísticas de desempeño de los modelos

Mejor desempeño	Modelo	Precisión
1	Bosque aleatorio	73,8%
2	Árbol de decisión	69,0%
3	Árbol de decisión condicional	68,4%
4	Regresión logística	65,2%

Mejor desempeño	Modelo	Precisión
1	Bosque aleatorio IEFIC 2010-2016	70,0%
2	Árbol de decisión condicional IEFIC 2010-2016	66,7%
3	Árbol de decisión IEFIC 2010-2016	65,8%

El modelo que presenta mejor desempeño es el bosque aleatorio, esto va en línea con la literatura, pues este modelo optimiza la iteración de particiones para seleccionar las variables que muestran mayor poder de discriminación y disminución del error, en ambos conjuntos de datos de entrenamiento tanto para 2010-2016 y 2017-2018 el indicador de precisión del modelo de bosque aleatorio es mejor 70% y 73,8% respectivamente. En segundo lugar, se encuentra el modelo de árbol condicional mostrando una precisión del 66,7% y 69% sobre los datos de validación.

Tabla 23 Selección de variables sugeridas por los modelos

Bosque aleatorio 2017-2018	**Bosque aleatorio 2010-2016**	**Árbol de decisión condicional 2017-2018**	**Árbol de decisión condicional 2010-2016**	**Árbol de decisión 2017-2018**	**Árbol de decisión 2010-2016**
Gasto telefono	Cuenta ahorro	Gasto telefono>=57.000	Cuenta ahorro>0	Gasto teléfono>58.000	Ingreso>1.140.697
Departamento	Gasto internet	Cuenta ahorro>0	Propietario vivienda>0	Cuenta de ahorros =1	Propietario vivienda = 1
Gasto internet	Nivel educativo	Gasto internet >76.000	Nivel educativo>6	Propietario vivienda = 1	Gasto internet >56.400
Mora	Gasto telefono	Propietario vivienda >0	Ingreso>1.283.000	Gasto internet >52.000	Cuenta de ahorros =1
Ingreso	Gasto mes total	Ingreso>1.475.000	Antigüedad>= 19	Precio vivienda>=71.000.000	Computador =1
Gasto mes total	Edad	Cantidad activos financieros>0	Gasto teléfono>=63.000	Ingreso>1.400.000	
Cantidad activos financieros	Propietario vivienda	Computador>0	Computador>0		
Cuenta ahorro	Gasto salud	Precio vivienda>70.000.000	Gasto internet>=56.000		
Computador	Gasto alimentos		Parentesco<=2		
Propietario vivienda	Parentesco				
Precio vivienda	Gasto recreación				
Edad	Pago mes seguro				
	Personas hogar				
	Antigüedad				
	Computador				

Fuente: Elaboración propia

Tomando como referente el modelo de bosque aleatorio se comparan las variables sugeridas por los otros modelos, se evidencia que hay alta recurrencia en el número de veces que aparecen las variables y el orden de posición en que aparecen. Así que se estima la mediana y se compara con la posición referente para estimar una posición global de la variable.

Tabla 24 Matriz de posición global de variables seleccionadas

Posición global	Variable
1	Gasto teléfono
2	Gasto internet
3	Cuenta ahorro
4	Ingreso
5	Gasto mes total
6	Propietario vivienda
7	Cantidad activos f
8	Computador
9	Precio vivienda
10	Edad

En la tabla 24 se calcula una matriz de posición global de las variables significativas que permiten predecir si una persona accede al crédito. Las variables que presentan mayor capacidad explicativa y de discriminación están relacionadas con los usuarios de servicios de pago de telefonía e internet, esto va en línea con la práctica, es la expansión natural del mercado del crédito hacia segmentos de mercado que pueden demostrar buen hábito de pago en servicios de suscripción recurrente y que no tienen experiencia previa financiera.

Esto va en línea con la literatura analizada sobre las barreas de acceso al crédito. Pues si un usuario tiene la limitación de no tener experiencia previa con el uso de productos de crédito para demostrar un buen comportamiento de pago, su desempeño de pago en telefonía e internet puede suplir este insumo de información.

Las variables de gasto teléfono, gasto internet y tener computador sugieren el tipo de usuario de los servicios financieros actuales, son clientes que esperan servicios digitales multiplataforma y que están conectados a las aplicaciones y a las facilidades móviles

financieras, así que valdría la pena en los trabajos futuros evaluar la inclusión financiera y su estrecha relación con la penetración móvil y la conectividad a internet en el país.

Los usuarios que tienen cuenta de ahorro representan para las entidades financieras, clientes objetivo de alto valor, esta variable sugiere capacidad de ahorro por parte del cliente, experiencia y relacionamiento con la banca, una oportunidad para realizar campañas de venta cruzada de productos de crédito a usuarios con una cuenta de ahorro activa con la entidad financiera.

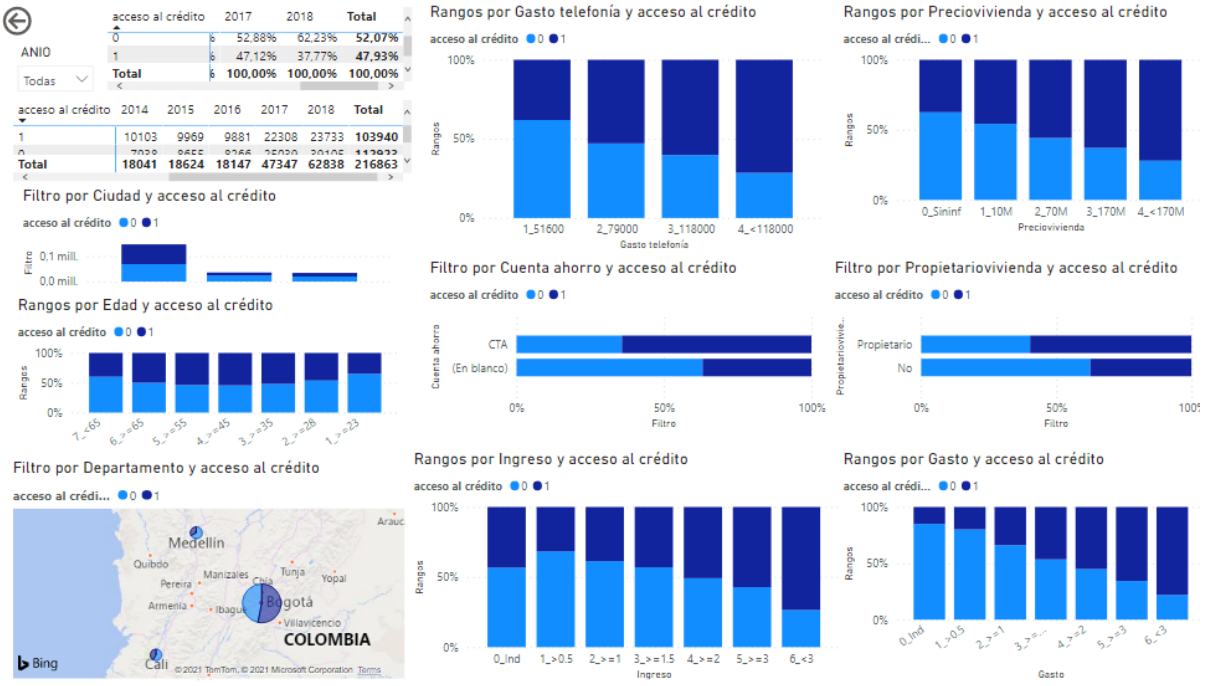
Como se observa en la literatura previa, las variables de ingreso, ser propietario de vivienda y contar con activos financieros son las principales características para demostrar el nivel de riqueza de los hogares, el ingreso refleja liquidez para hacer frente a los pagos contraídos y ser propietario de vivienda, el precio de la vivienda y la cantidad de activos financieros son indicadores del valor de las garantías y patrimonio del hogar.

La edad es una variable sociodemográfica relevante en diversos estudios sobre el acceso a servicios financieros ya que refleja hábitos y comportamientos financieros diferenciales por tramos de edad, se espera que las personas que sobrepasan los treinta años tengan mayor estabilidad económica en sus ingresos, experiencia financiera y un comportamiento estable en sus hábitos de pago y de gastos, la variable del valor de gasto mensual permite evaluar la capacidad de endeudamiento y de gasto de los hogares respecto al ingreso disponible.

6.5. Visualización de resultados

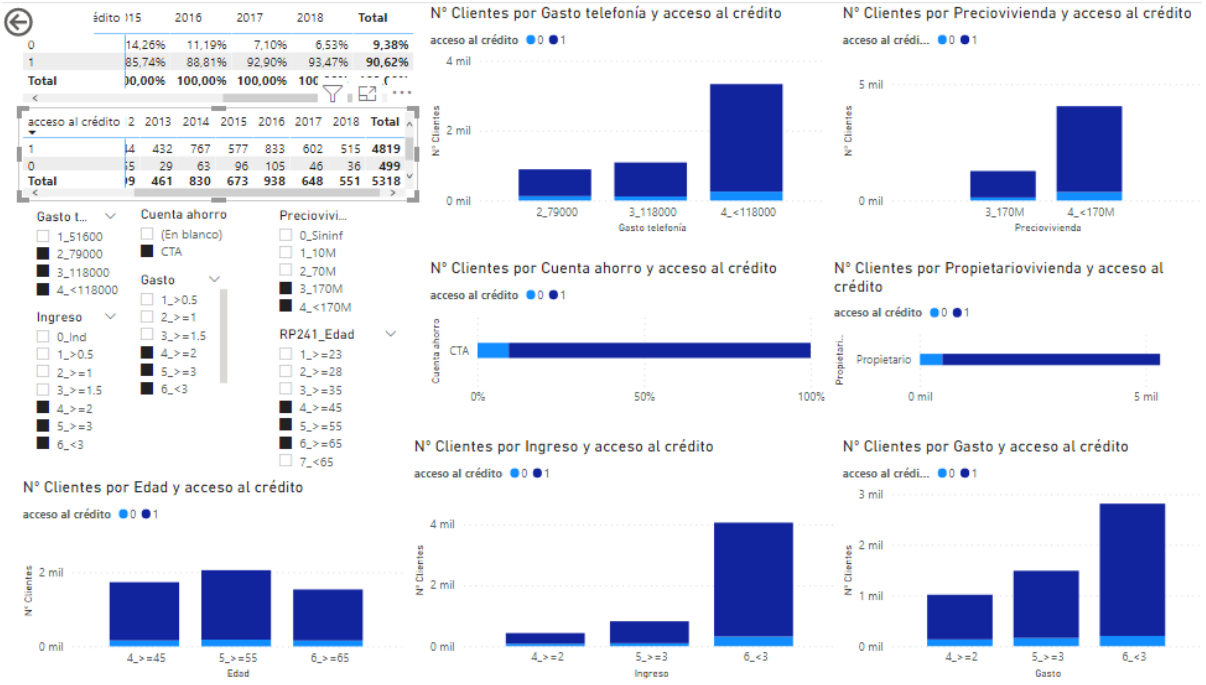
En la gráfica 29 se observa el perfil de la población total, se utilizan las principales variables explicativas brindadas por los modelos predictivos. Para los años analizados se tiene una tasa de acceso al crédito del 47,9% de la población. Las variables son monotónicas, es decir, la tasa de acceso al crédito se ordena de acuerdo con los puntos de corte definidos. Por ejemplo: a mayor nivel de ingreso mayor es la tasa de acceso al crédito, a mayor gasto en telefonía mayor tasa de acceso al crédito.

Gráfica 29 Perfil de la población total



Fuente: Elaboración propia. Microsoft Power BI.

Gráfica 30 Perfil alto de acceso al crédito



Fuente: Elaboración propia. Microsoft Power BI.

En la gráfica 30 se observa el perfil más alto de la población que accede al crédito, al filtrar o aplicar a cada variable significativa, los puntos de corte sugeridos por los árboles de decisión y el peso de la evidencia (WOE). La tasa acceso al crédito se incrementa a 90,6% de la población. Y la cantidad de clientes se acumulan en los rangos más elevados, el ingreso superior a 3 salarios mínimos, gasto superior a 3 salarios mínimos, el gasto en telefonía mayor a 118.000 pesos colombianos, el precio de la vivienda mayor a 170 millones de pesos colombianos y las edades entre los rangos de 45 a 55 años.

7. Conclusiones

En el análisis de caracterización y predicción de acceso al crédito formal en el sector financiero, se utilizó un enfoque teórico de modelos de descubrimiento de conocimiento a partir de bases de datos, el cual se desarrolla en dos fases: análisis descriptivo y análisis predictivo, en el descriptivo se caracteriza el cliente, es decir, se observa el comportamiento promedio del cliente en un instante de tiempo específico, de acuerdo con las variables relevantes indicadas a partir del análisis de correlación. Y con el análisis predictivo se utilizan tres técnicas supervisadas de clasificación.

En la selección del conjunto de datos abiertos se utiliza el modelo Meloda 4.0, para medir el nivel de reutilización de los datos abiertos y se elige la *Encuesta anual de carga financiera y educación financiera de los hogares*, ya que presenta un nivel de reutilización alto, y los aspectos con mejor desempeño son: el jurídico, los datos se pueden utilizar sin restricciones, y el modelo de datos cumple estándares normativos internacionales de calidad.

Para asegurar la calidad de los datos y se observaron los criterios inherentes de la norma ISO / IEC 25012, se concluye que el conjunto de datos seleccionado cumple en su totalidad con criterios de exactitud, consistencia y credibilidad. El diseño del estudio, la extracción de datos y el desarrollo de los modelos predictivos se desarrolla en software abierto, la herramienta Rstudio cloud que brinda ventajas sobre la multitud de funciones estadísticas documentadas, versatilidad en el espacio de la herramienta, tiempo de cómputo, conservación, reproducción y reutilización del material desarrollado.

A partir de los modelos de predicción de datos abiertos se observa un desempeño en el indicador de precisión estadístico cercano al 70% sobre los conjuntos de datos evaluados, desde un punto de vista de negocio este modelo logra demostrar el valor de los datos abiertos como un insumo para el desarrollo de modelos explicativos.

El uso de datos abiertos y la selección de variables no tradicionales de la banca lograron demostrar alta discriminación, esto es de utilidad en el desarrollo de modelos de score de acceso al crédito para usuarios que enfrentan barreras de entrada al mercado de crédito financiero: ingresos insuficientes, reportes negativos sobre sus hábitos de pago, y sobre todo exclusión por no tener historial crediticio previamente.

En este sentido, los resultados comparativos entre los diferentes modelos desarrollados permiten la generalización de las variables significativas, en primer lugar, las variables de gasto en servicios de telefonía e internet representan alto valor de interpretación en la práctica, para aplicar este modelo a usuarios que no tienen experiencia financiera previa, los hábitos de pago se reflejan en el uso de servicios recurrentes como la telefonía e internet. Además, el valor de la variable tenencia de computador, refuerza la hipótesis para un trabajo futuro de evaluar la estrecha relación entre inclusión financiera y conectividad móvil.

El conjunto de variables significativas va en línea con la teoría económica y financiera sobre el acceso al crédito, reflejan nivel de riqueza, garantía y liquidez de pago, capacidad de endeudamiento, hábitos de pago y experiencia financiera.

El resultado de esta investigación brinda percepciones claves para caracterizar y comprender la expansión del mercado del crédito más allá de los hábitos de pago, las variables sugieren el perfil actual de los usuarios, son usuarios que pueden utilizar servicios digitales financieros. Son usuarios que, aunque no cuenten con experiencia previa de crédito pueden demostrar su conducta de pago de telefonía e internet y tiene capacidad de endeudamiento y de gasto.

Y aquellos usuarios que ya tienen experiencia y relacionamiento con la banca mediante productos de ahorro, fondos de inversión, u otros, representan para las entidades financieras, clientes de alto valor y una oportunidad para realizar campañas de venta cruzada de productos de crédito a usuarios activos con la entidad financiera.

7.1 Trabajos futuros

- Continuar en la investigación de las barreras de acceso al crédito utilizando microdatos.
- Profundizar en el desarrollo de la hipótesis para evaluar la estrecha relación entre inclusión financiera, conectividad móvil y conectividad a internet.
- Diseñar un estimador de ingreso disponible a partir de datos abiertos, pues con este trabajo se deja evidencia que los microdatos abiertos capturan con gran capacidad un proxy de la situación financiera de los hogares histórica, su nivel de ingreso, capacidad de gasto y de endeudamiento.
- Promover el uso de datos abiertos con un enfoque de microdatos y el desarrollo de código abierto en herramientas de software libre para su posterior reutilización.

- Implementar una solución de analítica visual con datos abiertos para evaluar el acceso al crédito como una estrategia de seguimiento a los indicadores de inclusión financiera.

7.2 Relación del trabajo desarrollado con los estudios cursados

Este Trabajo final de máster en Gestión de Información se realiza en conjunto para las universidades: Escuela Colombiana de Ingeniería Julio Garavito y la Universitat Politécnica de Valencia (UPV), las asignaturas que sirvieron de base para aplicar los conocimientos adquiridos y desarrollar competencias en el ámbito de la gestión de información.

Por parte de la UPV:

- Explotación de datos masivos: Competencias aplicar técnicas para el uso, diseño y evaluación de fuentes y recursos de información digital, obtener, tratar e interpretar datos y emitir criterios en la evaluación y reflexión en el ámbito de la gestión de información.
- Analítica digital: Planificar, explotar y evaluar servicios de información digital, plantear indicadores a partir de herramientas web y comunicar la metodología y los resultados obtenidos de la naturaleza de los datos.
- Análisis de datos empresariales: Desarrollar un proyecto de analítica de negocio siguiendo las fases y los pasos especificados. Análisis y resolución de problemas desde criterios propios y especializados de la gestión de datos.

- Datos en la web: Comprender e integrar fuentes y metodologías para la gestión de la información y desarrollar productos y servicios a partir de la información en la web.
- Marco legal y deontológico de la información: Responsabilidad ética y pensamiento crítico. Emitir juicios en función de criterios legales y normativos en el ámbito de la gestión de información.

Por parte de la Escuela:

- Gestión de conocimiento: Comprender modelos de datos correspondientes a las necesidades de la organización, desarrollar metodológicamente un modelo de explotación de datos orientado a la optimización en la toma de decisiones organizacionales.
- Inteligencia de negocios: Identificar las fases de un proyecto de inteligencia de negocios, asegurar la calidad de los datos para su posterior explotación, integrar datos internos y externos para fortalecer las decisiones en la organización.
- El poder de la visualización de datos: Aplicar técnicas de visualización de información bajo la metodología de descubrimiento de información en bases de datos.

8. Bibliografía

Abella, A., Ortiz, M., & De Pablos, C. (2014). Obtenido de Meloda, métrica para evaluar la reutilización de datos abiertos: http://www.elprofesionaldelainformacion.com/contenidos/2014/nov/04.pdf

Abella, A., Ortiz, M., & De Pablos, C. (2017). *A model for the analysis of data-driven innovation and value generation in smart cities' ecosystems.* Obtenido de Cities, Volume 64, Pag. 47-53: https://www.sciencedirect.com/science/article/pii/S0264275116303845

Alfred, R. (2005). *Knowledge discovery: Enhancing data mining and decision support integration.* Obtenido de The university of York, qualifyind dissertation. United Kingdom.

Allue, I., Olivia, R., Guitart, M., Moresco, E., Ortega, A., Soria, L., & Terán, P. (2018). Obtenido de Machine Learning techniques in the modeling of credit risk admission [Técnicas de aprendizaje automático en el modelado de admisión de riesgo de crédito]: http://www.mat.ucm.es/congresos/mweek/XII_Modelling_Week/Informes/Report3.pdf

Asobancaria. (Septiembre de 2020:67). *Reporte de inclusión financiera 2019.* Obtenido de https://www.asobancaria.com/wp-content/uploads/2020/09/Copia-de-Informe_RIF_2019_compressed_compressed.pdf

Asociación bancaria del euro. (2017). *Data exploration opportunities in corporate banking: key concepts and applications. Open Banking Working Group [Oportunidades de exploración de datos en la banca corporativa: conceptos clave y aplicaciones].* Obtenido de https://www.abe-eba.eu/epaper/epaper-data_exploration_opportunities_in_corporate_banking/epaper/ausgabe.pdf

Banca de las oportunidades. (2017:45). *Estudio de demanda para analizar la inclusión financiera en Colombia Informe de resultados.* Obtenido de http://bancadelasoportunidades.gov.co/sites/default/files/2017-03/Estudio_demanda_para_analizar_inclusi%C3%B3n_financiera_en_colombia_1.pdf

Banco de la República. (2020). *Reporte de Estabilidad Financiera.* Bogotá.

Banco de la República. (2021). *Reporte de la Situación del crédito en Colombia.* Obtenido de Banco de la República: https://repositorio.banrep.gov.co/bitstream/item/31601335-1b45-4ea9-9297-1f73a4229894/RSSC%20-%20Publicacion.pdf?sequence=1

Bartual, C., García, F., Giménez, V., & Romero, A. (2012). *Journal of applied finance & banking. 2(6):1-13.* Obtenido de Credit risk analysis: reflections on the use of the logit model : http://hdl.handle.net/10251/60119

BBVA. (16 de Abril de 2018). *BBVA open mind.* Obtenido de Analítica de datos, inteligencia artificial y big data en la banca:

https://www.bbvaopenmind.com/economia/finanzas/analitica-de-datos-inteligencia-artificial-y-big-data-en-la-banca/
Breiman, L. (1999). *Random Forest--Random Features.* Obtenido de University of California, Berkeley: https://www.stat.berkeley.edu/~breiman/random-forests.pdf
Butaru, F. Q. (2016). Risk and risk management in the credit card industry. *Journal of Banking and Finance*, 218-239.
Calabrese, J., Esponda, S., Pasini, A., Boracchia, M., & Pesado, P. (Octubre de 2019). *Guía para evaluar calidad de datos basada en ISO/IEC 25012.* Obtenido de https://core.ac.uk/download/pdf/288490953.pdf
Cardona, & Paola. (2004). *Revista colombiana de estadística* . Obtenido de https://www.emis.de/journals/RCE/V27/V27_2_139Cardona.pdf
CEMLA. (2018). Decisiones financieras de los hogares e inclusión financiera: evidencia para América Latina y el Caribe. 133-163. Obtenido de Factores determinantes del ahorro formal e informal en Colombia: https://www.cemla.org/PDF/ic/ic-2016/ic-2016.pdf
Céspedes, N. (2017). *Banco central de reserva del Perú.* Obtenido de La demanda de crédito a nivel de personas: RCC conoce a ENAHO: https://www.bcrp.gob.pe/publicaciones/documentos-de-trabajo/dt-2017-009.html
CFPB. (2020). *Analysis: CFPB Complaints Surge During Pandemic, Led By credit report complaints*. Obtenido de https://uspirg.org/sites/pirg/files/reports/Pandemic%20CFPB%20Complaints%20Report%20Analysis_0.pdf
Chapman. (2000).
Colea, E. B. (2013). Who gets credit after bankruptcy and why? An information channel. *Journal of banking and finance*, 5101-5117.
DANE. (2003). Obtenido de COLOMBIA - Encuesta Nacional de Calidad de Vida - ENCV [Base de datos - Micro datos]: http://microdatos.dane.gov.co/index.php/catalog/186/get_microdata
DANE. (2016). *Colombia - Encuesta de carga financiera y educación financiera de los hogares IEFIC 2010-2016.* Obtenido de http://microdatos.dane.gov.co/index.php/catalog/470/get_microdata
DANE. (2018). Obtenido de Ficha metodológica Encuesta de carga financiera y educación financiera de los hogares- IEFIC: http://microdatos.dane.gov.co/index.php/catalog/626/data_dictionary
DANE. (13 de Agosto de 2019). *Departamento administrativo nacional de estadística.* Obtenido de Encuesta de Carga Financiera y Educación Financiera de los Hogares - IEFIC-2017-2018 [Base de datos microdatos]: http://microdatos.dane.gov.co/index.php/catalog/626#page=overview&tab=study-desc
DANEX. (2009). *Informe de avance primer trimestre Encuesta de carga financiera y educación de los hogares Convenio Dane- Banco de la República N° 23 DE 2009* . Obtenido de

https://www.dane.gov.co/files/investigaciones/boletines/carga_fin/Historicos_carga_fin.pdf
Deloitte. (2015). Obtenido de Ser o no Ser Digital, ¿es esa la cuestión? Cómo la Industria Financiera Latinoamericana debería abordar esta problemática: https://www2.deloitte.com/content/dam/Deloitte/cr/Documents/financial-services/estudios/151104-FS-Ser-o-no-Ser-Digital.pdf
Doherty, C., Camiña, S., White, K., & Orenstein, G. (2016). Obtenido de The path to predictive analytics and machine learning: https://www.tradeandindustrydev.com/industry/technology-r-d/path-predictive-learning-14870
Espino, C. (2017). Obtenido de Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/59565/6/caresptimTFG0117mem%C3%B2ria.pdf
Fayyad, e. (1996).
Fayyad, U. (Octubre de 1996). *Data mining and knowledge discovery: making sense out of data.* Obtenido de Data mining. IEEE: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=539013
FICO. (2017). *FICO/blog.* Obtenido de Uso de datos alternativos en el modelado de riesgo de crédito: https://www.fico.com/blogs/using-alternative-data-credit-risk-modelling
García, N., Gámez, M., & Alfaro, E. (2018). *Ensemble classification methods with applications in R.* Obtenido de https://onlinelibrary.wiley.com/doi/book/10.1002/9781119421566
Garg, A., Grande, D., Macias, G., Miranda, L., Sporleder, C., & Windhagen, E. (Abril de 2017). *Analytics in banking: Time to realize the value [Analítica en la banca: es hora de darse cuenta del valor].* Recuperado el 21 de Enero de 2020, de McKinsey & Company: https://www.mckinsey.com/~/media/McKinsey/Industries/Financial%20Services/Our%20Insights/Analytics%20in%20banking%20Time%20to%20realize%20the%20value/Analytics-in-banking-Time-to-realize-the-value.ashx
Gobierno de España. (2017). *Manual práctico para mejorar la calidad de los datos abiertos.* Obtenido de https://datos.gob.es/sites/default/files/doc/file/manual_practico_para_mejorar_la_calidad_de_los_datos_abiertos_1_0.pdf
Gujarati, D., & Porter, D. (2010). *Econometría básica.* McGrawHill.
Haque. (2017). *Predictive analytics of loans issuance and default using random forest in the online peer- to peer lending marketplace* . Obtenido de Universidad de Princeton.
Hardoon, D., & Schmueli, G. (2016). *Getting started with business analytics: insightful decision making [Introducción a la analítica de negocios: toma de decisiones inteligentes].*
Hernández, S. (2012). *Métodos de análisis de datos.* Obtenido de https://www.unirioja.es/cu/zehernan/docencia/MAD_710/Lib489791.pdf

Hothorn, T., Hornik, K., & Zeileis, A. (2006). *Unbiased Recursive Partitioning: A Conditional Inference Framework.* Obtenido de Journal of Computational and Graphical Statistics: Volume 15, Number 3, Pages 651–674.: https://eeecon.uibk.ac.at/~zeileis/papers/Hothorn+Hornik+Zeileis-2006.pdf

Hurley, M., & Adebayo, J. (2017). *Yale Journal of law and technology.* Obtenido de https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1122&context=yjolt

Iglesias, C. (Febrero de 2020). *Datos abiertos más allá del sector público: publicadores, motivaciones y modelos de colaboración.* Obtenido de https://datos.gob.es/sites/default/files/doc/file/toc_informe_1_-_datos_abiertos_mas_alla_de_los_gobiernos_final_0.pdf

Iregui, A., Melo, L., Ramírez, M., & Tribín, A. (2018). Crédito formal e informal de los hogares en Colombia. *Decisiones financieras de los hogares e inclusión financiera: evidencia para América Latina y el Caribe*.

KPMG. (Abril de 2017). Obtenido de El nivel de madurez digital del sector financiero en España: https://assets.kpmg/content/dam/kpmg/es/pdf/2017/04/nivel-madurez-digital-sector-financiero-espana-kpmg-funcas.pdf

Madrid, A., & Minetti, R. (2013). *Sharing information in the credit market: Contract-level evidence from U.S. firms.* Obtenido de doi:10.1016/j.jfineco.2013.02.007

Maglogiannis, Karpouzis, & Wallace. (2007). Supervised machine learning: a review of classification techniques. En *Emerging Artificial Intelligence Applications in Computer Engineering* (págs. 3-4). Amsterdam: IOS Press.

Mariño, J., Pacheco, D., & Segovia, S. (2018). *Banco de la República.* Obtenido de Carga Financiera - Informe especial de Estabilidad Financiera - Primer semestre 2018: https://www.banrep.gov.co/es/carga-financiera-informe-especial-estabilidad-financiera-primer-semestre-2018

MINTIC. (17 de Diciembre de 2019). *Ministerio de tecnologías de la información y las comunicaciones* . Obtenido de Datos abiertos Encuesta de Carga Financiera y Educación Financiera de los Hogares [Base de datos - Microdatos]: https://www.datos.gov.co/Estad-sticas-Nacionales/Encuesta-de-Carga-Financiera-y-Educaci-n-Financier/9xrr-i6kz

Molina, J., & García, J. (2006). *Técnicas de análisis de datos.* Obtenido de http://matema.ujaen.es/jnavas/web_recursos/archivos/weka%20master%20recursos%20naturales/apuntesAD.pdf

Moral, I. (2006). *Modelos de regresión: lineal simple y regresión logística.* Obtenido de https://revistaseden.org/files/14-CAP%2014.pdf

Murcia, A. (2007). Determinantes del acceso al crédito de los hogares colombianos. *Ensayos sobre Política Económica, 25*(55), 40-83. Recuperado el 15 de Marzo de 2020, de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-44832007000200003&lng=en&tlng=es.

Öhman, O., & Lundstrom, L. (2019). *Machine learning in credit risk: Evaluation of supervised machine learning models predicting credit risk in the financial sector.* Obtenido de http://www.diva-portal.se/smash/get/diva2:1360926/FULLTEXT01.pdf

Orellana, J. (2018). *Arboles de decisión y random forest.* Obtenido de Departamento de recursos hídricos y ciencias ambientales de la universidad. Universidad de Cuenca: https://bookdown.org/content/2031/

Pérez, C., & Santín, D. (2007). *Minería de datos: técnicas y herramientas.* Obtenido de https://books.google.es/books?id=wz-D_8uPFCEC&lpg=PA628&ots=Ti0Zyi7A3I&dq=CHAID%22%2C%20Kass%201980)&pg=PA628#v=onepage&q=CHAID%22,%20Kass%201980)&f=false

Pérez, J. (2017). *La regresión logística como modelo de predicción del riesgo crediticio en las organizaciones de la economía social y solidaria.* Obtenido de https://www.uv.mx/iiesca/files/2018/03/23CA201702.pdf

Reis, I., Baron, D., & Shahaf, S. (2019). *Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets.* Obtenido de https://iopscience.iop.org/article/10.3847/1538-3881/aaf101/pdf

Ruppert, T. (2018). *Visual Analytics to Support Evidence-Based Decision Making [Análisis visual para apoyar la toma de decisiones basada en evidencia]*. Obtenido de https://d-nb.info/1149824336/34

Sancho, F. (2019). *Aprendizaje inductivo: árboles de decisión.* Obtenido de http://www.cs.us.es/~fsancho/?e=104

Sarath, S. (2018). *Área bajo la curva ROC.* Obtenido de https://medium.com/@sarath13/area-under-the-roc-curve-explained-d056854d3815

SAS. (2019). *Analítica predictiva ¿Qué es y por qué es importante?* Obtenido de https://www.sas.com/es_es/insights/analytics/predictive-analytics.html

Singh, R., & Aggarwal , R. (2011). *Comparative Evaluation of Predictive Modeling techniques on credit card.* Obtenido de https://pdfs.semanticscholar.org/17f9/1346a5948b86ec2bbfc78675e9cfb83b311c.pdf

Superintendencia financiera de Colombia. (28 de Noviembre de 2019). *Empoderamiento del consumidor: el nuevo reto de la industria financiera digital.* Obtenido de Superintendencia financiera de Colombia: https://www.superfinanciera.gov.co/jsp/10102319

Superintendencia financiera de Colombia. (2020). Obtenido de https://www.superfinanciera.gov.co/inicio/consumidor-financiero/informacion-general/quejas-contra-entidades-vigiladas/datos-estadisticos-cifras/informacion-estadistica-anual-11129

Tang, L., Cai, F., & Ouyang, Y. (2019). *Applying a nonparametric random forest algorithm to asses the credit risk of the energy industry in China* . Obtenido de Technological forecasting and social change. Volume 144. July 2014. Pages 563-572: https://doi.org/10.1016/j.techfore.2018.03.007

Tello, M., Eslava, H., & Tobías, L. (2013). *Análisis y evaluación del nivel de riesgo en el otorgamiento de créditos financieros utilizando técnicas de minería de datos.* Obtenido de Revista visión electrónica: https://revistas.udistrital.edu.co/index.php/visele/article/view/4389/6755

Timarán, S., Hernández, L., Caicedo, S., Hidalgo, A., & Alvarado, J. (2016). *El proceso de descubrimiento de conocimiento en bases de datos.* Obtenido de Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas profesionales: https://ediciones.ucc.edu.co/index.php/ucc/catalog/download/36/40/230-1?inline=1

Universidad de los Andes. (2013). Obtenido de Encuesta longitudinal colombiana [Base de datos]: https://encuestalongitudinal.uniandes.edu.co/es/datos-elca/2013-ronda-2

University, D. (2014). *chapter 8: Bagging and random forest.* Obtenido de http://www2.stat.duke.edu/~rcs46/lectures_2015/random-forest/ch6_bagging.pdf

9. Anexos

Anexo 1. Características dependientes del sistema del modelo de calidad de datos ISO/ IEC 25012

Grupo	Característica	Descripción	Método de validación	Documento asociado	Concepto sobre los datos
Inherente y dependiente del sistema	Accesibilidad	Grado de acceso a los datos	Forma de acceso a los datos	Sitio web DANE	Descarga del sitio web del DANE
Inherente y dependiente del sistema	Conformidad	Los datos cumplen con estándares y normativas vigentes	Normatividad vigente nacional e internacional Ley 1581 de 2012 Decreto 1074 de 2015 Decreto 090 de 2018	Documentación de la Encuesta Base de datos anonimizada	El DANE recolecta los datos de la fuente en GEA un aplicativo móvil diseñado y a través del DMC que contiene las normas de validación que contribuyen a la calidad del dato.
Inherente y dependiente del sistema	Confidencialidad	Acceso por usuarios autorizados	Reglas de confidencialidad nacional DANE	Documentación de la Encuesta Base de datos anonimizada	Asignación de permisos a funcionarios del DANE que requieren acceso directo a microdatos sin anonimizar Anonimización de la base de

					datos, todas las variables relacionadas con identificación de personas y hogar son extraídas de la base de datos pública.
Inherente y dependiente del sistema	Eficiencia	Los datos pueden ser procesados y proporcionados con el rendimiento esperado	Tasa de respuesta y tiempo de procesamiento Cantidad de recursos	Los datos están publicados en el sitio web y permiten su descarga en formato xls en tiempo razonable	El tiempo de descarga por archivo es < 10 segundos.
Inherente y dependiente del sistema	Precisión	Los datos requieren valores exactos	Correspondencia referencial: Los datos mantienen una relación uno a uno enlazados a un recurso real, tienen etiquetas correctas y sin duplicados.	Cuestionario IEFIC Documentación de la encuesta	Los datos son recogidos mediante el cuestionario diseñado por el DANE, Posteriormente, pasa a un proceso de consistencia de la información real y verificación estadística.
Inherente y dependiente del sistema	Trazabilidad	Proporcionan un registro de acontecimientos que los modifican	Flujo de procesos que modifican los datos	Documentación de la encuesta	El procesamiento y edición de los datos se lleva a cabo por diferentes

					equipos técnicos del DANE
Inherente y dependiente del sistema	Comprensibilidad	Los datos son expresados en lenguaje y simbología apropiada para su interpretación adecuada	Conceptos y especificación de los atributos	Diccionario de variables	Se tiene el diccionario de cada una de las variables, con las siguientes características: ID VARIABLE, NOMBRE, ETIQUETA, TIPO: (Discreta, continua), FORMATO: (carácter o numérica), PREGUNTA
Dependiente del sistema	Disponibilidad	Facilidad para su obtención autorizada	Tiempo de respuesta del sistema cuando se recibe una petición de acceso	Los datos están publicados en el sitio web y permiten su descarga en formato xls en tiempo razonable	El tiempo de descarga por archivo es < 10 segundos.
Dependiente del sistema	Portabilidad	Capacidad de los datos para ser copiados, remplazados o eliminados al realizar un cambio de sistema.	Módulos que se transfieren a una nueva plataforma	Documentación de la encuesta Transmisión de datos a DANE Central	La transmisión de datos se realiza a través de FTP, basado en la arquitectura cliente-servidor. Datos de investigación se conservan

					en Briefcase, posteriormente en ORACLE, y posteriormente se consolidan en un proceso ETL con la herramienta Data integration de Pentaho
Dependiente del sistema	Recuperabilidad	Comprobación de preservar un nivel de operación en caso de fallos.			

Fuente: Elaboración propia. ISO / IEC 25012.

Anexo 2. Propuesta de agrupación de variables de la IEFIC

A.2.1 Descripción de variables de la IEFIC

Categoría DANE	**Descripción DANE**	**Observación identificada**	**Propuesta**
1. Información de la vivienda	Características de identificación de la vivienda	Algunas variables de esta categoría no están disponibles por anonimización	Utilizar las variables disponibles como: cantidad de hogares en la vivienda se combinan con otras categorías de variables de la vivienda.
2.Activos reales y deuda hipotecaria	El objetivo de esta sección es conocer si el hogar al momento de adquirir vivienda utilizó servicios financieros, tales como crédito hipotecario	Contiene variables que describen el crédito hipotecario.	Las variables de deuda hipotecaria se unen en una misma categoría con otras modalidades de crédito formal.
3.Activos reales	Cuantifica los activos en electrodomesticos que posee el hogar	Inventario de electrodomesticos del hogar no se utiliza.	Se utiliza la variable tener computador y se incluye en categpría de conectividad.
4.Consumo	Establecer los gastos básicos del hogar, y el ingreso destinado a ahorro	Varias variables no son obligatorias y no están pobladas.	Se deja la misma categoría y se excluyen variables que no están completas.
5.Activos y deuda hipotecaria	Nivel de endeudamiento, y que ha hecho para disminuirlo	Contiene variables de crédito hipotecario complementario Contiene variables de percepción del endeudamiento.	Las variables de deuda hipotecaria se unen en una misma categoría con otras modalidades de crédito formal. Las variables de percepción del endeudamiento se dejan en una categoría independiente.
6.Seguros y pensiones	Detectar el numero de hogares que cuentan con un seguro para cubrir la pérdida total o parcial de sus bienes y propiedades	Contiene variables de seguro. Estas variables no son obligatorias no están completas	Se incluyen dentro de las variables de activos financieras.

7.Educación financiera	Indagar sobre la educación financiera asociada al crédito y funcionamiento del mercado financiero.	Contiene preguntas de conocimiento financieros	No se utilizan porque se brinda prioridad a variables que reflejen patrones de comportamiento
8.Deuda no hipotecaria	Establece cuantas tarjetas de crédito tienen las personas del hogar, créditos con casas comerciales, préstamos de libre inversión, crédito con prestamistas, tiendas de barrio, cajas de compensación, y cómo estan pagando los créditos	Contiene variables de crédito formal y deuda informal	Se hace la diferencia de crédito formal y deuda informal, se crean dos categorías independientes.
9. Activos financieros	Determinar si el hogar tiene inversiones financieras, tales como acciones, fondos mutuos de inversión, cuentas de ahorro, pensiones, certificados de depósito.	Contiene variables de productos financieros de ahorro, inversiones, etc	Se deja la misma categoría
10. Percepción de carga financiera y restricciones al crédito	Indaga a cerca de la percepción de los hogares frente a su carga financiera y su actitud frente al endeudamiento.	Contiene variables de mora, endeudamiento y solicitudes de crédito	Se diferencian las variables de mora, endeudamiento y solicitudes de crédito y se crean categorías independientes.

Fuente: DANE

A.2.2 Propuesta de agrupación de variables

Propuesta de agrupación	Descripción	Ejemplo de variables	Cantidad de variables	Aspectos
Características del individuo	Información sociodemográfica de la persona que responde el cuestionario	1. Ingreso 2. Ciudad 3. Edad 4. Educación 5. Sexo	7	Sociodemográfica
Características del hogar	Información de la vivienda	1.total de personas en el hogar. 2. Vivienda propia 3. Precio vivienda 4.Tiene subsidio de vivienda 5.Tiene subsidio de vivienda de interés social	8	Nivel de Riqueza Garantía
Crédito formal	Información sobre créditos con el sector financiero formal, modalidad del crédito, cantidad, valor del crédito, capital, intereses, cuota y saldo	1. Adquirió la vivienda con crédito hipotecario 2.Tiene tarjetas de crédito 3.Prestamos de libre inversión 4.Crédito educativo 5.Crédito vehículo,	87	Adquisición de créditos formales

		embarcaciones o aviones. 6.Crédito maquinaria y equipo		
Deuda informal	Información sobre deudas con el sector informal: casas comerciales, gota a gota, tiendas, pariente o amigo, el valor de la deuda, capital, intereses, cuota y saldo.	1. Tiene deuda con casas comerciales 2.Tiene deudas gota a gota 3. Tiene deuda con tiendas 4. tiene deuda con parientes o amigos	40	Adquisición de créditos informales
Gastos	Rubros y conceptos de los gastos del hogar.	1.Gasto anual en educación 2. Gasto mensual 3. Percepción del gasto mensual respecto al ingreso 4.Gasta en su totalidad el ingreso 5.En que utiliza el ingreso que no gasta 6.Cómo cubre la diferencia cuando el gasto supera el ingreso.	42	Nivel de gasto
Activos financieros	Información sobre los activos financieros que tiene el individuo, monto de inversión y renta percibida anual	1.Tiene fondos de inversión 2.Tiene acciones 3.Tiene CDT 4. Tiene	23	Nivel de Riqueza y ahorros

		Cuenta de ahorro 5.Tiene pensiones voluntarias 6.Tiene sociedades 7.Tiene swaps		
Conectividad	Información sobre teléfono e internet	1.Tiene computador 2. Tiene TV	2	Conectividad
Endeudamiento	Percepción financiera sobre el endeudamiento	1. Nivel de deuda respecto al ingreso 2.Nivel de endeudamiento 3.Le gustaría disminuir su endeudamiento 4.Médidas para reducir su endeudamiento	4	Nivel de endeudamiento
Mora	Cantidad de veces que ha caído en mora en los últimos 12 meses	1. Número de moras en los últimos 12 meses	1	Morosidad
Solicitudes de crédito formal	Solicitudes de crédito y refinanciación de préstamos anteriores al sector financiero formal	1.Cantidad de solicitudes y refinanciación en los últimos dos años 2.Cantidad de solicitudes de crédito le han sido rechazadas totalmente 3.Motivo de rechazo	3	Restricción al crédito
Otras de poco interés	Identificación de la encuesta,		68	

	conocimiento financiero, electrodomésticos y otras que su nivel de registros vacíos es cercano al 70% como el detalle de saldo, cuota, plazo de los diferentes tipos de crédito.			
Total			285	

Fuente: Elaboración propia. DANE (2018).

Anexo 3. Análisis univariado de variables.

1. Características del individuo

Nombre	Tipo	Transformación
INGTOTOB	Continua	Validar outliers
Niveleducativo	Numérica	Categórica
Edad	Continua	Validar outliers
Parentesco	Numérica	Categórica
Sexo	Numérica	Categórica

```
   INGTOTOB          n_educativo        Edad          Parentesco        Sexo
Min.   :        0   Min.   :100.0   Min.   : 18.00   Min.   :1.000   Min.   :1.000
1st Qu.:   200000   1st Qu.:408.0   1st Qu.: 27.00   1st Qu.:1.000   1st Qu.:1.000
Median :   784350   Median :511.0   Median : 39.00   Median :2.000   Median :2.000
Mean   :  1232918   Mean   :500.1   Mean   : 41.39   Mean   :2.298   Mean   :1.537
3rd Qu.:  1338000   3rd Qu.:603.0   3rd Qu.: 53.00   3rd Qu.:3.000   3rd Qu.:2.000
Max.   :100544999   Max.   :999.0   Max.   :101.00   Max.   :9.000   Max.   :2.000
```

WOE – IV

```
$INGTOTOB
              INGTOTOB     N    Percent        WOE         IV
1           [0,199998] 85677 0.39507431 -0.2519834 0.02482472
2       [2e+05,599900] 21744 0.10026607 -0.4646021 0.04588746
3       [6e+05,799900] 22684 0.10460060 -0.2711388 0.05348812
4       [8e+05,992211] 21697 0.10004934 -0.1893205 0.05704952
5  [992333.33,1299999] 20813 0.09597303  0.1308524 0.05869491
6 [1300000,2099666.67] 22481 0.10366453  0.5575101 0.09046386
7 [2100000,1001083225] 21767 0.10037212  1.3514364 0.25417423

$P241
       P241     N    Percent         WOE         IV
1   [18,21] 20419 0.09415622 -0.66887259 0.04008581
2   [22,25] 22383 0.10321263 -0.23985536 0.04596632
3   [26,29] 20759 0.09572403  0.01783739 0.04599678
4   [30,34] 23149 0.10674481  0.17687482 0.04933980
5   [35,39] 20393 0.09403633  0.22957760 0.05429788
6   [40,45] 21968 0.10129898  0.27356261 0.06187448
7   [46,51] 22502 0.10376136  0.25181123 0.06845346
8   [52,57] 21031 0.09697828  0.18349541 0.07172203
9   [58,66] 22327 0.10295440  0.05986989 0.07209140
10 [67,108] 21932 0.10113297 -0.35799999 0.08482292
```

2. Características del hogar

Nombre	Tipo	Transformación
Personas en el hogar	Numérica	NA=0
Propietario de vivienda	Numérica	NA=0
Antigüedad	Numérica	NuevaNA=0 y Outliers=0
Precio vivienda	Continua	NA=0
Subsidio vivienda	Numérica	Categórica
Hogares en la vivienda	Numérica	

```
   P_hogar       Prop_vivienda      antiguedad      Precio_vivienda      subsidio_vivienda
Min.   : 0.000   Min.   :0.0000   Min.   : -1.000   Min.   :         0   Min.   :0.00000
1st Qu.: 2.000   1st Qu.:0.0000   1st Qu.:  0.000   1st Qu.:         0   1st Qu.:0.00000
Median : 3.000   Median :1.0000   Median :  1.000   Median :         8   Median :0.00000
Mean   : 3.011   Mean   :0.5115   Mean   :  9.752   Mean   :  79481628   Mean   :0.05262
3rd Qu.: 4.000   3rd Qu.:1.0000   3rd Qu.: 18.000   3rd Qu.: 100000000   3rd Qu.:0.00000
Max.   :12.000   Max.   :1.0000   Max.   :114.000   Max.   :9999999999   Max.   :9.00000

hogares_vivienda
Min.   :1.000
1st Qu.:1.000
Median :1.000
Mean   :1.038
3rd Qu.:1.000
Max.   :5.000
```

3. Crédito formal

Nombre	Tipo	Transformación
Cformal	Numérica	Nueva
Crédito vivienda	Numérica	Binomial, NA=0
Tiene más viviendas	Numérica	Binomial, NA=0
Tiene vehículo	Numérica	Binomial, NA=0
Tarjeta de crédito	Numérica	Binomial, NA=0
Crédito libre inversión	Numérica	Binomial, NA=0
Crédito de educación	Numérica	Binomial, NA=0
Crédito de vehículo	Numérica	Binomial, NA=0

```
   cformal       P2466_Creditovivienda P2483_Tienemasviviendas P2502_Tienevehiculo   P2540_TDC
Min.   :0.0000   Min.   :0.000         Min.   :1.000           Min.   :1.000       Min.   :1.000
1st Qu.:0.0000   1st Qu.:0.000         1st Qu.:2.000           1st Qu.:1.000       1st Qu.:1.000
Median :1.0000   Median :0.000         Median :2.000           Median :2.000       Median :2.000
Mean   :0.5437   Mean   :0.197         Mean   :1.902           Mean   :1.672       Mean   :1.652
3rd Qu.:1.0000   3rd Qu.:0.000         3rd Qu.:2.000           3rd Qu.:2.000       3rd Qu.:2.000
Max.   :1.0000   Max.   :9.000         Max.   :9.000           Max.   :9.000       Max.   :2.000
                                                                                   NA's   :3493
P2602_CreditoLibre P2734_Creditoeducacion P337_Creditovehiculo
Min.   :1.000      Min.   :1.000          Min.   :1.00
1st Qu.:2.000      1st Qu.:2.000          1st Qu.:1.00
Median :2.000      Median :2.000          Median :2.00
Mean   :1.871      Mean   :1.972          Mean   :1.52
3rd Qu.:2.000      3rd Qu.:2.000          3rd Qu.:2.00
Max.   :2.000      Max.   :2.000          Max.   :9.00
NA's   :3505       NA's   :3510           NA's   :61945
```

4. Deuda informal

Nombre	Tipo	Transformación
Cinformal	Numérica	Nueva
Debe casa comercial	Numérica	Binomial, NA=0
Debe gota a gota	Numérica	Binomial, NA=0
Debe tienda	Numérica	Binomial, NA=0
Debe amigos	Numérica	Binomial, NA=0

```
  cinformal       P2548_debecasacom P2633_debegota P2692_debetienda P2771_debeamigos
Min.   :0.00000   Min.   :1.000     Min.   :1.00   Min.   :1.000    Min.   :1.000
1st Qu.:0.00000   1st Qu.:2.000     1st Qu.:2.00   1st Qu.:2.000    1st Qu.:2.000
Median :0.00000   Median :2.000     Median :2.00   Median :2.000    Median :2.000
Mean   :0.09517   Mean   :1.996     Mean   :1.99   Mean   :1.975    Mean   :1.934
3rd Qu.:0.00000   3rd Qu.:2.000     3rd Qu.:2.00   3rd Qu.:2.000    3rd Qu.:2.000
Max.   :1.00000   Max.   :2.000     Max.   :2.00   Max.   :2.000    Max.   :2.000
```

5. Gastos

Nº	Nombre	Tipo	Transformación
1	Gasto educación	Continua	NA=0
2	Gasto alimentos	Continua	NA=0
3	Gasto vestuario	Continua	NA=0
4	Gasto agua	Continua	NA=0
5	Gasto luz	Continua	NA=0
6	Gasto gas	Continua	NA=0
7	Gasto celular	Continua	NA=0
8	Gasto servicio domestico	Continua	NA=0
9	Gasto recreación	Continua	NA=0
10	Gasto salud	Continua	NA=0
11	Gasto internet	Continua	NA=0

12	Gasto transporte	Continua	NA=0
13	Gasto manutención	Continua	NA=0
14	Gasto mes total	Continua	Nueva
15	Nivel de gasto	Numérica	Categórica
16	Gasto completo	Numérica	Binomial
17	Diferencia gasto ingreso	Categórica	Nueva

```
P2477_Gastoeducacion P2478_1_Gastoalimentos P2478_2_Gastovestuario P2478_3_Gastoagua
Min.   :         0   Min.   :        0      Min.   :       0       Min.   :       0
1st Qu.:         0   1st Qu.:   350000      1st Qu.:   17000       1st Qu.:   25000
Median :    300000   Median :   500000      Median :   50000       Median :   40000
Mean   :   3002101   Mean   :   589571      Mean   :   87129       Mean   :   52436
3rd Qu.:   2500000   3rd Qu.:   700000      3rd Qu.:  100000       3rd Qu.:   60000
Max.   :8000000000   Max.   :300000000      Max.   :87900000       Max.   :62500000
P2478_4_Gastoluz    P2478_5_Gastogas  P2478_6_Gastocelular P2478_7_Gastosserviciodomestico
Min.   :        0   Min.   :      0   Min.   :       0     Min.   :      0
1st Qu.:    25000   1st Qu.:  10000   1st Qu.:   10000     1st Qu.:      0
Median :    40000   Median :  18000   Median :   35000     Median :      0
Mean   :    59161   Mean   :  21716   Mean   :   60950     Mean   :  37637
3rd Qu.:    70000   3rd Qu.:  29000   3rd Qu.:   80000     3rd Qu.:      0
Max.   :100000000   Max.   :1100000   Max.   :42000000     Max.   :4000000
P2478_8_Gastoreecreacion P2478_9_Gastosalud P2478_10_Gastointernet P2478_11_Gastotransporte
Min.   :      0          Min.   :       0   Min.   :       0       Min.   :      0
1st Qu.:      0          1st Qu.:       0   1st Qu.:   20000       1st Qu.:      0
Median :  40000          Median :       0   Median :   58000       Median :      0
Mean   :  91610          Mean   :   60416   Mean   :   63930       Mean   :  22143
3rd Qu.: 100000          3rd Qu.:   30000   3rd Qu.:   90000       3rd Qu.:      0
Max.   :7000000          Max.   :15000000   Max.   :58000000       Max.   :4000000
P2478_12_Gastomanutencion   gasto_mest
Min.   :        0          Min.   :        0
1st Qu.:        0          1st Qu.:   635000
Median :        0          Median :   898000
Mean   :    51753          Mean   :  1198452
3rd Qu.:        0          3rd Qu.:  1340000
Max.   :360000000          Max.   :362104000
```

6. Activos financieros

	Nombre	Tipo	Transformación
Nº	CDT	Numérica	Binomial, NA=0
1	Monto anual ahorro	Continua	NA=0, validar Outliers
2	Cuenta ahorro	Numérica	Binomial, NA=0
3	Acciones	Numérica	Binomial, NA=0
4	Monto anual cta ahorros	Continua	NA=0, validar Outliers
5	Cotiza pensión	Numérica	Binomial, NA=0
6	Sociedades	Numérica	Binomial, NA=0
7	Otros activos	Numérica	Binomial, NA=0
8	Fondos de inversión	Numérica	Binomial, NA=0
9	Seguro	Numérica	Binomial, NA=0
10	Pago mensual seguro	Continua	NA=0, validar Outliers
11	Crédito caja compensación	Numérica	Binomial, NA=0

```
   P1239_CDT       P2579_Montoanualahorro P2584_Cuentaahorro P2819_Acciones    P2962_Montoanualcta
ahorros
 Min.   :0.00000   Min.   :       0       Min.   :0.0000     Min.   :0.00000   Min.   :        0
 1st Qu.:0.00000   1st Qu.:       0       1st Qu.:0.0000     1st Qu.:0.00000   1st Qu.:        0
 Median :0.00000   Median :       0       Median :1.0000     Median :0.00000   Median :        0
 Mean   :0.01432   Mean   :    5497       Mean   :0.5385     Mean   :0.02509   Mean   :   345612
 3rd Qu.:0.00000   3rd Qu.:       0       3rd Qu.:1.0000     3rd Qu.:0.00000   3rd Qu.:        0
 Max.   :1.00000   Max.   :40000000       Max.   :1.0000     Max.   :1.00000   Max.   :400000000
 P2964_Cotizapension P2967_Sociedades  P2970_Otrosactivos  P622_Fondosinversion P2506_Segurodevida
 Min.   :0.000000    Min.   :0.00000   Min.   :0.0000000   Min.   :0.00000      Min.   :0.0000
 1st Qu.:0.000000    1st Qu.:0.00000   1st Qu.:0.0000000   1st Qu.:0.00000      1st Qu.:0.0000
 Median :0.000000    Median :0.00000   Median :0.0000000   Median :0.00000      Median :0.0000
 Mean   :0.009483    Mean   :0.01302   Mean   :0.0009829   Mean   :0.01426      Mean   :0.2662
 3rd Qu.:0.000000    3rd Qu.:0.00000   3rd Qu.:0.0000000   3rd Qu.:0.00000      3rd Qu.:1.0000
 Max.   :1.000000    Max.   :1.00000   Max.   :1.0000000   Max.   :1.00000      Max.   :1.0000
 P408_Pagomesseguro  P2695_Ccajacompensacion
 Min.   :        0   Min.   :0.00000
 1st Qu.:        0   1st Qu.:0.00000
 Median :        0   Median :0.00000
 Mean   :    20459   Mean   :0.04676
 3rd Qu.:     5000   3rd Qu.:0.00000
 Max.   :100000000   Max.   :1.00000
```

7. Conectividad

Nº	Nombre	Tipo	Transformación
1	TV	Numérica	NA=0
2	Computador	Numérica	NA=0

```
  P2476_6_TV    P2476_9_Computador
Min.   :0.000   Min.   :0.0000
1st Qu.:1.000   1st Qu.:0.0000
Median :2.000   Median :1.0000
Mean   :1.916   Mean   :0.8427
3rd Qu.:2.000   3rd Qu.:1.0000
Max.   :9.000   Max.   :9.0000
```

8. Endeudamiento

Nº	Nombre	Tipo	Transformación
1	Percepción endeudamiento	Numérica	NA=0, categórica
2	Califica endeudamiento	Numérica	NA=0, categórica
3	Disminuye endeudamiento	Numérica	NA=0, binomial
4	Reduce endeudamiento	Numérica	NA=0, binomial

```
P2973_Pendeuda  P2974_Cendeuda  P2975_Dendeuda   P2976_Rendeuda
Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
1st Qu.:2.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
Median :3.000   Median :0.000   Median :0.0000   Median :0.0000
Mean   :2.963   Mean   :1.293   Mean   :0.3858   Mean   :0.2303
3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:0.0000
Max.   :4.000   Max.   :4.000   Max.   :1.0000   Max.   :1.0000
```

9. Mora

Nº	Nombre	Tipo	Transformación
1	Mora 12 m	Numérica	NA=0, outliers

```
P2977_Mora12m
Min.   : 0.0000
1st Qu.: 0.0000
Median : 0.0000
Mean   : 0.4791
3rd Qu.: 0.0000
Max.   :13.0000
```

10. Solicitudes

Nº	Nombre	Tipo	Transformación
1	Solicitud	Numérica	NA=0
2	Rechazo	Numérica	NA=0
3	Motivo de rechazo	Numérica	NA=0, categórica

```
P2978_Solicitud   P2979_Rechazo      P2980_Mrechazo
Min.   : 0.0000   Min.   : 0.00000   Min.   :0.0000
1st Qu.: 0.0000   1st Qu.: 0.00000   1st Qu.:0.0000
Median : 0.0000   Median : 0.00000   Median :0.0000
Mean   : 0.3137   Mean   : 0.05819   Mean   :0.1568
3rd Qu.: 0.0000   3rd Qu.: 0.00000   3rd Qu.:0.0000
Max.   :30.0000   Max.   :20.00000   Max.   :7.0000
```

Las estadísticas descriptivas de las muestras IEFIC 2017-2018 son:

```
   INGTOTOB            G_ING_SM
Min.   :        0   Min.   :0.000
1st Qu.:        0   1st Qu.:0.000
Median :   800000   Median :3.000
Mean   :  1087769   Mean   :2.146
3rd Qu.:  1254807   3rd Qu.:3.000
Max.   :100000000   Max.   :6.000

   P241edad          r_edad
Min.   : 18.00   Min.   :1.000
1st Qu.: 28.00   1st Qu.:2.000
Median : 41.00   Median :3.000
Mean   : 43.25   Mean   :3.211
3rd Qu.: 56.00   3rd Qu.:5.000
Max.   :108.00   Max.   :6.000

P232total_personas_h P2735_tot_h_en_vivienda
Min.   : 1.000       Min.   :1.00
1st Qu.: 2.000       1st Qu.:1.00
Median : 3.000       Median :1.00
Mean   : 2.939       Mean   :1.04
3rd Qu.: 4.000       3rd Qu.:1.00
Max.   :13.000       Max.   :7.00

P2439_propietario_hip P2461_precio_viv     P2462_subsidio_hip
Min.   :0.0000        Min.   :         0   Min.   :0.00000
1st Qu.:0.0000        1st Qu.:         0   1st Qu.:0.00000
Median :0.0000        Median :         0   Median :0.00000
Mean   :0.4385        Mean   :  50934738   Mean   :0.02227
3rd Qu.:1.0000        3rd Qu.:         0   3rd Qu.:0.00000
Max.   :1.0000        Max.   :5000000000   Max.   :1.00000

 Tiene_CFormal    Num_productos_Formal Credito_hipotecario
 Min.   :0.0000   Min.   :0.0000       Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000       1st Qu.:0.0000
 Median :0.0000   Median :0.0000       Median :0.0000
 Mean   :0.4368   Mean   :0.6516       Mean   :0.1433
 3rd Qu.:1.0000   3rd Qu.:1.0000       3rd Qu.:0.0000
 Max.   :1.0000   Max.   :5.0000       Max.   :1.0000
```

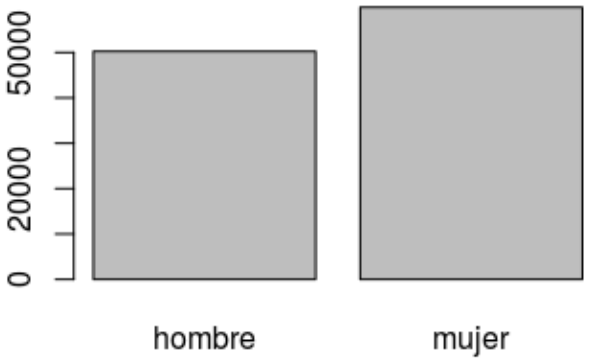
```
P2168_valor_hip      P2470_hip_act    P2986_cant_tdc
Min.   :         0   Min.   :0.0000   Min.   : 0.0000
1st Qu.:         0   1st Qu.:0.0000   1st Qu.: 0.0000
Median :         0   Median :0.0000   Median : 0.0000
Mean   :   5320167   Mean   :0.4313   Mean   : 0.3398
3rd Qu.:         0   3rd Qu.:0.0000   3rd Qu.: 0.0000
Max.   :2500000000   Max.   :4.0000   Max.   :10.0000

P354_tdc_actual      C_INF        cant_activos_financieros
Min.   :0.0000   Min.   :0.0000   Min.   :0.000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
Median :0.0000   Median :0.0000   Median :0.000
Mean   :0.6115   Mean   :0.0945   Mean   :0.318
3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.000
Max.   :4.0000   Max.   :1.0000   Max.   :7.000

P2976_reducir_endeuda   P2977_mora      P2978_cant_solicitudes
Min.   :0.0000        Min.   : 0.0000   Min.   : 0.0000
1st Qu.:0.0000        1st Qu.: 0.0000   1st Qu.: 0.0000
Median :0.0000        Median : 0.0000   Median : 0.0000
Mean   :0.1456        Mean   : 0.5057   Mean   : 0.3534
3rd Qu.:0.0000        3rd Qu.: 0.0000   3rd Qu.: 0.0000
Max.   :1.0000        Max.   :99.0000   Max.   :99.0000
                      NA's   :1314      NA's   :1314

P2979_rechazo_solicitudes
Min.   : 0.00
1st Qu.: 0.00
Median : 0.00
Mean   : 0.49
3rd Qu.: 0.00
Max.   :99.00
NA's   :98810
P2477_gasto_educacion_anual gasto_mensual_total
Min.   :        0           Min.   :       0
1st Qu.:        0           1st Qu.:       0
Median :        0           Median :       0
Mean   :  1714492           Mean   :  494327
3rd Qu.:   800000           3rd Qu.:  817000
Max.   :200000000           Max.   :33660000

P2476_televisor P2476_computador
Min.   :0.000   Min.   :0.000
1st Qu.:1.000   1st Qu.:0.000
Median :2.000   Median :1.000
Mean   :1.692   Mean   :0.653
3rd Qu.:2.000   3rd Qu.:1.000
Max.   :8.000   Max.   :8.000
```

```
P3035S1_mp_tarjetadebito P3035S2_transferenciasweb
Min.   :1                Min.   :1
1st Qu.:1                1st Qu.:1
Median :1                Median :1
Mean   :1                Mean   :1
3rd Qu.:1                3rd Qu.:1
Max.   :1                Max.   :1
NA's   :59660            NA's   :103877
P3035S3_pagos_moviles_telefono P3035S4_tarjetaprepago
Min.   :1                      Min.   :1
1st Qu.:1                      1st Qu.:1
Median :1                      Median :1
Mean   :1                      Mean   :1
3rd Qu.:1                      3rd Qu.:1
Max.   :1                      Max.   :1
NA's   :108260                 NA's   :109444

P3035S5_transporte P3035S6_cheque   tiene_telefono  P2478_Gastotelefono
Min.   :1          Min.   :1        Min.   :0.000   Min.   :      0
1st Qu.:1          1st Qu.:1        1st Qu.:1.000   1st Qu.:   6000
Median :1          Median :1        Median :1.000   Median :  30000
Mean   :1          Mean   :1        Mean   :0.786   Mean   :  48904
3rd Qu.:1          3rd Qu.:1        3rd Qu.:1.000   3rd Qu.:  61000
Max.   :1          Max.   :1        Max.   :1.000   Max.   :4219166
NA's   :61361      NA's   :109840
tiene_internet   P2478_Gasto_internet
Min.   :0.0000   Min.   :      0
1st Qu.:1.0000   1st Qu.:  22000
Median :1.0000   Median :  70000
Mean   :0.7957   Mean   :  70991
3rd Qu.:1.0000   3rd Qu.: 110000
Max.   :1.0000   Max.   :1115000
                 NA's   :16
```

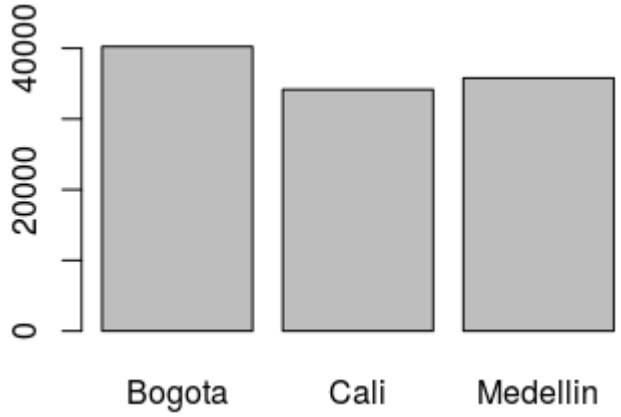
Al observar las medidas de tendencia central para las variables binarias, discretas y continuas, se observa que

Las variables categóricas se representan a continuación mediante tablas de frecuencia

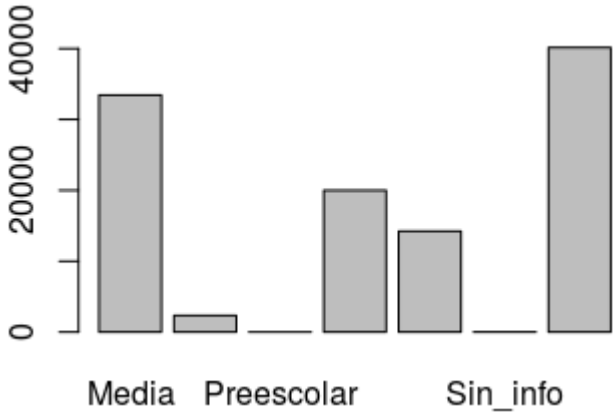
```
       Frequency Percent Cum Percent
hombre     50237   45.59       45.59
mujer      59948   54.41      100.00
Total     110185  100.00
```



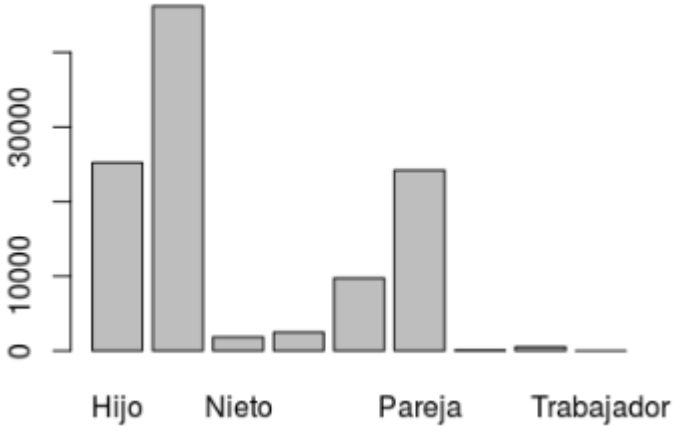
	Frequency	Percent	Cum Percent
Bogota	40237	36.52	36.52
Cali	34162	31.00	67.52
Medellin	35786	32.48	100.00
Total	110185	100.00	



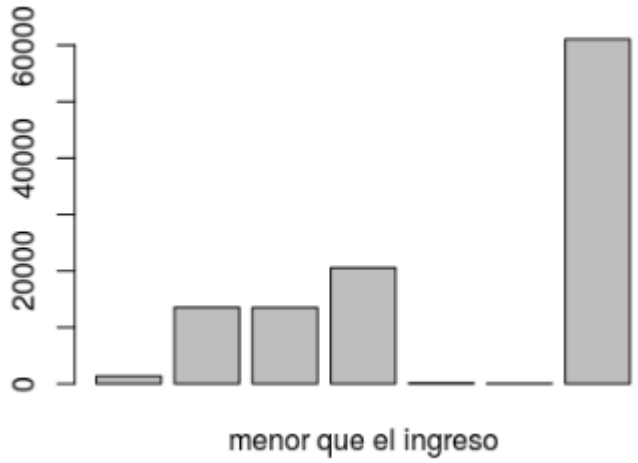
	Frequency	Percent	Cum Percent
Media	33444	3.035e+01	30.35
ninguno	2337	2.121e+00	32.47
Preescolar	8	7.261e-03	32.48
Primaria	19998	1.815e+01	50.63
Secundaria	14234	1.292e+01	63.55
Sin_info	23	2.087e-02	63.57
Universitaria	40141	3.643e+01	100.00
Total	110185	1.000e+02	



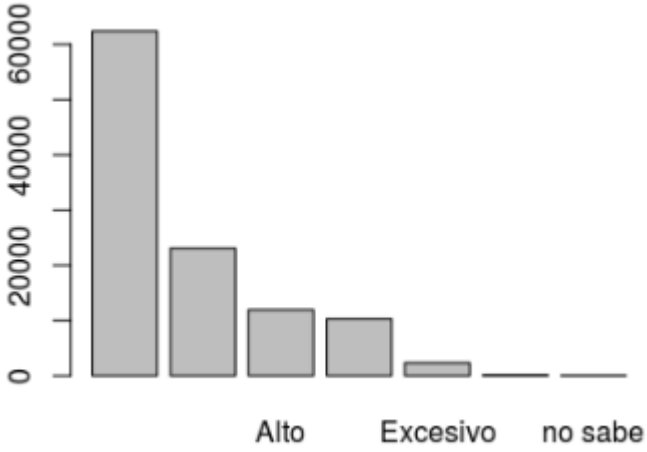
	Frequency	Percent	Cum Percent
Hijo	25207	2.288e+01	22.88
Jefe	46170	4.190e+01	64.78
Nieto	1817	1.649e+00	66.43
nopariente	2469	2.241e+00	68.67
Otropariente	9728	8.829e+00	77.50
Pareja	24183	2.195e+01	99.45
Pensionista	95	8.622e-02	99.53
Serdomestico	507	4.601e-01	99.99
Trabajador	9	8.168e-03	100.00
Total	110185	1.000e+02	



	Frequency	Percent	Cum Percent
	1314	1.19254	1.193
igual que el ingreso	13539	12.28752	13.480
mayores que el ingreso	13515	12.26573	25.746
menor que el ingreso	20544	18.64501	44.391
no informa	146	0.13250	44.523
no sabe	72	0.06534	44.589
no tiene deudas	61055	55.41135	100.000
Total	110185	100.00000	

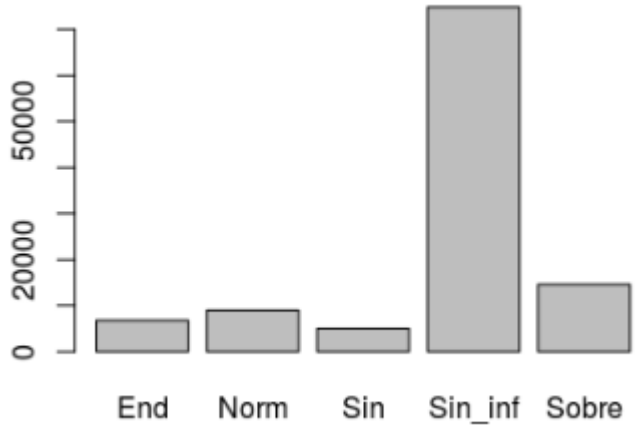


	Frequency	Percent	Cum Percent
	62369	56.60389	56.60
Adecuado	23084	20.95022	77.55
Alto	11910	10.80909	88.36
Bajo	10304	9.35155	97.71
Excesivo	2307	2.09375	99.81
no informa	146	0.13250	99.94
no sabe	65	0.05899	100.00
Total	110185	100.00000	

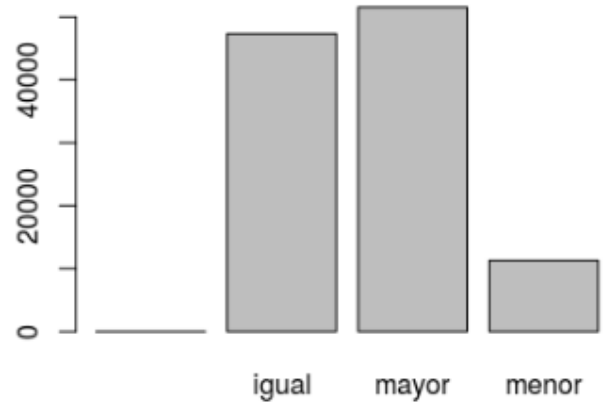


	Frequency	Percent	Cum Percent
	108437	98.41358	98.41
endeudamiento excesivo	255	0.23143	98.65
garantias	224	0.20329	98.85
ingreso suficiente	372	0.33761	99.19
no responde	26	0.02360	99.21
no sabe	34	0.03086	99.24
otro motivo	111	0.10074	99.34
reportes negativos	380	0.34487	99.69
sin motivo	204	0.18514	99.87
situacion laboral	142	0.12887	100.00
Total	110185	100.00000	

	Frequency	Percent	Cum Percent
End	6779	6.152	6.152
Norm	8982	8.152	14.304
Sin	5005	4.542	18.846
Sin_inf	74837	67.919	86.766
Sobre	14582	13.234	100.000
Total	110185	100.000	



	Frequency	Percent	Cum Percent
	16	0.01452	0.01452
igual	47326	42.95140	42.96592
mayor	51524	46.76136	89.72728
menor	11319	10.27272	100.00000
Total	110185	100.00000	

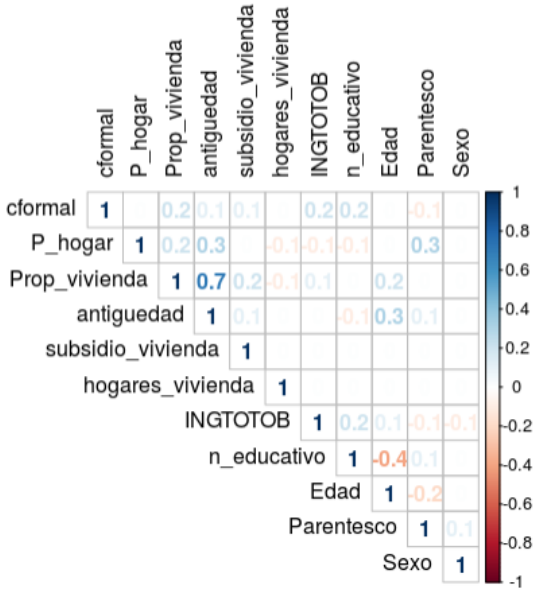


Anexo 4. Matrices de correlación.

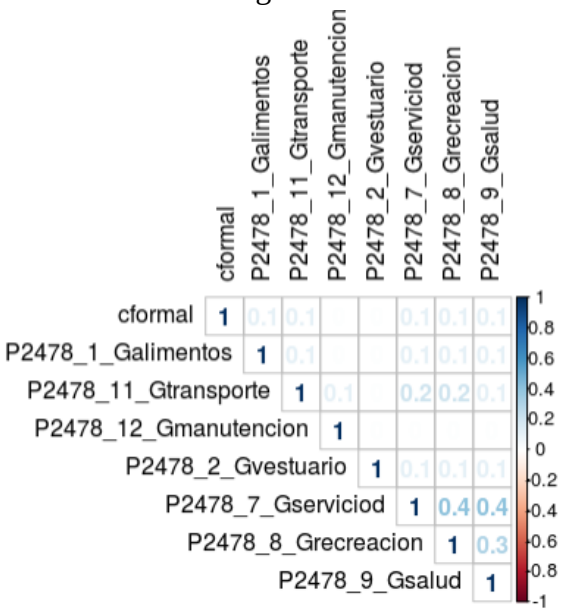
Variables significativas según matrices de correlación IEFIC 2010-2016

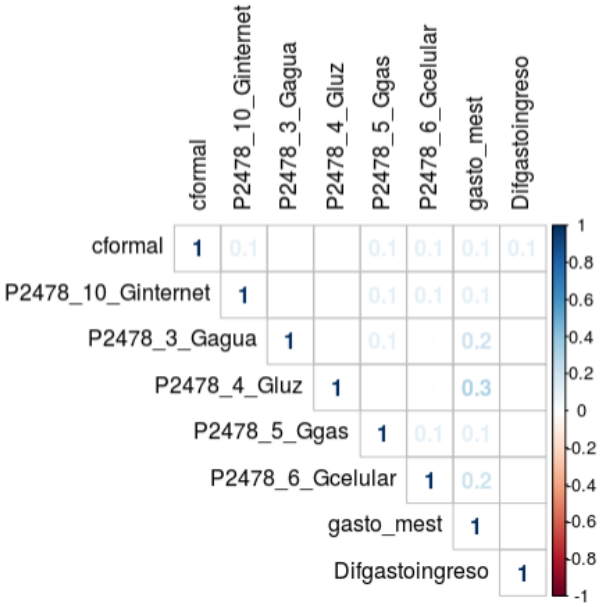
Nº	Variable	Índice de correlación
1	Calificaciónendeudamiento	0,5
2	Disminuye endeudamiento	0,5
3	Reduce endeudamiento	0,3
4	Ingreso	0,2
5	Nivel educativo	0,2
6	Propietario vivienda	0,2
7	Cuenta de ahorros	0,2
8	Computador	0,2
9	Televisor	0,2
10	Mora12m	0,2
11	Solicitudes de crédito	0,2
12	Antigüedad de la vivienda	0,1
13	Subsidio de vivienda	0,1
14	Acciones	0,1
15	Sociedades	0,1
16	Fondos de inversión	0,1
17	Cotizar pensión	0,1
18	Seguro	0,1
19	Gasto de internet	0,1
20	Gasto de celular	0,1
21	Gasto mes total	0,1
22	Diferencia gasto e ingreso	0,1
23	Gasto alimentos	0,1
24	Gasto transporte	0,1
25	Gasto servicio domestico	0,1
26	Gasto recreación	0,1
27	Gasto salud	0,1
28	Percepción endeudamiento	-0,3

Características del individuo y del hogar:

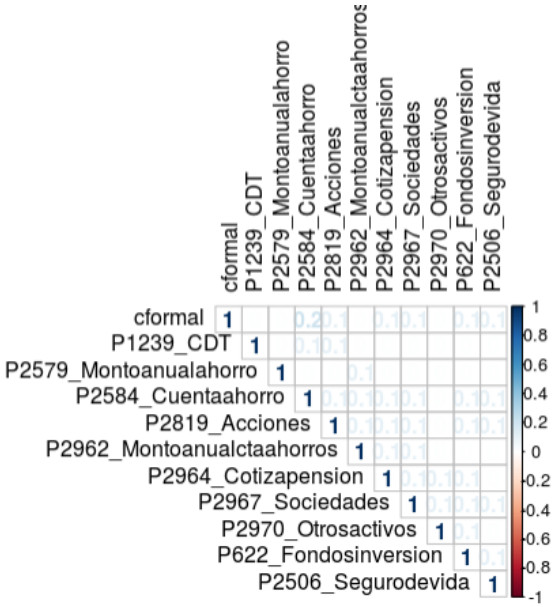


Variables de gasto:

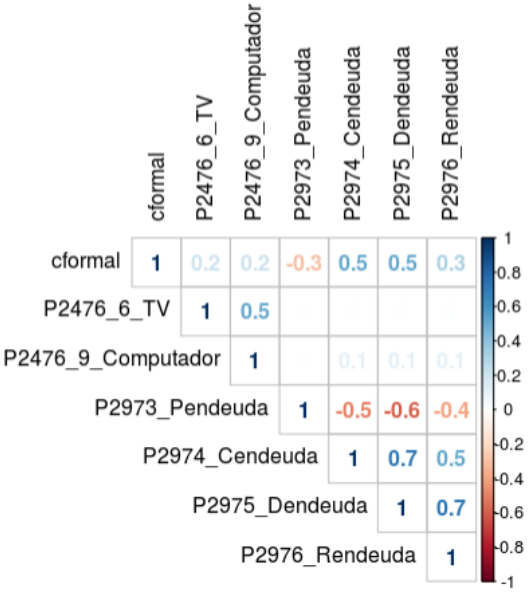




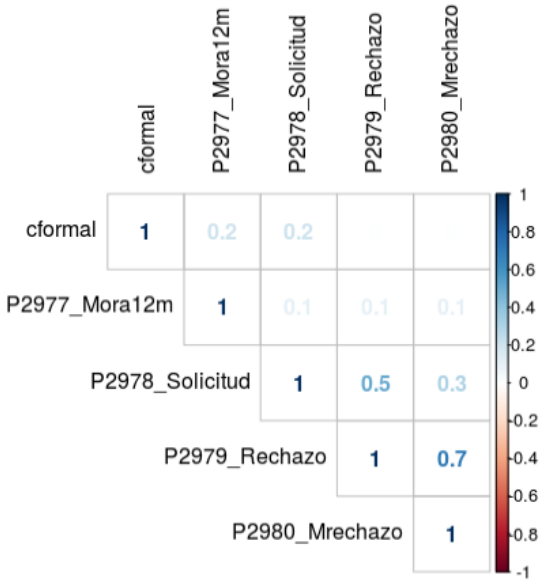
Variables de activos financieros



Variables de conectividad y endeudamiento:



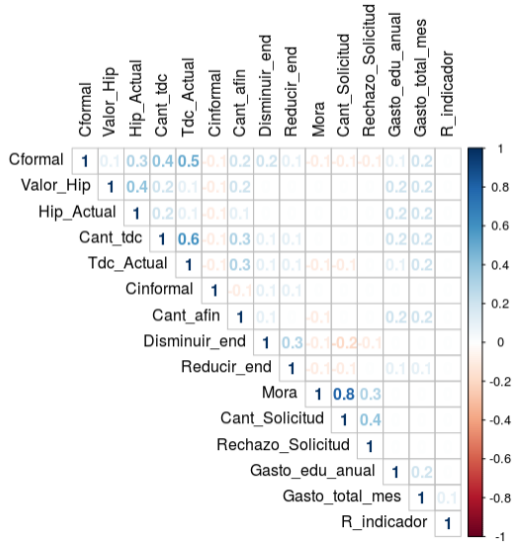
Variables de morosidad y solicitudes de crédito:

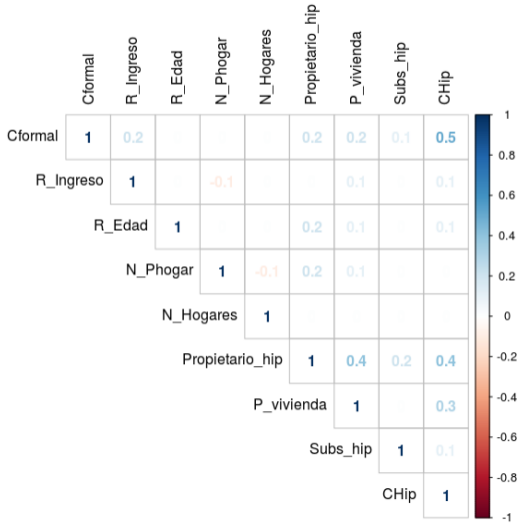


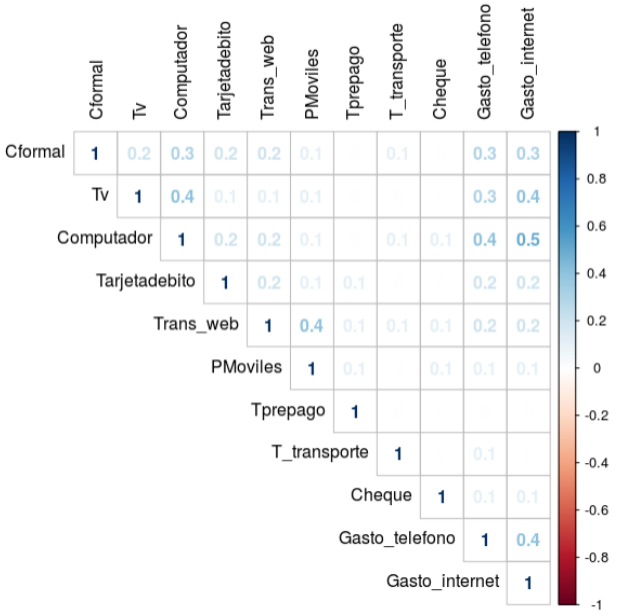
V

Variables significativas según matrices de correlación IEFIC 2017-2018

Variable	Índice de correlación
Gasto teléfono	0,3
Computador	0,3
R_Ingreso	0,2
Propietario Hip	0,2
P_Vivienda	0,2
Tarjetadebito	0,2
Transferencias web	0,2
Tv	0,2
Cant_afin	0,2
Gasto_total_mes	0,2







Anexo 5. Desarrollo de modelos predictivos

A5.1 Árbol de decisión 2010-2016

	CP	nsplit	rel error	xerror	xstd
1	1.149211e-01	0	1.0000000	1.0000000	0.004291871
2	1.147180e-01	1	0.8850789	0.9107373	0.004246210
3	1.435245e-02	2	0.7703608	0.7703947	0.004113410
4	5.839144e-03	4	0.7416559	0.7430100	0.004078315
5	3.266536e-03	6	0.7299777	0.7307224	0.004061543
6	3.046510e-03	8	0.7234446	0.7260849	0.004055045
7	2.990093e-03	9	0.7203981	0.7232753	0.004051064
8	1.658655e-03	14	0.7041162	0.7093629	0.004030847
9	1.049353e-03	18	0.6961952	0.7041500	0.004023054
10	9.816532e-04	21	0.6930472	0.7044208	0.004023462
11	8.970280e-04	22	0.6920655	0.7024914	0.004020550
12	8.631778e-04	25	0.6893575	0.7019498	0.004019729
13	7.785526e-04	27	0.6876312	0.7020852	0.004019934
14	7.447025e-04	28	0.6868526	0.7024237	0.004020447
15	6.770022e-04	31	0.6846185	0.7031345	0.004021522
16	6.093020e-04	32	0.6839415	0.7036761	0.004022340
17	4.908266e-04	33	0.6833322	0.7043870	0.004023411
18	4.739016e-04	35	0.6823506	0.7040146	0.004022850
19	4.231264e-04	41	0.6788301	0.7036423	0.004022289
20	3.723512e-04	43	0.6779839	0.7035069	0.004022084
21	3.131135e-04	48	0.6761221	0.7040485	0.004022901
22	3.046510e-04	52	0.6748697	0.7047255	0.004023920
23	2.708009e-04	54	0.6742604	0.7055717	0.004025191
24	2.031007e-04	58	0.6731772	0.7065534	0.004026662
25	1.861756e-04	60	0.6727710	0.7086182	0.004029741
26	1.692506e-04	64	0.6720263	0.7085844	0.004029691
27	1.421705e-04	79	0.6686074	0.7091937	0.004030596
28	1.354004e-04	85	0.6676935	0.7090244	0.004030344
29	1.015503e-04	99	0.6656963	0.7093291	0.004030797
30	9.308781e-05	106	0.6649854	0.7092952	0.004030746
31	9.026696e-05	110	0.6646131	0.7094645	0.004030997
32	8.462528e-05	113	0.6643423	0.7095322	0.004031098
33	6.770022e-05	117	0.6640038	0.7092614	0.004030696

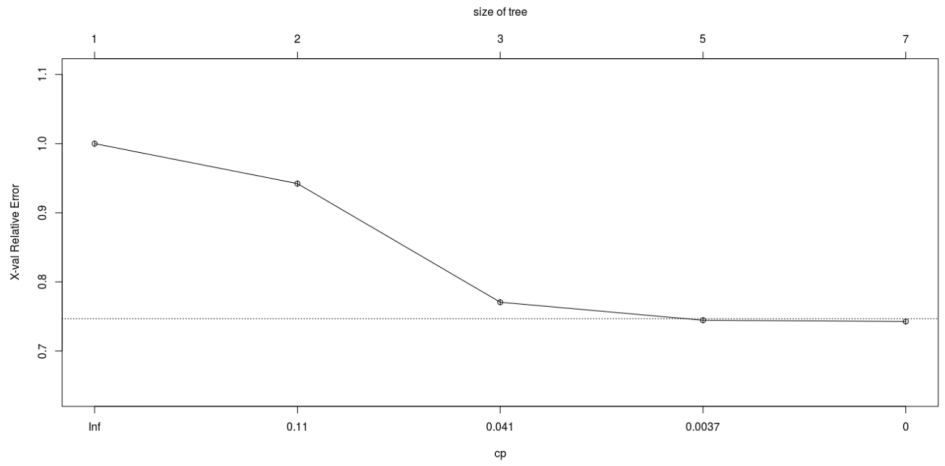
```
Call:
rpart(formula = cformal ~ ., data = train, method = "class",
    control = control.poda)
  n= 64809

          CP nsplit rel error    xerror        xstd
1 0.114921129      0 1.0000000 1.0000000 0.004291871
2 0.114718029      1 0.8850789 0.9530499 0.004271505
3 0.014352447      2 0.7703608 0.7709024 0.004114032
4 0.005839144      4 0.7416559 0.7491368 0.004086439
5 0.003362444      6 0.7299777 0.7294699 0.004059797
6 0.001658655     11 0.7127141 0.7228014 0.004050389
7 0.001320154     13 0.7093968 0.7175547 0.004042853
8 0.000000000     14 0.7080766 0.7164376 0.004041233

Variable importance
               Ingresomes                 antiguedad  P2439_Propietariovivienda         P2584_Cuentaahorro
                       22                         17                         16                          8
       P2478_10_Ginternet         P10_niveleducativo                 gasto_mest                P2476_6_TV
                        7                          6                          6                          3
       P232_Personashogar P2962_Montoanualctaahorros                  P241_Edad         P2476_9_Computador
                        3                          3                          2                          2
       P2478_7_Gserviciod         P2478_1_Galimentos     P2579_Montoanualahorro           P2478_6_Gcelular
                        1                          1                          1                          1
      P2478_8_Grecreacion            P242_Parentesco
                        1                          1
```

```
n= 64809

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 64809 29542 1 (0.4558318 0.5441682)
   2) Ingresomes< 1140697 44689 20647 0 (0.5379847 0.4620153)
     4) P2439_Propietariovivienda< 0.5 22576  7896 0 (0.6502481 0.3497519)
       8) P2584_Cuentaahorro< 0.5 12429  3281 0 (0.7360206 0.2639794) *
       9) P2584_Cuentaahorro>=0.5 10147  4615 0 (0.5451858 0.4548142)
        18) P2476_9_Computador< 0.5 4483  1755 0 (0.6085211 0.3914789) *
        19) P2476_9_Computador>=0.5 5664  2804 1 (0.4950565 0.5049435) *
     5) P2439_Propietariovivienda>=0.5 22113  9362 1 (0.4233709 0.5766291)
      10) P2478_10_Ginternet< 56400 9574  4699 0 (0.5091916 0.4908084)
        20) P2584_Cuentaahorro< 0.5 6072  2612 0 (0.5698287 0.4301713) *
        21) P2584_Cuentaahorro>=0.5 3502  1415 1 (0.4040548 0.5959452) *
      11) P2478_10_Ginternet>=56400 12539  4487 1 (0.3578435 0.6421565) *
   3) Ingresomes>=1140697 20120  5500 1 (0.2733598 0.7266402) *
```

Indicador del costo de complejidad: Identificación del número de árboles óptimo

El costo de complejidad del árbol permite identificar el número de nodos necesarios para obtener el error más bajo posible con el costo de complejidad cercano a cero.

```
             CP nsplit rel error    xerror        xstd
1 0.1149211292      0 1.0000000 1.0000000 0.004291871
2 0.1147180286      1 0.8850789 0.9422179 0.004265660
3 0.0143524474      2 0.7703608 0.7705978 0.004113659
4 0.0009478031      4 0.7416559 0.7445332 0.004080349
5 0.0000000000      6 0.7397603 0.7426715 0.004077861
```

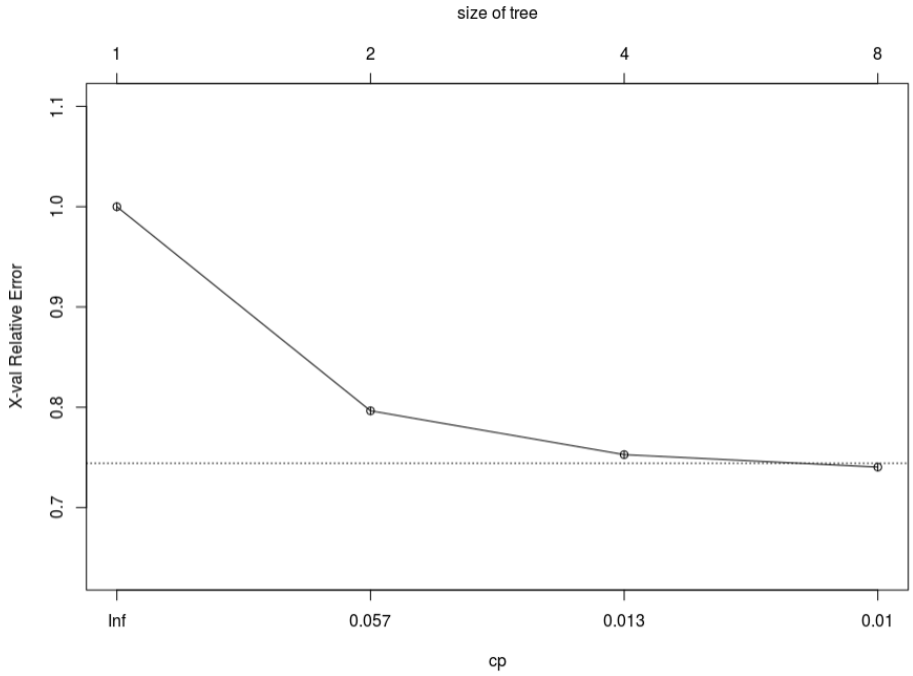


A5.2 Árbol de decisión 2017-2018

```
arbol_1 <- rpart(formula = Cformal ~., data = train, method = 'class')
rpart.plot(arbol_1, extra = 110)
```

```
n= 77130

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 77130 33604 0 (0.5643200 0.4356800)
   2) Gasto_telefono< 57500 53792 18497 0 (0.6561385 0.3438615)
     4) Tarjetadebito< 0.5 31957  8627 0 (0.7300435 0.2699565) *
     5) Tarjetadebito>=0.5 21835  9870 0 (0.5479734 0.4520266)
      10) Gasto_internet< 51600 9136  3178 0 (0.6521454 0.3478546) *
      11) Gasto_internet>=51600 12699  6007 1 (0.4730294 0.5269706)
        22) P_vivienda< 71000 10016  4836 0 (0.5171725 0.4828275)
          44) Ingreso< 1419206 6976  3045 0 (0.5635034 0.4364966) *
          45) Ingreso>=1419206 3040  1249 1 (0.4108553 0.5891447) *
        23) P_vivienda>=71000 2683   827 1 (0.3082370 0.6917630) *
   3) Gasto_telefono>=57500 23338  8231 1 (0.3526866 0.6473134)
     6) Tarjetadebito< 0.5 9790  4828 1 (0.4931563 0.5068437)
      12) Propietario_hip< 0.5 4703  1827 0 (0.6115246 0.3884754) *
      13) Propietario_hip>=0.5 5087  1952 1 (0.3837232 0.6162768) *
     7) Tarjetadebito>=0.5 13548  3403 1 (0.2511810 0.7488190) *
```

Indicador del costo de complejidad: Identificación del número de árboles óptimo



A5.3 Árbol de decisión condicional 2010-2016

```
Model formula:
Cformal ~ Departamento + Gasto_total_mes + P_vivienda + N_Phogar +
    Mora + Sexo + R_ging + Tarjetadebito + Ingreso + R_ging_cal +
    Parentesco + PMoviles + Motivo + Rechazo_Solicitud + Gasto_telefono +
    Tv + N_Hogares + Tprepago + Gasto_internet + T_transporte +
    Computador + R_Edu + Trans_web + Subs_hip + Cheque + R_indicador +
    Cant_afin + Edad + Cinformal + Gasto_edu_anual + Propietario_hip
```

```
Fitted party:
[1] root
|   [2] Gasto_telefono <= 57000
|   |   [3] Cant_afin <= 0
|   |   |   [4] Computador <= 0
|   |   |   |   [5] P_vivienda <= 70000: 0 (n = 21695, err = 21.4%)
|   |   |   |   [6] P_vivienda > 70000: 0 (n = 2218, err = 42.2%)
|   |   |   [7] Computador > 0
|   |   |   |   [8] Propietario_hip <= 0: 0 (n = 9814, err = 30.9%)
|   |   |   |   [9] Propietario_hip > 0: 0 (n = 8303, err = 47.0%)
|   |   [10] Cant_afin > 0
|   |   |   [11] Gasto_internet <= 57900
|   |   |   |   [12] Computador <= 0: 0 (n = 3635, err = 36.4%)
|   |   |   |   [13] Computador > 0: 1 (n = 1787, err = 49.0%)
|   |   |   [14] Gasto_internet > 57900
|   |   |   |   [15] Propietario_hip <= 0: 1 (n = 3171, err = 48.8%)
|   |   |   |   [16] Propietario_hip > 0: 1 (n = 3169, err = 33.0%)
|   [17] Gasto_telefono > 57000
|   |   [18] Tarjetadebito <= 0
|   |   |   [19] Propietario_hip <= 0
|   |   |   |   [20] Gasto_telefono <= 103000: 0 (n = 3325, err = 34.2%)
|   |   |   |   [21] Gasto_telefono > 103000: 0 (n = 1378, err = 50.0%)
|   |   |   [22] Propietario_hip > 0
|   |   |   |   [23] Gasto_internet <= 67300: 0 (n = 1184, err = 48.1%)
|   |   |   |   [24] Gasto_internet > 67300: 1 (n = 3903, err = 34.3%)
|   |   [25] Tarjetadebito > 0
|   |   |   [26] Gasto_internet <= 76000
|   |   |   |   [27] Ingreso <= 1475000: 1 (n = 2349, err = 45.7%)
|   |   |   |   [28] Ingreso > 1475000: 1 (n = 1120, err = 22.2%)
|   |   |   [29] Gasto_internet > 76000
|   |   |   |   [30] Propietario_hip <= 0: 1 (n = 4157, err = 28.7%)
|   |   |   |   [31] Propietario_hip > 0: 1 (n = 5922, err = 15.0%)

Number of inner nodes:    15
Number of terminal nodes: 16
```

A5.4 Árbol de decisión condicional 2017-2018

```
Model formula:
cformal ~ P2964_Cotizapension + P2476_6_TV + P2480_Gcompleto +
    P2695_Ccajacompensacion + P2962_Montoanualctaahorros + P10_niveleducativo +
    P2478_6_Gcelular + P2483_Tienemasviviendas + P2819_Acciones +
    P2771_debeamigos + P622_Fondosinversion + P2439_Propietariovivienda +
    cinformal + P1239_CDT + P2478_11_Gtransporte + P2967_Sociedades +
    P2633_debegota + P2579_Montoanualahorro + P2479_Nivelgasto +
    P232_Personashogar + P408_Pagomesseguro + Difgastoingreso +
    P241_Edad + gasto_mest + P35_Sexo + P2476_9_Computador +
    Ingresomes + P2478_8_Grecreacion + P2735_hogaresvivienda +
    P2478_1_Galimentos + antiguedad + P2478_10_Ginternet + P2478_7_Gserviciod +
    P2692_debetienda + P2478_9_Gsalud + P2506_Segurodevida +
    P2548_debecasacom + P2584_Cuentaahorro + P2462_subsidiovivienda +
    P242_Parentesco + P2970_Otrosactivos
```

```
Fitted party:
[1] root
|   [2] P2584_Cuentaahorro <= 0
|   |   [3] P2439_Propietariovivienda <= 0
|   |   |   [4] P2478_6_Gcelular <= 59000
|   |   |   |   [5] P242_Parentesco <= 2
|   |   |   |   |   [6] P2476_9_Computador <= 0: 0 (n = 4310, err = 26.5%)
|   |   |   |   |   [7] P2476_9_Computador > 0: 0 (n = 3079, err = 36.8%)
|   |   |   |   [8] P242_Parentesco > 2: 0 (n = 3750, err = 15.4%)
|   |   |   [9] P2478_6_Gcelular > 59000: 0 (n = 3337, err = 38.4%)
|   |   [10] P2439_Propietariovivienda > 0
|   |   |   [11] P2478_10_Ginternet <= 56000
|   |   |   |   [12] antiguedad <= 18: 1 (n = 3403, err = 47.9%)
|   |   |   |   [13] antiguedad > 18: 0 (n = 3469, err = 37.2%)
|   |   |   [14] P2478_10_Ginternet > 56000
|   |   |   |   [15] antiguedad <= 19: 1 (n = 4648, err = 32.5%)
|   |   |   |   [16] antiguedad > 19: 1 (n = 3806, err = 45.6%)
|   [17] P2584_Cuentaahorro > 0
|   |   [18] P2439_Propietariovivienda <= 0
|   |   |   [19] P2478_6_Gcelular <= 63000
|   |   |   |   [20] P2476_9_Computador <= 0: 0 (n = 5285, err = 41.7%)
|   |   |   |   [21] P2476_9_Computador > 0
|   |   |   |   |   [22] Ingresomes <= 1002700: 0 (n = 3433, err = 45.6%)
|   |   |   |   |   [23] Ingresomes > 1002700: 1 (n = 3210, err = 35.5%)
|   |   |   [24] P2478_6_Gcelular > 63000: 1 (n = 5219, err = 31.6%)
|   |   [25] P2439_Propietariovivienda > 0
|   |   |   [26] P10_niveleducativo <= 602
|   |   |   |   [27] antiguedad <= 19: 1 (n = 5338, err = 26.7%)
|   |   |   |   [28] antiguedad > 19: 1 (n = 4342, err = 38.0%)
|   |   |   [29] P10_niveleducativo > 602
|   |   |   |   [30] Ingresomes <= 1283333.33333: 1 (n = 3005, err = 24.7%)
|   |   |   |   [31] Ingresomes > 1283333.33333: 1 (n = 5175, err = 11.3%)
```

A5.5 Bosque aleatorio IEFIC 2017-2018

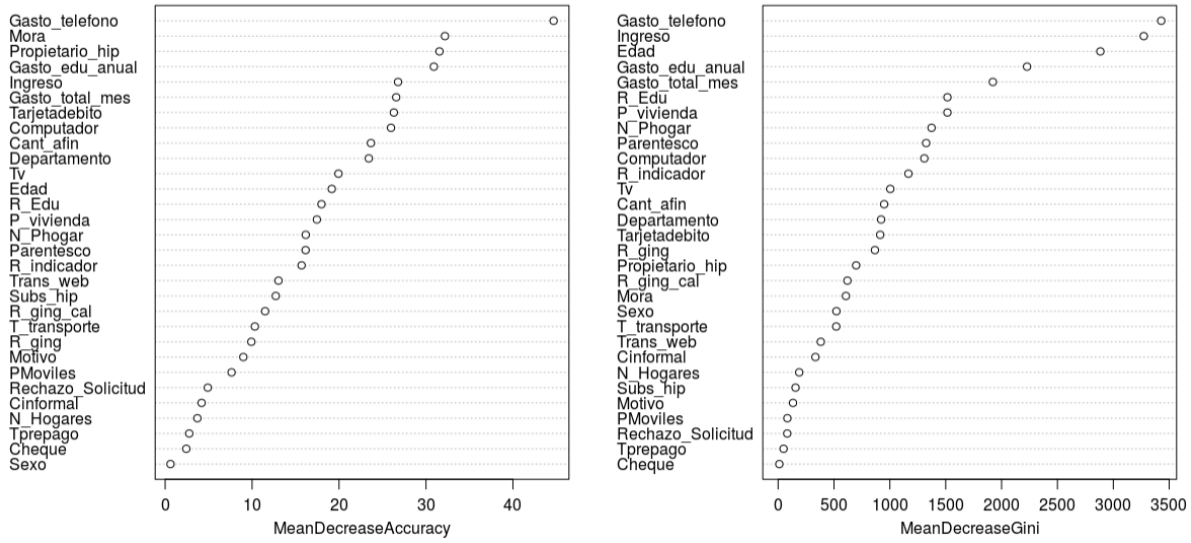
```
#####################
# Random Forest
#####################

install.packages("randomForest")
library("randomForest")


mrf <- randomForest(Cformal ~ ., data = train, importance = TRUE, prOximity=TRUE,
                    na.action=na.roughfix)


mrf <- randomForest(Cformal ~ ., data = train, ntree = 50, mtry = 6, importance = TRUE, prOximity=TRUE,
                    na.action=na.roughfix)
```

```
Call:
 randomForest(formula = Cformal ~ ., data = train, ntree = 50,      mtry = 6, importance = TRUE, prOxi
mity = TRUE, na.action = na.roughfix)
               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 6

        OOB estimate of  error rate: 27.72%
Confusion matrix:
      0     1 class.error
0 35035  8491   0.1950788
1 12891 20713   0.3836150
```

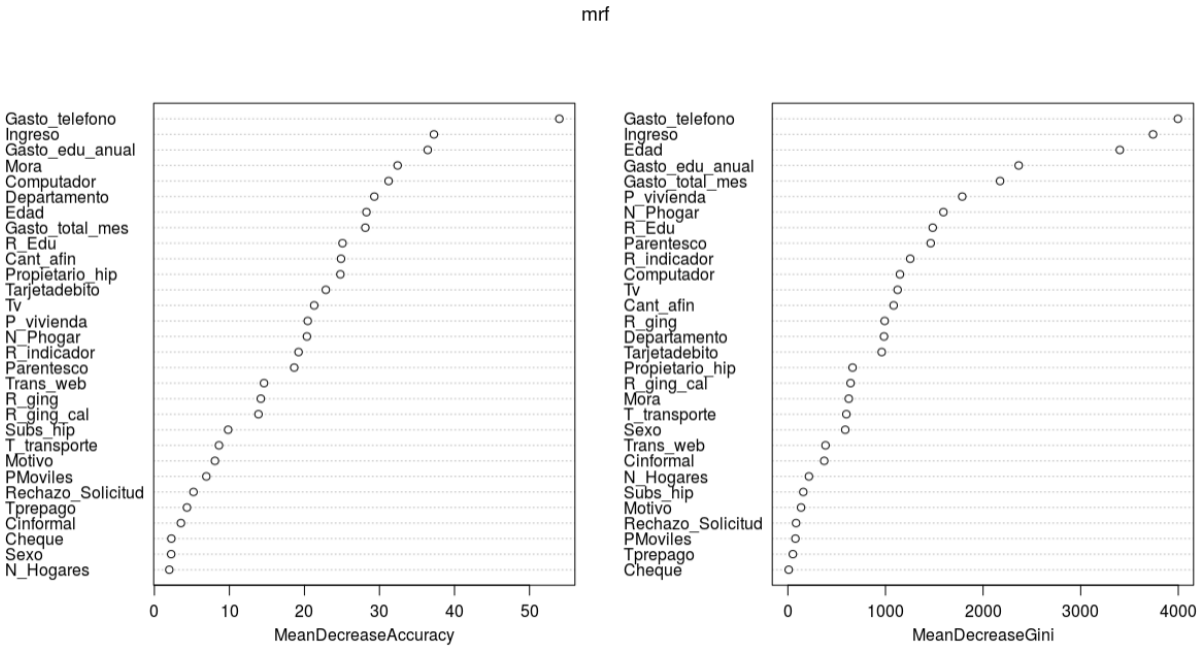
```
Call:
 randomForest(formula = Cformal ~ ., data = train, ntree = 55,      mtry = 4, importance = TRUE, prOxi
mity = TRUE, na.action = na.roughfix)
               Type of random forest: classification
                     Number of trees: 55
No. of variables tried at each split: 4

        OOB estimate of  error rate: 28.06%
Confusion matrix:
      0     1 class.error
0 35518  8008   0.1839820
1 13631 19973   0.4056362
```

```
Call:
 randomForest(formula = Cformal ~ ., data = train, ntree = 60,      mtry = 5, importance = TRUE, prOxi
mity = TRUE, na.action = na.roughfix)
               Type of random forest: classification
                     Number of trees: 60
No. of variables tried at each split: 5

        OOB estimate of  error rate: 27.72%
Confusion matrix:
      0     1 class.error
0 35408  8118   0.1865092
1 13264 20340   0.3947149
```

mrf



Mrt=55


mrf
Gasto_telefono
Ingreso
Gasto_edu_anual
Mora
Computador
Departamento
Edad
Gasto_total_mes
R_Edu
Cant_afin
Propietario_hip
Tarjetadebito
Tv
P_vivienda
N_Phogar
R_indicador
Parentesco
Trans_web
R_ging
R_ging_cal
Subs_hip
T_transporte
Motivo
PMoviles
Rechazo_Solicitud
Tprepago
Cinformal
Cheque
Sexo
N_Hogares
0
10
20
30
40
50
MeanDecreaseAccuracy
Gasto_telefono
Ingreso
Edad
Gasto_edu_anual
Gasto_total_mes
P_vivienda
N_Phogar
R_Edu
Parentesco
R_indicador
Computador
Tv
Cant_afin
R_ging
Departamento
Tarjetadebito
Propietario_hip
R_ging_cal
Mora
T_transporte
Sexo
Trans_web
Cinformal
N_Hogares
Subs_hip
Motivo
Rechazo_Solicitud
PMoviles
Tprepago
Cheque
0
1000
2000
3000
4000
MeanDecreaseGini

Mrt=60