**Comportamiento de la movilidad de productos de consumo masivo en Estados Unidos durante el periodo julio 2021 a mayo 2022.**

**Carlos Alberto Barrantes Brenes**

**Sabrina**

**Universidad de Costa Rica**

**Programa de Posgrado de Estadística**

**RESUMEN**

En el presente artículo se realizó una clasificación utilizando modelos estadísticos y de aprendizaje de máquina automático de la serie agregada de los tarjetahabientes del Sistema Financiero Nacional de Costa Rica con base en el año 2020. El conjunto de datos, la población de estudio presenta un desbalance en la serie variable a predecir “buen pagador?” por lo que se tienen que recurrir a una transformación de datos desde un punto de vista de muestreo para equilibrar el conjunto de datos antes de aplicar las técnicas estadísticas. Al finalizar esta estimación, se seleccionan el modelo de reducción del sesgo en modelos lineales generalizados de respuesta binomial y el modelo ensamblaje de apilamiento como los modelos con un mejor pronóstico de los datos y con mejores indicadores de desempeño. Se concluye que el modelo de aprendizaje automático de apilamiento es el mejor modelo por los indicadores de desempeño obtenidos y por los valores pronosticados y por mejorar la detección de clientes que no cuentan con la capacidad adquisitiva para hacer frente a sus deudas.

**INTRODUCCIÓN**

La crisis financiera del 2008 resaltó la importancia de la protección del consumidor financiero, con el fin de lograr la estabilidad del sistema financiero a nivel mundial. Según el Banco Mundial (2012), el incremento en el uso de servicios financieros ha dado lugar a la regulación financiera para proteger y empoderar a los consumidores. En Costa Rica, la temática de las finanzas personales y la protección financiera ha tomado relevancia en los últimos 5 años, evidenciado en proyectos de ley que apuntan hacia la regulación de los proveedores de servicios financieros y la inclusividad por medio de la educación financiera en grupos estratégicos de la población.

Para el caso costarricense, los proyectos de ley propuestos a partir de 2017 están enfocados en el accionar de los proveedores de servicios financieros, dejando de lado la implementación de medidas preventivas, más allá de la educación financiera, para guiar a este grupo etario a tomar mejores decisiones crediticias.

A nivel gubernamental, el Ministerio de Economía, Industria y Comercio (MEIC) y la Superintendencia General de Entidades Financieras (SUGEF), desde finales del 2010, han realizado estudios de mercado con el fin de registrar la cantidad de tarjetas de crédito y débito que poseen los costarricenses, su nivel de endeudamiento y la transparencia de las entidades bancarias con los clientes. A pesar de estas publicaciones, no se han gestionado investigaciones de teoría de las perspectivas y economía conductual más allá de lo descriptivo.

Complementado a lo anterior, el curso de las finanzas personales en Costa Rica se ha deteriorado rápidamente. Según datos del Gobierno de Costa Rica (2019), en los últimos cinco años, el endeudamiento de los costarricenses ha aumentado un 10% hasta alcanzar 1.2 billones de colones siendo los más afectados funcionarios, mujeres y jóvenes.

Los datos anteriores están directamente relacionados con los datos proporcionados por el MEIC (2019) con respecto al saldo en tarjetas de crédito de los costarricenses. Esta deuda, deflactada con la inflación, ha aumentado un 66% desde mediados del 2015, llegando hasta un balance de 1.3 billones de colones y viene aunado a un aumento de 900 mil tarjetas de crédito colocadas en el mercado entre el mismo periodo.

De acuerdo con comunicados del Gobierno de Costa Rica, las personas jóvenes entre 15 a 24 años, son uno de los grupos más afectados por el endeudamiento. Adicionalmente, según la Encuesta de Actualidades, publicada por la Escuela de Estadística de la Universidad de Costa Rica (2017), el 60% de los jóvenes encuestados, cuya edad oscilaba entre los 18 y 29 años, tenía algún tipo de deuda. La investigación también reveló que el 25% de los encuestados estaba endeudado con tarjetas de crédito.

Debido a lo anterior, la temática crediticia ha provocado el accionar institucional en Costa Rica en dos vertientes, la primera impulsada en el 2013, enfocada a grupos vulnerables de la sociedad a través de la Estrategia Nacional de Educación Financiera (ENEF) y segundo, por medio de la creación proyectos de ley para la Protección al Consumidor Crediticio en 2017 y Protección al Consumidor Financiero en el 2019 (Proyecto de ley N.° 20.425 y 12.666 respectivamente) y más recientemente, por medio de la Ley de Usura (ley N.° 9859).

Por consiguiente, es importante generar investigaciones con que analicen el comportamiento del sistema financiero con la finalidad de detectar anomalías o irregularidades que afecten a la ciudadanía en general, pero especialmente, a las personas con un menor acceso a recursos financieros y legales. De acuerdo con Ramos (2017) y la OCDE (2017), las estadísticas evidencian el importante crecimiento del endeudamiento de la población joven; población que oscila entre los 18 y 29 años, y se caracteriza por poseer un ingreso bajo, variable y escaso conocimiento financiero sobre instrumentos de crédito, ahorro e inversión a la hora de insertarse en el mercado laboral; lo que provoca que sea una población principalmente vulnerable.

Uno de los beneficios de las tarjetas de crédito estan en que presentan la facilidad de comprar un bien o servicio en el corto plazo e ir pagando el monto del principal en cuotas en el largo plazo con diferentes modalidades de pago. Sin embargo, antes de otorgar a los clientes estos medios de pago, las instituciones crediticias por medio de modelos crediticios o de aprendizaje de máquinas determinan si el cliente va a tener los medios para hacer frente a las compras realizadas.

Identificar y discriminar a los tarjetahabientes como buenos o malos pagadores por sus cualidades es el tema a central del presente artículo. Por lo tanto, el artículo se estructura primeramente en explicar la metodología de los modelos de aprendizaje de máquina y del modelo crediticio seleccionados para la predicción de la serie de corte transversal. Posteriormente, se detallarán los resultados obtenidos para cada uno de los modelos y su relación con la coyuntura económica costarricense para una mejor contextualización de los hallazgos. Por último, se darán reflexiones acerca de los hallazgos y recomendaciones para futuros artículos académicos.

Debido a lo anterior, se establecen los siguientes los siguientes objetivos para el siguiente artículo.

**Objetivo General**

* Clasificar a tarjetahabientes por medio de sus características socio económicas y laborales utilizando e información de tarjetas de crédito en colones del Sistema Financiero Nacional como base a marzo del año 2020

**Objetivo específico:**

* Determinar el mejor modelo crediticio con la información de tarjetas de crédito en colones del Sistema Financiero Nacional como base a marzo del año 2020
* Determinar el mejor modelo de aprendizaje de máquinas utilizando como información los datos agregados de las tarjetas de crédito en colones del Sistema Financiero Nacional como base a marzo del año 2020.

**MATERIALES Y MÉTODOS.**

Una tarjeta de crédito (o débito) es una tarjeta plástica con la que el tarjetahabiente puede pagar bienes y servicios. Estas medio de pagos están asociados a una entidad financiera o por una local comercial supervisado por el BCCR, CONNASIF, SUGEF y MEIC, este último por medio de la Ley de Promoción de la Competencia y Defensa Efectiva del Consumidor.

Las tarjetas de crédito han sido durante mucho tiempo uno de los métodos más populares para realizar pagos y acceder a préstamos no garantizados en las últimas décadas, representando en Estados Unidos el 37 por ciento de las compras de los consumidores por valor en dólares en 2021. (McKinsey & Company, 2022)

En Costa Rica y según el Banco Central de Costa Rica (BCCR) en el su informe de Medios de Pago Estadísticas 2020, el valor de los pagos efectuados con tarjetas de crédito fue de representó ₡3.8 billones en el año 2020, representando el 51,1% a pagos realizados con tarjetas en circulación o más del 10% del PIB. Es así como existe un sistema nacional, la Superintendencia General de Entidades Financieras (SUGEF), que tiene como objetivo velar por la estabilidad, la solidez y el funcionamiento eficiente del sistema financiero nacional, con estricto apego a las disposiciones legales y reglamentarias y de conformidad con las normas, directrices y resoluciones que dicte la propia institución, todo en salvaguarda del interés de la colectividad. (SUGEF, 2022).

La variable de estudio para el presente artículo es una variable dicotómica que indica si el tarjetahabiente presenta las características socioeconómicas y laborales para ser considerado un “buen pagador” o, en otras palabras, si tiene las capacidades de hacer frente a una deuda bajo el parecer de la entidad.

**Conceptualización**

El presente artículo se desarrollará utilizando cinco variables independientes relacionadas al gasto gubernamental en educación y una variable poblacional con el fin de explicar cuál es el un efecto sobre sobre la desigualdad de ingresos, específicamente sobre el índice de Gini que será la variable dependiente en Costa Rica en el periodo 1999 al 2014.

Los datos de la base de datos son de corte transversal y fueron recabados para marzo del año 2020 sobre información de tarjetas de crédito en colones del Sistema Financiero Nacional. El saldo de la tarjeta es representado es un monto agregado en colones a precios corrientes, con efecto inflacionario, y las otras variables secundarias son el género, edad, estado civil, código de profesión, código de sector de trabajo, límite de tarjeta de crédito y coeficiente de solvencia. La cantidad de casos de entrenamiento corresponde a 14 500 casos mientras que la cantidad de casos nuevos corresponde a 353.

Las variables seleccionadas de las bases de datos se definen en el cuadro 1.

**Cuadro 1. Definición de variables**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Abreviatura en R** | **Definición** |
| Género | genero | Indica si el tarjetahabiente esta registrado como hombre y mujer. |
| Edad | edad | Edad del tarjetahabiente. |
| Estado Civil | estado.civil | Estado civil de la persona según el registro civil. |
| Código Profesion | cod.prof | Código de profesión según la entidad bancaria |
| Codg Sector trab. | cod.sector | Área profesional en la que se desenvuelve laboralmente el tarjetahabiente. |
| Limite Tarjeta Crédito | limite.tarj | Límite de crédito en colones otorgado al tarjetabiente. Este sería su límite global porque los datos son agregados. |
| Saldo Tarjeta | saldo | Saldo monetario del tarjetahabiente en las cuentas bancarias de la entidad financiera. |
| Coeficiente solvencia | solvencia | La solvencia se asocia con la capacidad económica que tiene un tarjetahabiente para enfrentarse a sus obligaciones, en relación con su patrimonio existente. |
| Plazo (meses) | plazo | Plazo de pago hábil para la tarjeta de crédito. |
| Destino | destino | Tarjeta destinada para el consumo |
| Buen pagador? | estado | Clasificador de la entidad bancaria para determinar si el tarjetahabiente es considerado un buen pagador o no. |

Fuente: Sistema Financiero Nacional (2020).

Es importante mencionar que todas las variables del artículo de referencia fueron incorporadas a esta investigación, sin embargo, las variables “Destino” y “Plazo (meses)” fueron omitidas debido que son variables con la misma observación, “Consumo” y “66 meses” correspondientemente, y por lo tanto no aportan gran significancia estadística a los modelos predictivos.

A su vez, incluir estas variables podría conducir a resultados sesgados porque interfiere en que tan bien se puede usar cada variable independiente para predecir la variable dependiente en un modelo estadístico, lo anterior por problemas de multicolinealidad.

Como se mencionó en la introducción, en una primera instancia se aplicará modelos de clasificación por medio de una serie de modelos de aprendizaje automático (Machine learning) y aprendizaje profundo (deep learning) con un ajuste intención en los hiper parámetros con el objetivo discriminar acertadamente un buen pagador de otro.

Según MIT (s.f) el aprendizaje automático es un subcampo de la inteligencia artificial, que se define ampliamente como la capacidad de una máquina para imitar el comportamiento humano inteligente. Los sistemas de inteligencia artificial se utilizan para realizar tareas complejas de una manera similar a la forma en que los humanos resuelven problemas. Por otra parte, el aprendizaje profundo es un campo perteneciente a la inteligencia artificial cuyo objetivo es el estudio y construcción de sistemas informáticos capaces de “aprender” a partir de la experiencia.

Para el presente artículo, se aplicaron los siguientes modelos de clasificación: modelos binomiales, redes neurales, modelo de árboles de decisión, modelo discriminante lineal, modelos de árboles aleatorios unidos por medio de bagging, métodos boosting y, por último, modelos de apilamientos. Los modelos seleccionados por sus características fue el modelo de apilamiento y modelo de árboles aleatorios (sin bagging).

Cada uno de los modelos de aprendizaje de máquinas será sometido a través de indicadores de desempeño para determinar cuál modelo es más confiable prediciendo, tiene resultados más precisos y minimiza el error en la predicción si se compara con los resultados del test. Los principales indicadores de desempeño que se utilizarán directa e indirectamente son a) matriz de confusión b) métrica de Precisión c) sensibilidad d) falsos positivos e) precisión positiva.

Los indicadores de desempeño de las predicciones del modelo de aprendizaje de máquina se aplicarán con el fin de determinar el mejor modelo basado en los objetivos planteados. Para el presente artículo, es importante determinar con exactitud la precisión de los resultados y particularmente minimizar los falsos negativos porque podrían representar un riesgo financiero para las entidades bancarias.

* Precisión: Brinda información sobre qué tan correctamente el modelo ha detectado los resultados positivos. Es la relación entre los verdaderos positivos y el total de positivos.
* Verdadero positivo (VP): Se detecta que el cliente es buen pagador y efectivamente lo es.
* Verdadero Negativo (VN): Se detecta que el cliente no es buen pagador y efectivamente no cuenta con los medios para pagar sus deudas.
* Falso Positivo (FP): Se detecta que el cliente es buen pagador, pero realmente no cuenta con los medios para pagar sus deudas. (escenario más crítico).
* Falso Negativo (FN): Se detecta que el cliente no es buen pagador, pero realmente si lo es.

Posteriormente, se analizan los resultados por medio de la curva ROC (curva de característica operativa del recepto) es un gráfico que muestra el rendimiento de un modelo de clasificación en todos los umbrales de clasificación. La curva se gráfica contrastando los falsos positivos (eje x) versus la precisión positiva (eje y) en diferentes umbrales donde la expectativa de los resultados es obtener más falsos positivos y más precisión positiva.

A partir de la curva de ROC se obtiene el Área bajo la curva ROC (AUC). El AUC mide toda el área bidimensional por debajo de la curva ROC completa (cálculo integral) de (0,0) a (1,1). El AUC proporciona una medición agregada del rendimiento en todos los umbrales de clasificación posibles. Una forma de interpretar el AUC es como la probabilidad de que el modelo clasifique un ejemplo positivo aleatorio más alto que un ejemplo negativo aleatorio.

En la práctica se busca tener un indicador entre 0.6 y 0.9, ya que más bajo de 0.6 podría indicar que el modelo no es predictivo ni confiable cuando ingresen nuevos datos al modelo predictivo. En la práctica, aunque ideal, no es usual que los modelos registren valores mayores a 0.9 porque puede generar sospechas de sobreajuste de los datos.

Otro indicador que se aplicará para la toma de decisión de cuál modelo de aprendizaje será utilizado es elKolmogorov-Smirnov (KS). El KS es la máxima diferencia entre las distribuciones acumuladas relativas de las clases, ordenadas de menor a mayor según la probabilidad predicha por el modelo de clasificación. Se esperan valores de KS entre 20 y 70, donde un valor de 20 indica un pobre nivel predictivo y valores superiores a 70 indican sobreajuste.

Asimismo, se realizó una estrategia para calibrar los hiper parámetros de cada uno d ellos modelos de aprendizaje automático. Primero, se estableció un rango para cada uno de los hiper parámetros de importancia y se comenzó a calcular y contrastar gráficamente el AUC con el fin de maximizar sus resultados. Estos cálculos se realizaron por medio de validación cruzada utilizando un

**Modelos de aprendizaje de máquina seleccionados:**

El modelo de apilamiento es una técnica de aprendizaje automático que combina varios modelos básicos para producir un modelo predictivo óptimo. Según Wang y Hao (2011), el aprendizaje en conjunto mejora el rendimiento al crear y agregar múltiples métodos de aprendizaje diferentes con estrategias específicas. La serie construida de modelos son aprendices base, y el método de agregación de aprendices base es estrategia de integración (stacking).

La figura 1 muestra los dos tipos de métodos de ensamblaje, el paralelo y el secuencial. Para el caso del presente artículo, se utilizó el método paralelo que mezcla diferentes modelos de aprendizaje de máquina paralelamente y, al mismo tiempo, la validación cruzada y repeticiones para optimizar la calibración de los hiper parámetros.

**Diagram

Description automatically generatedFigura 1. Tipos de modelos de ensamblaje.**

Fuente: Wang y Hao, 2011.

Los modelos de aprendizaje de máquina que se utilizaron en este proceso son 7: árboles de decisión, K vecinos más cercanos, máquinas vectoriales de soporte, árboles aleatorios, modelos de árboles aleatorios unidos por medio de bagging, modelo lineal generalizado y, por último, AdaBoost, abreviatura de *Adaptive Boosting* (en inglés). A continuación, se explicará cada uno de estos modelos seleccionados:

El conjunto de datos se divide en dos grupos, conjunto de entrenamiento (80% de las muestras) y conjunto de prueba (20% de las muestras), que se utilizan para entrenar el modelo y evaluar el rendimiento, respectivamente. En particular, para garantizar la comparabilidad entre diferentes experimentos, el conjunto de entrenamiento y el conjunto de prueba son los mismos en diferentes clasificadores.

**Árboles de decisión**

Según Wang y Hao (2011), uno de los mejores métodos de aprendizaje de máquina en el área del de la calificación crediticia son los árboles de decisión debido a sus similitudes de clasificación porque al razonamiento humano y es fácil de entender. La idea detrás del DT es resumir una serie de reglas de árboles de decisión de un conjunto de datos con características y etiquetas dividiendo nodos sobre características específicas y usar un diagrama de árbol para presentar el algoritmo de estas reglas.

Para el caso de árboles de decisión, la variable respuesta es categórica y, por el tipo de método de aprendizaje utilizado, estos métodos se pueden resumir en la técnica llamada CART, que significa en inglés Classification and Regression Trees, por lo que buscar generar reglas de decisión para clasificar los datos en las categorías con las que se cuenta.

Una nueva muestra se puede clasificar simplemente usando el árbol de decisión existente y el principio básico es hacer coincidir las características correspondientes y las condiciones relacionadas continuamente hasta llegar a un nodo final o un “hoja”. Es así como se puede descomponer este procedimiento en dos etapas, la construcción del árbol y la poda del mismo.

El orden de las variables para ir colocándolas en el árbol está relacionado con la búsqueda de un árbol más puro, ósea con nodos más completos. Las variables pueden aparecer varias veces en el árbol y el objetivo de la poda busca que no queden nodos tan pequeños, casos individuales o reglas muy repetitivas.

**Máquinas vectoriales de soporte** (SVM, por sus siglas en inglés).

SVM es un clasificador de gran margen que se utiliza para resolver problemas de clasificación binaria y es un método de aprendizaje automático para clasificar datos encontrando el hiperplano óptimo. Específicamente, para lograr la máxima generalización, SVM intenta encontrar el margen de clasificación máximo en el conjunto de datos de entrenamiento como límite de decisión y separa los datos en dos categorías (préstamos buenos y malos). El marco resultante se puede utilizar para estimar la categoría de nuevas muestras.

Los autores Trevor, Tibshirani y Friedman (2008) explican que las máquinas de soporte vectorial buscan encontrar funciones que sean límites lineales en el espacio y separar de la mejor forma las clases con las características que se tienen. A diferencia de análisis discriminante, va a buscar dividir al máximo los dos grupos.

**Redes Neuronales**

Las redes neuronales son un medio de hacer aprendizaje automático, en el que una computadora aprende a realizar alguna tarea mediante el análisis de ejemplos de entrenamiento (IBM Education, 2019).

Modelada libremente en el cerebro humano, una red neuronal consta de miles o incluso millones de nodos de procesamiento simples que están densamente interconectados. La mayoría de las redes neuronales de hoy en día están organizadas en capas de nodos y son "feed-forward", lo que significa que los datos se mueven a través de ellas en una sola dirección. Un nodo individual puede estar conectado a varios nodos en la capa debajo de él, desde donde recibe datos, y a varios nodos en la capa superior, a los que envía datos. (IBM Education, 2019)

A cada una de sus conexiones entrantes, un nodo asignará un número conocido como "peso". Cuando la red está activa, el nodo recibe un elemento de datos diferente, un número diferente, sobre cada una de sus conexiones y lo multiplica por el peso asociado. Luego suma los productos resultantes, dando un solo número. Si ese número está por debajo de un valor de umbral, el nodo no pasa ningún dato a la siguiente capa. Si el número excede el valor de umbral, el nodo se "dispara", lo que en las redes neuronales actuales generalmente significa enviar el número, la suma de las entradas ponderadas, a lo largo de todas sus conexiones salientes. (Zhou, 2019)

Cuando se entrena una red neuronal, todos sus pesos y umbrales se establecen inicialmente en valores aleatorios. Los datos de entrenamiento se envían a la capa inferior, la capa de entrada, y pasan a través de las capas siguientes, multiplicándose y sumando de maneras complejas, hasta que finalmente llegan, radicalmente transformados, a la capa de salida. Durante el entrenamiento, los pesos y los umbrales se ajustan continuamente hasta que los datos de entrenamiento con las mismas etiquetas arrojan resultados similares de manera consistente. (Zhou, 2019)

**Construcción por bloques.**

Las Neuronas son la unidad básica de una red neuronal. Una neurona toma entradas, hace algunas operaciones matemáticas con ellas y produce una salida. A continuación, en la figura 1 se muestra una neurona de 2 entradas:

**Figura 2. Entrada y Salido de una Neurona.**

Diagram

Description automatically generated

Fuente: Zhou, 2019

Los pasos en la elaboración la creación de una neurona se ejemplifica en el siguiente cuadro:

**Cuadro 2. Pasos para la creación de una neurona.**

|  |  |
| --- | --- |
| **Pasos** | **Fórmula** |
| Cada entrada se multiplica por un peso |  |
| Posteriormente, todas las entradas ponderadas se suman junto con un sesgo *b*: | ) + ) + *b* |
| Finalmente, la suma se pasa a través de una función de activación: | + + *b)* |

Fuente: MIT, 2017

La función de activación queda a criterio de investigador y depende de la naturaleza de los datos, pero una función comúnmente utilizada es la función sigmoide , siendo x el valor el valor resultante en el último paso ejemplificado anteriormente. Por ejemplo, si y = *f* (7), el resultado es 0.999, siendo este valor “feed-forward” que alimentará al modelo posteriormente.

**Combinar neuronas en una red neuronal**

Una red neuronal puede tener cualquier cantidad de capas con cualquier cantidad de neuronas en esas capas. La idea básica sigue siendo la misma: alimente la (s) entrada (s) hacia adelante a través de las neuronas en la red para obtener la (s) salida (s) al final.

Las redes neuronales artificiales (ANN) se componen de capas de un nodo, que contienen una capa de entrada, una o más capas ocultas y una capa de salida. Cada nodo, o neurona artificial, se conecta a otro y tiene un peso y un umbral asociados. Si la salida de cualquier nodo individual está por encima del valor de umbral especificado, ese nodo se activa y envía datos a la siguiente capa de la red. De lo contrario, no se transmiten datos a la siguiente capa de la red. (IBM Education, 2019)

Esta red tiene 2 entradas, una capa oculta con 2 neuronas y una capa de salida con 1 neurona (). Una capa oculta es cualquier cada entre la primera capa y el resultado y entre más más complejo sea el modelo, pueden existir muchas capas ocultas. (MIT, 2017)

En conclusión, una red neuronal no es más que un grupo de neuronas conectadas entre sí como se ejemplifica en la figura 2.

**Figura 3: Red Neuronal Grupo de neuronas conectadas.**

**Diagram

Description automatically generated**

Fuente: Zhou, 2019

Para realizar una validación de los resultados, se comparará el para determinar si la predicción realizada por el modelo y sus variables fueron realizadas correctamente. Entre más cercano sea este valor a 0, mejor será la predicción.

La principal diferencia entre la regresión y una red neuronal es el impacto del cambio en un solo peso. Cuando se realiza una regresión lineal simple se puede cambiar un peso sin afectar las otras entradas en una función, sin embargo, cuando se utiliza el método de redes neuronales, dado que la salida de una capa pasa a la siguiente capa de la red por medio del método “feed-forward”, es decir, un solo cambio puede tener un efecto en cascada sobre las otras neuronas de la red.

**Modelos de árboles aleatorios unidos por medio de bagging.**

El algoritmo de embolsado (Bootstrap Aggregating) se utiliza para mejorar la precisión del modelo en problemas de regresión y clasificación. La construcción de múltiples modelos a partir de subconjuntos separados de datos de trenes y la construcción de un modelo final agregado y más preciso es el concepto básico del algoritmo de embolsado. El embolsado es uno de los métodos de aprendizaje de conjuntos.

**Adaptive Boosting (Adaboost)**

Por último, el método adaboost se utilizar junto con muchos otros tipos de algoritmos de aprendizaje para mejorar el rendimiento. La razón por la cual va a ser utilizado en el presente artículo es debido a que se utiliza para modelos stacking sequenciales y busca sacar provecho de las dependencias entre modelos.

La salida de los otros algoritmos de aprendizaje ("aprendices débiles") se combina en una suma ponderada que representa la salida final del clasificador potenciado. Por lo general, AdaBoost se presenta para la clasificación binaria, aunque se puede generalizar a múltiples clases o intervalos acotados en la línea real.

Según Trevor, Tibshirani y Friedman (2008), después de entrenar un clasificador en cualquier nivel, ada-boost asigna peso a cada elemento de entrenamiento. Al elemento mal clasificado se le asigna mayor peso para que aparezca en el subconjunto de entrenamiento del siguiente clasificador con mayor probabilidad. Después de entrenar cada clasificador, el peso también se asigna al clasificador en función de la precisión y es así como al clasificador más preciso se le asigna mayor peso para que tenga más impacto en el resultado final.

En segunda instancia, para los modelos estadísticos, se utilizaron tres modelos derivados de los modelos lineal generalizados: el modelo lineal generalizado clásico, modelo de reducción del sesgo en modelos lineales generalizados de respuesta binomial (BRGLM, por sus siglas en inglés) y, por último, la derivación de los modelos logísticos para casos extraños por parte de los autores Gary King y Langche Zeng. El modelo seleccionado fue BRGLM.

Todos estos modelos serán examinados previamente para verificar supuestos de normalidad, homocedasticidad, detección de valores extremos y de influencia general por medio de DFFITS y Distancia de Cook, correlación entre variables para detectar multicolinealidad y será depurada de ser necesario por criterios de información por medio de los métodos Akaike, Bayes y selección hacia atrás.

Asimismo, los coeficientes del modelo final serán analizado para determinar la significancia de sus coeficientes con un nivel de confianza del noventa y cinco por ciento y su posible efecto en la interpretación de las variables.

En el caso de la selección de las variables, se inicia con un modelo completo con todos los predictores y posteriormente es sometido a diferentes pruebas donde se elimina el predictor que tenga el valor de p más alto al valor establecido por el modelo. Posteriormente, se vuelve a ajustar el modelo y se elimina el siguiente predictor con el mismo procedimiento y, este es completado cuando ningún predictor tiene un valor de p mayor al alfa seleccionado.

El criterio de información BIC penaliza con a los modelos con más variables como se observa en la ecuación 2 y tiende a preferir modelos más pequeños en comparación con el AIC. Este proceso de selección lo realiza con el objetivo del balance entre el ajuste de los datos y el tamaño del modelo.

**Ecuación 1:** AIC = -2 *log-likelihood* + 2p

**Ecuación 2:** BIC = -2 *log-likelihood* + *p* log(*n*)

**Selección de la muestra**

Los autores Gary King y Langche Zeng en su publicación argumentan que, si bien la estrategia habitual es el muestreo aleatorio en los modelos estadísticos, esto podría generar resultados no favorables para muestras no balanceadas. Es así como ellos recomiendan, pero no limitan, métodos de muestreo estratificado endógeno o basado en la elección. Para el presente artículo, se seleccionó una muestra basada en la elección para las observaciones con la variable explicativa con minoría y se obtuvo una muestra aleatoria para el resto de las observaciones.

Asimismo, se estandarizo las variables limite tarjeta crédito y saldo tarjeta con la finalidad de mitigar el efecto de la varianza en el modelo estadístico. Al igual que los modelos de aprendizaje de máquina, las variables plazo y destino fueron removidas del modelo.

**Agrupación de variables categóricas**

Las variables código de trabajo y código de sector presentan la limitan de que son variables categóricas con muchos valores únicos. Desde un punto de vista de un modelo mixto que contiene variables numéricas y categóricas, esto presenta un problema porque la interpretabilidad es limitada e incumple con el supuesto estadístico de parsimonia. Debido a lo anterior, se procede a agrupar estas variables según el clasificador del INEC del 2011 del Manual de Clasificación de Agrupaciones.

**Cuadro 3. Clasificación de ocupaciones de costa rica (COCR-2011)**

|  |  |
| --- | --- |
| **ID** | **Grupos Mayores** |
| 1 | Directores y Gerentes. |
| 2 | Profesionales científicos e intelectuales. |
| 3 | Técnicos y profesionales de nivel medio. |
| 4 | Personal de apoyo administrative. |
| 5 | Trabajadores de los servicios y vendedores de comercios y mercados. |
| 6 | Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros. |
| 7 | Oficiales, operarios y artesanos de artes mecánicas y de otros oficios. |
| 8 | Operadores de instalaciones y máquinas y ensambladores. |
| 9 | Ocupaciones elementales. |

Fuente: INEC, 2011.

**RESULTADOS**

**Análisis exploratorio de los datos agregados de los tarjetahabientes.**

A continuación, se inicia el apartado realizando un análisis exploratorio de los datos de corte transversal con el objetivo de caracterizar la población. En el cuadro 4 se observan las características principales de la población, donde se destaca el desequilibrio de la calificación de los tarjetahabientes en buenos pagadores y malos pagadores.

El 92% de los tarjetahabientes son considerados buenos pagadores y el restante 8%. En ámbito bancario, esta disparidad entre la cantidad de tarjetahabientes en cada grupo de estudio es común, especialmente si es cliente ya se encuentra registrado en el sistema financiero nacional por medio de una tarjeta de crédito en colones. Un principio fundamente de las entidades bancarias es minimizar la cantidad de personas que solicitan un crédito y posteriormente no cuentan con la capacidad adquisitiva para hacer frente a la deuda.

**Cuadro 4. Características principales de la población según la clasificación del pagador y su género, 2020.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Género** | **Cantidad de tarjetahabientes** | **Promedio Limite Tarjeta Crédito** | **Promedio Saldo Tarjeta** | **Promedio Coeficiente solvencia** |
| **Bueno** | **13 334** | **2 104 411** | **818 028** | **44%** |
| Femenino | 4 032 | 1 692 235 | 651 874 | 44% |
| Masculino | 9 302 | 2 283 070 | 890 048 | 44% |
| **Malo** | **1 166** | **1 756 493** | **564 643** | **36%** |
| Femenino | 344 | 1 518 822 | 458 519 | 34% |
| Masculino | 822 | 1 855 956 | 609 055 | 37% |

Fuente: Sistema Financiero Nacional, 2020.

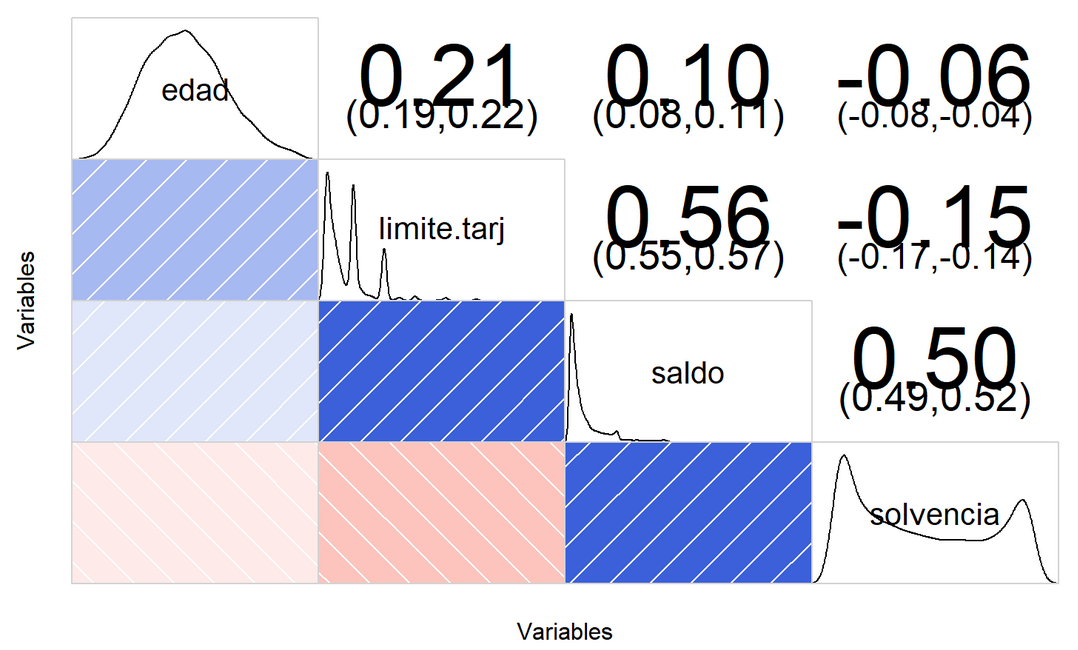
Asimismo, se observa los tarjetahabientes clasificados buenos pagadores tiene en promedio más límite de crédito, más saldo en la tarjeta y más coeficiente de solvencia, por lo que se determinará que entre más grandes sean estas variables, mejores serán las posibilidades para el tarjetahabiente de ser considerado buen pagador.

Si bien en el estudio hay más hombres que mujeres, 10 124 hombres en comparación a 4 376 mujeres, la proporción de malos pagadores para ambos géneros es de un 8% y el promedio del coeficiente de solvencia es de un 44%. Sin embargo, en las otras categorías como promedio de límite de crédito y saldos en tarjeta de crédito si se observa una diferencia entre ambos grupos. Debido a que la variable principal del estudio es “Buen pagador” y no existe una diferencia significativa a un 5% entre los géneros, se decide mantener la variable dentro del estudio.

Con respecto a la agrupación de variables por sector, inicialmente se utilizaron todos los grupos mayores derivados de la Clasificación De Ocupaciones de Costa Rica (COCR-2011) como se observa en el cuadro 1 de los anexos. Para el análisis de aprendizaje de máquinas, mantuvo todos los grupos mayores para aprovechas sus capacidades de clasificación, sin embargo, para el modelo estadístico, se utilizó solo los tres grupos con mayor cantidad de tarjetahabientes y se agrupó los demás grupos en una categoría nueva llamada “Otros”.

Dentro de COCR, el 40% de las personas de la población labora en profesiones catalogadas grupos “Profesionales científicos e intelectuales”, posteriormente el 18% de los trabajadores laboran en “Servicios y vendedores de comercios y mercados”, seguidamente con un 17% se encuentra el “Personal de apoyo administrativo” y los restantes 6 grupos se representan un 24%. Con respecto a la variable código del sector de trabajo, el 50% de la población labora como empleado asalariado en el sector privado, seguidamente el 35% labora en el sector público y los restantes 20 sectores de trabajo corresponden al 15% de la población. Al igual que los grupos, se especifican estos dos sectores dentro del modelo estadístico y los demás 20 sectores son agrupan en una categoría llamada “Otros”.

Para conocer la relación de las variables, se procedió a calcular las correlaciones de Pearson de las variables explicativas del estudio. Se denota una relación relativamente débil en todas las variables, siendo la correlación entre el saldo y el límite de la tarjeta la más directa y significativa. Seguidamente, la solvencia presenta una relación positiva con el saldo, pero una relación negativa con el límite de la tarjeta. Por último, se observa una relación directa y positiva entre la edad y las variables límite de tarjeta y saldo.

**Gráfico 1. Correlaciones de Pearson de las variables del estudio.**

Fuente: Sistema Financiero Nacional, 2020.

La edad promedio de la población es de 44 años tanto para hombres y mujeres y este promedio se mantuvo constante cuando se realizan tablas de contingencia con las demás variables del estudio. Por otra parte, en el cuadro 2 de los anexos se muestra la cantidad de tarjetahabientes según su estado civil. Esta variable presenta 7 categorías y el 95% de los tarjetahabientes se encuentran concentrado en dos de ellas, estado civil casado y soltero.

Trevor, Tibshirani y Friedman (2008), recomiendan que los valores faltantes de las bases de datos de aprendizaje de máquina no sean tratados, cuando exista la posibilidad, y por lo tanto, sean desechados del modelo de estudio. Ellos argumentan que al realizar modelos de predicción de capacidad de pago y clasificación crediticia, es importante analizar los datos por cliente para ir construyendo perfiles basado en su historial de compra, saldo en la tarjeta y valor promedio de las transacciones para determinar eficientemente si un cliente va a ser frente a sus deudas.

Asimismo, cuando se aplica aprendizaje de máquina por medio de stacking, y como se mencionó en el apartado de materiales y métodos, la metodología y la formulación tienen el objetivo de buscar con certeza los casos donde hay más confianza de predicción basado en las variables del estudio. A estos casos les asigna un peso en la modelación matemática y cuando comienza una nueva iteración, estos valores que contienen una mayor ponderación, son utilizados posteriormente para la predicción de los casos.

Es así como los autores mencionan que los casos con datos faltantes si bien se pueden mantener en el estudio, se les va a asignar un peso muy inferior a los casos con datos completos. Es así, como se toma la decisión de apartarlos antes de aplicar la modelación estadística y los métodos de aprendizaje de máquinas. En total, la población para de 14 500 datos a 14 424.

**Modelo de aprendizaje de máquinas**

Según Dumitrescu et al. (2021), las clases desequilibradas impiden la clasificación en primera instancia de las variables de estudios porque algunos clasificadores pueden centrarse demasiado en la clase mayoritaria y descuidar el grupo minoritario (de interés). Por lo tanto, pueden exhibir un buen desempeño general a pesar de identificar mal al grupo minoritario, es decir, los tarjetahabientes que incumplen con sus deudas.

Es así como el primer paso es realizar un ajuste de la población sin los valores faltantes y para el caso de los modelos de aprendizajes se seleccionó la técnica de sobre muestreo de minorías sintéticas (SMOTE, por sus siglas en inglés). Según Microsoft (2021), SMOTE es una técnica estadística para mitigar el desequilibrio en el conjunto de datos y resuelve los inconvenientes estadísticos que presentan el sobre muestreo o la duplicación de la clase minoritaria. En su lugar, SMOTE tiene como objetivo sintetizar nuevos casos a partir de observaciones existentes.

El objetivo del algoritmo es tomar muestras del conjunto de características para cada clase objetivo y genera nuevas observaciones que combinan características del caso objetivo con características de sus vecinos. Para el presente artículo, el SMOTE fue calibrado en aumentar un 200% el tipo de tarjetahabiente clasificado como “malo” y disminuir un 200% el “bueno”. Este enfoque aumenta las funciones disponibles para cada clase y hace que las muestras sean más generales.

A partir de esto enfoque, la cantidad de observaciones que se ingresan a los modelos de aprendizaje de máquina fue de 6 503, siendo 3 716 clientes clasificados como “bueno” y 2 787 clasificado como “malo”. A partir de esta muestra se comenzó el proceso de calibración de los hiper parámetros de cada uno de los modelos que se mencionaron en el apartado de materiales y métodos. Cada una de estas calibraciones fueron procesadas por un método de validación cruzada con diez iteraciones con la finalidad de buscar consistencia en los resultados y los resultados se observan en el apartado de los anexos.

**Modelo Stacking**

Como se mencionó en el apartado de materiales y métodos, el método de aprendizajes de máquina seleccionado fue el método stacking que incluye los cinco modelos con mejores resultados. Estos modelos fueron introducidos al modelo stacking con sus calibraciones correspondientes y tienen el objetivo de mejorar los indicadores de desempeño. A continuación, en el cuadro 5 se muestra una tabla resumen de los cinco modelos calibrados y sus respectivos hiper parámetros.

Estos parámetros fueron seleccionados a partir de la curva de ROC y el área bajo la curva (AUC). El objetivo de calibrar utilizando como base el AUC es porque muestra el desempeño global del modelo de clasificación en todos los umbrales de clasificación.

**Cuadro 5. Modelos apilados con sus respectivos hiper parámetros.**

|  |  |
| --- | --- |
| **Modelo** | **Hiper parámetros** |
| Árboles de decisión | Complejidad = 0.0065  Profundidad máxima = 3 |
| SVM | Función de activación: Radial |
| Redes Neuronales | Cantidad de iteraciones (maxit) = 150  Capas ocultas = 10  Rango = 0.02 |
| Bosques aleatorios | Máximo nodos = 201  Máximas variables aleatorias (mtry) = 2 |
| Adaboost | Numero de árboles (iter). = 201  Parámetro de contracción (nu) = 0.1 |

**Fuente**: Sistema Financiero Nacional, 2020.

Estos modelos por sí solos obtuvieron resultados aceptables en comparación con los indicadores de desempeño propuestos en el apartado de materiales y métodos. Sus resultados se observan en el cuadro 6.

**Cuadro 6. Resultados de los modelos calibrados en los hiper parámetros según los indicadores de desempeño**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | **Precisión** | **Verdadero positivo (VP)** | **Verdadero Negativo (VN):** | **Falso Positivo (FP):** | **Falso Negativo (FN):** | **AUC** | **KS** |
| Árboles de decisión | 0.89 | 0.94 | 0.27 | 0.72 | 0.059 | 0.71 | 38 |
| SVM | 0.77 | 0.93 | 0.13 | 0.86 | 0.06 | 0.7 | 30 |
| Redes Neuronales | 0.92 | 0.96 | 0.13 | 0.86 | 0.03 | 0.63 | 12 |
| Bosques aleatorios | 0.84 | 0.93 | 0.597 | 0.402 | 0.068 | 0.86 | 58 |
| Adaboost | 0.87 | 0.94 | 0.28 | 0.71 | 0.05 | 0.85 | 57 |

**Fuente**: Sistema Financiero Nacional, 2020.

Se observa que los valores de precisión están, en promedio, por encima del 80% indicando que los resultados son aceptables. Sin embargo, el indicador falso positivos es alarmantemente alto para todos los modelos. Esto indica que en la gran mayoría de casos se indica que el cliente es buen pagador, pero realmente no cuenta con los medios para pagar sus deudas. (escenario más crítico). Es así como nace la necesidad de realizar un método que busque disminuir este escenario lo más posible para evitar

El modelo stacking, por medio de la librería estadística Caret en R, se agrega la cantidad de algoritmos que se van a incorporar al ensamblaje con el objetivo de encontrar la mejor combinación lineal de varios modelos de clasificación o regresión, utilizando la regresión lineal.

Los resultados del método Stacking son los siguientes:

**Cuadro 7. Resultados de los modelos stacking según los indicadores de desempeño**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | **Precisión** | **Verdadero positivo (VP)** | **Verdadero Negativo (VN):** | **Falso Positivo (FP):** | **Falso Negativo (FN):** | **AUC** | **KS** |
| Stacking  5 modelos | 0.97 | 0.93 | 0.65 | 0.27 | 0.052 | 0.87 | 65 |

**Fuente**: Sistema Financiero Nacional, 2020.

Se observa una mejora considerable en la precisión del modelo y una reducción de los falsos positivos, indicando que el método de stacking cumplió el objetivo de encontrar la mejor combinación línea de los modelos seleccionados.

**Modelo estadístico**

En el gráfico 2 se muestra la importancia relativa de las 5 variables con mayor ponderación según los criterios de información. Esto nos da una idea de que la variable solvencia es la que aporta más al modelo general.

**Gráfico 2. Variables según el criterio de información**

**Chart, bar chart, histogram

Description automatically generated**

**Fuente**: Sistema Financiero Nacional, 2020.

Por otra parte, se sometió el modelo lineal generalizado a los procedimientos de selección de variables por el método “Hacia atrás”, AIC y BIC con el fin de determinar cuáles era posibles de eliminar. Esto procedimiento se utilizó como referencia porque como mencionan diferentes académicos, el criterio de experto es valioso en esta eliminación o selección de variables por su sustento teórico.

El método “Hacia atrás” y “AIC” recomendaron eliminar las variables “estado civil” y “código del sector” por su significancia, valor AIC y devianza. El método con BIC, método que toma en cuenta la cantidad de variables del modelo, propuso mantener únicamente dos variables el “límite de la tarjeta” y la “solvencia”, coincidiendo con los resultados obtenidos en el gráfico 2. Sin embargo, para mantener la congruencia y poder realizar una comparación con los modelos de aprendizaje de máquina, se decidió únicamente eliminar las variables “estado civil” y “código del sector”.

**Cuadro 8. Resultados del modelo de reducción del sesgo en modelos lineales generalizados de respuesta binomial.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coeficientes** | **Estimate** | **Std. Error** | **z value** | **Pr(>|z|)** |
| (Intercept) | 1.57 | 0.16 | 9.63 | 0 |
| Limite.tarj | 0.31 | 0.05 | 5.77 | 0.00 |
| Saldo | -0.10 | 0.07 | -1.47 | 0.14 |
| Genero masculino | -0.10 | 0.07 | -1.47 | 0.14 |
| Solvencia | 0.96 | 0.14 | 7.02 | 0.00 |
| INEC- Intelectuales y científicos | 0.33 | 0.08 | 3.92 | 0.00 |
| INEC-OTRAS | 0.18 | 0.11 | 1.59 | 0.11 |
| INEC-Personal Apoyo | 0.42 | 0.10 | 4.14 | 0.00 |
| INEC-Prof. nivel med | 0.35 | 0.11 | 3.31 | 0.00 |
| Edad | 0.01 | 0.00 | 2.34 | 0.02 |

**Fuente**: Sistema Financiero Nacional, 2020.

Se observa que las todas las variables exceptuando saldo y el género e INEC-OTRAS son variables estadísticamente significativas con una alfa de 0.05. Es decir, el modelo indica que el beta o coeficiente asociado a esas tres variables es igual a 0 y se podrían eliminar del modelo. Es digno de destacar que el género haya sido una de las variables seleccionadas porque anteriormente en el artículo se mencionó que la proporción de variables no balanceadas era proporcional entre hombres y mujeres.

Por su parte, la variable saldo en la cuenta bancaria de la persona también fue seleccionada como no significativa. Se podría argumentar que si bien una persona puede tener mucho saldo en su cuenta bancaria, el monto que realmente importa es el saldo de la deuda bancaria que tienen que hacer frente. En el presente artículo, esta variable se mantendrá en el modelo, al igual que las otras, pero se recomendará en las conclusiones un criterio de experto.

Un aumento de una unidad en la variable predictora “límite de crédito”  
está asociado con un cambio promedio de 0.31 en las probabilidades logarítmicas (log odds) de la variable de respuesta “buen pagador” tomando un valor de 1. Esto significa que valores más altos del “límite de crédito” están asociados con una menor probabilidad de que el tarjetahabiente sea un mal pagador.

El gráfico 3 muestra los resultados de la validación de los resultados. Se muestra todo es espectro de resultados en el eje x pero, como se mencionó al inicio de este apartado, el desequilibrio de las variables “bueno” con respecto a “malo” corresponde a un 92%. Con este resultado, y tomando como base el eje x, se obtiene como valor de precisión un 85%, sin embargo, los otros indicadores no son favorables en comparación a los resultados del modelo de aprendizaje de máquina apilamiento.

Chart, line chart

Description automatically generated**Gráfico 3. Resultados de la validación cruzada del modelo según los indicadores de desempeños.**

**Fuente**: Sistema Financiero Nacional, 2020.

**CONCLUSIONES**

Se llega a la conclusión que el modelo de apilamiento, stacking, es el modelo que mejor predice los datos por medio de los indicadores de desempeño. Este modelo es intensivo y tiene la característica de mejorar la predicción de los datos a partir de una combinación línea y diferentes repeticiones. A continuación, se muestra los resultados de las predicciones, concordando con lo observado en el cuadro 4.

**Cuadro 9. Valores pronosticados por el método estadístico de apilamiento y las características de las variables más relevantes.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Buen Pagador?** | **Tarjetahabiente** | **Promedio Limite Tarjeta Crédito** | **Promedio Saldo Tarjeta** | **Promedio Coeficiente solvencia** |
| bueno | 318 | 2 103 447 | 852 402 | 45% |
| malo | 35 | 1 912 993 | 519 400 | 33% |
| Resumen | 353 | 2 084 564 | 819 385 | 44% |

**Fuente**: Sistema Financiero Nacional, 2020.

Para futuras investigaciones, se recomienda explorar diferentes librerías de R para aplicar el concepto de meta-aprendiz por medio de los métodos de apilamiento y conducir más exploración teórica para un abordaje, específicamente en el tema de clasificación crediticia, con un mayor fundamente teórico.

**BIBLIOGRAFÍA**

Banco Mundial (2012). Buenas Prácticas para la Protección al Consumidor Financiero (p. 13).

Ramos, V. (2017). Con poca educación financiera, deuda de jóvenes crece en un 19% en tres años. El Financiero. Recuperado de: <https://www.elfinancierocr.com/finanzas/con-poca-educacion-financiera-deuda-de-jovenes/COU27MSINNFSLMCZFQEHVFOAMI/story/>.

Organización para la Cooperación y el Desarrollo Económico. (2017). OECD reviews of labour market and social policies (Primera publicación). Recuperado de: http://dx.doi.org/10.1787/9789264282773-e.

Chacón, N.A. y Mora, D.E. (2015). *Los derechos del consumidor financiero en la Nueva Arquitectura Financiera Internacional: su aplicación en Costa Rica*. [Tesis para optar por el grado de Licenciatura en Derecho]. Universidad de Costa Rica.

Presidencia de la República de Costa Rica. (2019). Gobierno lanza Estrategia Nacional de Educación Financiera. Recuperado de: <https://www.presidencia.go.cr/comunicados/2019/01/gobierno-lanza-estrategia-nacional-de-educacion-financiera/>

Ministerio de Economia, Industria Y Comercio de Costa Rica. (2019). *Estudios de tarjeta*. Ministerio de Economia, Industria y Comercio de Costa Rica - Índice. https://www.meic.go.cr/web/284/estudios/tarjetas-de-credito-y-debito/tarjetas-credito-debito.php

McKinsey & Company. (2022). *Reinventing credit cards: Responses to new lending models in the US*. <https://www.mckinsey.com/industries/financial-services/our-insights/reinventing-credit-cards-responses-to-new-lending-models-in-the-us>

Superintendencia General de Entidades Financieras. (2022). *Website*. index. <https://www.sugef.fi.cr/sugef/objetivos_funciones.aspx>

Sistema Costarricense de Información Jurídica. (1995). *Reglamento de Tarjetas de Crédito y Débito N° 35867-MEIC*. Procuraduría General de la República-Servicios en Línea. <https://www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_texto_completo.aspx?param1=NRTC&nValor1=1&nValor2=67564&nValor3=94010&strTipM=TC>

Massachusetts Institute of Technology: MIT. (2021). *Machine learning, explained*. MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Banco Central de Costa Rica. (2021). *Medios de Pago Estadísticas 2020*. BCCR - Inicio. https://www.bccr.fi.cr/sistema-de-pagos/DocEstadisticas/Informe-estadistico-medios-de-pago-2020.pdf

Instituto Nacional de Estadísticas y Censos. (2011). *Clasificación de Ocupaciones de Costa Rica 2011*. Sistema Nacional de Estadística. <https://sistemas.inec.cr/sitiosen/sitiosen/FrmOcupacion.aspx>

Wang, G., & Hao, J. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1). https://doi.org/10.1016/j.eswa.2010.06.048

IBM Education. (2019). *What are Neural Networks?*. Ibm.com. <https://www.ibm.com/cloud/learn/neural-networks>.

Zhou, V. (2019). *Machine Learning for Beginners: An Introduction to Neural Networks*. Medium. <https://towardsdatascience.com/machine-learning-for-beginners-an-introduction-to-neural-networks-d49f22d238f9>.

MIT. (2017). *Explained: Neural networks*. MIT News | Massachusetts Institute of Technology. https://news-mit-edu.translate.goog/2017/explained-neural-networks-deep-learning-0414?\_x\_tr\_sl=en&\_x\_tr\_tl=es&\_x\_tr\_hl=en-US&\_x\_tr\_pto=nui.

Hastie Trevor, Tibshirani Robert, Jerome Friedman. (2008) The element of the statistical learning. Stanford, California. Springer Editorial.

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2020). Machine learning or econometrics for credit scoring: Let's get the best of both worlds. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.3553781>

Microsoft. (2021). Machine Learning - Smote. Microsoft Learn: Build skills that open doors in your career. https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/smote

# ANEXOS:

**Cuadro 1. Cantidad de tarjetahabientes, promedio de del límite de crédito y promedio del saldo de la tarjeta según variable “Buen Pagador” y COCR-2011. Año 2022**

|  |  |  |  |
| --- | --- | --- | --- |
| **Buen Pagador y clasificaciones de ocupación del INEC** | **Clientes** | **Promedio Limite Tarjeta Crédito** | **Promedio Saldo Tarjeta** |
| **Bueno** | **13 334** | **2 104 411** | **818 028** |
| Profesionales científicos e intelectuales | 5 333 | 2 571 692 | 987 790 |
| Trabajadores de los servicios y vendedores de comercios y mercados | 2 395 | 2 470 098 | 920 291 |
| Personal de apoyo administrativo | 2 335 | 1 560 688 | 641 222 |
| Técnicos y profesionales de nivel medio | 1 974 | 1 474 681 | 604 803 |
| Oficiales, operarios y artesanos de artes mecánicas y de otros oficios | 371 | 1 034 403 | 459 471 |
| Directores y Gerentes | 330 | 1 683 301 | 700 443 |
| Ocupaciones elementales | 261 | 1 745 435 | 603 346 |
| Operadores de instalaciones y máquinas y ensambladores | 236 | 1 064 827 | 457 322 |
| Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros | 99 | 2 304 572 | 782 367 |
| **Malo** | **1 166** | **1 756 493** | **564 643** |
| Profesionales científicos e intelectuales | 419 | 2 208 255 | 718 619 |
| Trabajadores de los servicios y vendedores de comercios y mercados | 260 | 2 141 414 | 694 520 |
| Personal de apoyo administrativo | 187 | 994 407 | 289 489 |
| Técnicos y profesionales de nivel medio | 167 | 1 077 580 | 345 594 |
| Oficiales, operarios y artesanos de artes mecánicas y de otros oficios | 55 | 1 071 259 | 438 101 |
| Directores y Gerentes | 34 | 1 974 353 | 603 572 |
| Operadores de instalaciones y máquinas y ensambladores | 23 | 1 364 257 | 596 167 |
| Ocupaciones elementales | 17 | 1 899 353 | 125 234 |
| Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros | 4 | 2 605 000 | 1 097 696 |
| **Resumen** | **14 500** | **2 076 433** | **797 652** |

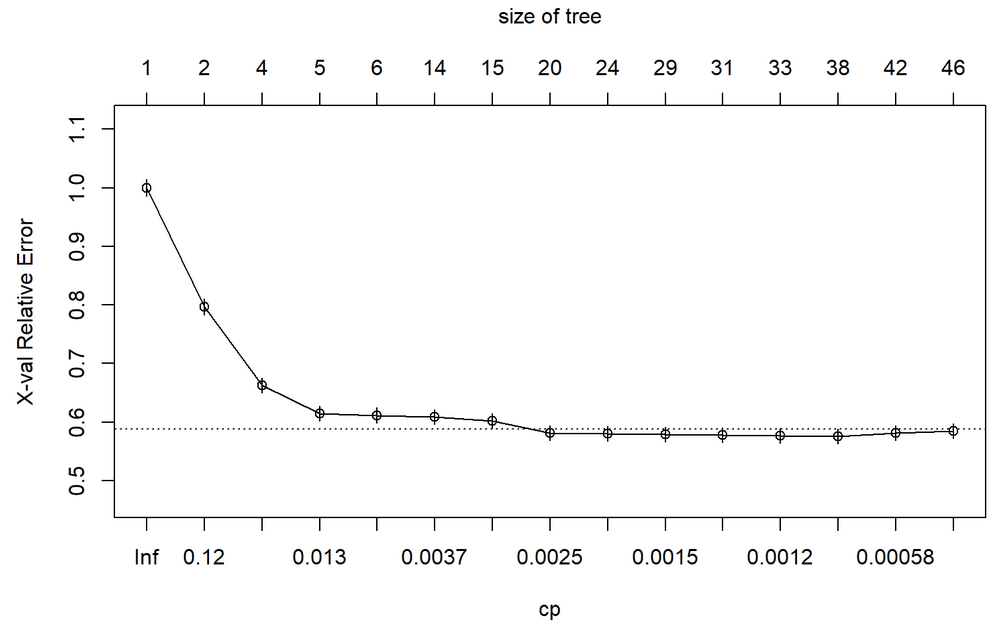
**Fuente:** Sistema Financiero Nacional (2020) e INEC (2011).

**Cuadro 2. Cantidad de tarjetahabientes por estado y género según su estado civil.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Estado** | **Genero** | **Casado** | **Divorciado** | **Religioso** | **Separado** | **Soltero** | **Unión libre** | **Viudo** |
| **bueno** | Femenino | 1904 | 257 | 0 | 40 | 1755 | 17 | 54 |
| Masculino | 5973 | 245 | 2 | 23 | 2981 | 48 | 20 |
| **malo** | Femenino | 170 | 22 | 0 | 5 | 142 | 0 | 3 |
| Masculino | 487 | 22 | 0 | 2 | 298 | 8 | 3 |

**Fuente:** Sistema Financiero Nacional (2020) e INEC (2011).

**Gráfico 1. Calibración del hiper parámetro de complejidad (cp) según el tamaño de árbol.**

**.**

**Fuente:** Sistema Financiero Nacional (2020)

**Gráfico 2. Calibración del hiper parámetro de complejidad (cp) basado en el AUC.**

**Chart

Description automatically generated**

**Fuente:** Sistema Financiero Nacional (2020)

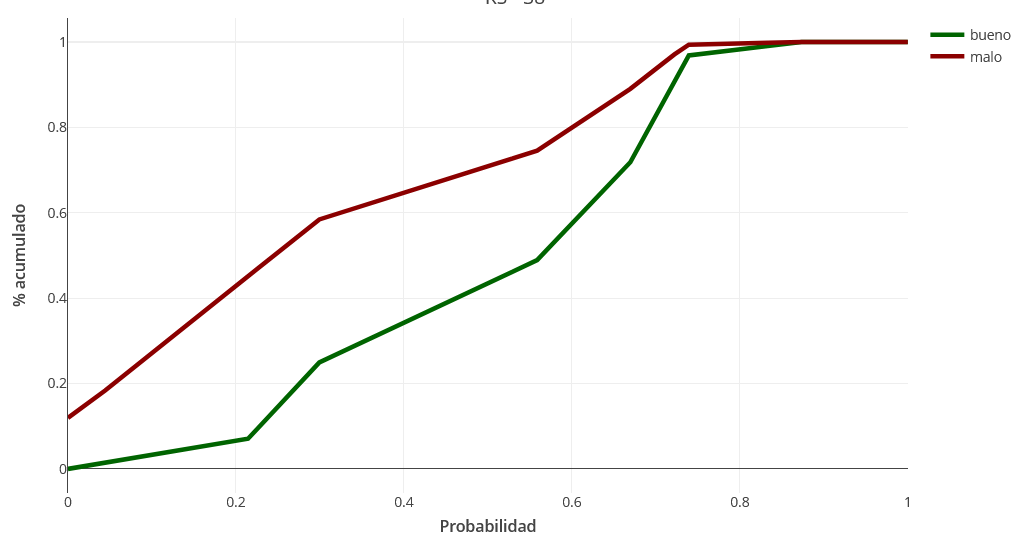
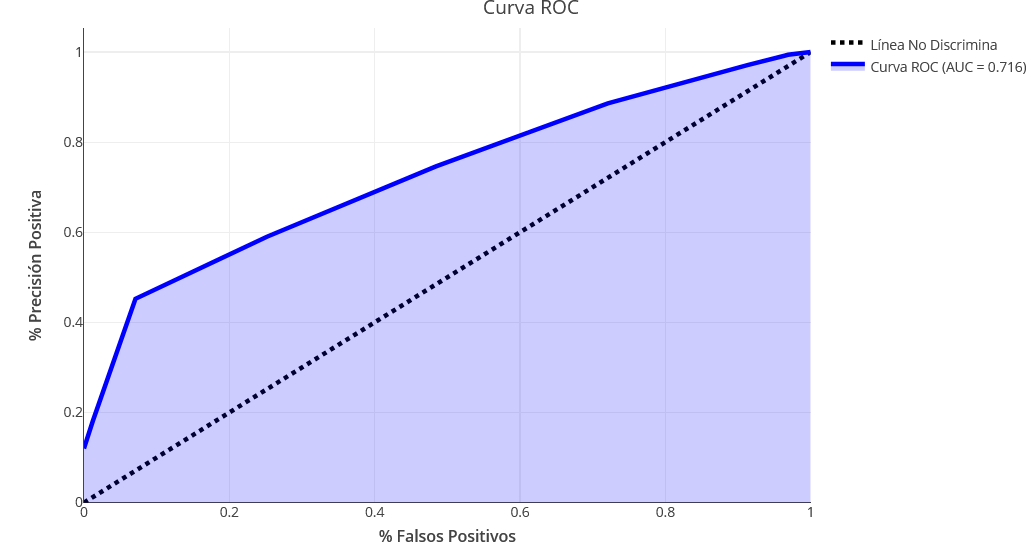
**Gráfico 3. Calibración de la profundidad máxima de los árboles**

**Chart, line chart

Description automatically generated**

**Fuente:** Sistema Financiero Nacional (2020)

**Gráfico 4. Curvas de ROC y KS del modelo árboles de decisión**



**Fuente:** Sistema Financiero Nacional (2020)

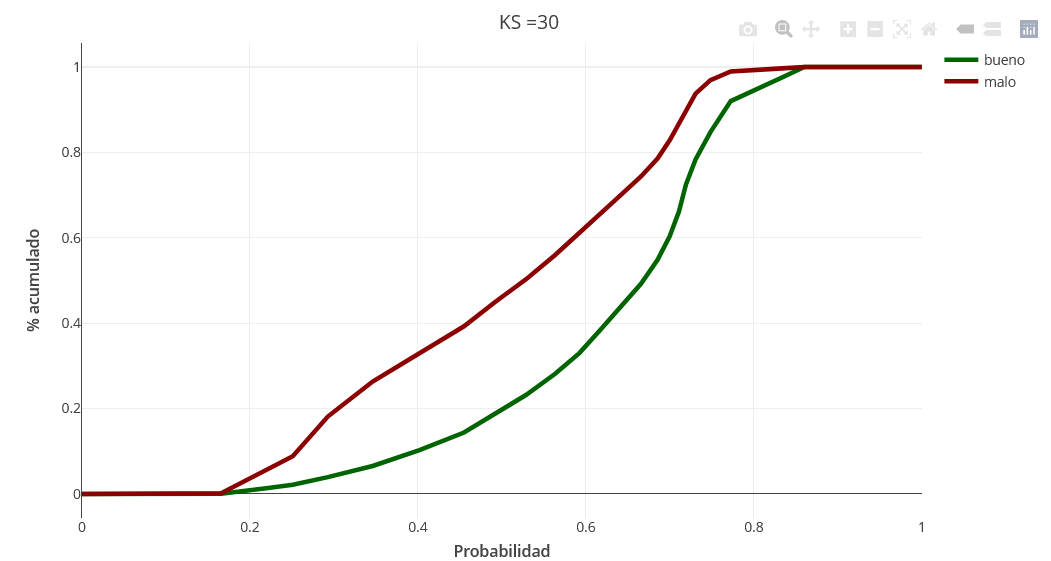
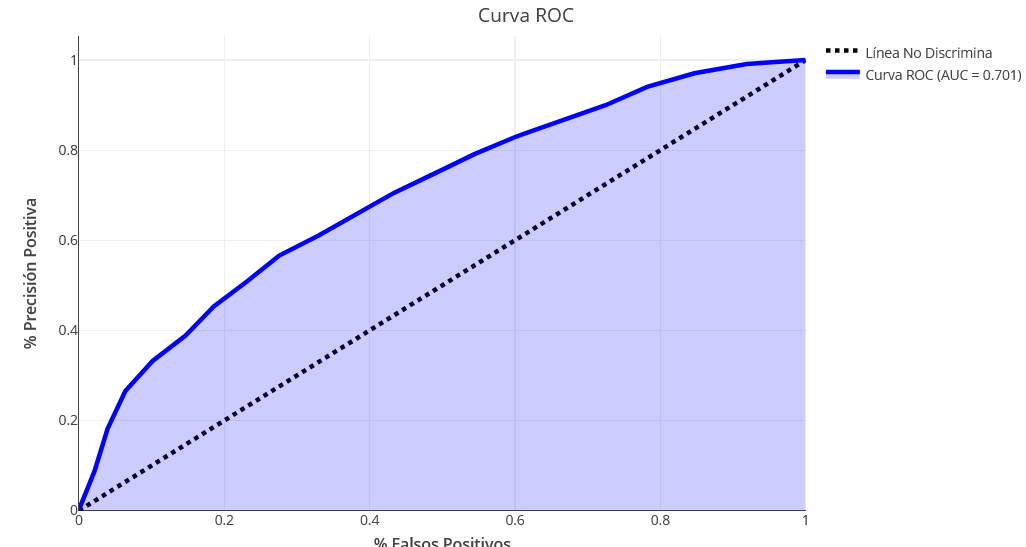
**Gráfico 5. AUC del modelo máquinas vectoriales de soporte (SVM) según el AUC**

**Chart, bar chart

Description automatically generated**

**Fuente:** Sistema Financiero Nacional (2020)

**Gráfico 6. Curvas de ROC y KS del modelo máquinas vectoriales de soporte (SVM) según el AUC**



**Fuente:** Sistema Financiero Nacional (2020)

**Gráfico 7. Selección del número de capas ocultas en** **el modelo de redes neuronales según AUC.**

**Chart, line chart

Description automatically generated**

**Fuente:** Sistema Financiero Nacional (2020)

**Gráfico 8. Selección del rango en el modelo de redes neuronales**

**Chart, line chart

Description automatically generated**

**Fuente:** Sistema Financiero Nacional (2020)

**Gráfico 9. Curvas de ROC del modelo de redes neuronales**

**Chart, histogram

Description automatically generated**

**Fuente:** Sistema Financiero Nacional (2020)

**Chart, line chart

Description automatically generatedGráfico 10. Calibración de las variables aleatorias del modelo de bosques aleatorios**

**Fuente:** Sistema Financiero Nacional (2020)

**Gráfico 11. Calibración y selección de la cantidad de árboles del modelo bosques aleatorios según el AUC.**

**Chart, line chart

Description automatically generated**

**Fuente:** Sistema Financiero Nacional (2020)

**Gráfico 12. Curvas de ROC del modelo bosques aleatorios según el AUC.**

**Chart, line chart

Description automatically generated**

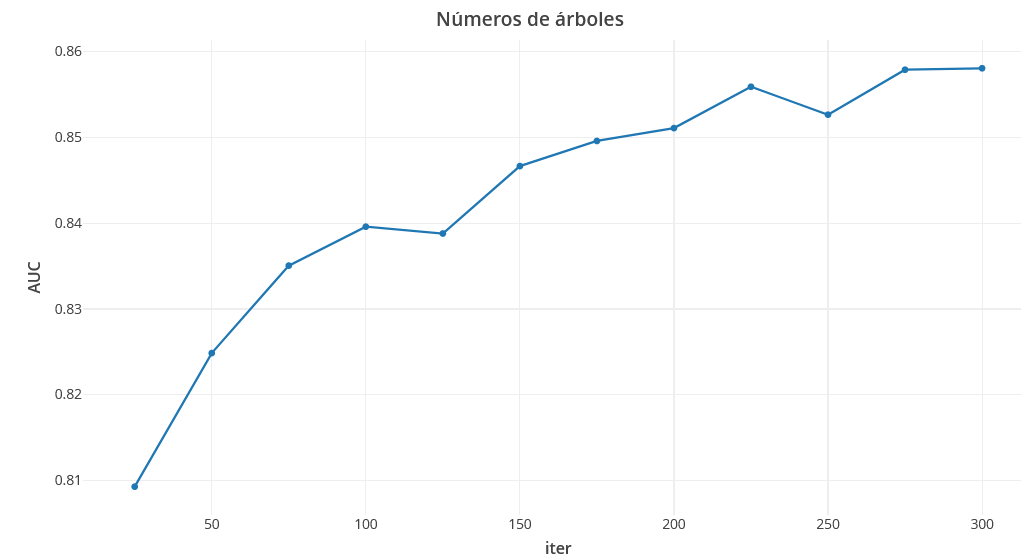
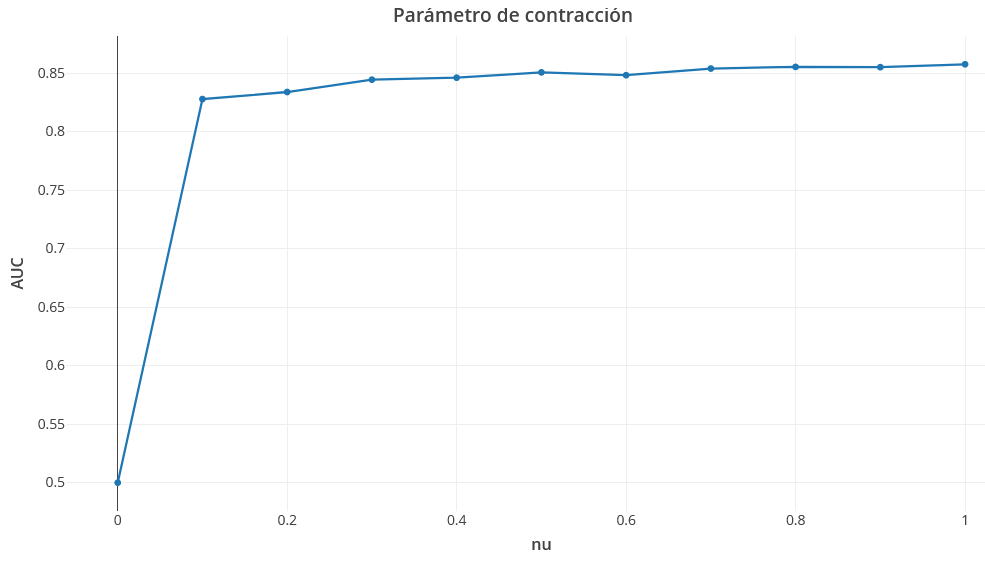
**Fuente:** Sistema Financiero Nacional (2020)

**Gráfico 12. Curvas KS del del modelo bosques aleatorios según el AUC.Chart, line chart

Description automatically generated**

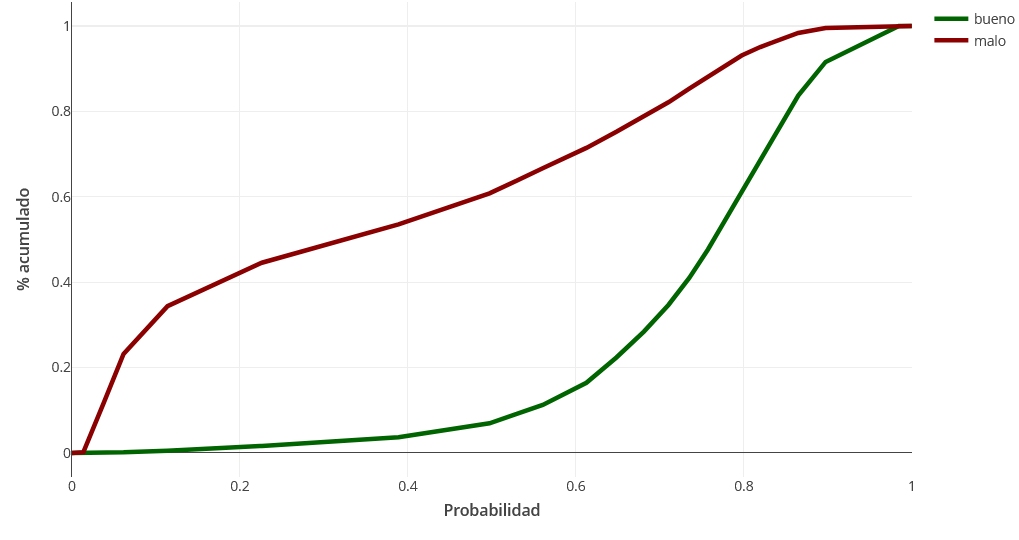
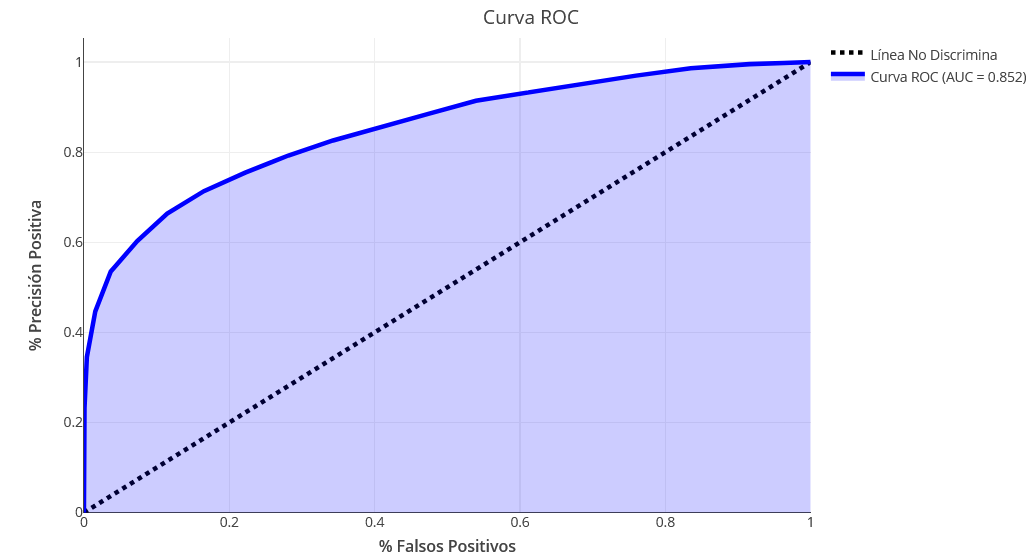
**Fuente:** Sistema Financiero Nacional (2020)

**Gráfico 13. Calibración del parámetro de contracción y el número de árboles en el modelo Adaboost.**



**Fuente:** Sistema Financiero Nacional (2020)

**Gráfico 14. Curvas de ROC y KS del modelo Adaboost según AUC.**



**Fuente:** Sistema Financiero Nacional (2020)