Universidad De Bogotá Jorge Tadeo Lozano

Maestría En Ingeniería Y Analítica De Datos

Identificación Predictiva de Fallos POS A Través del Uso de Algoritmos de Aprendizaje Automático

Presenta:

Carlos Huérfano Lenis

Director:

Sebastián Zapata Ramírez

Bogotá D.C. Abril de 2020

Identificación Predictiva de Fallos POS A Través del Uso de Algoritmos de Aprendizaje Automático (Abril de 2020)

Carlos Huérfano Lenis. Maestría en Ingeniería y Analítica de Datos. Universidad Jorge Tadeo Lozano.

Resumen - El proyecto tuvo la finalidad de aplicar algoritmos de Aprendizaje Automático para lograr la identificación predictiva de fallas en los POS (Puntos de venta o Point of sale por sus siglas en inglés) de la red de Credibanco, lo que permitió a la entidad mantener su red activa (core del negocio) y orientar positivamente su modelo operativo para beneficiarse con la mejora los índices de experiencia de cliente.

En el presente trabajo se pretendió elaborar un modelo que identificara la presencia de patrones de comportamiento anómalo en los dispositivos, analizando variables transaccionales, de software y de hardware, para asociarlos de esta manera a las fallas en los POS que se encontraban en producción.

Con el apoyo de la librería de Python, Sklearn, se propuso la generación del modelo, se suplió la necesidad de predecir las afectaciones en los POS, que en definitiva es lo que termina por deteriorar la usabilidad de la red.

Como resultado de este trabajo se alcanzó un 73% de Precisión además de tener un 32% de Sensibilidad, por otra parte, se identificaron anomalías que podían ser trabajadas de forma diferente por el área de operaciones de la compañía. Adicionalmente, se logró un cambio en la forma de abordar las incidencias predictivas en los datáfonos, lo que permitió una disminución en los costos asociados a las fallas y en consecuencia maximizó la rentabilidad de estos comercios.

Índice de Términos - Algoritmo, Aprendizaje Automático, Aprendizaje Supervisado, Arquitectura flexible, AUC, Código único, Comercio, Datáfono, Datamart, Dataset, Dispositivo, Establecimiento, Machine learning, Medio de Acceso, Precisión, POS (Point Of Sale), Recall, ROC, Sensibilidad, Terminal, Transacción, Versionamiento.

Abstract: The project had the purpose to apply Machine Learning algorithms, to achieve the predictive identification of failures in the Credibanco network POS, which allowed the entity to maintain its active network (core of the business), positively guide its operating model and benefit from improved customer experience rates.

This research wants to achieve the development of a model that identified the presence of anomalous behavior patterns in the devices, analyzing transactional, software and hardware variables, to associate them in this way to the failures in the POS that were in production.

With the support of the python library, sklearn, the construction of the model was built, thus satisfying the need to predictive the effects on the POS, which ultimately is what ends up deteriorating the usability of the network.

The result of this study, 73% of Precision was achieved. In addition, results shows of 32% recall indicator, otherwise anomalies were identified that could be worked differently by the company's Operations area, such that a change in the way to address the predictive incidences in the Dataphone, which allowed a decrease in the costs associated with the failures and consequently maximized the profitability of these businesses.

Key Words: Algorithm, AUC, Supervised Learning, flexible architecture, Unique code, Company, Dataphone, Datamart, Dataset, Device, Commercial establishment, Machine learning, Means of Access, POS, Precision, Recall, Terminal, Transaction, Versions.

I. INTRODUCCIÓN

Según Colombia Fintech en su publicación titulada "Colombia: Así se beneficiaría el país al aumentar uso de pagos electrónicos" [1] afirma, “que el país incrementaría entre 1.8 y 2.9 puntos en el PIB si logra aumentar en un 25% los pagos por medios electrónicos, además de generar cerca de 91 mil empleos”. Una de las empresas que viene trabajando en este frente es Credibanco, una institución colombiana vigilada por la Superintendencia Financiera fundada hace 47 años y con una amplia experiencia en la administración y desarrollo de sistemas de pago. Esta empresa está encargada de promover los pagos electrónicos en el país, con el propósito de sustituir el uso del dinero en efectivo, fomentando la formalización e inclusión financiera de los comercios [2].

Teniendo en cuenta el informe de cierre de año de Credibanco [2], se puede evidenciar que para el 2018 el 13.6% de las compras se hicieron con pagos por medios electrónicos,

su red participó en el 53.64% de las operaciones a través de sus datafonos. Además, en este reporte concluye que los datáfonos son el segundo medio de pago preferido en Colombia para realizar operaciones compras, ya que los datáfonos permiten relacionar de forma electrónica a los compradores, los comercios y las entidades financieras para hacer intercambio de transacciones de forma presente. Asimismo, este dispositivo permite la interconectividad entre estos tres actores, con la finalidad de admitir que una transacción por medio de un elemento de pago electrónico sea aprobada por el banco emisor (el que emitió la tarjeta que está realizando el pago) y abonada a un banco adquirente (entidad financiera a la cual está suscrito el comercio para el abono de los pagos).

Credibanco es la red que permite a través de los POS (Point of sale) conectar el comprador con su banco y el comercio con el suyo para poder hacer efectiva la transacción, para el caso en desarrollo, que hace referencia a la venta presente [2].

Lo descrito anteriormente, pone en evidencia el objetivo principal de la empresa, que se resume en tener disponible el medio de acceso y facilitar la interconectividad, manteniendo siempre activa su red. En ese sentido, se tomaron como objeto de análisis y desarrollo los algoritmos de Aprendizaje Automático para la identificación predictiva de fallas en los datáfonos, que son la principal forma de generar transacciones en esta entidad y la segunda más utilizada en el país para realizar operaciones comerciales [2].

Hoy la entidad cuenta con cerca de 200 mil datáfonos operativos a nivel nacional, y presenta una tasa de fallos (cociente entre número de fallas del mes y número de datáfonos al cierre del mes) del 7%; se reciben un promedio de 37 millones de transacciones al mes por los datáfonos y los montos de las transacciones ascienden a billones de pesos en cada periodo [2].

La identificación actual de los fallos se hace esperando el reporte del cliente a través de alguno de los canales disponibles, posteriormente se realiza un primer diagnóstico de forma telefónica guiando al usuario y en caso de no lograr la solución por este medio, se envía un técnico de medios de acceso al lugar afectado y se hace la reparación o el cambio del dispositivo según sea el caso. Según Credibanco [2] este proceso toma un promedio de 48 horas (después del reporte, y se estima que un cliente se demora 32 horas en comunicar la anomalía), por supuesto, el datáfono, no está disponible para realizar el proceso transaccional en este tiempo, lo cual sugiere una pérdida de 384 mil transacciones promedio por mes.

Por lo anterior, el objetivo principal es poder predecir cuándo va a fallar el próximo datáfono en un comercio y de esta forma, desplegar un proceso operativo antes de que el POS deje de funcionar, evitando así la pérdida transaccional. Con esta identificación predictiva se prevé generar una reducción de los costos operativos, dado que se disminuye el volumen de llamadas y con esto todo el despliegue que se requiere para la atención telefónica, incrementando la satisfacción del cliente y en consecuencia una mayor fidelización con efecto directo en la disminución de la deserción de clientes (Churn rate). Además, maximiza los ingresos de Credibanco porque se mantiene la red activa evitando pérdida de transacciones, de otra parte, beneficia de forma colateral, los actores que esta empresa interconecta, pues los tarjetahabientes seguirán su consumo sin percances beneficiando al banco emisor, al comercio y al banco adquirente.

II. MARCO TEÓRICO

La importancia de las decisiones, organizacionalmente hablando, radican en la capacidad de modificar el rumbo que se le dé a la compañía, por la elección de una posible solución a una situación problemática. Las decisiones que se toman son susceptibles de ser mejoradas con la ayuda del análisis [3]. El proceso de análisis mencionado por Cabañete [3] afirma que podrá tener un grado de dificultad, el cual estará marcado por la existencia o no de aspectos mensurables directa o indirectamente relacionados con los atributos a analizar.

Adicionalmente, el mismo autor indica que el combustible que alimenta el motor decisor es la información y enumera 4 actividades básicas para poder involucrarlas en su proceso decisorio: seleccionar y recolectar datos, guardar y recuperar datos, procesar y razonar datos para convertirlos en información y por último identificar la importancia y significación de la información obtenida [3].

La información es uno de los activos potencialmente más valiosos de una empresa. El valor real de la información depende de cómo es gestionada, del tiempo que se emplea en procesarla y traducirla en lanzamiento de productos o servicios, y de en qué medida se utiliza eficiente y cualitativamente es mejor que la de las empresas competidoras [4].

Una de las limitantes en el proceso de toma de decisiones es la capacidad humana, si bien son los humanos los que identifican y deciden sobre un problema, su capacidad cada vez se ve más limitada debido a los volúmenes y variedad de datos con los que se enfrentan, lo que dificulta su proceso de razonamiento y análisis y en consecuencia la toma de decisiones se ve afectada ya sea por oportunidad como sugiere [4] o por idoneidad en la elección.

Por otro lado, Conesa [5] en su libro denominado “Introducción a la Inteligencia Empresarial” recalca de forma contundente la manera en que la computación personal cambió la manera de administrar los datos y lo beneficioso que fue analizarlos en conjunto para poder identificar patrones y comportamientos que ayudaran a las empresas a tomar sus decisiones.

No obstante, en un mundo cada vez más globalizado, todos tienen “igualdad de condiciones” para acceder a los recursos de capital, tecnológicos, informativos, sin embargo, el factor

clave que marca la diferencia entre unas y otras es [puede ser] la capacidad de tomar decisiones de calidad. Aprovechando cada uno de los recursos y avances que se han conseguido [4].

El uso de teorías estadísticas y sus bases matemáticas ha hecho que el uso de Aprendizaje Automático en la actualidad haya tenido una gran acogida por los científicos, académicos y empresarios [6]. Adicionalmente, el poder computacional y el incremento exponencial de la tecnología ha permitido desarrollar modelos (algoritmos) que manejan gran cantidad de datos en un mínimo tiempo, por lo cual se ha podido aplicar la teoría expuesta por estas dos ciencias con benéficos resultados.

Sin lugar a duda, una de las líneas que han generado gran furor está delimitada por el uso de estos algoritmos para la predicción de eventos [7]. Y es que uno de los problemas más antiguos de la ciencia ha sido el encontrar la explicación a los fenómenos y con esta identificación poder predecir comportamientos para finalmente tomar decisiones precisas y oportunas [10].

Diferentes definiciones se han dado, dentro de ellas encontramos que el Aprendizaje Automático es una rama en evolución de los algoritmos computacionales diseñados para emular la inteligencia humana al aprender del entorno [11]. Aprendizaje Automático se refiere a programar computadoras para optimizar un criterio de rendimiento utilizando datos de ejemplo o experiencias pasadas [6], lo cual se complementa con la necesidad de generar aprendizaje en algunos casos donde no es posible escribir un programa de computadora para resolver un problema dado [7].

Sin embargo, la combinación entre las ciencias estadísticas y matemáticas más la computación otorgan la capacidad que hoy tiene el Aprendizaje Automático, a tal punto que se definen dos roles específicos en el trabajo del doctor Alpaydin [6], el primero define la creación de algoritmos con soporte matemático y la capacidad de poder procesarlos y almacenarlos en grandes volúmenes de datos, y, la segunda, una vez entrenado el algoritmo corresponde a la computación poder procesar de forma adecuada y óptima en espacio y tiempo la gestión de los datos, y el rol de las ciencias estadísticas y matemáticas se hacen cargo de la precisión de la predicción, teniendo las dos el mismo nivel de importancia [6].

El Aprendizaje Automático es objeto de un constante e intensivo estudio y desarrollo. Actualmente se distinguen tres ramas, el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo [9].

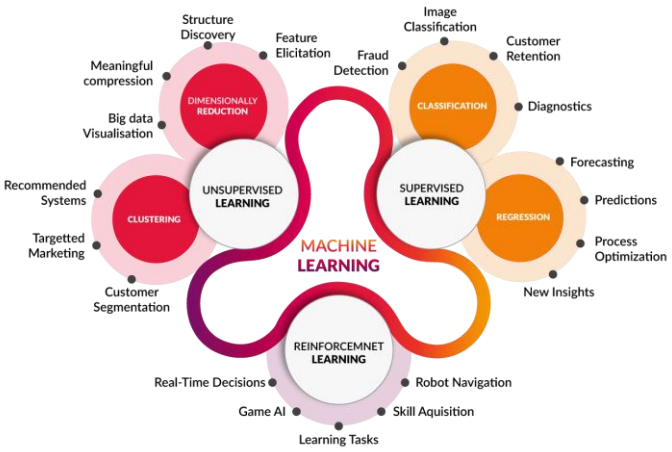

Fig. 1 Aprendizaje automático, ramas [9].

Tomando un aparte de la definición de Aprendizaje Automático, donde el aprendizaje está dado por la experiencia o por ejemplos de datos [6] surge el aprendizaje supervisado, el cual consiste en tomar una serie de entradas X y unas salidas Y para entrenar el modelo con los algoritmos estadísticos ofrecidos según la estructura de los datos, para que luego, cuando el sistema se encuentre con una nueva entrada X pueda decidir sobre su salida Y [8].

En síntesis, los algoritmos de este grupo trabajan con etiquetas, se intenta encontrar una función que, según las variables de entradas, sea capaz de identificar los patrones y asignar una etiqueta óptima para proporcionar adecuadamente una salida. Estos algoritmos “aprenden” de la información histórica etiquetada con la que se entrena y “predice” el valor de salida al encontrar el patrón de clasificación cuando ingresa un nuevo dato.

Bajo esta metodología se resuelven dos tipos de problemas; el primero se conoce como clasificación, que es el proceso de asignación de categoría a la muestra de datos de entrada. Ejemplos de uso: predicción de si una persona está enferma o no, detección de transacciones fraudulentas o como es el caso que se está abordando, que datáfonos se van a dañar en un determinado tiempo. El segundo se conoce como regresión, que es proceso de predicción de un valor numérico continuo para la muestra de datos de entrada. Ejemplos de uso: evaluar el precio de la TRM, pronosticar el precio de los alimentos en determinada época del año, pronosticar la temperatura [9].

Dentro del método de clasificación, existen diferentes algoritmos que ayudan a generar la mejor precisión en la predicción, dependiendo la estructura de los datos, iniciando por algoritmos que atienden modelos lineales hasta los modelos inspirados biológicamente como las redes neuronales, redes neuronales profundas y Deep Learning [10].

Los algoritmos de clasificación están divididos en dos grupos; los algoritmos de clasificación binaria (ver Fig. 2), en el cual las etiquetas corresponden a sólo dos valores (1 y 0, falla y no falla, a y b). Los algoritmos multiclase (ver Fig. 3) a los que se les puede asignar diferentes valores a las etiquetas y

el modelo de clasificación arrojará tantas clases como etiquetas se hayan ingresado [12].

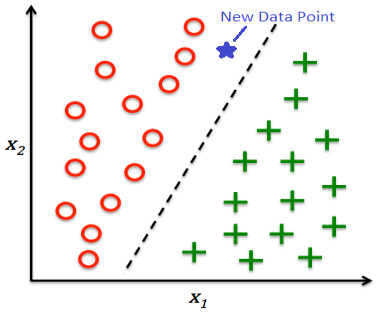

Fig. 2 Clasificación Binaria [12]

Dada la característica de los datos, no todos los modelos de clasificación serán útiles para generar una separación óptima que permita una adecuada precisión en la clasificación, por este motivo la selección del mejor algoritmo que pueda modelar con mejor precisión los datos será, tal vez, el primer gran problema al que se enfrente quien se haga cargo de diseñar modelos supervisados de clasificación. Actualmente existen modelos que miden el rendimiento de estos algoritmos, pero en la práctica se hace probando el rendimiento de cada uno de ellos sobre el mismo conjunto de datos [12].

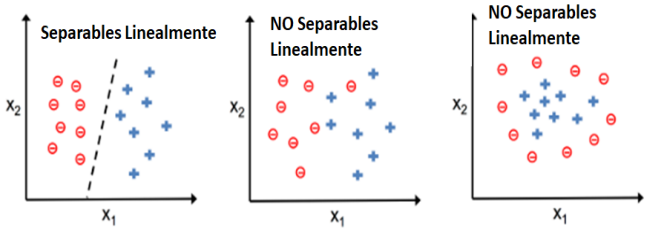

Fig. 3 Tipos de Conjuntos de datos que impactan en la selección del algoritmo [12].

En la Figura 3, se ejemplifican 3 tipos de conjuntos de datos, existen algoritmos diseñados para trabajar mejor en cada uno de estos casos, sin embargo, estará muy influido por los datos disponibles, número de características y ejemplos, las diferentes clases, y si son o no linealmente separables [12].

Muchos de los problemas de clasificación no son separables linealmente, esto hace que los algoritmos no converjan en la actualización de los pesos mientras están siendo entrenados [9]. Profundizaremos en los algoritmos que posteriormente serán utilizados para resolver el problema propuesto.

La definición de aplicación de estos modelos vienen atadas a las necesidades propias del estudio y la relación entre el entendimiento del modelo (por parte de los humanos) y la precisión del mismo es inversamente proporcional, por lo tanto, a mayor precisión menor entendimiento; por ejemplo en el manejo de las redes neuronales profundas, donde conocemos el resultado y su precisión, pero no, el sin fin de neuronas que se activan con cada entrada al sistema y su justificación matemática que da cuenta del ajuste del modelo. Mientras que un modelo lineal de mínimos cuadrados es fácilmente explicable, pero son muy pocos las actividades de la vida real en las que la precisión resultante es alta [13].

Algunos Algoritmos De Predicción

Regresión Logística

Este algoritmo también conocido como regresión logit, clasificación de máxima entropía (MaxEnt) o clasificador log-lineal [10] resulta ser de gran importancia dado que predice la probabilidad de que una muestra pertenezca a una clase determinada, permitiendo interpretar la relación que tienen las variables independientes sobre la dependiente [14]. Este algoritmo se basa en la proporción de las probabilidades, donde

$$\frac{P}{1 - P}$$

Siendo p la probabilidad positiva del evento
De aquí se puede deducir la función logit.

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

La función logit es simplemente el logaritmo de la función impar. Debido a que la motivación principal es predecir la probabilidad de que una muestra pertenezca a una clase determinada se halla la inversa de la función logit, y se denomina frecuentemente la función sigmoide.

$$y = \frac{1}{1 + e^{-f(X)}}$$

Donde f(X) es una función analítica en X (función que puede expresarse como una serie de potencias convergentes) como lo indica Greene [14]. Por lo anterior, el modelo de regresión logística es igual a la red de capa simple. Adicionalmente, la derivada de la función es continua y puede ser usada fácilmente en propagación hacia atrás.

Árboles de decisión:

Son un método de aprendizaje supervisado no paramétrico utilizado para la clasificación y la regresión. El objetivo es crear un modelo que prediga el valor de una variable objetivo mediante el aprendizaje de reglas de decisión simples inferidas de las características de los datos [10].

"Un árbol de decisión tiene unas entradas las cuales pueden ser un objeto o una situación descrita por medio de un conjunto de atributos y a partir de esto devuelve una respuesta la cual en últimas es una decisión que es tomada a partir de las entradas" [15]

Un árbol empieza a realizar una serie de validaciones a medida que se avanza por las hojas y según la decisión tomada

va entregando la respuesta. El árbol de decisión contiene nodos internos, nodos de probabilidad, nodos hojas y arcos.

"Un nodo interno realiza la validación algún valor de una de las propiedades. Un nodo de probabilidad indica que debe ocurrir un evento aleatorio de acuerdo con la naturaleza del problema. Un nodo hoja representa el valor que devolverá el árbol de decisión y finalmente las ramas brindan los posibles caminos que se tienen de acuerdo con la decisión tomada" [15]

Los árboles de decisión constan de: nodos de decisión representados por un cuadrado, nodos de probabilidad representados por un círculo y ramas o alternativas representadas por líneas o una línea cruzada por otras dos, para notar que es una decisión rechazada (ver fig. 4).

Símbolo	Nombre	Descripción
	Nodo de decisión	Indica que hay una decisión que se debe tomar
	Nodo de probabilidad	Muestra varios resultados inciertos
	Ramificación de alternativas	Cada línea de ramificación indica un posible resultado
	Alternativa rechazada	Muestra una alternativa o un resultado que no se debe tener en cuenta

Fig 4 Elementos de un árbol de decisión. Resumen propio basado en las definiciones de [15]

Al igual que todos los algoritmos de clasificación, se basa en las características de los datos de entrenamiento, el árbol de decisión "aprende" una serie de factores para inferir las etiquetas de clase de los ejemplos [15]. El nodo de comienzo es la raíz del árbol, y el algoritmo dividirá de forma iterativa el conjunto de datos en la característica que contenga la máxima ganancia de información, hasta que los nodos finales (hojas) sean puros.

De la adecuada selección de hiper parámetros dependerá el rendimiento del algoritmo convertido en modelo.

Máxima profundidad:
Es la mayor longitud desde la raíz a las hojas. Una gran profundidad puede causar sobreajuste, y pequeña profundidad puede causar subajuste. Para evitar sobreajuste, se "poda" el árbol de decisión estableciendo un hiper parámetro con la máxima longitud. Un pequeño número de muestras caerá en sobreajuste, mientras que un gran número de muestras caerá en subajuste.

Máximo número de muestras:
Cuando se corta un nodo, se puede tener el problema de conseguir 99 muestras en uno de los cortes y 1 muestra en el otro, lo que sería un mal uso de los recursos, para evitarlo, podemos establecer un máximo para el número de muestras que permitimos para cada hoja. Esto se puede especificar como un entero o como un número flotante.

Máximo número de características:
Muy frecuentemente se tienen muchas características (columnas) para construir un árbol. En cada corte, se tiene que hacer revisar todo el conjunto de datos en cada una de las características, lo que resulta ser muy costoso. Una posible solución a este problema es limitar el número de características que se buscan en cada corte. Si este número es suficientemente alto, es probable que encontramos una buena característica entre aquellas que buscamos (aunque pueda no ser la perfecta). Sin embargo, si no es tan alto como el número total de características, la velocidad de los cálculos se elevará de manera significativa.

En general, como lo evidencia Sklearn [10], el costo del tiempo de ejecución para construir un árbol binario balanceado es:

$$O(n_{samples} n_{features} \log(n_{samples}))$$

Y tiempo de consulta:

$$O(\log(n_{samples}))$$

Aunque el algoritmo de construcción de árboles intenta generar árboles equilibrados, no siempre estarán equilibrados. Suponiendo que los subárboles permanecen aproximadamente equilibrados, el costo en cada nodo consiste en buscar a través de

$$O(n_{features})$$

Para encontrar la característica que ofrece la mayor reducción en entropía. Esto tiene un costo en cada nodo de

$$O(n_{features} n_{samples} \log(n_{samples}))$$

Lo que lleva a un costo total sobre los árboles completos (sumando el costo en cada nodo) de

$$O(n_{features} n_{samples}^2 \log(n_{samples}))$$

La librería Sklearn [10] nos da a conocer su perspectiva sobre lo que consideran ventajas de los árboles de decisión donde sobresalen su facilidad de interpretación, dada su estructura analítica no se requiere de una preparación de datos avanzada, el costo de cómputo es logarítmico, puede analizar datos numéricos y categóricos, también es útil para análisis sobre procesos multiclases, es de fácil despliegue, incluso utilizando lenguajes de base de datos SQL.

Según Sklearn [10], dentro de las desventajas que pone en evidencia la librería están, la tendencia a generar sobre-ajuste (overfitting) y las técnicas de poda que usan los algoritmos de este tipo no están disponibles aún en la librería. Otro aspecto sobresaliente refiere que se sugiere tener equilibrado el conjunto de datos para que el rendimiento y variabilidad del modelo sean adecuadas.

Bosques Aleatorios (Random Forest):

Los árboles de decisión tienden a presentar problemas cuando el número de características (columnas) es grande, en la mayoría de los casos tiende a sobre ajustarse, por lo cual el

problema inicial a resolver eleva su nivel de complejidad. El problema se logra solventar seleccionando cada columna de forma aleatoria y realizando árboles de decisión para cada conjunto de columnas (Ver Fig. 5), de esa forma se desarrolla un algoritmo de agrupación de aprendizaje que combina una serie de modelos más débiles para crear otro más robusto [15].

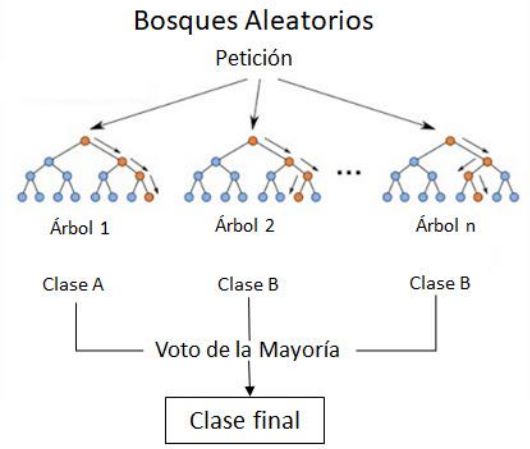

Fig 5 Explicación funcional de un bosque aleatorio [12]

Cada árbol del conjunto se construye a partir de una muestra extraída con reemplazo (es decir, una muestra de bootstrap) del conjunto de entrenamiento. Adicionalmente, se divide cada nodo durante la construcción de un árbol, la mejor división se encuentra en todas las características de entrada o en un subconjunto aleatorio del tamaño máximo de las características. El objetivo de tener dos de estas dos fuentes de aleatoriedad es disminuir la varianza del estimador del bosque. Los árboles de decisión individuales suelen presentar una gran variación y tienden a sobre ajustarse.

La aleatoriedad inyectada en los bosques produce árboles de decisión con errores de predicción algo desacoplados. Al tomar un promedio de esas predicciones, algunos errores pueden cancelarse. Los bosques aleatorios logran una variación reducida al combinar diversos árboles, a veces a costa de un ligero aumento en el sesgo. En la práctica, la reducción de la varianza a menudo es significativa, por lo que se obtiene un mejor modelo general [10].

El algoritmo realizará los siguientes pasos:

- Diseñar una muestra de arranque de tamaño n.
- Desarrollar un árbol de decisión desde la muestra de arranque. En cada nodo habrá características seleccionadas aleatoriamente sin reemplazamiento y el nodo se cortará maximizando la ganancia de información.
- El proceso previo se repetirá K veces.
- Agregar la predicción hecha para cada árbol, asignando la etiqueta de clase por votación mayoritaria.

La principal ventaja de este método es que normalmente no se necesita podar el bosque aleatorio, el modelo es suficientemente resistente al ruido. Sin embargo, es mucho menos interpretable que los árboles de decisión.

El único hiper parámetro que necesita ser ajustado es el número de árboles K. Normalmente, cuanto más grande es K, mejor es el rendimiento del modelo, pero en contraparte se incrementa drásticamente el esfuerzo de computación y, por tanto, el coste [12]. Además, se debe tener en cuenta que los resultados dejarán de mejorar significativamente más allá de un número crítico de árboles [10].

Rendimiento de los modelos:

Si bien el tiempo y el espacio en términos computacionales son importantes, ya que hacen referencia a la velocidad de gestión e identificación y a la capacidad de procesamiento y almacenamiento, la medición del rendimiento de la predicción finalmente constituye el factor decisivo el despliegue o no de un modelo. Para lo anterior existen artefactos que dan cuenta de la precisión de acierto, entre ellos se encuentran la matriz de confusión, que indica de forma matricial las clasificaciones correctas e incorrectas que tuvo un modelo en su ejecución.

	Predicted	
	Positive	Negative
Actual True	TP	FN
Actual False	FP	TN

Fig. 6 Matriz de confusión de la clasificación binaria [19]

La matriz de confusión o matriz de error expuesta en la Figura 6, es una tabla que permite la identificar el rendimiento de un algoritmo de aprendizaje supervisado (en el aprendizaje no supervisado generalmente se denomina matriz coincidente). Cada fila de la matriz representa la clase predicha, mientras que cada columna representa la clase real o viceversa [17]. El nombre se deriva del hecho de que hace que sea fácil ver si el sistema está confundiendo dos clases etiquetando incorrectamente una como otra.

La matriz es un tipo especial de tabla de contingencia, con dos dimensiones ("real" y "prevista"), y conjuntos idénticos de "clases" en ambas dimensiones (cada combinación de dimensión y clase es una variable en la tabla de contingencia) [17]. Adicionalmente, se pueden entrever diversas relaciones que generan métricas y estas a su vez serán útiles para medir el rendimiento del modelo en términos de lo que se esté buscando para resolver la situación problemática (Ver Fig. 7).

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	True positive	False positive, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	False negative, Type II error	True negative	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Fig. 7 Matriz de confusión con métricas [18].

El Accuracy es la medida por excelencia de los modelos, así lo establece Microsoft [19] debido a su fácil interpretación y adaptabilidad a problemas de clasificación tanto binarios como multiclase, esta medida es muy útil siempre y cuando los modelos estén balanceados [20]. Es decir, las variables de predicción sean proporcionalmente similares.

Otra medida es la Precisión, que responde a la pregunta ¿Cuántos positivos predichos son realmente positivos?,

Precisión = True Positive / (True Positive + True Negative)

En el ejemplo de las transacciones no se predijo ningún positivo, por lo tanto, la precisión sería igual a 0. Esta métrica es valiosa cuando se trata de tener mayor seguridad sobre la predicción. Para el caso en desarrollo será fundamental, dado que del total de datafonos los que presentan fallas equivalen a cerca del 4%, este desequilibrio sumado a el costo operativo que representa la presencia de falsos positivos hace de la precisión un indicador clave para los resultados del modelo.

La sensibilidad (Recall), es una medida que permite responde a la pregunta ¿de todos los positivos reales cuántos encontró en el modelo? De acuerdo con esto, la sensibilidad se vuelve importante cuando el objetivo del modelo es encontrar la mayor cantidad de positivos [17].

Recall = True Positive / (True Positive + False Negative)

Uno de los principales intereses de los modelos está basado en mantener la mejor Sensibilidad y precisión posibles (esto es, tener la mayor cantidad de positivos reales entregando la menor cantidad de falsos positivos). La medida que ayuda, sin lugar a duda, a identificar esta relación es el F1Score, que corresponde a la media armónica entre la precisión y la Sensibilidad, donde su mayor y mejor valor es 1 y su menor y peor valor es 0 [17]. Preexiste un problema con la medida F1score, y corresponde a que da igual peso a la precisión y a la Sensibilidad, y en algunos problemas es necesario que predomine uno de los dos [20].

Existe una medida adicional que permite inspeccionar la tasa de verdaderos positivos contra la de falsos positivos es la curva de Características Operativas del Receptor (ROC) y el valor del Área Bajo la Curva (AUC) correspondiente [16]. Mientras el área bajo la curva se acerque a 1 mayor será la precisión y el rendimiento del clasificador. Por el contrario, mientras más se acerque a 0,5 el valor del área menor será la precisión y el rendimiento del clasificador (Ver Fig.8)

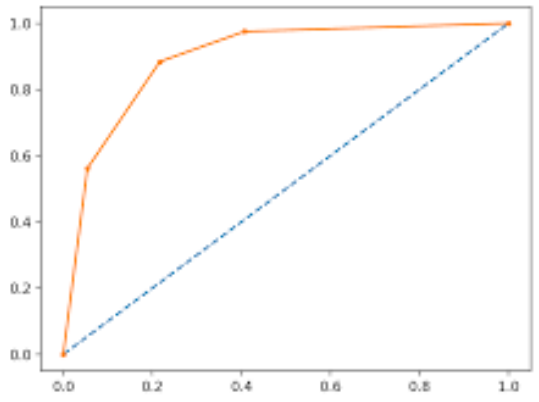

Fig 8 Resultados de la evaluación de clasificación binaria con ROC y AUC [20].

Seleccionar un modelo de aprendizaje automático perfecto, es en parte ciencia y en parte arte, en las empresas, los KPI de rendimiento del modelo juegan un papel importante y decisivo, sin embargo aparecen actores que afectan la elección del modelo, el costo económico en recursos y el costo en tiempo de desarrollo y puesta en producción resultan ser los más importantes al momento de seleccionarlos, sin embargo, en esta nueva ola, la interpretabilidad de los modelos resulta ser crucial en el mundo real; la transición entre lo descriptivo y lo predictivo en las áreas administrativas y operativas aún no está resuelto del todo, y los impactos de modelos que afecten áreas de este tipo requieren, en general, una explicación a nivel gerencial que definirá la inversión o no en el proyecto [19].

Diferentes clases de modelos son buenos para modelar diferentes tipos de patrones subyacentes en los datos. Por lo anterior, diferentes autores coinciden en la necesidad de probar diferentes modelos, configurar sus hiper parámetros, tal que se pueda ver el resultado de estos, bajo las métricas de interés para resolver el problema y de allí se seleccionen el o los de mejor rendimiento y profundizar más en ellos [20].

La evaluación de tiempos entre el entrenamiento y la predicción, atados a los resultados del KPI de elección es un factor más para tener en cuenta al momento de realizar la selección adecuada del modelo a desarrollar. También lo es la relación entre la necesidad de poder explicar el modelo o por el contrario es aceptada una mayor complejidad en el mismo. Su escalabilidad, puesta en producción, limitación de recursos son aspectos que definitivamente impactan la decisión y no deben ser obviados al momento de la selección.

III. ESTADO DEL ARTE

En la actualidad son muchas las industrias que se han volcado al uso de Machine Learning para mejorar sus procesos y toma de decisiones. Dentro de ellas según las referencias de SAS [21] están:

Servicios Financieros:

Las entidades financieras a nivel mundial utilizan la tecnología Machine Learning para dos fines principales: predicción y prevención del fraude. También para identificar oportunidades de inversión. Actualmente realizan estos algoritmos para identificar clientes con perfiles de alto riesgo o bien utilizar el cíber vigilancia para detectar signos de advertencia de fraude [21].

Gobiernos:

Áreas gubernamentales como seguridad y los servicios públicos a nivel mundial han despertado su interés particular en machine learning porque tienen múltiples fuentes de datos de las que se pueden extraer nuevo conocimiento. Por ejemplo, el análisis de datos de sensores identifica formas de incrementar la eficiencia y ahorrar dinero. Asimismo, el aprendizaje basado en máquina puede ayudar a detectar fraude y minimizar el robo de identidad [21]. En Colombia la Fiscalía General de la Nación presentó al Fiscal Watson, esta entidad en conjunto con IBM y su herramienta de AI utilizan algoritmos para la predicción de delitos y la identificación de procesos que estaban por fuera del su control [22].

Salud:

El Machine Learning es una tendencia en rápido crecimiento en la industria de atención a la salud, gracias a la aparición de dispositivos y sensores de vestir que pueden usar datos para evaluar la salud de un paciente en tiempo real. Así mismo, la tecnología puede ayudar a expertos médicos a analizar datos para identificar tendencias o banderas rojas que puedan llevar a diagnósticos y tratamientos mejorado [21].

Marketing:

Las páginas Web que generan recomendaciones para la compra de artículos según los gustos de los clientes basados en su historial de compras, hacen uso del Machine Learning para analizar su historial de compras y así hacer promoción de otros artículos que son afines a los intereses de sus compradores [21].

Petróleo y gas:

Atender a las preguntas de cómo encontrar nuevas fuentes de energía. Generar estrategias para el análisis de minerales del suelo. Poder predecir fallos de sensores de refinerías. El número de casos de uso del Machine Learning en esta industria es vasto y continúa creciendo [21].

Transporte:

Analizar datos para identificar patrones y tendencias es clave para la industria del transporte, que se sustenta en hacer las rutas más eficientes y anticipar problemas potenciales para incrementar la rentabilidad. Los aspectos de análisis y modelado de datos del machine learning son herramientas importantes para las compañías de mensajería, transporte público y otras organizaciones de transporte [21].

De otra parte, **Deep Learning** [23] se perfila como el método predilecto para la generación de predicciones, sobre todo en demanda, fraude y fallos, por su mayor capacidad en la fusión de los datos, a tal punto que Gartner predice que para 2020 será un factor determinante en este aspecto.

IV. PREGUNTA DE INVESTIGACIÓN

Para Credibanco, empresa vinculada al sector financiero de bajo valor, se desarrolló un modelo de aprendizaje supervisado, que diera respuesta a la pregunta ¿cómo utilizar algoritmos de machine learning para predecir los fallos de datáfonos de su red para evitar la pérdida transaccional, y, maximizar de esta manera los ingresos de la compañía?

Con el fin de tal cometido, se utilizaron variables que daban cuenta de las características lógicas y físicas de la terminal, el comportamiento transaccional diario realizado por ella y la frecuencia de reporte de fallas realizada por los clientes de la entidad. Adicionalmente, y atendiendo lo identificado en el marco teórico, se corrieron diferentes modelos con el objetivo de poder analizar el rendimiento de estos, y así determinar el más adecuado para profundizar en la configuración de hiper parámetros y ajuste de KPI's. En este sentido la data tuvo 3 momentos, la base 1 para entrenamiento, la base dos para validación y la base tres para testing.

Para efectos de cumplir los requerimientos de la empresa, la precisión (Verdaderos positivos / verdadero positivo + falso positivo) fue primordial, dado que el despliegue operativo podría ser más costoso que esperar a que el cliente mismo reportara la falla. Así pues, el modelo financiero arrojó un punto de equilibrio, que llamamos esfuerzo, en 3:1, esto es de cada 4 predicciones tan solo 1 podría ser falso positivo, o en otras palabras el modelo cumpliría los requisitos operativos y financieros superado el 75% de precisión. Adicionalmente, la operación requería identificar la Sensibilidad entregada por el modelo, para así calcular los impactos colaterales de la gestión, tales como satisfacción de clientes, disminución de llamadas, incremento de recomendación, productividad de técnicos en calles entre los más importantes.

Los modelos probados en todos los casos fueron entrenados con el 70% de la información, el 30% restante fue para validación y una base con 633.811 registros fueron procesados para el testing.

Los resultados que se presentan a continuación:

Regresión Logística:

En principio el modelo se utilizó por su fácil interpretación, además de poder identificar la importancia de las variables que se había introducido en la definición de la base para ejecutar el modelo, se podía ver también la linealidad o no de los datos.

En Validación, se alcanzó una Sensibilidad del 4.9% de los datos con una precisión del 74.3%, En Testing sus resultados fueron 4.87% de Sensibilidad y 69.18% de precisión.

Árbol de decisión:

Este modelo fue seleccionado por la facilidad que ofrece en interpretación y en puesta en producción, sin embargo, es muy sensible a la configuración de hiper parámetros, siendo la selección de la profundidad del árbol, el número de hojas permitido y la tasa de aprendizaje los de mayor acierto al momento de entregar la precisión [15]. En primera instancia, con una profundidad de 7 niveles, 100 hojas y una tasa de aprendizaje de 0.1 (el estándar del algoritmo) el modelo entregó para la base de validación una Sensibilidad de 10.1% y una precisión del 76.7%. Por otra parte, en la base de testing los resultados de Sensibilidad ascendieron a 12% y de precisión al 70.15%.

Bayesiano:

Dada la dependencia del resultado de falla con las características físicas o lógicas de la terminal afectada, se puso a prueba el modelo Bayesiano ofrecido por la biblioteca Sklearn de Python. Se alcanzó una Sensibilidad del 18.3% en validación y de 18.24% en test, con una precisión del 81.5% y 70.38% respectivamente. En principio lograba superar el umbral definido de precisión, además maximizó la Sensibilidad 3 veces por encima de lo arrojado en el modelo lineal. Fue candidato a potencializar bajo la configuración de hiper parámetros y afinación de variables.

Random Forest:

El modelo fue seleccionado con el objetivo de maximizar el resultado del árbol de decisión, que no fueron del todo erráticos. En validación su resultado en Sensibilidad ascendió a 28% y en test registró 32%, por su parte, la precisión registró 88.9% en validación y 70.4% en testing. Con estos resultados el modelo fue escogido para el desarrollo práctico de identificación de fallas en la compañía.

	RESULTADOS VALIDACIÓN 1.525.621 Registros		RESULTADOS TESTING 633.811 Registros	
	Precisión	Sensibilidad	Precisión	Sensibilidad
Reg. Logística	74,30%	4,90%	69,18%	4,87%
Árbol de Decisión	76,70%	10,10%	70,15%	12,02%
Bayesiano	81,50%	18,30%	70,38%	18,24%
Random Forest	88,90%	28,03%	70,49%	32,01%

Tabla 1 Resultados modelos probados en validación y testing. Selección modelo Random Forest. Fuente: Elaboración propia

V. OBJETIVOS

A. OBJETIVO GENERAL

Generar un modelo utilizando algoritmos de Aprendizaje Automático que permita identificar con alta precisión los fallos de datafonos de Credibanco de forma predictiva en un umbral de tiempo adecuado para la óptima gestión operativa.

B. OBJETIVOS ESPECÍFICOS

- Estudiar los modelos de predicción de Aprendizaje Automático que permitan identificar los fallos en los datáfonos.
- Construir un modelo de Aprendizaje Automático que permita identificar con alta precisión los fallos en los datafonos de Credibanco.
- Proponer un esquema de acción para los fallos de los datáfonos según los resultados obtenidos por el modelo.

VI. MODELO

La metodología de aplicación del modelo se basó en la ejecución de 7 pasos de los 8 propuestos por SUNQ [24]:

- Construcción de la base y definición de variables.
- Realizar y presentar análisis descriptivo de los datos
- Identificar correlaciones entre las variables independientes.
- Aplicar el modelo de Machine Learning seleccionado
- Evaluar los rendimientos del modelo.
- Comunicar los resultados del modelo seleccionado.
- Ejecutar una prueba en conjunto con área de operaciones.

Construcción De La Base y Definición De Variables

Para la construcción de la base de datos a modelar fue necesario obtener conexión a siete (7) sistemas propios de la empresa, consultar y almacenar información de 41 tablas con un total de 598 millones de registros, esto obligó a realizar la construcción de 23 trabajos (ETL's) en la herramienta de integración de datos (también conocidas como herramientas ETL) de IBM llamada DataStage, suministrada por la entidad para los procesos de extracción (ver Figura 9).

Para el almacenamiento fue designado un espacio en un motor de base de datos analítico, también licenciado por IBM, conocido comúnmente como Netezza o, para IBM, Pure Data for Analytics (PDA). Como resultado de este proceso de integración, se construyó una base con 193 variables, 1 etiqueta y 9 campos destinados a la identificación de los registros, para un total de 203 atributos y 5.085.407 registros.

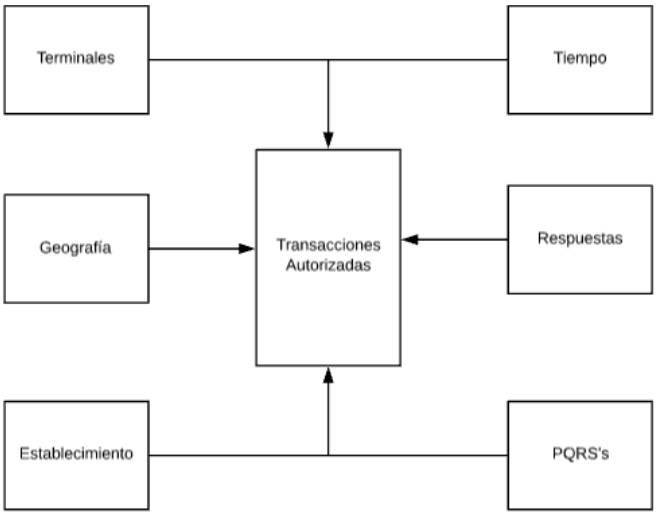

Fig 9 Modelo Dimensional Alto Nivel, construcción base para modelo de identificación de fallas Fuente: Elaboración propia

Las variables independientes se distribuyeron en 7 grupos de esta manera:

- Variable Geográfica (1 variable incluida al modelo): corresponde a la oficina de Credibanco más cercana al municipio donde se encuentra ubicada la terminal físicamente.
- Variable Caracterización (1 variable incluida al modelo): corresponde al grupo económico asignado por Credibanco, al que pertenece el comercio que está haciendo uso de la terminal física.
- Variable Software (2 variables incluidas al modelo): Identifica la versión de software descargada en el datáfono en el momento de la evaluación de la terminal y el browser sobre el cual corre la aplicación de control remoto de terminales que tiene la entidad.
- Variable Hardware (3 variables incluidas al modelo): En la parte física se evalúa la marca, el modelo y la tecnología que soporta el datáfono que se está evaluando.
- Variable Tiempo (5 variables incluidas al modelo): En esta variable se incluyó el número de días transcurridos desde el evento hasta el día de evaluación. Los eventos asociados a la base fueron, antigüedad del comercio con Credibanco, antigüedad de la terminal desde su creación lógica, antigüedad del datáfono físico desde su compra, antigüedad del datáfono en el comercio que lo tiene en uso actualmente, antigüedad de la versión de software que tiene descargado el dispositivo.
- Variable de reporte de fallas (6 variables incluidas al modelo): Incluye el número de soportes solicitados por la terminal en los últimos 3 meses, y el número de soportes solicitados a nivel de comercio en los últimos 3 meses.
- Variable transaccional (175 variables incluidas al modelo): a cada terminal se le vincularon las transacciones totales, transacciones aprobadas, transacciones negadas y los tiempos de respuesta asociados a cada una de ellas, de los últimos 14 meses. Lo propio se hizo para las transacciones de los últimos 14 días.

Variable	Categórica	Numérica	Total
Grupo Económico	1		1
Geografía	1		1
Software	2		2
Hardware	3		3
Tiempo		5	5
Reporte de falla		6	6
Transaccional		175	175
Total	7	186	193

Tabla 2 Clasificación de las variables construidas para el modelo de predicción de fallas. Fuente: Elaboración propia

Una vez identificadas las variables que se iban a incluir en el modelo, se procedió a construir el conjunto de datos. Para cumplir este objetivo, la idea fue, seleccionar días de diferentes meses de forma aleatoria (puntos de observación) y evaluar las terminales que estaban presentes ese día en la empresa y obtener las 193 variables asociadas a dicho punto de observación. Este proceso se hizo 30 veces, con días al azar entre octubre de 2018 y junio de 2019, en promedio cada día fueron evaluadas 169.500 terminales.

La etiqueta o variable dependiente construida, identificaba los datáfonos que fallaron entre los días 7 y 37 después del día evaluado. Esta etiqueta fue construida de tal forma que solamente informara si la terminal había presentado o no reporte de fallo durante ese lapso. Por tanto, la solución al problema sugería la construcción de un modelo de clasificación binario.

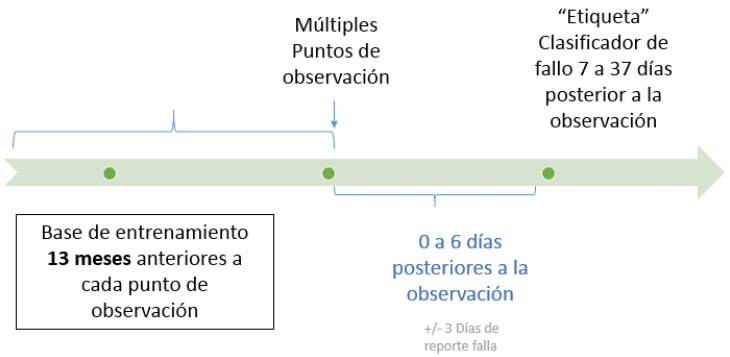

Fig 10 Diseño del conjunto de datos para construcción del modelo de fallos POS para Credibanco Fuente: Elaboración propia

Realizar Y Presentar Análisis Descriptivo de los Datos e Identificar Correlaciones Entre Las Variables Independientes

Con la base objeto del modelo construido, se dio inicio al análisis univariado, con el objetivo de evaluar la calidad del conjunto de datos, además de identificar la composición de las variables seleccionadas, su distribución (y frecuencia en caso de las variables categóricas) y sus medidas de tendencia central. Debido a que el trabajo de construcción de la base fue definido en el modelo dimensional previamente elaborado, los errores de calidad en los datos fueron mitigados en este proceso, evitando pasar valores nulos o que vulneraran las estructuras requeridas en cada campo, por lo tanto, para este paso no fue necesario eliminar registros asociados a fallas de calidad en el conjunto de datos evaluado, pero si fue trascendental para poner a prueba la efectividad del DataMart construido para la extracción de la base.

Los resultados de análisis univariado permitieron corroborar y precisar el desbalanceo con el que se contaba (ver Tabla2), pues se identificó que tan solo el 6.9% de la base tenía etiqueta de falla, por lo cual era importante evaluar la estrategia para minimizar el impacto en la elaboración del algoritmo, acoger el indicador de rendimiento del modelo más adecuado y este hallazgo también fue decisivo para ratificar la selección del modelo que se iba a trabajar, por que ofrecía robustez al momento de trabajar con datos desequilibrados.

TITLE	FL_FALLA
count	5.085.407
mean	0.069

Tabla 3 Aparte del análisis univariado, etiqueta de falla, evidencia del desequilibrio de la base Fuente: Elaboración propia.

Dentro del proceso analítico, y con la idea de identificar la relación entre los fallos y las variables seleccionadas se procedió a efectuar un análisis bivariado, para este proceso se ejecutaron dos procedimientos. El primero buscaba hallar la participación entre la etiqueta y las variables independientes, y fue decisivo para identificar variables dependientes fuertemente asociadas a la etiqueta, candidatas a demás a ser las variables fuertes del modelo (ver tabla 3).

TECNOLOGIA_TERMINAL

FL_FALLA	DIAL	DIAL-LAN	EFT-PIN PAD	GPRS	MPOS	PROTOTIPO	RPF	WIFI	ALL
NO	227	2,725,152	38,041	1,546,234	373,613	135	26,438	24,673	4,734,513
SI	18	169,142	3,022	174,884	2,387	-	139	1,302	350,894
All	245	2,894,294	41,063	1,721,118	376,000	135	26,577	25,975	5,085,407
All PORC	7.35%	5.84%	7.36%	10.16%	0.63%	0.00%	0.52%	5.01%	6.90%

Tabla 4 Ejemplo análisis bivariado, etiqueta de falla vs Tecnología de terminal, se identifica que la tecnología con más fallas es GPRS, seguida de EFT-PINPAD. Fuente: Elaboración propia – Datos análisis bivariado

El segundo análisis, en este mismo aspecto, fue de correlación de variables dependientes, con el objetivo de identificar colinealidad entre ellas y determinar si podría reducirse la base de datos en dimensiones en caso de tener que recurrir a un modelo sensible a la dependencia entre variables. Resultado de este proceso se identificó la fuerte correlación entre las variables transaccionales a nivel diario, y la baja correlación de las fallas reportadas un mes atrás con las demás variables.

A continuación, se ve un mapa de calor con el índice de correlación de variables, donde el verde oscuro indica la más alta correlación y el rojo la ausencia de correlación (ver Figura 11).

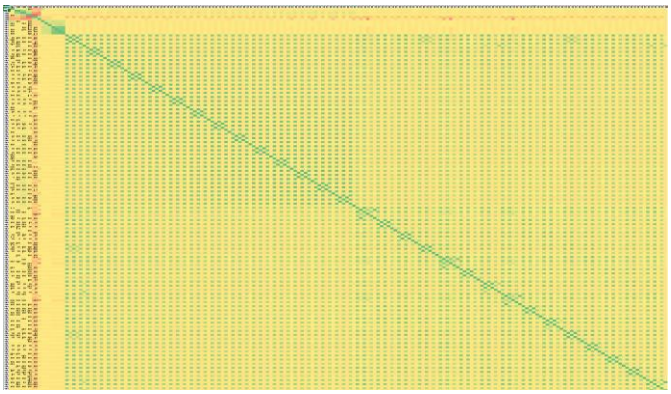

Fig 11 Análisis de correlación. Verde altamente correlacionados, rojos No correlacionados. Fuente: Resultados obtenidos a través de Python

Aplicar El Modelo de Aprendizaje Automático Seleccionado y Evaluar Rendimiento.

Una vez analizados los datos, identificadas las condiciones que ofrecía la base y ratificado el modelo que se iba a trabajar, se dio paso a la construcción de este.

Con el apoyo de la librería de scikit-learn se procedió a realizar la construcción del algoritmo, teniendo en cuenta que el éxito de lo que restaba de trabajo estaba basado en la adecuada selección de parámetros, bajo el entendido de que la base ya se encontraba disponible y el análisis de esta arrojaba una buena calidad en los datos.

Las librerías trabajadas fueron seleccionadas para la manipulación de la data tales como NumPy y Pandas, para el manejo de gráficos sobre los resultados que se iban dando como Matplotlib, la presentación de métricas como Confusion matrix, roc_curve, auc, recall score, precision: score, f1_scores, make classification y la del algoritmo seleccionado para la construcción del modelo RandomForestClassifier:

```
In [1]: import matplotlib.pyplot as plt
        import numpy as np
        import pandas as pd
        from sklearn.metrics import confusion_matrix
        from sklearn.model_selection import train_test_split
        from sklearn.metrics import roc_curve, auc
        from sklearn.metrics import recall_score, precision_score, f1_score
        from sklearn.ensemble import RandomForestClassifier
        from sklearn.datasets import make_classification
```

Fig 12 Librerías utilizadas para la construcción del modelo
Fuente: Resultados obtenidos a través de Python

De otra parte, las variables categóricas fueron transpuestas para manejarlas como variables Dummy. Variables seleccionadas para transponer:

'DESC_MCC','NM_BROWSER_AF','NM_VERSION_ACTUAL','NM_SECCIONAL_COMERCIAL','TECNOLOGIA_TERMINAL','CD_MODELO_AGRUPADO','NM_MARCA_TERMINAL'.

Paso seguido, se inició con la configuración de los parámetros del clasificador. Si bien se muestra que alcanzó los mejores resultados, fue necesario iterar cerca de 9 con estos parámetros hasta encontrar el de mayor efectividad.

- N_estimators: Este parámetro hace referencia al número de árboles que se deben evaluar en el bosque aleatorio, para el caso se inició con 80 árboles y fue incrementándose en cada iteración hasta finalizar con 250.
- Max_depth: Aquí se configura la profundidad máxima de cada árbol, dada la teoría de los árboles, su profundidad no debe ser tan elevada dado que los modelos suelen converger rápidamente y sería un desperdicio de recursos hacerlo con profundidades muy amplias. Para el caso en mención se inició con una profundidad de 12 hojas y se finalizó con 25.
- Min_samples_leaft: se configura el número mínimo de muestras que se requieren para poder decidir sobre la hoja en la que se está. Por defecto llega 1, este parámetro solo se configuró 4 veces, en las 9 iteraciones, es un parámetro opcional pero sí que ayuda a suavizar el modelo. Se inició con el valor por defecto que es 1 y se finalizó con 10
- Class_weight: quizás fue el más importante y sensible parámetro a configurar, pues es el que permitía identificar el balanceo de la etiqueta de falla. El objetivo era escoger para la etiqueta de falla el peso ideal para balancear las muestras. Se inició con 14.28 que correspondía al número de datáfonos presentes para que se reportara una falla según el análisis de balanceo inicial (si el 6.9% de los datafonos presentaban falla, por cada 14.28 datafonos 1 fallaría) y se finalizó con 20.5, esto debido a que fue necesario devolverse a la base original e identificar los datáfonos que reportaban varias fallas, eliminando esta duplicidad se determinó que la tasa de fallos sin duplicidad de terminales reportadas era de 4.8% por tanto la participación ideal del balanceo debía rondar en 20.5.
- Criterion: se seleccionó Entropía por encima de Gini, aunque fueron probados los dos, sin embargo, la ganancia de información ofrecida por la Entropía generó un incremento de 4 puntos porcentuales en Sensibilidad, tal incremento fue decisivo para la selección del parámetro.

```
clf = RandomForestClassifier(n_estimators=250, max_depth=25, min_samples_split=10,
                             random_state=0, class_weight = {1:20.5},criterion='entropy' )
```

Fig 13 Configuración final de parámetros Random Forest.
Fuente: Resultados obtenidos a través de Python

Comunicar Los Resultados Del Modelo Seleccionado.

Tal como se ha mencionado durante la construcción del modelo, dos indicadores que califican su rendimiento son de suma importancia Sensibilidad y Precisión. Por tanto, sin descuidar los otros indicadores, estos dos fueron decisivos para tomar como válido el modelo y llevarlo a la etapa de presentación.

La Base de entrenamiento correspondía al 70% de la base total. Para efectos de Prueba se tomó el 30% restante. Los resultados que se presentarán a continuación corresponden a la base de Prueba.

		Real	
		No falla	Si falla
Predicción	No falla	1453426	40147
	Si falla	5863	26185

```
accuracy test 0.92
f1_score 0.5344
Recall V2 0.3947
Precision V2 0.81
AUC 0.9797
```

Fig 14 Matriz de confusión e indicadores, resultados Test con 30% base de entrenamiento. Fuente: Resultados obtenidos a través de Python

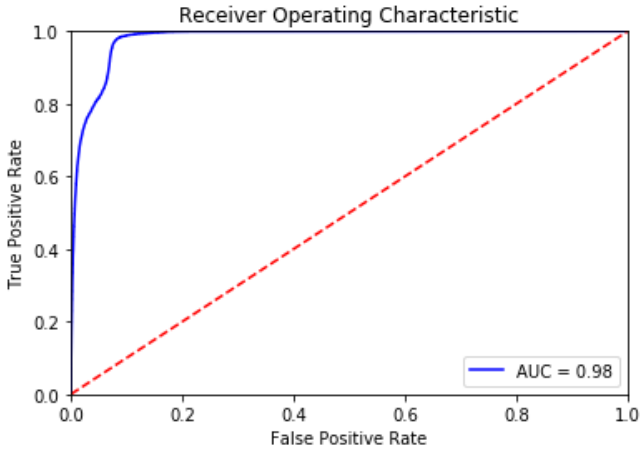

Fig. 15 Curva ROC y AUC, resultados Test con 30% base de entrenamiento. Fuente: Resultados obtenidos a través de Python

Con estos resultados de prueba con el 30% restante de la base con la cual fue entrenado el modelo y los parámetros configurados como se informó en el aparte anterior se lograron resultados viables para las pretensiones de la compañía pues se identificó el 39% del total de fallos de la compañía y la relación de falsos positivos fue 4.4:1, sobre el 3:1 mínimo requerido. Paso seguido, se procedió a generar una nueva base con 603.811 registros, con periodos diferentes a los incluidos en la base inicial para hacer una segunda prueba con data nueva, este proceso tenía dos objetivos, el primero y principal medir la estabilidad del rendimiento del modelo y el segundo poner a prueba la generación de los datos del conjunto de datos construido para las puestas en producción.

Los resultados de la nueva muestra, totalmente independiente fueron:

		Real	
		No falla	Si falla
Predicción	No falla	608455	15101
	Si falla	3032	7223

```
accuracy test 0.95
f1_score 0.611
Recall V2 0.32
Precision V2 0.704
AUC 0.9723
```

Fig 16 Matriz de confusión e indicadores, resultados Test con nueva base. Fuente: Resultados obtenidos a través de Python

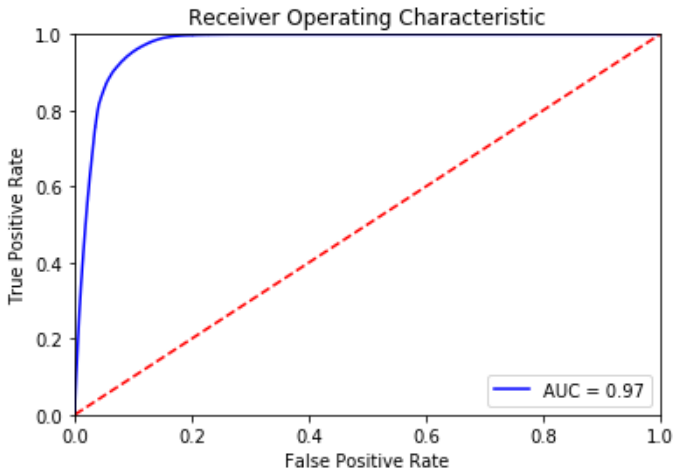

Fig 17 Curva ROC y AUC, resultados Test con nueva base Fuente: Resultados obtenidos a través de Python

Los resultados de la nueva base pusieron en evidencia una pérdida porcentual de 9 puntos en la precisión y de 7 en la Sensibilidad, lo que nos dejó en 3,38:1 en falsos positivos, cifra que fue aceptada por la gerencia de operaciones para enviar el primer piloto a terreno, con el objetivo de identificar cómo trabajar operativamente con los clientes y medir su satisfacción frente a este proceso.

Ejecutar Una Prueba En Conjunto Con Área De Operaciones.

Para la ejecución de la prueba piloto con el área de operaciones se generó una nueva base el día 29 de octubre de 2019, para esta prueba se seleccionaron de forma aleatoria 2.803 registros que habían sido etiquetados por el modelo como datáfonos con probabilidad de fallo mayor al 50%. Esta base se dividió en dos partes. 1.000 registros fueron enviados al call center para agendar una visita de mantenimiento preventivo y 1.803 fueron dejados en evaluación para identificar si reportaban falla en los siguientes 30 días según la definición del modelo.

Los resultados de los 1.803 registros evidenciaron consistencia en la prueba de la segunda base hecha en el proceso de construcción y evaluación, arrojando una precisión del 73% sobre el 70.4% identificado (ver Figura 16) demostrando estabilidad en los resultados del último testing realizado. Adicionalmente se pudo identificar la consistencia en la probabilidad arrojada por el modelo, pues entre más alta era más fallos reales fueron presentados.

Rangos	Casos enviados	Reporte falla	% precisión
50% - 68%	158	111	70%
66% - 68%	67	41	61%
68% - 70%	96	63	66%
70% - 72%	136	101	74%
72% - 74%	175	122	70%
74% - 76%	194	141	73%
76% - 78%	251	168	67%
78% - 80%	254	192	76%
80% - 82%	236	178	75%
82% - 84%	144	116	81%
84% - 86%	67	56	84%
86% - 88%	21	17	81%
88% - 90%	4	3	75%
Total	**1803**	**1309**	73%

Tabla 5 Base piloto fallos POS – corte 29-10-2019 finalizado el 27-11-2019. Fuente: Resultados producción del modelo sobre el conjunto de observación.

Con respecto a los 1.000 registros enviados a operaciones para su gestión en terreno, se logró hacer la visita técnica al 87% de los informados, de ellos el 12% reincidió reportando falla durante los 30 días siguientes, el aprendizaje de este proceso fue tal que llevó a realizar cambios en los procesos operativos para la gestión de alertas predictivas de fallas POS.

VII. CONCLUSIONES

El proceso de construcción de este modelo incentivó varias partes de la empresa que facilitaron el tiempo, los recursos y la información para poder llevarlo a cabo.

La principal ganancia fue el cambio del modelo operativo para la gestión predictiva de alertas, pues en este proceso se puso en evidencia que no todos los fallos debían ser llevados a la parte técnica en terreno, sino que una gestión remota de actualización de parámetros o de software era suficiente para que los comercios no presentaran un daño.

Este cambio representó la disminución del 13% de las visitas (897 visitas al mes) teniendo en cuenta que el costo promedio por visita oscila en $70.000 se generó un ahorro estimado en 758 millones de pesos anuales para la compañía; este efecto colateral de gran impacto para los costos de la empresa no había sido tenido en cuenta en la construcción inicial del proyecto, sin embargo, en términos de eficiencia representaron un ahorro del 5% del total de gastos de la empresa. Por lo tanto, es concluyente decir que estos tipos de modelos disruptivos, una vez son trabajados en equipo e involucrando los stakeholders adecuados generan cambios de impacto para las organizaciones.

La construcción de estos modelos requiere creatividad y paciencia, encontrar parámetros adecuados para un conjunto de datos no siempre va a ser igual, requiere conocer los datos, poder evaluarlos en diferentes contextos y con diferentes parámetros para identificar las ganancias en su rendimiento y luego si ponerlos en producción. Luego, si es un producto pensado en el cliente, ver cómo va cambiando su perspectiva y satisfacción por la novedad es el pago de tan complejo proceso.

Definitivamente existen muchas técnicas, lenguajes, programas y tecnologías para el desarrollo de estos procesos, sin embargo la clave está en el enfoque, en seguir la línea de lo que se está trabajando para llegar al resultado, pues muchas veces la gente aconsejaba probar otros métodos, modelos o lenguajes, pero creo que se logró un muy buen resultado con el actual, por supuesto con toda la disposición de poder mejorarlo, pero con una salida a producción exitosa y unos cambios importantes en la compañía.

REFERENCES

[1] C. García. (2019, Oct 20). Colombia: Así se beneficiaría el país al aumentar uso de pagos electrónicos [Online]. Available: https://www.colombiafintech.co/novedades/colombia-asi-se-beneficiaria-el-pais-al-aumentar-uso-de-pagos-electronicos.

[2] Credibanco (2019, Mar 3). Informe de gestión 2018. [Online]. Available: https://www.credibanco.com/file/912/download?token=raN2YDQP.

[3] A. Cabañete. Toma de decisiones: Análisis y entorno organizativo. Editorial UPC, Barcelona. ISBN: 978-84-8301-184-3. 1997.

[4] L. Méndez del Rio. Más allá del Bussines Intelligence: 16 experiencias de éxito. Editorial Gestión 2000. Barcelona. ISBN: 978-84-96612-10-5. 2006.

[5] J. Conesa, J curto. Introducción a la Inteligencia Empresarial. Editorial UOC. Barcelona. ISBN: 978-84-9788-886-8. 2010.

[6] E. Alpaydin. Introduction to Machine Learning second edition. Massachusetts Institute of Technology.2010.

[7] E. Alpaydin. Introduction to Machine Learning third edition. Massachusetts Institute of Technology.2014.

[8] I.H. Witten, E. Frank, and M. Hall.. Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). DOI:https://doi.org/0120884070, 9780120884070. 2011

[9] K. Krzyk. (2018, Jul 25) Coding Deep Learning for Beginners. [Online]. Available: https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d.

[10] Supervised learning — scikit-learn 0.20.0 documentation. Scikit-learn.org, 2018. http://scikit-learn.org/stable/supervised_learning.html#supervised-learning.

[11] T. Michell. The Discipline of Machine Learning. Carnegie Mellon. Jul 2006.

[12] V. Roman. (2019, Mar 27). Aprendizaje Supervisado: Introducción a la Clasificación y Principales Algoritmos. [Online]. Available: https://medium.com/datos-y-ciencia/aprendizaje-supervisado-introducci%C3%B3n-a-la-clasificaci%C3%B3n-y-principales-algoritmos-dadee99c9407.

[13] J. T. Palma y R. Marín. Inteligencia Artificial Técnicas, métodos y aplicaciones. Mc Graw Hill. Primera Edición. 2008.

[14] W. H. Greene. Econometric Analysis. Prentice Hall. Quinta Edición. 2010.

[15] L. Rokach y O. Maimond. Data Mining With Decision Trees Theory and Applications. World Scientific. Segunda Edición. 2015.

[16] T. M. Mitchell. The Discipline of Machine Learning. Mach. Learn. Aprendizaje automático: Qué es y por qué es importante. Sas.com, 2018.

[17] P. Flash y M.Kull. Precision-Recall-Gain Curves: PR Analysis Dine Right. 2014.

[18] V. Sofia. "Confusion Matrix-based Feature Selection.". 2011. [Online]. Avalaible : https://www.researchgate.net/profile/Atsushi_Inoue/publication/220833227_Page_Ranking_Refinement_Using_Fuzzy_Sets_and_Logic/links/54b743480cf24eb34f6e9e80.pdf#page=126. Google Academic. P. 120.

[19] Microsoft. Evaluación del rendimiento de un modelo en Machine Learning. [Online]. Available: https://docs.microsoft.com/es-es/azure/machine-learning/studio/evaluate-model-performance. 2017.

[20] T. Fawcett. An introduction to ROC analysis. Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, CA 94306, USA. 2005.

[21] Aprendizaje automático: Qué es y por qué es importante. Sas.com. [Online]. Available: https://www.sas.com/es_co/insights/analytics/machine-learning.html. 2018.

[22] FM, L. ¿Cómo funciona el nuevo sistema de inteligencia artificial de la Fiscalía?. Lafm.com.co. [Online]. Available: https://www.lafm.com.co/judicial/como-funciona-el-nuevo-sistema-de-inteligencia-artificial-de-la-fiscalia. 2018.

[23] Group, I. El deep learning será fundamental en las predicciones de demanda, fraude y fallo. Ituser.es, [Online]. Available:. https://www.ituser.es/estrategias-digitales/2017/09/el-deep-learning-sera-fundamental-en-las-predicciones-de-demanda-fraude-y-fallo. 2018

[24] Zemsania. Pasos para desarrollar un proyecto de Machine Learning. [Online]. Available: https://www.zemsania.com/recursos-zemsania/whitepapers/DTS/Machine_learning.pdf. 2018