ISBN: 978-628-96613-1-6. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2025.1.1.1827

Development of a predictive model to estimate disbursements and unit costs of sale in a financial institution using *Machine Learning*

Nicole Natividad-Lopez, Bachiller[1], Oscar Miranda-Castillo, Doctor[2]
[1,2]Pontificia Universidad Católica del Perú, Perú, nicole.natividad@pucp.edu.pe, miranda.oe@pucp.pe

***Abstract*– The accurate estimation of sales and their respective costs is essential for strategic decision-making in companies, so optimizing its calculation through advanced techniques is key to improving financial management. This research applies machine learning algorithms in the prediction of sales of financial products and unit costs of sale, using classification and regression models based on Random Forest. Through data preprocessing and hyperparameter optimization with GridSearchCV, a significant improvement in the accuracy of predictions is achieved, reducing time and effort compared to traditional methods.**

***Keywords*-- Machine learning, prediction, costs, Random Forest, financial automation.**

Desarrollo de un modelo predictivo para estimar desembolsos y costos unitarios de venta en una entidad financiera mediante *Machine Learning*

Nicole Natividad-Lopez, Bachiller[1], Oscar Miranda-Castillo, Doctor[2]
[1,2]Pontificia Universidad Católica del Perú, Perú, nicole.natividad@pucp.edu.pe, miranda.oe@pucp.pe

***Resumen*– La estimación precisa de ventas y sus respectivos costos es fundamental para la toma de decisiones estratégicas en las empresas, de manera que optimizar su cálculo mediante técnicas avanzadas es clave para mejorar la gestión financiera. Esta investigación aplica algoritmos de machine learning en la predicción de venta de productos financieros y costos unitarios de venta, utilizando modelos de clasificación y regresión basados en Random Forest. A través del preprocesamiento de datos y la optimización de hiperparámetros con GridSearchCV, se logra una mejora significativa en la precisión de las predicciones, reduciendo el tiempo y esfuerzo respecto a métodos tradicionales.**

***Palabras clave*-- Machine learning, predicción, costos, Random Forest, automatización financiera.**

I. Introducción

El costo de venta se refiere al costo directo asociado con la venta de productos o servicios. En el contexto de una entidad financiera, este costo incluye aspectos como la remuneración de los empleados directamente involucrados en las ventas, los costos de procesamiento de transacciones y otros gastos relacionados con la prestación de servicios financieros; además, puede incluir comisiones pagadas a agentes de ventas, costos de marketing y publicidad directamente vinculados con la promoción de productos financieros, entre otros [1].

Actualmente, el área de Finanzas de la entidad financiera ABC, denominada de esta manera por motivos de confidencialidad, se encarga de la actualización mensual del modelo de costo de venta de los productos de manera granular, es decir, el costo de venta de los productos según los distintos aspectos estratégicos de la entidad, como canal de venta o monto desembolsado. Dicho proceso es muy operativo respecto a las demás actividades que se realizan en el área de Finanzas y su mayor debilidad es el tiempo que consume para su actualización.

Por otro lado, los pronósticos permiten a los gerentes orientar la estrategia y tomar decisiones informadas sobre aspectos cruciales del negocio, como ventas, gastos, ingresos y asignación de recursos. Cuando se realizan de manera adecuada, los pronósticos proporcionan una ventaja competitiva y puede ser determinante entre resultados exitosos y fallidos. Además, los pronósticos financieros precisos son esenciales para desarrollar planes operativos y de personal que mejoren la eficiencia del negocio [2]. En particular, cabe mencionar que los pronósticos financieros se consideran una parte vital del proceso de planificación financiera de las entidades financieras, ya que les permite mantener un sistema financiero saludable y una economía eficiente [3].

La integración del aprendizaje automático (ML) y la inteligencia artificial (IA) en los pronósticos financieros marca una evolución significativa en el sector financiero, ya que surge de la necesidad de herramientas de análisis más precisas y eficientes. El panorama tradicional de los pronósticos financieros, alguna vez dominado por métodos estadísticos convencionales, se ha visto revolucionado con la llegada de sofisticadas técnicas de aprendizaje automático e inteligencia artificial, que ofrecen una precisión y eficiencia sin precedentes en la predicción de las tendencias de los datos [4]. En ese sentido, el aprendizaje automático demuestra ser una nueva manera de poder realizar proyecciones financieras, sin que esta pueda tener algún inconveniente por la cantidad de variables que se pueda utilizar o por el poder computacional que se pueda necesitar para entrenar al modelo [5].

Entonces, dado que se cuenta con data histórica del costo de venta de la entidad financiera ABC, se propuso desarrollar un modelo predictivo del costo de venta usando algoritmos de aprendizaje automatizado. A su vez, el modelo predictivo permite realizar un presupuesto de costo de venta granular, lo cual facilita tomar decisiones para mejorar la gestión financiera de los costos.

En la actualidad, el volumen de datos históricos disponible para una empresa supera la capacidad humana de procesarlos, interpretarlos y tomar decisiones efectivas basadas en ellos. En este contexto, la inteligencia artificial permite realizar estas tareas complejas a través de algoritmos avanzados.

De lo mencionado, considerando el acceso a la tecnología y la información de la entidad financiera ABC, se planteó investigar el comportamiento de los costos de una entidad financiera a través del desarrollo de un modelo con algoritmos de *Machine Learning*, propuesta que permitió automatizar su proceso de costeo actual.

II. Marco Teórico

A. *Minería de datos*

La minería de datos (*Data Mining*) es un proceso clave en el descubrimiento de patrones y conocimiento útil a partir de grandes volúmenes de datos [6]. La metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) es un enfoque estándar ampliamente utilizado para el desarrollo de

proyectos de minería de datos y ciencia de datos, que consta de las siguientes fases [7]:

1) Comprensión del negocio: Definición de los objetivos y necesidades del negocio.
2) Comprensión de los datos: Exploración y evaluación de la calidad de los datos.
3) Preparación de los datos: Limpieza, transformación y selección de variables relevantes.
4) Modelado: Aplicación de algoritmos de *machine learning* para la predicción.
5) Evaluación: Validación de los modelos según métricas de desempeño.
6) Despliegue: Implementación del modelo en un entorno de producción.

Estas fases fueron utilizadas como base metodológica en esta investigación para garantizar un proceso estructurado y adaptable en la construcción de los modelos predictivos.

B. Machine Learning

Machine learning o aprendizaje automático es el subconjunto o rama de la inteligencia artificial enfocada en el desarrollo de modelos y algoritmos que pueden aprender y tomar decisiones sin programación explícita; la característica que la distingue es su capacidad para aprender de los datos y adaptarse a ellos [8] [9]. En un contexto financiero, debido a su enfoque numérico, la rama de la inteligencia artificial más relevante ha sido el aprendizaje automático. Las técnicas de aprendizaje automático e inteligencia artificial brindan importantes beneficios a los responsables de tomar decisiones financieras en términos de nuevos enfoques para modelar y realizar predicciones a partir de los datos financieros [10].

C. Modelos de Machine Learning

La naturaleza de la variable a predecir y el tipo de datos determinan qué modelo es adecuado. Usualmente se elige el mejor modelo, considerando métricas de desempeño para comparar entre sí a los modelos, y luego este se optimiza para obtener los mejores resultados para un problema en particular.

En la Tabla I se presentan los tipos de modelos de aprendizaje automático comúnmente empleados para desarrollar predicciones [8].

TABLA I
TIPOS DE MODELOS DE MACHINE LEARNING

Tipos de aprendizaje	Tipo de modelo
Supervisado	Regresión Lineal Simple y Múltiple
	Regresión Logística
	Naïve Bayes
	Árbol de decision
	Bosque aleatorio
	Máquinas de Vectores de Soporte (SVM)
	K vecinos más cercanos (KNN)
	Redes neuronales (ANN)
No supervisado	Análisis de Componente Principal (PCA)
	K-Means Clustering

En el aprendizaje supervisado se entrenan modelos con datos etiquetados que permiten hacer predicciones sobre datos no vistos o futuros, lo que significa que el resultado deseado se produce mientras se entrena el modelo; en consecuencia, los modelos pueden aprender cómo las características de entrada se relacionan con las etiquetas de salida. El término "supervisado" se refiere a un conjunto de ejemplos de entrenamiento (entradas de datos) donde las variables de salida deseadas (etiquetas) ya son conocidas. Este aprendizaje está compuesto por dos categorías: regresión, cuya variable de salida es continua, y clasificación de datos, cuya variable de salida es discreta [8] [9] [11].

D. Métricas de evaluación

Las métricas de evaluación en el aprendizaje supervisado son fundamentales para cuantificar la calidad y el rendimiento de los modelos. Asimismo, estas permiten comprender si un modelo está cumpliendo con sus objetivos y ofrecen información clave para identificar problemas como sobreajuste (*overfitting*) o subajuste (*underfitting*).

Las métricas utilizadas en la presente investigación incluyen la métrica de sensibilidad en clasificación, así como el coeficiente de determinación (R^2) y el error cuadrático medio (MSE) en regresión.

1) Sensibilidad (*Recall*): Evalúa la capacidad del modelo para identificar correctamente los casos positivos, es decir, casos en los que el valor predicho coincide con el valor real.
2) Coeficiente de determinación (R^2): Indica qué porcentaje de la variabilidad total de la variable objetivo es explicada por las variables independientes en el modelo.
3) Error cuadrático medio (MSE): Mide el promedio de los cuadrados de los errores, es decir, las diferencias entre los valores observados y los valores predichos.

II. METODOLOGÍA

Esta investigación es de tipo exploratoria y descriptiva, puesto que se exploraron diversas técnicas de *machine learning* para identificar la más adecuada para el problema específico, y se describieron las características del modelo final; además, esta investigación se enfoca en un período específico, analizando datos históricos de la entidad ABC desde enero del 2023 hasta julio del 2024 para predecir comportamientos futuros. Cabe mencionar que para la presente investigación la variable dependiente u objetivo es el costo de venta unitario de los productos, la cual conceptualmente es el costo de venta total entre la cantidad de operaciones desembolsadas del producto determinado. En ese sentido, si en un determinado mes no hubo operaciones de algún producto, su costo unitario será cero.

En la Fig. 1 se presenta el esquema del *workflow* de la presente investigación, el cual está compuesto por tres etapas: extracción, entrenamiento y validación. Asimismo, dentro de este esquema se encuentran las 5 primeras fases de la metodología CRISP-DM mencionadas previamente.

En la primera etapa, se inicia con la comprensión del negocio, la cual permite alinear el modelo con los objetivos comerciales de la empresa. Esto garantiza que el proyecto de *machine learning* no solo tenga valor técnico, sino también un impacto tangible en los resultados del negocio. Además, en esta etapa se recopiló la base de datos de la empresa y se analizó para entenderla y definir los objetivos del proyecto. A partir de ello, se llevó a cabo la preparación de los datos, proceso que implica la limpieza de datos, identificación de anomalías, descubrimiento de patrones y organización la información de acuerdo a las necesidades.

En la segunda etapa, una vez que los datos están procesados, se implementaron los algoritmos de *machine learning* para desarrollar el modelo predictivo, ya sean de regresión o clasificación.

En la tercera etapa, se validó el modelo mediante diversas pruebas para asegurarse de que refleje adecuadamente la realidad. Finalmente, se evaluó si el modelo es efectivo y útil para predecir el costo de venta unitario de los productos de la empresa.

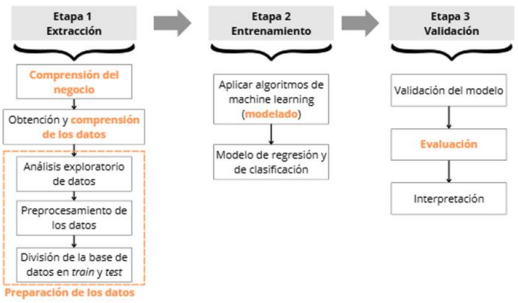

Fig. 1 *Workflow* de la investigación.

La propuesta de mejora se desarrolló usando el lenguaje de programación *Python*, el cual es uno de los lenguajes más usados para desarrollar modelos de *machine learning* y cuenta con librerías que facilitan el uso de algoritmos complejos, además de ser un lenguaje de fácil aprendizaje y uso.

A. Fase 1: Comprensión del negocio

Para la adecuada comprensión del negocio, se elaboró la recopilación de la información de la situacional actual de la entidad financiera ABC. En ese sentido, se empezó por conocer la empresa: identificar los hechos más relevantes de la compañía a través de su desarrolló histórico, describir la estructura de trabajo, visión, misión, datos estratégicos, modelo de negocio, cadena de suministro y mapa de procesos.

B. Fase 2: Comprensión de los datos

Los datos fueron proporcionados por la entidad financiera ABC con el objetivo de desarrollar un modelo predictivo del costo de venta unitario de sus productos. Debido a la variabilidad del costo de venta unitario y a la gran cantidad de valores iguales a 0 en la base de datos, se decidió dividir dicha base en dos bases y generar así dos modelos predictivos: uno de clasificación, de manera que se aprovechen los registros con valor de costo 0 para predecir si se venderán o no ciertos productos financieros, y uno de regresión, que considere solo los registros cuya variable objetivo tenga valores distintos a 0 que permita predecir el costo de venta de los productos financieros.

C. Fase 3: Preparación de los datos

Como parte de la tercera fase, se analizaron las variables y la estructura de los datos recolectados. Fig. 1 muestra que en esta etapa se realizó el análisis exploratorio de datos (EDA), preprocesamiento de los datos, y la división y preparación de los dos *data sets* tanto para el modelo de clasificación como para el de regresión.

El análisis exploratorio de datos (EDA) facilita a los analistas y científicos de datos a comprender la estructura y distribución de los datos. A través de visualizaciones y análisis descriptivos, se identifican patrones, tendencias y relaciones entre las variables. El EDA antes del preprocesamiento ayuda a diagnosticar problemas que luego guían las estrategias de preprocesamiento, como la selección de técnicas de imputación o transformación de variables; esto se debe a que realizar este análisis antes del preprocesamiento permite evaluar la calidad de los datos crudos, lo cual es importante para identificar cualquier inconsistencia o característica que deba corregirse durante el preprocesamiento, como valores faltantes, duplicados o atípicos [12].

En primer lugar, se realizó el análisis univariante para la variable objetivo de las dos bases de datos. Por un lado, dado que la variable objetivo para el modelo de clasificación es una variable discreta dicotómica, es decir, es una variable categórica de dos clases, se analizará el porcentaje de cada clase en la base de datos. Fig. 2 presenta que la clase “0” tiene una mayor proporción en la base (81.14%), en otras palabras, la base está desbalanceada ya que no se tiene una proporción similar entre las dos clases de la variable objetivo. En ese sentido, es importante elegir modelos que puedan manejar este desequilibrio o ajustar los modelos para que consideren el desbalance en la variable objetivo; por ello, se trabajó con los siguientes modelos: Regresión Logística, Naïve Bayes, K-NN, Árbol de decisión y Bosque aleatorio.

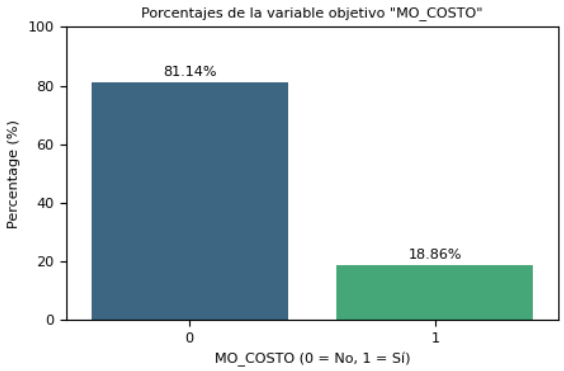

Fig. 2 Proporción de la variable objetivo - Clasificación.

Por otro lado, debido a que la variable objetivo para el modelo de regresión es una variable numérica, se realizó el análisis univariante de esta a través de gráficas como un histograma y un *boxplot*. La Fig. 3 muestra que la distribución de la variable objetivo parece estar sesgada hacia la derecha, es decir, la mayoría de los datos están en un rango relativamente pequeño, con una gran cantidad de datos acumulándose en los valores cercanos al valor mínimo que es 12.5, y hay algunos valores que superan los 2,500 o 5,000, y los valores más extremos (mayores de 10,000) son menos frecuentes; esto sugiere que hay una fuerte asimetría positiva en la variable objetivo. De lo mencionado, cuando la variable objetivo en un problema de regresión tiene asimetría positiva, en otras palabras, cola larga hacia la derecha, algunos modelos y estrategias específicas pueden manejar mejor esta situación; por ello, además de estandarizar la variable objetivo, se trabajó con los siguientes modelos: Regresión Lineal Múltiple, Árbol de decisión, Bosque aleatorio y Redes neuronales.

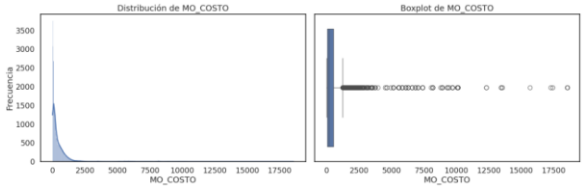

Fig. 3 Histograma y *Boxplot* de la variable objetivo - Regresión.

La ingeniería de variables, es decir, la transformación y codificación de las variables, es uno de los aspectos más críticos para el éxito del aprendizaje automático [13]. En particular, las variables categóricas deben ser transformadas en una representación numérica antes de ser introducidas en los modelos de *machine learning*. Las técnicas de codificación, como el *Label Encoding* y el *One-Hot Encoding*, son formas comunes de transformar datos categóricos; cabe mencionar que las variables categóricas pueden causar problemas si no se procesan adecuadamente, ya que muchos modelos *de machine learning* no pueden manejar texto directamente [14]. Asimismo, la ingeniería de características tiene un impacto directo en el rendimiento del modelo, de manera que este proceso no solo hace que los datos sean "legibles" para los algoritmos, sino que también pueden mejorar la capacidad predictiva al proporcionar una mejor representación de la información presente en las características originales [15].

Como se observa en la Fig. 4, la base de datos proporcionada por la empresa contiene variables categóricas cuyos valores no son números enteros sino texto. En ese sentido, para los modelos propuestos, se optó por codificar las variables categóricas presentes en las dos bases de datos usando el método de *Label Encoding*.

	CO_RANGO	NU_PERI_MES	MES	CO_CANAL	TI_CLIENTE	MO_COSTO		NIVEL_0	CO_RIESGO	MOD_ATENCION
0	4	202301	1	AD	NUEVO	0	Capital		RA	EM
1	5	202301	1	AD	NUEVO	0	Capital		RA	EM
2	6	202301	1	AD	NUEVO	0	Capital		RA	EM
3	7	202301	1	AD	NUEVO	0	Capital		RA	EM
4	8	202301	1	AD	NUEVO	0	Capital		RA	EM

Fig. 4 Estructura de los datos.

Otro paso importante del preprocesamiento de datos es la normalización, ya que muchos algoritmos de aprendizaje automático requieren que la variable objetivo esté en una determinada escala para un rendimiento óptimo. Cuando se trabaja con variables que presentan asimetría positiva (sesgo hacia la derecha), es importante aplicar ciertas transformaciones antes de la estandarización. La transformación logarítmica es una de las más comunes para corregir este tipo de asimetría ya que ayuda a reducir el sesgo al comprimir la escala de valores, lo que lleva a una distribución más normal. Esta transformación es útil para cumplir con los supuestos de normalidad que muchos modelos estadísticos requieren; en adición, este tipo de transformación es efectiva para mitigar el impacto de los valores atípicos. Cabe mencionar que los datos con valores extremos pueden distorsionar significativamente las medidas de tendencia central (como la media) y las medidas de dispersión (como la varianza) [16].

En ese sentido, para el modelo de regresión, dado que la variable objetivo numérica, costo de venta unitario de los productos, presenta asimetría positiva como se observó en la Fig. 3, se aplicó una transformación logarítmica a dicha variable antes de estandarizarla para reducir la asimetría. Después de la transformación, se evaluó nuevamente la distribución de los datos para validar que la normalidad haya mejorado antes de proceder con la estandarización.

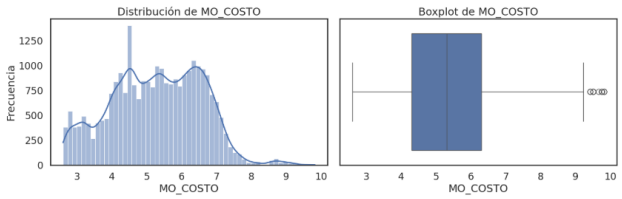

Fig. 5 Histograma y *Boxplot* de la variable objetivo transformada - Regresión.

La Fig. 5 muestra que la distribución de la variable objetivo luego de transformarla usando el método logarítmico ha reducido considerablemente tanto su asimetría positiva como la cantidad de valores atípicos. Después de transformar la variable, se procedió con la estandarización, cuyo resultado se observa en la Fig. 6; este proceso se realiza para escalar los datos a una escala comparable, generalmente con media 0 y desviación estándar 1, esto es importante para que los modelos de *machine learning*, especialmente los que utilizan distancias (como KNN, etc.), no se vean sesgados por las escalas de las variables

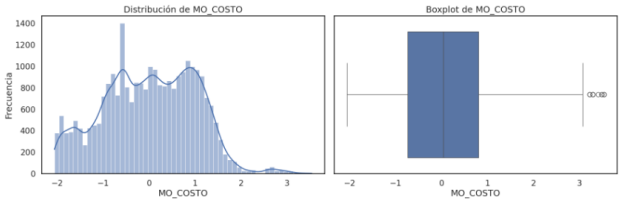

Fig. 6 Histograma y *Boxplot* de la variable objetivo estandarizada-Regresión.

Finalmente, se realizó la división de las bases de datos en entrenamiento y prueba, esta es una práctica esencial en *machine learning* para evaluar el rendimiento del modelo en datos no vistos.

D. Fase 4: Modelado a través de algoritmos de machine learning

Siguiendo la metodología CRISP-DM, se entrenaron los modelos propuestos para clasificación y regresión, de manera que se optimizaron los hiperparámetros mediante *GridSearchCV*. Asimismo, se compararon diferentes configuraciones de parámetros para seleccionar el modelo con mejor desempeño.

E. Fase 5: Evaluación de los modelos de machine learning

Con respecto al modelo de clasificación, para el presente trabajo de investigación, se tiene como objetivo principal maximizar la sensibilidad, es decir, detectar correctamente todos los casos positivos de desembolso. La Fig. 7 muestra que el modelo *Random Forest* (80.65%) es el modelo más sensible, es decir, es el mejor para identificar correctamente los casos en los que sí habrá desembolso. Además, la Regresión Logística (68.32%) sigue como una alternativa, mientras que Naïve Bayes (18.38%) tiene un rendimiento notablemente bajo en esta métrica, lo que indica que pierde muchos casos positivos reales. Cabe mencionar que Naïve Bayes y Regresión Logística tienen un rendimiento limitado, especialmente en sensibilidad y F1 Score, por lo que no son recomendados. Además, el Árbol de Decisión muestra resultados intermedios y podría ser una opción en contextos donde la interpretación del modelo es crítica.

	Modelo	Umbral	Exactitud	Precisión	Sensibilidad	F1 Score
0	Regresión Logística	0.5	0.661640	0.315544	0.683274	0.431716
1	Naive Bayes	0.5	0.784984	0.359662	0.183368	0.242898
2	KNN	0.5	0.909139	0.803030	0.684960	0.739311
3	Árbol de Decisión	0.5	0.822752	0.564597	0.252107	0.348569
4	Random Forest	0.5	0.760147	0.427140	0.806518	0.558495

Fig. 7 Métricas de evaluación - Clasificación.

Por otro lado, para que un modelo de regresión tenga buen rendimiento, su valor de MSE ideal es 0 o valores cercanos a 0, ya que esta es una métrica que mide el promedio de los errores cuadrados, es decir, la diferencia al cuadrado entre los valores reales y los valores predichos por el modelo. Además, tendría un valor ideal de R^2 igual a 1, lo que indica que el modelo explica el 100% de la variabilidad de los datos; un valor de R^2 cercano a 0 sugiere que el modelo no explica bien la variabilidad y que podría ser inapropiado para el conjunto de datos. La Fig. 8 muestra que el modelo de redes neuronales es el mejor modelo para este conjunto de datos, dado su rendimiento superior tanto en el MSE como en el R^2. Si bien el árbol de decisión también ofrece un buen desempeño, tiene un MSE más alto que las redes neuronales. Cabe destacar que el modelo *Random Forest* es una opción sólida y parece ser un buen compromiso entre complejidad y rendimiento; la regresión lineal múltiple tiene el peor rendimiento, lo que sugiere que un modelo más complejo puede ser necesario para capturar las relaciones en los datos de manera efectiva.

Estos resultados indican que, para predicciones más precisas, se recomienda optar por modelos más complejos, como redes neuronales o árboles de decisión, en lugar de enfoques lineales más simples.

	Modelo	MSE	R2
0	Regresión Lineal Múltiple	0.531561	0.473389
1	Árbol de Decisión	0.124135	0.877021
2	Random Forest	0.235732	0.766463
3	Redes neuronales	0.086045	0.914756

Fig. 8 Métricas de evaluación - Regresión.

F. Optimización de los modelos de machine learning

GridSearchCV es una técnica de optimización exhaustiva utilizada en aprendizaje automático para ajustar los hiperparámetros de un modelo. Esta técnica funciona explorando todas las combinaciones posibles de un conjunto de valores predefinidos para los hiperparámetros y selecciona la combinación que maximiza el rendimiento del modelo en una métrica específica, por ejemplo, precisión, F1-score, R2, entre otras. Además, esta técnica ajusta adecuadamente los hiperparámetros, como la regularización en regresión o los criterios de división en árboles de decisión, de manera que ayuda a encontrar el balance ideal entre el ajuste del modelo y su capacidad de generalización [17]. Este método se aplicó tanto para los modelos de clasificación como para los modelos de regresión.

En la Fig. 9 se observa que el modelo *Random Forest* sigue siendo el que obtiene el mayor resultado en la métrica sensibilidad.

	Modelo	Umbral	Exactitud	Precisión	Sensibilidad	F1 Score
0	Regresión Logística	0.5	0.814085	0.545994	0.068927	0.122401
1	Naive Bayes	0.5	0.784984	0.359662	0.183368	0.242898
2	KNN	0.5	0.912310	0.898490	0.601798	0.720808
3	Árbol de Decisión	0.5	0.917524	0.799082	0.750140	0.773838
4	Random Forest	0.5	0.760147	0.427140	0.806518	0.558495

Fig. 9 Métricas de los modelos de clasificación optimizados.

Por otro lado, respecto a los modelos de regresión, la Fig. 10 muestra que el modelo *Random Forest* muestra el mejor rendimiento en términos de R^2 (93.2%), lo que indica que es el modelo más preciso para predecir la variable objetivo. Además, el Árbol de Decisión también muestra un buen rendimiento, con un MSE más bajo que las Redes Neuronales y un R^2 ligeramente más alto (92.7%). Cabe mencionar que la Regresión Lineal Múltiple es el modelo menos efectivo, con un R^2 bajo y un MSE relativamente más alto (0.53).

En ese sentido, los modelos basados en árboles (Árbol de Decisión y *Random Forest*) parecen ser más efectivos para este conjunto de datos en particular, especialmente el *Random Forest*, que es el modelo más robusto y preciso de los cuatro.

	Modelo	MSE	R2
0	Regresión Lineal Múltiple	0.531561	0.473389
1	Árbol de Decisión	0.073420	0.927264
2	Random Forest	0.068419	0.932219
3	Redes Neuronales	0.082154	0.918611

Fig. 10 Métricas de los modelos de regresión optimizados.

Cabe mencionar que el algoritmo *Random Forest* o Bosque Aleatorio es ampliamente utilizado en el aprendizaje automático, especialmente dentro del área de la minería de datos, debido a su efectividad en diversas aplicaciones. Una de las principales fortalezas de este algoritmo es su versatilidad, ya que puede aplicarse tanto en modelos de regresión como de clasificación; asimismo, este algoritmo proporciona resultados de alta calidad sin requerir un ajuste exhaustivo de los hiperparámetros, y destaca por su claridad y facilidad de interpretación [18].

En adición, el Bosque Aleatorio reduce el sesgo y la variabilidad al combinar múltiples árboles de decisión. De esta manera, supera el riesgo de sobreajuste propio del algoritmo de árbol de decisión, un problema importante en el aprendizaje automático ya que se busca desarrollar modelos que mantengan un buen desempeño incluso con datos nuevos [18].

IV. Resultados

Se identificó como mejor modelo de clasificación al *Random Forest*, ya que obtuvo el mayor resultado en la métrica *recall* o sensibilidad, la cual mide la capacidad del modelo para identificar correctamente los casos donde sí habrá desembolsos. Además, este modelo permite maximizar la sensibilidad ajustando hiperparámetros o configurando el umbral de clasificación. Asimismo, el modelo de regresión que obtuvo los mejores resultados tanto en R^2 como en MSE fue *Random Forest*.

El beneficio más importante generado por esta propuesta es la planificación de asignación de recursos, dado que al tener un presupuesto del costo de venta unitario mensual de los productos se puede planificar de manera óptima a qué productos asignarle adecuadamente un costo. Además, se genera un ahorro importante del recurso tiempo, el cual es crucial en particular en las áreas de finanzas. Un indicador importante para cuantificar el impacto es la cantidad de días que toma la actualización y revisión del modelo de costos tradicional. De la Fig. 11, considerando un periodo de doce meses consecutivos, se observa que como mínimo ha transcurrido diez días para que el modelo tradicional se cierre, y como máximo se tiene que ha tomado veintinueve días cerrar el modelo. Asimismo, se observa en la figura que, en los últimos 3 meses, los días tuvieron una tendencia creciente, lo cual afecta negativamente retrasando la generación de reportes que necesitan esta información.

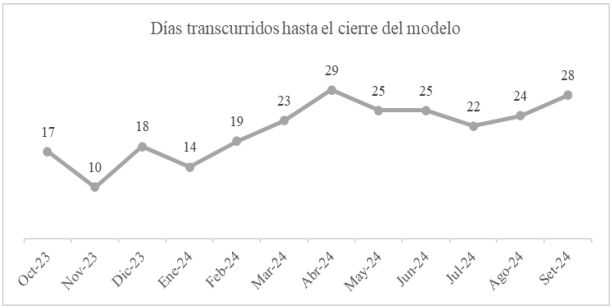

Fig. 11 Días transcurridos hasta el cierre del modelo tradicional.

TABLA II
EVALUACIÓN DE DÍAS TRANSCURRIDOS HASTA EL CIERRE DEL MODELO DE COSTOS

AS IS (días)	TO BE (días)	Variación (Δ%)
21.17	**4.5**	**-78.7%**

Por otro lado, la generación de predicciones con el modelo de *machine learning* tiene una duración de 4 a 5 días, de manera que se reduce en 78% aproximadamente el tiempo que implica obtener los costos unitarios mensuales de los productos financieros. Esto representa una mejora significativa ya que antes de la mejora el proceso duraba en promedio 21 días.

V. Conclusiones

El desarrollo de modelos predictivos utilizando técnicas de *machine learning*, como *Random Forest*, ha permitido abordar de manera eficiente y precisa dos aspectos clave: la clasificación de casos de desembolso de productos y la predicción del costo de venta unitario en la entidad financiera

ABC. Además, el uso de técnicas de optimización, como *GridSearchCV*, en el desarrollo de los modelos predictivos permitió alcanzar un desempeño óptimo tanto en clasificación como en regresión.

En la clasificación, *Random Forest* sobresalió debido a su alta sensibilidad, minimizando falsos negativos, lo cual es fundamental para evitar la pérdida de oportunidades comerciales y optimizar la gestión de recursos.

Para regresión, el preprocesamiento de la variable objetivo fue crucial para mejorar el rendimiento del modelo: se aplicó una transformación logarítmica para reducir la asimetría positiva y una estandarización para garantizar una mejor distribución, lo que potenció la capacidad predictiva del modelo. Como resultado, *Random Forest* también se destacó en regresión por su capacidad para capturar relaciones no lineales en los datos, reflejado en un mayor R^2 y menor MSE, lo que garantiza estimaciones confiables para el costo de venta unitario. Cabe mencionar que la optimización ajustó los hiperparámetros clave del modelo de *Random Forest*, maximizando su sensibilidad en clasificación y, mejorando el R^2 y reduciendo el MSE en regresión.

En adición, este enfoque predictivo automatizado representa una mejora significativa frente al método tradicional basado en Excel, no solo en términos de precisión, sino también en la eficiencia operativa. El tiempo requerido para realizar análisis complejos se ha reducido considerablemente, eliminando riesgos asociados a errores humanos y mejorando la capacidad de procesar grandes volúmenes de datos.

De lo mencionado, se concluye que esta propuesta no solo fortalece la toma de decisiones estratégicas, sino que también proporciona un marco escalable y replicable para futuros proyectos de análisis predictivo en la entidad financiera ABC.

VI. Recomendaciones

Se recomienda ampliar los conocimientos en otros lenguajes de programación, como R, que ofrece herramientas específicas para análisis estadístico y modelado avanzado.

Se sugiere también desarrollar modelos en tiempo real, es decir, sistemas de predicción en línea que puedan adaptarse dinámicamente a los nuevos datos a medida que se generan.

Además, se podría considerar aplicar modelos predictivos a otros contextos financieros, pero es importante ajustar el enfoque según las particularidades de cada área. Por ejemplo, en fondos mutuos, se debe adaptar el modelo predictivo para capturar la volatilidad del mercado; en inversión, para predecir rendimientos y riesgos; en seguros, para estimar reclamaciones y primas; y en gestión de riesgos, para identificar y mitigar riesgos financieros. Por lo tanto, se recomienda evaluar cómo estos modelos pueden adaptarse a las particularidades de cada área financiera, ajustando los parámetros según el tipo de análisis requerido en cada contexto; esta evaluación no solo ayudará a mejorar la precisión de cada modelo, sino también a asegurar que estos sean relevantes y útiles para diferentes aplicaciones dentro del sector financiero.

Agradecimientos

Los autores, Nicole Natividad Lopez y Oscar Miranda Castillo, agradecen a la Pontificia Universidad Católica del Perú por el apoyo recibido en la formación académica, especialmente en los temas de *machine learning* y sus algoritmos, conocimientos esenciales para la realización de esta investigación. Asimismo, se extiende un agradecimiento a la entidad ABC por proporcionar acceso a su información histórica, lo que permitió aplicar dichos conocimientos y optimizar procesos clave.

Referencias

[1] Zendesk. (2023, mayo 1). ¿Qué es costo de venta? Calcúlalo en 4 pasos prácticos. https://www.zendesk.com.mx/blog/que-es-costo-de-venta/.

[2] Boffa, J. (2023). AI Assisted Business Analytics Techniques for Reshaping Competitiveness. https://doi.org/10.1007/978-3-031-40821-2.

[3] Islam, U., & Masum, A. K. M. (2022). Forecasting of Bank Performance using Hybrid Machine Learning Techniques. International Conference on Innovations in Science, Engineering and Technology (ICISET), 433-438. https://ieeexplore-ieee-org.ezproxybib.pucp.edu.pe/stamp/stamp.jsp?tp=&arnumber=9775833&tag=1.

[4] Olubusola, O., Zamanjomane Mhlongo, N., Obinna Daraojimba, D., Olusola Ajayi-Nifise, A., & Falaiye, T. (2024). Machine learning in financial forecasting: A U.S. review: Exploring the advancements, challenges, and implications of AI-driven predictions in financial markets. https://doi.org/10.30574/wjarr.2024.21.2.0444.

[5] Bellido-Zea, C., Villalobos-Meneses, B., Alfaro-Rodriguez, C., Grados-Espinoza, A., Gomero-Ostos, N., Hoyos-Rivas, F., & Ramirez-Veliz, F. (2023). Machine Learning applied to Projected Financial Statements (PFS). https://doi.org/10.18687/LACCEI2023.1.1.1420.

[6] Han, J., Kamber, M., & Pei, J. (2011). Data Mining. Concepts and Techniques (3.a ed.). The Morgan Kaufmann Series in Data Management Systems. https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf.

[7] Vorhies, W. (2016, julio 26). CRISP-DM – a Standard Methodology to Ensure a Good Outcome. Tech Target - Data Science Central. https://www.datasciencecentral.com/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome/.

[8] Singh, P., Mishra, A. R., & Garg, P. (Eds.). (2024). Data Analytics and Machine Learning (Vol. 145). Springer Nature Singapore. https://doi.org/https://doi.org/10.1007/978-981-97-0448-4.

[9] Organización para la Cooperación y el Desarrollo Económicos. (2021). Artificial Intelligence, Machine Learning and Big Data in Finance Opportunities, Challenges and Implications for Policy Makers. 1-72. https://www.oecd.org/finance/financial-markets/Artificial-intelligence-machine-learning-big-data-in-finance.pdf.

[10] Aziz, S., Dowling, M., Hammami, H., & Piepenbrink, A. (2022). Machine learning in finance: A topic modeling approach. European Financial Management, 28(3), 744-770. https://doi.org/10.1111/EUFM.12326.

[11] Raschka, S., & Mirjalili, V. (2019). Python machine learning : machine learning and deep learning with python, scikit-learn, and TensorFlow 2 (3.a ed.).

[12] Van der Laan, M. J., & Rose, S. (2011). Targeted learning: Causal inference for observational and experimental data. Springer Science & Business Media. https://doi.org/10.1007/978-1-4419-9782-1.

[13] Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. communications of the acm, 55, 1-10. https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf.
[14] Harris, D., & Harris, S. (2012). Digital Design and Computer Architecture. https://edisciplinas.usp.br/pluginfile.php/7910542/mod_resource/content/1/Digital%20Design%20and%20Computer%20Architecture.pdf.
[15] Zheng, A., & Casari, A. (2018). Feature Engineering for Machine Learning. O'Reilly Media, Inc. https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/.
[16] Gnanadesikan, R., Kettenring, J., & Telephone, B. (1972). ROBUST ESTIMATES, RESIDUALS AND OUTLIER DETECTION WITH MULTIRESPONSE DATA. Biometrics, 28(1), 81. https://doi.org/10.2307/2528963.
[17] scikit-learn. (2024). Tuning the hyper-parameters of an estimator. https://scikit-learn.org/stable/modules/grid_search.html.
[18] Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. Babylonian Journal of Machine Learning, 2024, 69–79. https://doi.org/10.58496/bjml/2024/007