Trabajo Práctico Nº 1. Aprendizaje Automático

Detección del perfil de riesgo crediticio de clientes tomadores de préstamos personales en una entidad financiera, con árboles de decisión.

Comisión II

Grupo Nº: 18

MORENO, Federico Nicolás

PERINI, Sofía Clara

PICCHETTI, Bianca

Buenos Aires, 30 de Mayo de 2019

Facultad de Ciencias Exactas y Naturales

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Aprendizaje Automático

Maestría en Explotación de Datos
y Descubrimiento del Conocimiento
Universidad de Buenos Aires

Resumen

Se presentan en el siguiente informe los resultados obtenidos durante el proceso de desarrollo de un modelo predictivo para una entidad financiera, que busca clasificar a sus potenciales clientes tomadores de préstamos en “cumplidores” o “deudores” de acuerdo a determinadas características personales, condiciones del crédito e información de contacto brindada. Para el aprendizaje automático del modelo se utiliza un conjunto de datos que consta de ejemplos de clientes de ambas clases.

Luego de efectuar pruebas con diversos modelos y técnicas de aprendizaje automático supervisado, se obtuvo el más adecuado para clasificar el perfil del cliente: un árbol de decisión de altura 8 calculado con *InformationGain*. Este último fue el que mejor desempeño demostró en los análisis y validaciones realizadas.

Basándose en este mejor modelo se pudo concluir que la variable con más peso a la hora de clasificar el comportamiento de pago de un cliente es la cantidad de cuotas del préstamo. Le siguen en importancia el interés mensual y la edad, y posteriormente la zona de residencia y el ingreso formal del cliente. Contrario a lo esperado (hipótesis), no se observa un gran peso para la clasificación de las características personales del cliente, como género, puesto laboralo información de contacto.

En particular, se observa que los clientes que solicitan préstamos cuya cantidad de cuotas es menor a un año suelen ser pagadores mientras que los perfiles deudores se corresponden con préstamos a devolver a más de un año de plazo. Un comportamiento deudor podría estar asociado a la demanda de un crédito a un plazo mayor al promedio y por un monto menor o igual al promedio, lo que podría implicar que se trata de una persona con menor posibilidad / disposición a pagar mensualmente.

Por último, se proponen algunas líneas de investigación a futuro, que incluyen análisis más específicos, enfocadosa características personales del cliente y/o de las condiciones del préstamo, por separado.

1. Introducción

Uno de los principales inconvenientes que suelen enfrentar las empresas financieras ala hora de otorgar préstamos a sus clientes es el riesgo crediticio. Este último consiste en “el riesgo de pérdida de capital causado por la falta de pago en tiempo y forma por parte de un acreedor” (Econlink, 2014).Este incumplimiento de las obligaciones crediticias, ya sea por falta total de pago, pago parcial o pago fuera de término, es una preocupaciónfundamental de las entidades financieras dado que ocasiona pérdidas indeseadas. Es por ello, que habitualmente los bancos y demás instituciones financieras suelen hacer exhaustivos análisis de susclientes y su capacidad de pago al momento de otorgar créditos y definir los montos a otorgar a cada uno.

1.1. Objetivos

En este sentido, se propone como objetivo del presente trabajo identificar el perfil de cliente deudor o cumplidor de una pequeña empresa financiera. Se intentará predecir si un cliente supone

un riesgo crediticio (deudor) o no para dicha empresa,en base a una serie de datos vinculados a las características del crédito otorgado, como también atributos personales del cliente, entre otros.

Según la política de la empresa financiera objeto de estudio, los clientes corresponden a personas que trabajan en relación de dependencia con más de un año de antigüedad y que al momento de requerir un préstamo personal no poseen embargos judiciales ni informes en el veraz. Es decir, la entidad toma clientes que al momento de otorgarles crédito no presentan mora con otras instituciones financieras. Todos ellos deben pasar por un filtro de solvencia y la cuota a otorgar no puede superar el 30% del recibo de sueldo.

La predicción del perfil de cliente (deudor o cumplidor) será la función objetivo que guiará esta investigación. Un cliente se considera deudor cuando se encuentra en situación de mora. Esta última se define como el retraso o la falta de pago con respecto a las obligaciones fijadas contractualmente. De acuerdo con la política de la empresa, se considera retraso a partir de los 160 días del vencimiento de una cuota.

De tal forma, el trabajo se encuentra dividido en tres secciones, la primera corresponde a la presente introducción e incluye el objetivo del trabajo, la preparación de los datos y la metodología de investigación. Se determina lo que se va a investigar, incluyendo una serie de hipótesis acerca de los atributos que orientarán el análisis a los efectos de encontrar los determinantes del riesgo crediticio (cliente deudor).

En la segunda sección, se detallarán los principales resultados de la investigación, producto de la aplicación de las técnicas indicadas en la metodología, comparando la *Accuracy* (exactitud) de cada una para lograr el objetivo buscado.

En la última parte, se hará hincapié en las principales conclusiones, corroborando el cumplimiento (o no) de las hipótesis planteadas sobre los atributos a fin de contribuir a discernir cuáles son los determinantes del riesgo crediticio que ayudaron a entrenar el modelo para predecir los perfiles de clientes deudores.

En este sentido, se plantean las siguientes hipótesis de investigación vinculadas a características personales del cliente o préstamo, y que se intentarán corroborar a lo largo del presente trabajo:

- Las mujeres representan menor riesgo crediticio que los hombres.
- El barrio en el que residen los clientes tiene incidencia en la capacidad de pago.
- Los préstamos con mayores intereses no necesariamente son los que presentan más mora.
- Si el cliente brinda más información de contacto (varios teléfonos, mail, datos del jefe), existe menos probabilidad de que sea moroso. Se supone que teniendo mayor cantidad de información de contacto es más difícil que incumpla su obligación de pago.
- Los clientes que trabajan de “porteros” tienen mayor probabilidad de cumplir sus obligaciones que los demás con otros puestos de trabajo.

1.2. Selección y preparación de Datos

Los análisis llevados a cabo en la presente investigación se realizan a partir de una base de datos de alrededor de 1300 observaciones (casos, ejemplos) correspondientes a préstamos otorgados a 903 clientes durante el período 2016-2018, propiedad de la empresa financiera objeto de

estudio. Cabe aclarar que los clientes de la base de datos se clasifican como deudores si cumplen la condición de morosidad (citada previamente) a Abril del 2019, fecha en que se inició este trabajo.

El *dataset* original cuenta con 40 atributos numéricos y categóricos correspondientes a información sobre las condiciones del préstamo (monto, cuotas, interés, fecha) y características del cliente (ingresos, edad, sexo, zona de residencia, puesto laboral, información de contacto), entre otros. No obstante, previo al procesamiento de la base, debieron efectuarse una serie de transformaciones, a fin de preparar/acondicionar los datos para el análisis, entre los que cabe mencionar:

Anonimización de datos

Se eliminaron los datos de nombres y direcciones de los clientes para mantener la anonimidad de los individuos.

Ingeniería de atributos

Se optó por reducir el número de atributos a considerar ya que la base cuenta con 40 atributos, muchos de los cuales son dependientes entre sí. En tal sentido, se recortaron aquellas variables que estuvieran altamente correlacionadas (ver comparación de la correlación previa y posterior a la eliminación de los atributos con alta correlación, en los *heatmaps* de las *Figuras A1 y A2* del Anexo II), aquellas que no aportaran información adicional al análisis y todas las relacionadas con el comportamiento de pago dado que son redundantes con nuestra variable objetivo. Se muestran en la *Tabla A1* del Anexo I las variables originales y el motivo de su eliminación.

Con respecto al atributo “Fecha de Entrega” del préstamo, se excluyó del análisis el año 2019, ya que no parece ser comparable con los años anteriores, por tratarse de un período incompleto (solo existe información sobre los primeros meses) que podría introducir ruido.

Dada la existencia de múltiples clases en algunos de los atributos categóricos (puesto laboral específico y localidad de residencia), se agruparon por clases más generales que contienen varias de las categorías específicas (puesto laboral genérico y zonas de residencia) a fin de facilitar el análisis y posterior empleo de método de aprendizaje automático.

Algunos atributos numéricos o categóricos se convirtieron en binarios, como por ejemplo si el cliente aportó a la institución financiera (o no) información personal como su teléfono celular o mail para poder contactarse. En estos casos, lo importante es saber si existe información de contacto adicional y no específicamente el número o dirección de que se trata.

Se eliminó un conjunto de casos correspondientes a un cliente particular con un comportamiento atípico, por considerarse *outlier* según recomendación de un experto en la materia.

Luego de efectuar esta ingeniería de atributos, la base nueva consiste en un set de 1278 observaciones y 18 atributos, 12 de los cuales resultan categóricos (incluye variables binarias) y 6 numéricos. La lista de variables presentes en el *dataset* de trabajo se observan en la *Tabla A1*.

Detección de datos faltantes

Durante el análisis exploratorio del *dataset* se descubrió que existen atributos con datos faltantes, por tal motivo se hicieron algunas pruebas con un árbol de altura 3 para ver con cuál base convenía trabajar. En una primera prueba se eliminaron todos los casos con datos faltantes y en una segunda instancia, se examinó la posibilidad de imputar la moda para el atributo "Sucursal", que era el que mayor porcentaje de datos faltantes tenía (11%) y se eliminaron sólo los restantes.

En la primera prueba, se eliminan aproximadamente 25% de los datos (quedando 979 registros) y en la segunda sólo un 15% (1101 registros). Al comparar ambos experimentos, se observó que el segundo caso tenía una *Accuracy* levemente mayor, pero principalmente se notó que la cla-

sificación para "deudores" (clasificación de mayor interés para la investigación) tiene *precision* y *recall* más altas, por tanto *f1-score* más alto también que en el primer caso, es decir, que este segundo modelo clasifica mejor el target sobre el conjunto de datos originales, motivo por el cual se definió que este último *dataset* era el más apropiado para trabajar.

Adicionalmente, pudo notarse que mientras en el primer árbol el "puesto laboral" era significativo para explicar la clasificación del *target*, es decir, se encontraba entre los nodos de decisión más altos del árbol; en el segundo caso, donde se imputó la moda en sucursales, esa variable pasó a ser insignificante, ni siquiera aparece en el árbol. Ahora bien, los atributos interés mensual (clasificador principal), cantidad de cuotas, ingreso formal, edad y teléfono particular de contacto permanecen en ambos modelos como clasificadores relevantes de la función objetivo.

En las *Figuras A3 y A4* del Anexo II se muestran los diagramas de los árboles mencionados.

Balanceo de datos

Al efectuar las pruebas anteriores se advirtió que la clasificación de deudores en e l*dataset* parecía presentar *recall* baja (aunque resultaba mayor que en la primera prueba efectuada) mientras los pagadores tenían *recall* alta. En este sentido, se estudió si eso respondía a que la base de datos estaba desbalanceada, y se pudo notar que existíaun 75% más de pagadores que de deudores y, por lo tanto, la muestra estaba desbalanceada. Sólo un 36,6% de los datos correspondían a la clase deudor (403 casos) y un 63,4% eran clientes pagadores o cumplidores (698 casos).

Dado que trabajar con un conjunto de entrenamiento desbalanceado implica que se optimizarán los resultados para el grupo de individuos con mayor frecuencia en el conjunto, en este caso los clientes pagadores, se procedió al balanceo del conjunto de datos. Para ello, se recortó una muestra aleatoria del conjunto de datos de la clase de mayor frecuencia, para que el conjunto de datos resultante con el que se entrenara el modelo estuviera balanceado y los resultados se optimicen para ambas clases.

Luego de balancear el *dataset*, si bien la *Accuracy* se redujo, mejoraron tanto el *recall* como la precisión y *f1-score* para los deudores (*target*), reduciéndose tales indicadores para la clase de clientes pagadores. Lo cual indica que en el anterior conjunto de datos realmente se encontraban los resultados optimizados para los pagadores, y en este nuevo balanceado, los resultados son óptimos para ambos.

El árbol resultante del balanceo cambió con respecto a los anteriores (*Figura A5* del Anexo II). Mientras que en el *dataset* desbalanceado el principal clasificador era "Interés Mensual", al balancear el conjunto de datos pasó a ser "Cantidad de Cuotas". En los niveles siguientes se observan características personales del cliente (como ingreso formal, edad, puesto laboral y teléfonos) y características del crédito (interés mensual, capital entregado). La principal y más evidente diferencia que se observa en el diagrama es lo mencionado previamente: está equilibrada la clasificación entre cumplidor y deudor, tal como mostraron las métricas calculadas.

Por lo expuesto, el *dataset* final que se utilizó para la experimentación con distintas metodologías de inferencia inductiva, cuenta con 806 casos correspondientes a 18 atributos, de los cuales exactamente la mitad (ejemplos) corresponden a cada categoría del *target*: deudor y pagador. No obstante, cabe aclarar que hasta este momento los conjuntos de datos y pruebas efectuadas se utilizaron sólo para limpiar y balancear la base de datos con que se busca trabajar en la presente investigación. Las etapas de partición, experimentación y evaluación que se detallan en la metodología a continuación, se efectúan a partir de este *dataset* final en la búsqueda del mejor modelo para predecir la variable objetivo (el cliente deudor).

1.3. Metodología de investigación

De acuerdo con lo detallado previamente, se dispone de información de la calificación de clientes cumplidores y deudores (no riesgosos y riesgosos), en función de si han cumplido sus obligaciones crediticias o no, según la definición de mora previamente mencionada. Con estos datos, se puede emplear un método de aprendizaje automático supervisado para la inferencia inductiva de aquellos que potencialmente podrían ser riesgosos. La construcción del modelo se realiza en un contexto de aprendizaje supervisado ya que se dispone de un conjunto de datos anotados donde se conocen tanto los datos como su clasificación.

Por tanto, como parte de la etapa de experimentación se utilizaron dos algoritmos distintos que permiten analizar el caso: por un lado, se emplearon Árboles de Decisión y por otro, Naive Bayes. El primer método se ha presentado en varias investigaciones como adecuado para la predicción de probabilidades de incumplimiento crediticio, por su capacidad de discriminación, estabilidad en el tiempo y por tratarse de una herramienta de fácil entendimiento. En este caso, se plantea como herramienta para la gestión del riesgo crediticio que impida llevar a la institución financiera a situaciones de insolvencia (Cardona Hernández, 2004).

En efecto, luego de particionar los datos en conjuntos de entrenamiento, validación y test, se experimentó con distintos hiperparámetros para el primer algoritmo. Las proporciones que se utilizaron para particionar el conjunto de desarrollo y test, y luego entrenamiento y validación fueron 80% y 20%, respectivamente.

Utilizando la técnica de *RandomSearch*, se hicieron pruebas con varias alturas de árbol, y distintas medidas de efectividad de los atributos para clasificar a las instancias (*Gini, InformationGain*). Asimismo, se analizó la presencia de datos faltantes, y se procedió a su imputación cuando fue necesario. Por último, se analizó la tolerancia al ruido del modelo, induciéndolo con una función para distintos porcentajes de ruido. Una vez finalizada la etapa de experimentación y encontrado el modelo más adecuado para aproximar la función objetivo se procedió al testeo del mismo. Se utilizaron distintas medidas de performance como ROC AUC, Cross Validation y *Accuracy* para seleccionar el modelo más adecuado.

Adicionalmente, se utilizó la técnica *GridSearch* para definir las mejores condiciones para un árbol de decisión. Los hiperparámetros buscados fueron profundidad del árbol y criterio de clasificación; este modelo se comparó con los previamente obtenidos y se testeó con el conjunto de datos separado para tal fin.

Por último, se contrastó el resultado de los mejores algoritmos resultantes de las técnicas planteadas con el de Naive Bayes, siempre en la búsqueda del mejor método para clasificar la función objetivo y analizar los atributos que puedan explicar mejor la función objetivo.

2. Resultados

2.1. Etapa de Experimentación

Con el objeto de avanzar en la selección del mejor algoritmo para clasificar el *target* (perfil del cliente), se detallan a continuación algunos de los principales resultados producto de experimentos con distintos algoritmos, parámetros y datos (inducción de ruido o datos faltantes).

2.1.1. Árboles de decisión

2.1.1.1. RandomSearch

Para encontrar la mejor combinación de atributos, algoritmos e hiperparámetros, se explora un espacio de búsqueda, usando varias técnicas para medir el desempeño de cada combinación. En este caso se explorarán opciones y combinaciones al azar. Al terminar, se elegirá la combinación con mejor desempeño, y se utilizará el mejor modelo de esta etapa para realizar otras pruebas (introducción de ruido o datos faltantes).

Prueba con distintas medidas de performance

Luego de entrenar un árbol de decisión con altura 3 y el resto de los hiperparámetros con su valor en default, se estimó la performance del modelo utilizando 5-*fold Cross Validation*, y se calcularon medidas de performance para cada *Fold* sobre el conjunto de entrenamiento y el de validación.

La capacidad de clasificación del algoritmo se refiere a su habilidad para distinguir clientes deudores y cumplidores, en este caso se utilizaron dos métricas distintas para evaluar esa performance. Por un lado, la *Accuracy* que indica la proporción de datos que se clasificó correctamente con el modelo, tanto de deudores como cumplidores; y por otro, la curva ROC que mide la sensibilidad (ratio de verdaderos positivos) frente a la especificidad (ratio de falsos positivos) para un sistema clasificador según se varía el umbral de discriminación, representada por el área bajo la curva (AUC, por sus siglas en inglés).

Como se puede notar en la *Tabla 1*, con ambas métricas los resultados de desempeño (capacidad de clasificación) del modelo son mayores para el conjunto de entrenamiento e inferiores para el conjunto de validación, en todos los casos. Asimismo, se observa cómo va modificándose la performance en cada *Fold* ya que se toman distintas porciones del conjunto de datos para validación y entrenamiento. En promedio, la medida ROC AUC devuelve valores más elevados de performance que la *Accuracy* y se aprecia que los desvíos son mayores para los conjuntos de validación que para los de entrenamiento.

Tabla 1: Comparación de métricas por *Fold*, promedio y desvío estándar.

Folds	Accuracy Training	Accuracy Validation	ROC AUC Training	ROC AUC Validation
Fold 1	0.641	0.620	0.638	0.634
Fold 2	0.650	0.581	0.650	0.650
Fold 3	0.641	0.504	0.705	0.571
Fold 4	0.637	0.566	0.637	0.562
Fold 5	0.640	0.625	0.641	0.621
Promedio	0.642	0.579	0.654	0.595
Desvío Estándar	0.005	0.044	0.026	0.035

Fuente: Elaboración propia.

Prueba con distintos hiperparámetros

Adicionalmente, se estudió la eficacia del modelo alterando sus hiperparámetros: por un lado, se modificaron las alturas del árbol de decisión y por otro, las medidas de impureza y efectividad de los atributos para clasificar el target.

Según las pruebas realizadas, se pudo comprobar que el modelo que mejor clasifica la función objetivo es el árbol de decisión de altura 3. Tanto utilizando como medida de impureza y efectividad al Índice de *Gini* y *Gini Gain*, respectivamente, como Entropía e *InformationGain*, la performance es similar. En este sentido, tales modelos presentan el mejor desempeño en validación y menor desvío, bajo todas las métricas de *performance* calculadas (*Accuracy*, ROC AUC y Validación Cruzada) (*Tabla 2*).

Los restantes modelos (de altura mayor) tienen mejor desempeño en el conjunto de entrenamiento que el árbol de altura 3, y menor en el conjunto de validación. Esto es así porque puede darse sobreajuste a mayor profundidad del árbol, con lo cual, sirve para explicar muy bien el conjunto de entrenamiento, pero no el total. No servirían para realizar generalizaciones como se busca con este trabajo. En contraste, para los árboles de altura 3 la performance es apenas inferior en validación que en *training, es decir,* la diferencia no es tan significativa entre ambos conjuntos como ocurre en los demás modelos, lo que determina que el desvío sea menor para estos árboles.

Como se aprecia en la *Tabla 2*, el árbol de altura 3 presenta significativamente mayor performance en validación con *5-fold cross validation*. No resulta tan clara la selección de este modelo como mejor cuando se observa sólo la *Accuracy* o ROC AUC en validación porque presenta similar desempeño, o incluso inferior, que el árbol de altura 6, calculado con criterio de *InformationGain*.

Tabla 2. Comparación de Modelos, según medidas de performance para distintas alturas de árbol y medida de efectividad de atributos para clasificar el target.

Modelo	Accuracy Training	Accuracy Validation	ROC AUC Training	ROC AUC Validation	5-Folds CV Training	5-Folds CV Validation	Promedio	Desvío
Arbol altura 3 Gini	0.662	0.589	0.662	0.586	0.563	0.597	0.610	0.038
Árbol altura 6 Gini	0.759	0.581	0.758	0.579	0.571	0.504	0.625	0.098
Árbol altura sin límite Gini	1.000	0.527	1.000	0.523	0.536	0.566	0.692	0.218
Arbol altura 3 Info Gain	0.662	0.589	0.662	0.586	0.557	0.597	0.609	0.040
Árbol altura 6 Info Gain	0.724	0.597	0.723	0.596	0.569	0.543	0.625	0.072
Árbol altura sin límite Info Gain	1.000	0.543	1.000	0.536	0.532	0.535	0.691	0.219

Fuente: Elaboración propia.

Computando el promedio de las tres métricas, se obtiene que los mejores modelos serían los árboles de menor altura. En general, se observa que a medida que crece en altura se reduce la performance. Sin embargo, para elegir el mejor modelo se tuvieron presentes varias cuestiones. En principio, dado que con validación cruzada se reduce el riesgo de una mala partición de los datos, porque toma distintos subconjuntos para validar el modelo (entrenamiento y validación) y se obtiene un promedio de los resultados de las distintas iteraciones, podría ser de las tres la medida más robusta para determinar la eficacia del algoritmo para clasificar el *target*. En este caso indicaría que el mejor modelo es el árbol con altura 3 utilizando *Gini* o *InformationGain*.

No obstante, para terminar de definir la elección, se buscó corroborar esta decisión analizando la efectividad del modelo para predecir el cliente deudor, que es el que mayor interés reviste a esta investigación. De esta forma, se verificó que los árboles de altura 3 con*Gini* e *InformationGain*, no sólo presentan mejor desempeño general, sino que también presentan mayor *recall*. Dichos algoritmos tienen métricas de performance muy similares, y resultan los que mejor clasifican

la función objetivo particularmente para la clase deudora (*target* = 1). Se muestran los valores en la*Tabla 3*.

Tabla 3. Comparación de Medidas de performance para el set de validación según distintos Modelos.

Medida	3 Gini Val	6 Gini Val	Ultd. Gini Val	3 IG Val	6 IG Val	Ultd. IG Val
Accuracy	0.589	0.581	0.527	0.589	0.597	0.543
Recall (1)	0.620	0.606	0.563	0.620	0.606	0.606
Precision (1)	0.629	0.623	0.571	0.629	0.642	0.581
ROC-AUC	0.586	0.579	0.523	0.586	0.596	0.536
5-folds CV	0.597	0.504	0.566	0.597	0.543	0.535

Fuente: Elaboración propia.

Por los motivos expuestos y de acuerdo con el análisis de todas las métricas calculadas, los árboles de decisión de altura 3 calculados mediante el criterio de *Gini* e *InformationGain* parecen ser los modelos más adecuados para clasificar la variable objetivo. En particular, se definió la utilización de *Gini* para las próximas pruebas ya que tiene una *performance* promedio levemente superior, un desvío algo inferior, y se supone no debería cambiar el análisis con cualquiera de los dos algoritmos que se seleccione. Se muestra el diagrama de este árbol en la *Figura A6* del Anexo II.

Las principales diferencias del árbol de decisión obtenido mediante ambos criterios es que define distintos clasificadores para los nodos de decisión principales. En el caso de *Gini*, el atributo "Interés Mensual" se presenta como clasificador principal, mientras que con *InformationGain* es la "Cantidad de cuotas" (Ver *Figura A7*en Anexo II con este segundo árbol). Si bien este último resulta también de los principales clasificadores con *Gini*, aparecen algunos atributos con mayor relevancia como clasificadores con *Gini* que no figuran en el árbol calculado con *InformationGain*, como, por ejemplo, "Puesto Laboral" o "Ingreso Formal" del cliente. Asimismo, ambos coinciden en la importancia de la "Sucursal" y el "Capital Entregado". Ya no figuran como atributos relevantes para clasificar ni la "Edad" ni el "Teléfono particular" de contacto como aparecían en las pruebas iniciales.

2.1.1.2. Tratamiento de datos faltantes

Si bien los árboles de decisión son robustos a problemas con los datos (ya sea faltantes o ruido), ya que se usan todos los datos de entrenamiento para tomar decisiones basadas en estadísticas que permiten refinar las hipótesis, como parte de la etapa de experimentación con el *dataset* de desarrollo se buscó ejecutar una función para generar datos faltantes y luego otra para introducir ruido en distintos porcentajes, a fin de evaluar la respuesta del modelo (ver esto último en punto 2.1.1.3.).

En esta prueba, luego de generar datos faltantes al modelo, que van de 0% hasta un 80%, se procedió a imputar por la moda del atributo o moda de clase, para ver el impacto que tal situación tiene sobre el algoritmo utilizado (ver*Figuras1 y 2*).

Los resultados difieren según se trate de imputación por moda del atributo o moda de clase. En ambos casos se observa que a medida que crece el porcentaje de datos imputados por la moda del atributo o de clase, el tamaño del árbol varía. En varios casos se reduce el tamaño de 15 nodos de decisión (tamaño original del árbol sin faltantes) a 13, o incluso a 11, especialmente para los casos en que mayor porcentaje de datos imputados existe, que corresponden a los mayores por-

centajes de datos faltantes incorporados previamente. Esta situación se aprecia en mayor medida en la segunda prueba (moda de clase).

Por otro lado, al estudiar el comportamiento de la performance, medido por *Accuracy*, frente a datos faltantes imputados, los resultados de ambas pruebas tienen un comportamiento peculiar. En la prueba de imputación por moda del atributo, parece que se produce una mejora en la *accuracy* a medida que se incorpora mayor cantidad de datos. Lo contrario ocurre al estudiar el desempeño de los árboles en la muestra imputada por moda de clase. Parece haber una tendencia subyacente a la reducción de la *Accuracy* respecto del árbol original, a medida que se incrementa el número de valores imputados. Este resultado resulta interesante si se piensa que al introducir la moda del atributo se introduce la moda de la clase de mayor frecuencia para todos los datos. Esto se traduce en una mayor capacidad para clasificar (sesgada en favor de esa clase). Caso contrario cuando se introduce la moda de clase, al haber mayor cantidad de datos de cada una, resulta más difícil clasificar y definir a qué clase pertenece.

Figura 1. Tamaño y *Accuracy* del árbol según porcentaje de faltantes reemplazados por moda.

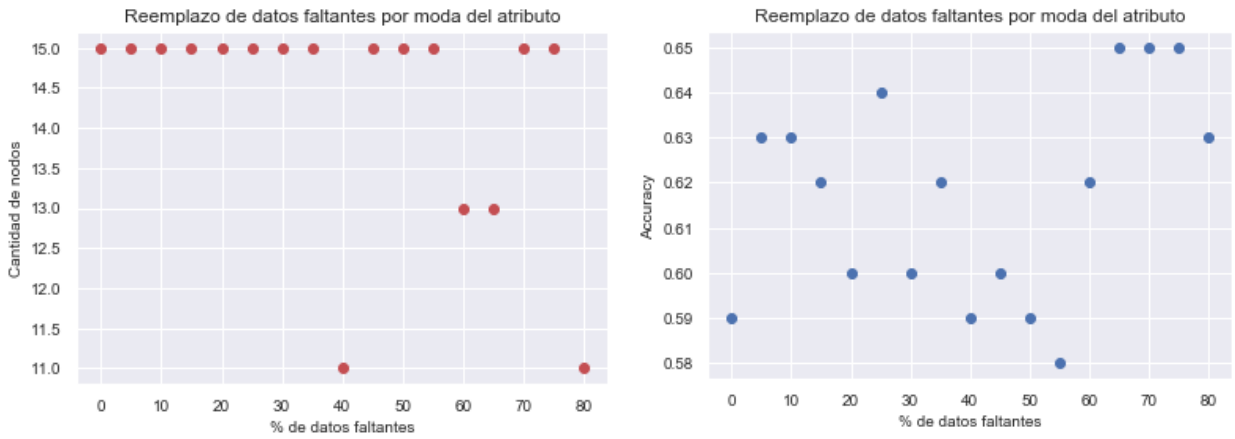


Figura 2. Tamaño y *Accuracy* del árbol según porcentaje de faltantes reemplazados por moda de clase.

Por otro lado, cuando se imputan valores de moda del atributo en un porcentaje elevado (en este caso se tomó el ejemplo del 80%), se observa que el árbol cambia significativamente su forma y contenido. No sólo se reduce su tamaño sino que cambian los atributos clasificadores. La variable "Interés mensual" (principal clasificador del árbol original) desaparece directamente del nuevo árbol y aparecen otros atributos como clasificadores relevantes (Ingreso Formal y Edad), aunque se mantienen otros como la "Cantidad de cuotas". Se muestra el diagrama de este árbol en la *Figura A8* del Anexo II.

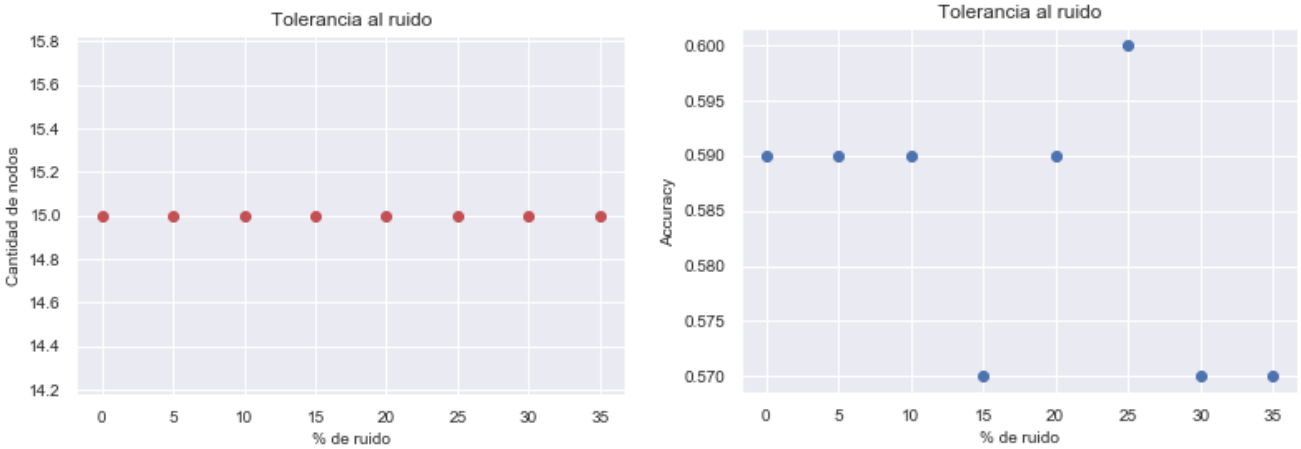
En el caso de la moda de clase sucede algo similar. Tomando el mismo ejemplo que en el caso anterior, es decir, el árbol que clasifica el conjunto de datos con mayor porcentaje de imputados por moda de clase, se obtiene un resultado donde el clasificador principal del árbol original (Interés mensual) desaparece y gana importancia la "Edad" del cliente, permaneciendo el "Importe de cuota" y luego el "Ingreso formal" del cliente como atributos relevantes. Se muestra el diagrama de este árbol en la *Figura A9* del Anexo II.

2.1.1.3. Tolerancia al ruido

Por otra parte, se evaluó la introducción de ruido al atributo que mejor clasifica el conjunto de datos, es decir, al atributo de la raíz del árbol de decisión original (Interés Mensual) a fin de evaluar qué ocurre con el modelo y la capacidad de clasificación de los atributos. En términos generales, los resultados obtenidos son consistentes con lo esperable por la robustez del modelo.

En efecto, se puede notar que el algoritmo es útil para clasificar los datos aún frente a presencia de ruido, por lo expuesto en el punto anterior (utiliza estadísticas en función de los datos disponibles para tomar decisiones en cada nodo), sosteniéndose inmutable el tamaño del árbol en 15 nodos de decisión, como el árbol original sin ruido(ver *Figura3*).

Figura 3. Tamaño y *Accuracy* del árbol según porcentaje de ruido introducido.



Adicionalmente, se percibe una reducción de desempeño (*Accuracy)*, a medida que se introduce mayor cantidad de ruido al conjunto de datos. Lo cual tiene sentido ya que al complejizar el conjunto de datos (a medida que se introduce más ruido), se reduce la capacidad de clasificación del *target* porque es más difícil identificar patrones subyacentes.

Por otro lado, se verifican modificaciones en la capacidad de los atributos para clasificar el *target (Figura A10*del Anexo II*)*. El "Interés mensual" deja de ser el clasificador principal, pasando a un segundo lugar y el primer puesto lo ocupa la "Cantidad de cuotas". Algo similar a lo observado con la imputación de datos, ganan relevancia otros atributos como "Edad" y "Trimestre de entrega"

del préstamo, que no aparecían en el árbol original, desplazando algunas de las variables clasificadoras originales como la “Sucursal”, el “Importe de cuota” y el “Capital entregado”.

2.1.1.4. GridSearch

En contraste con el *RandomSearch* empleado previamente, en este punto se propuso encontrar la mejor combinación de atributos, algoritmos e hiperparámetros, a través de G*ridSearch*, es decir, planteando opciones y explorando todas las combinaciones posibles.

De acuerdo con esta técnica el mejor árbol sería de altura 8 y calculado con criterio de entropía. No obstante, al evaluar el desempeño de este modelo con el set de validación y observar las distintas métricas utilizadas previamente, el resultado obtenido fue peor que con la técnica anterior. Es decir, el desempeño de este algoritmo fue peor para clasificar el target que el árbol de altura 3 con Gini calculado por *RandomSearch* (ver *Tabla 5* comparativa de modelos en el siguiente punto).

No obstante, es de destacar que, salvo el “Interés mensual”, que desaparece de este árbol de decisión, en los primeros nodos se encuentran las mismas variables que en el árbol de altura 3 con Gini, a saber: cantidad de cuotas, importe de la cuota, sucursal, capital entregado, ingreso formal y puesto laboral. Aunque con diversa relevancia entre los nodos principales y secundarios, se repiten los atributos que resultaban clasificadores significativos en el modelo seleccionado con la técnica anterior, las únicas variables nuevas introducidas por este árbol son la edad del cliente y el trimestre de entrega del préstamo entre los nodos inferiores (Ver *Figura A11* del Anexo II).

2.1.2. Naive Bayes

Por último, se empleó el modelo de *Naïve Bayes* (NB), que es un clasificador probabilístico que permite predecir la distribución de probabilidades de un conjunto de clases, es decir, brinda la probabilidad de ocurrencia de diferentes resultados posibles de un experimento. Utiliza la regla de Bayes con una suposición ingenua acerca de la independencia condicional de los atributos: supone que las probabilidades de los atributos son independientes dada la clase.

Comparando la performance sobre el conjunto de validación con el árbol de decisión (AD) de mejor desempaño según las pruebas anteriores (de altura 3 con *Gini*), se notó que NB tiene mejor performance para prácticamente todas las métricas calculadas. Se aprecia mayor *Accuracy* y ROC-AUC como también significativamente mayor *recall* (mejor desempeño para clasificar el cliente deudor). El árbol de decisión (AD) presenta mejor desempeño sólo para la métrica de validación

Tabla 4. Comparación de Medidas de Performance con AD 3 *Gini* y NB sobre conjuntos de entrenamiento y validación.

Medida	3 Gini Train	3 Gini Val	NB Train	NB Va
Accuracy	0.662	0.589	0.538	0.65
Recall (1)	0.648	0.620	0.672	0.76
Precision (1)	0.659	0.629	0.523	0.65
ROC-AUC	0.662	0.586	0.540	0.63
5-folds CV	0.563	0.597	0.551	0.59

Tabla 5. Comparación de Medidas de Performance con AD 3 *Gini*, AD 8 IG y NB sobre conjunto validación.

Medida	3 Gini Val	8 IG Val	NB Val
Accuracy	0.589	0.566	0.651
Recall (1)	0.620	0.606	0.761
Precision (1)	0.629	0.606	0.659
ROC-AUC	0.586	0.561	0.639
5-folds CV	0.597	0.504	0.590

cruzada, que resulta apenas superior al valor de NB para el conjunto de validación (Ver *Tabla 4*).

Adicionalmente, se contrastó con el mejor modelo de G*ridSearch* y se verificó que NB presenta mejor desempeño, para todas las métricas analizadas. Es decir, que de los tres modelos, mientras NB resulta el mejor evaluando *Accuracy, recall* y ROC-AUC, el AD de altura 3 con *Gini* lo es para *5-folds cross validation* (CV). Por su parte, el árbol de altura 8 con *InformationGain* no parece tener mejor desempeño que los restantes bajo ninguna de las métricas (Ver *Tabla 5*).

En función de lo anterior, dependiendo la métrica que se analice difiere el resultado, por tanto, habrá que utilizar el conjunto de Test para determinar cuál es el mejor clasificador para la variable objetivo. Hasta ahora NB y AD de altura 3 parecen las opciones más adecuadas si se considera CV como medida más robusta.

2.2. Etapa de Evaluación

Según los resultados obtenidos de la etapa de experimentación, se seleccionaron los mejores modelos clasificadores de la variable objetivo (cliente deudor), que servirán para contrastar con el conjunto Test, separado inicialmente para este fin. Ahora bien, luego de la evaluación de estos algoritmos, se obtuvieron resultados que difieren de lo esperable de acuerdo con lo observado en los análisis sobre el conjunto de validación.

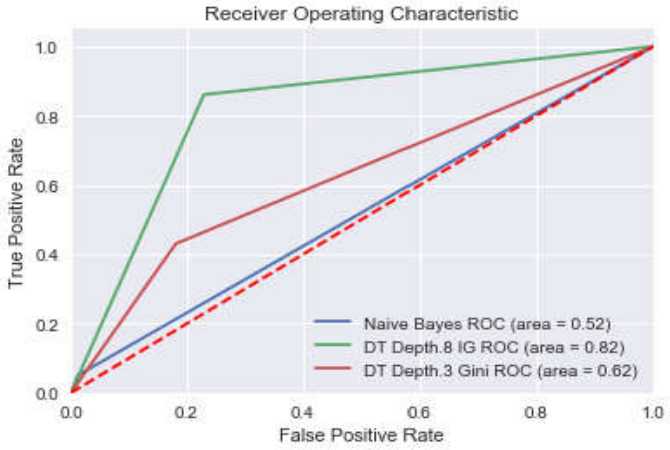
Tabla 6. Comparación de Medidas de Performance con AD 3 *Gini*, AD 8 IG y NB sobre el conjunto Test.

Medida	3 Gini Test	8 IG Test	NB Test
Accuracy	0.593	0.815	0.531
Recall (1)	0.481	0.861	0.051
Precision (1)	0.603	0.782	0.800
ROC-AUC	0.590	0.816	0.519
5-folds CV	0.481	0.501	0.586

La evaluación sobre el conjunto de Test modifica las conclusiones a las que se había arribado con el conjunto de validación, ya que la performance de los distintos modelos cambia sustancialmente. Por un lado, el árbol de decisión de altura 8 con *InformationGain*, que no aparecía como mejor modelo bajo ninguna métrica calculada sobre el conjunto de validación, en este caso aparece como el mejor para la mayoría de las métricas excepto *cross validation*. En este último caso el algoritmo de NB pasa a ser el mejor clasificador, quedando el árbol de altura 3 sin sobresalir en ninguna de las medidas de desempeño.

Se destaca la performance del árbol de decisión obtenido mediante la técnica de G*ridSearch* por su alto rendimiento para clasificar el perfil de cliente deudor. No así el modelo de NB que registra un muy mal desempeño.

Figura 4. Comparación de Curvas ROC con AD 3 Gini, AD 8 IG y NB.



Fuente: Elaboración propia.

Por último, como se puede ver en la *Figura 7*, las curvas ROC muestran los resultados de la performance de los tres modelos seleccionados por orden de eficacia, desde el mejor modelo (árbol de altura 8) más alejado de la diagonal principal del eje de coordenadas, al de menor (Naive Bayes), muy próximo a la mencionada diagonal.

De acuerdo a la validación efectuada con el conjunto de Test, el modelo

más eficaz para clasificar la variable objetivo parece ser el árbol de altura 8 con *InformationGain*. En tal sentido, las variables que mayor relevancia tienen para clasificar el target serían: por el lado de las condiciones del préstamo, la cantidad de cuotas, el interés mensual y capital entregado, y por el lado de las características personales del cliente, la edad, el ingreso formal y la zona de residencia. En niveles más bajos aparecen atributos como el trimestre de entrega, el puesto laboral y el teléfono alternativo de contacto. Si bien figura la nacionalidad como separador entre los primeros nodos, deja solo un grupo afuera que es el de menor frecuencia en el *dataset*, motivo por el cual, no se toma en cuenta la relevancia de este último atributo (Ver *Figura A12 y A13* con los últimos árboles entrenados con el conjunto de desarrollo).

3. Conclusiones

Luego de la realización de diversos árboles de decisión, se observa que el atributo que más separa los casos de estudio para su clasificación en deudores o pagadores es: la cantidad de cuotas (según el árbol de mejor desempeño: altura 8 IG). Le siguen en importancia, el interés mensual y la edad, y posteriormente, la zona de residencia y el ingreso formal del cliente.

Contrario a lo esperado (hipótesis), no se observa un gran peso para la clasificación de las características personales del cliente, como género, puesto laboral o información de contacto. En cambio, la edad, el ingreso formal y la zona de residencia podrían brindar información sobre el comportamiento de pago (Ver *Figura A13*, árbol de altura 8, el mejor obtenido).

En particular, se observa que los clientes que solicitan préstamos a una cantidad de cuotas menor a un año suelen ser pagadores. Para los casos de cuotas entre 10 y 18 meses (menores al año y medio), el monto de capital entregado, el ingreso formal del cliente y su edad figuran como determinantes del comportamiento de pago. Para ingresos mayores a 6.800 y por montos de préstamo mayores a 6.500, los clientes menores de 50 años parecen tener mayor probabilidad de ser deudores.

En contraste con lo anterior, se pudo notar que los perfiles deudores se encuentran en relación con préstamos a devolver en cuotas a más de un año de plazo (cantidad de cuotas superior a 10 meses). En términos del préstamo, esto significa que tanto el interés como el nivel de endeudamiento, también serán mayores.

Con cuotas por plazos superiores al año y medio, se aprecia que las variables de mayor relevancia para explicar el riesgo crediticio de un cliente resultan la zona de residencia y el ingreso del cliente, como también el monto de capital otorgado y el momento del año en que se otorgó. En estos casos, de acuerdo con los resultados del modelo, cuando los clientes residen en capital federal o zona norte, el monto del préstamo es menor a 16 mil pesos, siendo su ingreso formal menor a 20 mil pesos y, obteniendo el crédito en el primer semestre del año, parece haber mayor probabilidad de que sea deudor.

Este último caso podría tratarse de un cliente que demanda a un plazo mayor del promedio (14 meses) y pide un monto menor o igual al promedio (13 mil pesos). Dado un mismo monto de préstamo demandado, devolverlo a mayor cantidad de cuotas significa menor importe de las mismas. Si bien esto último no se puede verificar ni deducir del modelo empleado, podría implicar que se trata de una persona con menor posibilidad / disposición a pagar mensual y que, por tal motivo, podría tratarse de un potencial deudor.

Sería de interés para futuras investigaciones realizar un modelo de clasificación basado sólo en características personales del cliente y otro sólo en características del préstamo. De esta mane-

ra la empresa financiera podría identificar los rasgos de mayor relevancia que hacen al perfil del cliente y definir los datos personales a solicitar para efectuar una mejor selección de clientes cumplidores, y evitar problemas de riesgo crediticio. También podría resultar útil analizar algún tipo de cliente en particular.

Por ejemplo, en las hipótesis se había planteado el caso del cliente “portero” dado su buen comportamiento de pago, según lo indicado por expertos. Ahora bien, con los resultados del modelo, no se pudo verificar ni rechazar esa hipótesis. En este sentido, sería interesante efectuar un análisis de tales clientes para identificar patrones de perfiles cumplidores, que contribuyan a una mejor selección de futuros clientes, reduciendo el riesgo crediticio.

Por otro lado, para realizar un análisis más exhaustivo de las condiciones de préstamo, podrían fijarse tipos de préstamo a otorgar con sus respectivos atributos de interés, y así poder segmentar mejor en clases de préstamo que podrían inducir a un comportamiento deudor o cumplidor del cliente. Un ejemplo podría ser establecer intervalos o rangos para los préstamos y agruparlos de modo que se caractericen por montos bajos, medios y altos. Asimismo, podría incluirse como variable de estudio la relación cuota-ingreso, ya que el comportamiento deudor parece incluir a ambas y al unificarla en un mismo atributo podrían verificarse ambas.

El presente análisis permitió identificar cuáles son las principales variables que explican el comportamiento crediticio de los clientes, como también sentar bases para futuros análisis. Se proponen algunas líneas de investigación a futuro, que implican análisis más específicos o más enfocados, que deberán ser evaluadas por la entidad financiera de acuerdo a sus intereses.

4. Bibliografía

Econlink (06 de Ene de 2014). *"El Riesgo Crediticio"*. [en linea] Dirección URL: https://www.econlink.com.ar/riesgo-crediticio (Consultado el 30 de Abr de 2019)

Cardona Hernández, P.A. (2004). *Aplicación de árboles de decisión en modelos de riesgo crediticio*. Revista Colombiana de Estadística Volumen 27 No 2. Págs. 139 a 151. Diciembre 2004

Anexo I. Tablas

Tabla A1: Variables del dataset original y motivo de su eliminación.

Dataset Original	Tipo	Dataset final	Motivo de eliminación
ID	Numérica		Se decidió tratarlo como "casos" (préstamos) y no como "clientes".
Sucursal	Categórica	Sucursal	
Año de Entrega	Numérica		No se dispone de una serie tan grande como para encontrar patrones anuales.
Fecha Entrega	Fecha		Se opta por la variable "Trimestre".
Mes de Entega	Numérica		
Trimestre Entrega	Numérica	Trimestre Entrega	-
Cancelado	Categórica binaria		Se elimina por ser redundante con la variable "Target".
Cuotas Sin Vencer	Numérica		
Cuotas Adeudadas	Numérica		
Deuda a la fecha	Numérica		
Capital Entregado	Numérica	Capital Entregado	-
Cantidad Cuotas	Numérica	Cantidad Cuotas	-
Importe Cuota	Numérica	Importe Cuota	-
Interes	Numérica		Correlacionada con "Interés Mensual".
Interes Mensual	Numérica	Interes Mensual	-
Última Fecha Pago	Fecha		Se eliminan todas las variables relacionadas con el pago del crédito.
Año de Pago	Numérica		
Mes de Pago	Numérica		
dia de Pago	Numérica		
Trimestre de Pago	Numérica		
Localidad	Texto		Se reduce a categorías en "Zona de Residencia".
Código Postal	Texto		
Zona de Residencia	Categórica	Zona de Residencia	-
Fecha Nacimiento	Fecha		Redundante con "Edad".
Edad	Numérica	Edad	-
Nacionalidad	Categórica	Nacionalidad	-
Informa celular	Categórica binaria	Informa celular	-
Informa tel particular	Categórica binaria	Informa tel particular	-
Informa Email	Categórica binaria	Informa Email	--
Informa tel alternativo	Categórica binaria	Informa tel alternativo	-
Informa nombre jefe	Categórica binaria	Informa nombre jefe	-
Puesto Laboral (Categoría)	Categórica	Puesto Laboral	-
Puesto laboral (Nombre)	Texto		Se reduce a categorías en "Puesto Laboral (Categoría)"
Nombre Empresa	Categórica binaria		No aporta información extra.
Dirección Laboral	Texto		
Código Postal Laboral	Texto		
Informa tel laboral	Categórica binaria	Informa tel laboral	-
Ingreso Formal	Numérica	Ingreso Formal	-
Sexo	Categórica binaria	Sexo	-
Target	Categórica binaria	Target	-

Anexo II. Figuras

Figura A1: Correlación de atributos del dataset original.

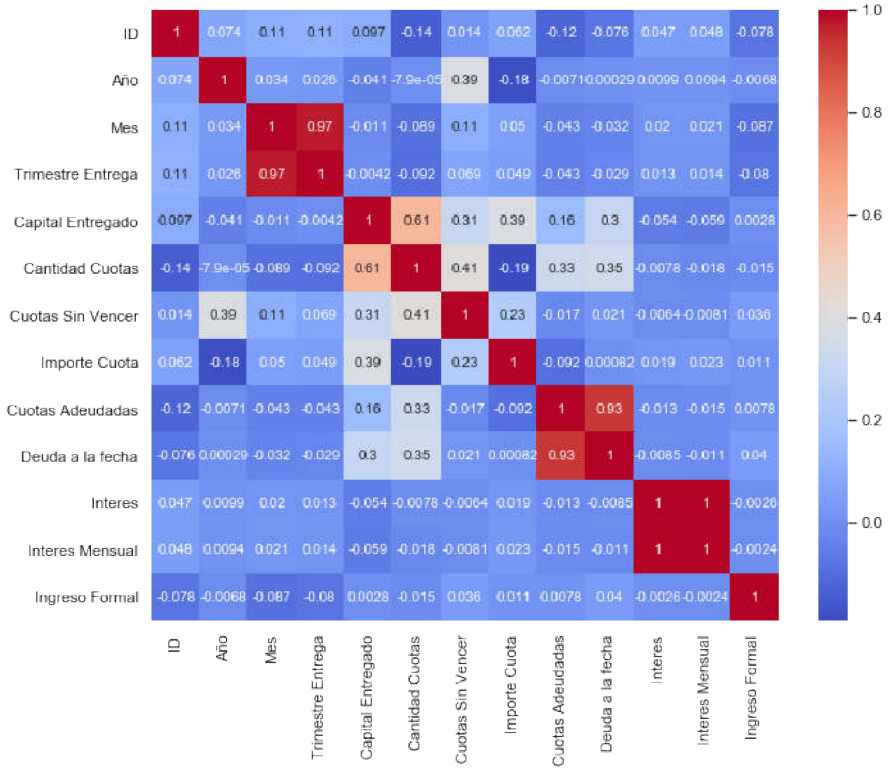


Figura A2: Correlación de atributos del dataset final.

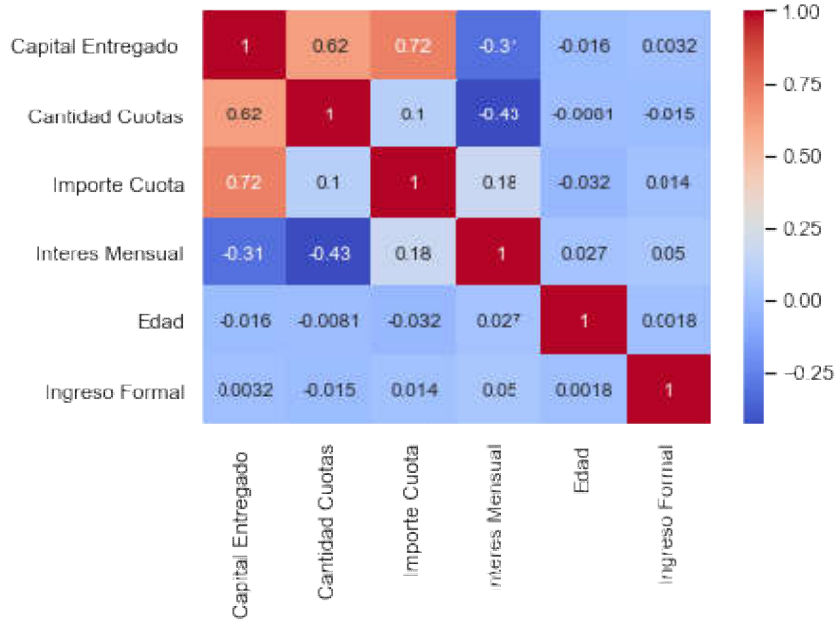


Figura A3. Árbol de Decisión de altura 3, con 25% de datos borrados

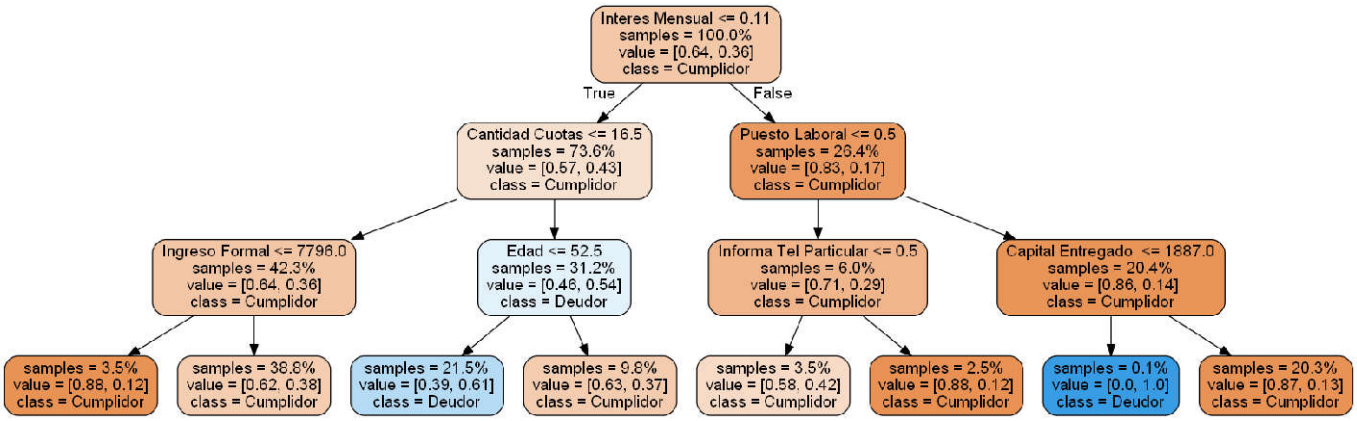


Figura A4. Árbol de Decisión de altura 3, con 11% de datos imputados por moda en "Sucursal"

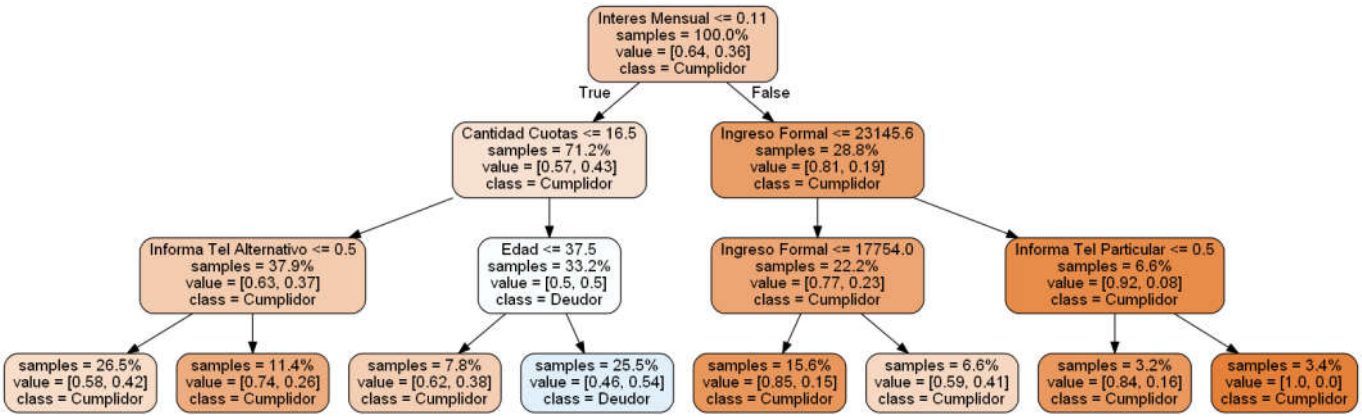


Figura A5. Árbol de Decisión de altura 3, Balanceado.

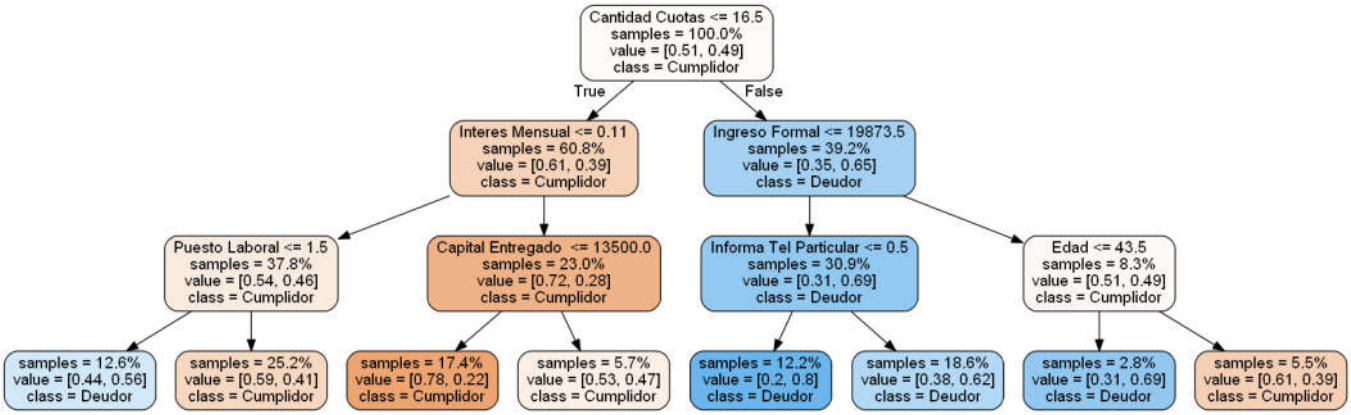


Figura A6. Árbol de Decisión de altura 3, criterio *Gini*Gain.

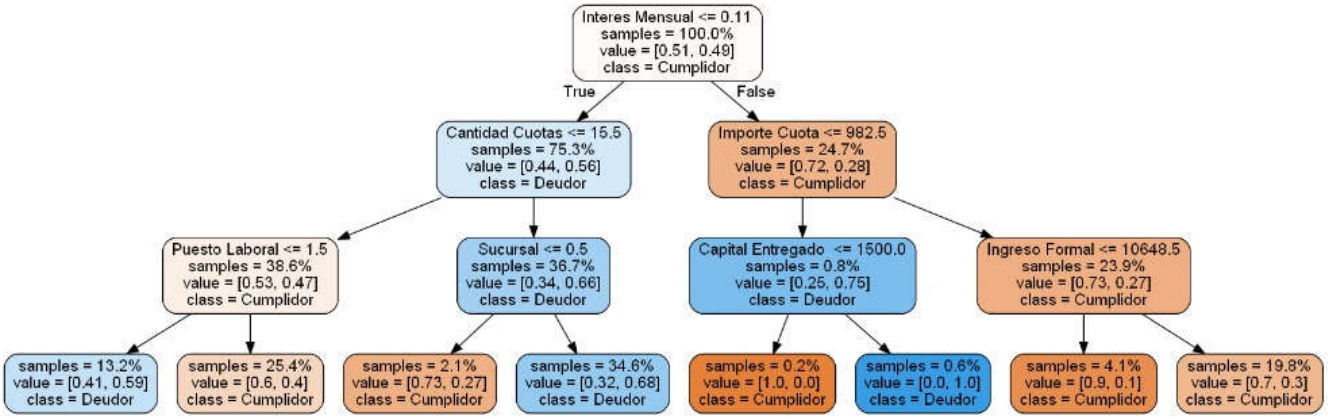


Figura A7. Árbol de Decisión de altura 3, criterio *InformationGain*.

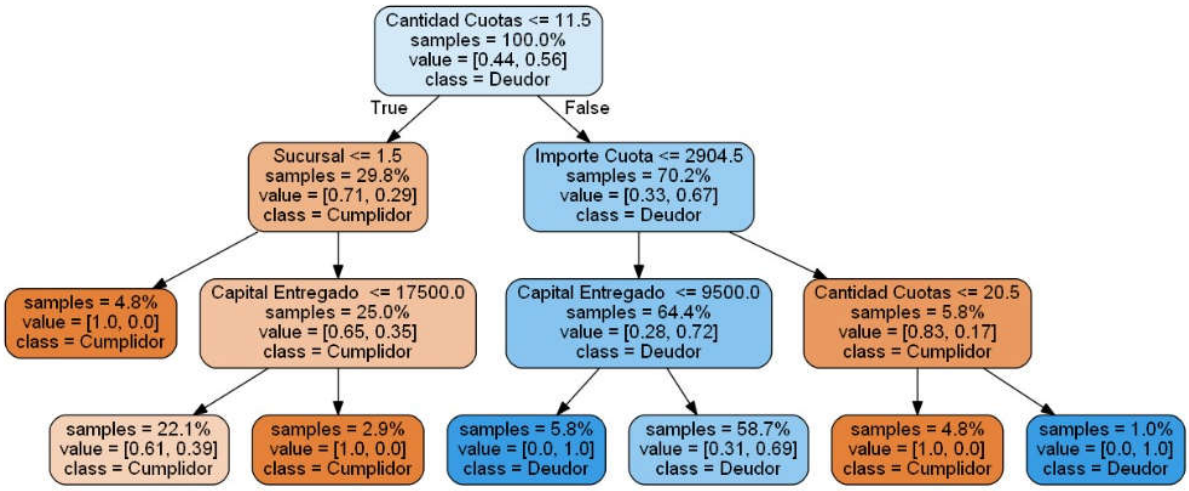


Figura A8. Árbol de Decisión de altura 3 con imputación de moda del atributo sobre el 80% de los datos.

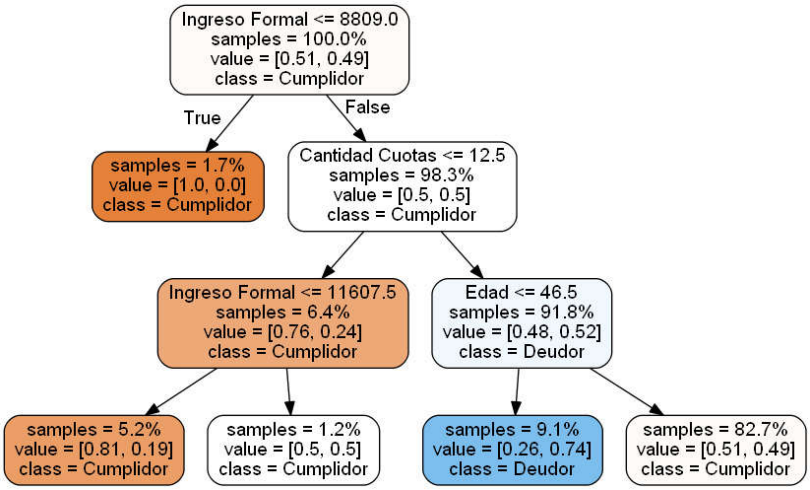


Figura A9. Árbol de Decisión con imputación de moda de clase sobre el 80% de los datos.

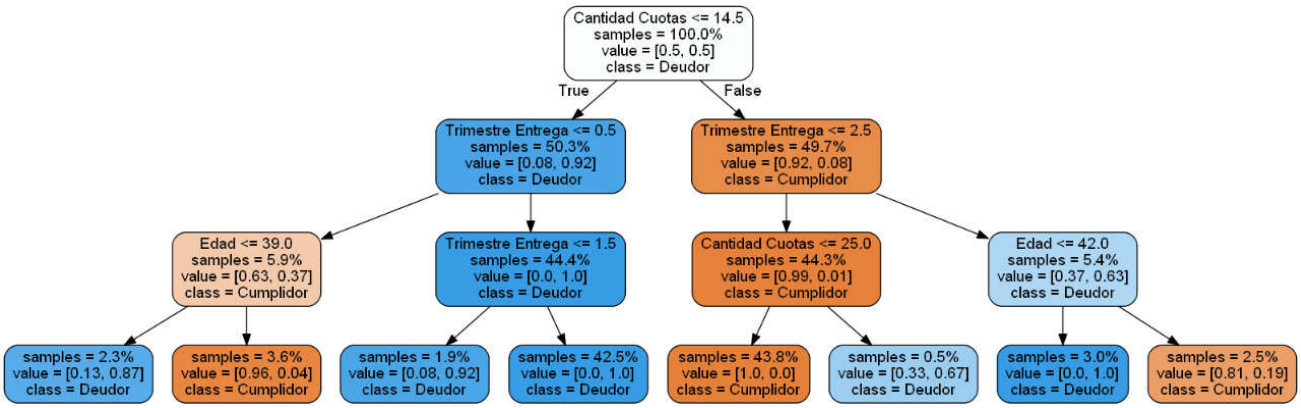


Figura A10. Árbol de Decisión Ruidoso

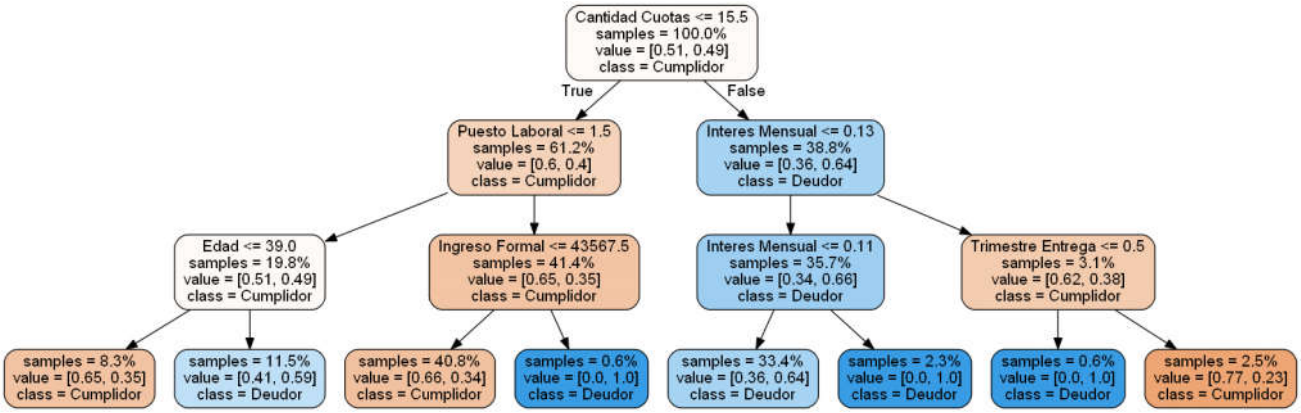


Figura A11. Árbol de Decisión de altura 8 con *InformationGain* entrenado con el conjunto de entrenamiento

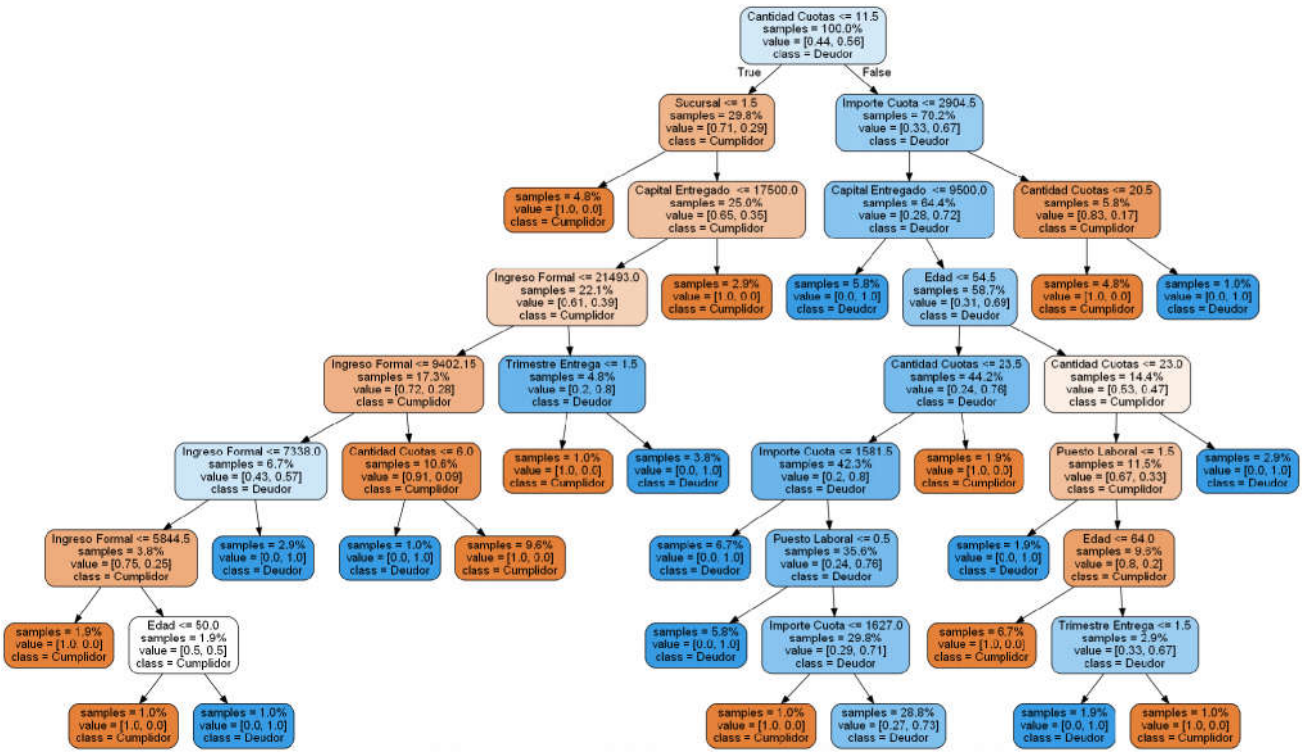


Figura A12. Árbol de Decisión de altura 3 con Gini entrenado con el conjunto de Desarrollo

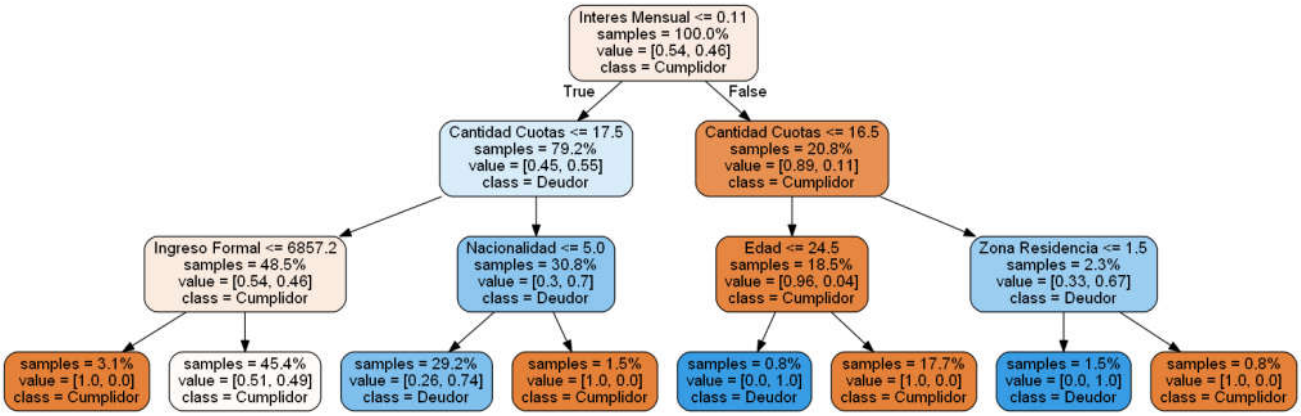


Figura A13. Árbol de Decisión de altura 8 con *InformationGain* entrenado con el conjunto de Desarrollo