Estimación de modelos de clasificación para la identificación de causas de insatisfacción de clientes basado en datos del NPS

Lizeth Tatiana Cabiativa Aranguren

Asesor

Rafael Roberto Ruiz Escocia

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Ciencia de Datos y Analítica

2025

Dedicatoria

Este proyecto aplicado está dedicado a Dios, que me dio la guía, fortaleza y constancia desde el día uno que inicie los estudios en la especialización, he sentido su compañía en los momentos buenos y su amparo y respaldo en los momentos de confusión, miedo o duda. A mis padres Nelson y Olga y a mi hermano Sebastián quienes siempre me han brindado su apoyo y cariño, son mi inspiración e impulso para confrontar de la mejor manera posible cada nuevo desafío en el día a día.

Agradecimientos

Gracias infinitas a mis padres, que, con su amor incondicional, su compañía y su sabiduría, hacen que todos los días me esfuerce por ser una mejor persona a nivel personal y profesional, gracias por siempre escucharme, apoyarme y acompañarme durante este proceso académico. También le doy gracias a la universidad Nacional Abierta y a Distancia por abrirme las puertas de la institución, permitirme ampliar mis conocimientos y habilidades en pro de mi desarrollo profesional.

Resumen

Este proyecto aborda la estimación de modelos de clasificación para la identificación de causas de insatisfacción de clientes basado en datos del Net Promotor Score (NPS), una métrica para medir la satisfacción de los clientes. La dificultad para detectar causas específicas de insatisfacción, la falta de priorización en los planos de acción y la ausencia de recomendaciones concretas representan desafíos para la empresa analizada. Durante la ejecución del proyecto se prepararon datos del NPS, se entrenaron varios modelos de machine learning, incluyendo regresión logística, Random Forest, Gradient Boosting Machines (GBM) y Support Vector Machines (SVM), y se propusieron recomendaciones basadas en los resultados del análisis con el fin de ayudar a la empresa a mejorar la satisfacción del cliente, optimizar procesos internos y generar mayor valor para los usuarios.

Palabras clave: NPS, regresión logística, random forest, GBM y SVM.

Abstract

This project focuses on estimating classification models to identify the causes of customer dissatisfaction based on Net Promoter Score (NPS) data; a metric used to measure customer satisfaction. The difficulty in detecting specific causes of dissatisfaction, the lack of prioritization in action plans, and the absence of concrete recommendations represent challenges for the analyzed company. During the project's execution, NPS data was prepared, several machine learning models were trained, including logistic regression, Random Forest, Gradient Boosting Machines (GBM) and Support Vector Machines (SVM), and recommendations were proposed based on the analysis results to help the company improve customer satisfaction, optimize internal processes, and generate greater value for users.

***Keywords*:** NPS, regression logistics, random forest, GBM y SVM.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Descripción del Problema

Actualmente, la empresa enfrenta altos niveles de insatisfacción en los clientes, ya que según los resultados de NPS, el 27.5% de los usuarios que se comunican con el equipo de servicio al cliente manifiestan inconformidad del servicio recibido, sin embargo al no tener un proceso adecuado de análisis de información se presenta dificultad para identificar las causas específicas de la insatisfacción, falta de priorización adecuada de los planes de acción de mejora, y ausencia de recomendaciones claras y planes de acción efectivos. Estas deficiencias dificultan la mejora del servicio y la calidad de los productos, afectando negativamente la satisfacción del cliente.

Implementar un modelo de clasificación basado en datos permitiría estructurar y categorizar de manera más precisa los motivos de insatisfacción de los clientes, facilitando la identificación de patrones y causas raíz de la insatisfacción. Esto contribuiría a generar recomendaciones más específicas y planes de acción priorizados para los equipos responsables de la mejora de CX (Customer Experience), y del desarrollo de productos y servicios, optimizando así la calidad del servicio y la satisfacción del cliente.

Justificación

Trabajo para una empresa de comercio electrónico que funciona por medio de aplicación web, esta empresa tiene como propósito generar experiencias extraordinarias para los usuarios, con el fin de que ellos hagan uso de los productos que hay en ella y que recomienden la aplicación a otras personas, este objetivo se mide por medio de la métrica de NPS, a partir de los resultados y de conocer la opinión de los usuarios se toman las decisiones de activar, mantener o retirar productos, actualizar funciones de la aplicación, contratar o desvincular personal, extender el alcance en el mercado, entre otros temas.

Este modelo de clasificación le ayudará a la empresa encontrar la causa raíz de la insatisfacción de los clientes, ya sea por problemas con los productos, servicios o el equipo de CX (Customer Experience), adicional ayudará a mejorar la experiencia del cliente, mejorar los procesos internos y seguir generando valor para los usuarios, ya que con un análisis más estructurado y automatizado, la empresa podrá mejorar la toma de decisiones estratégicas, optimizar los procesos internos y desarrollar planes de acción más efectivos, lo anterior no solo contribuirá a elevar la satisfacción del cliente y fortalecer la relación con los usuarios, sino que también permitirá una gestión más eficiente de los recursos, asegurando que las iniciativas de mejora generen un impacto significativo en la experiencia del cliente y en el crecimiento del negocio.

Objetivos

Objetivo General

Estimar modelos de machine learning que logren identificar las causas de insatisfacción de los clientes utilizando datos de NPS (Net Promoter Score, métrica para medir la satisfacción de los clientes), con el fin de clasificar la experiencia de los clientes y evaluar la calidad del servicio prestado.

Objetivos Específicos

Realizar un análisis exploratorio de la información de NPS del segundo trimestre de 2024, evaluando las tendencias y patrones en los datos de satisfacción del usuario, así como identificando posibles factores asociados a la insatisfacción.

Entrenar modelos de machine learning de clasificación, utilizando los datos de NPS, para identificar y categorizar las principales causas de insatisfacción de los clientes.

Evaluar el desempeño de los modelos de aprendizaje automático implementados, empleando métricas de precisión, sensibilidad y balance, con el objetivo de alcanzar resultados mayores al 80% en cada una de ellas, esto con el fin de garantizar su efectividad en la clasificación de la experiencia del cliente y la calidad del servicio de atención al cliente. El umbral se establece considerando que existe un porcentaje de error manual en las respuestas de los usuarios en las encuestas de NPS, lo que puede generar equivocación en la clasificación para los modelos. Además, al tratarse de los primeros modelos de clasificación utilizados en la empresa, este porcentaje se ha definido como un estándar aceptable.

Marco de Referencia

Estado del Arte

Para las empresas hoy en día es muy importante su capacidad para satisfacer las expectativas de los clientes, es por ello por lo que el uso de la métrica de Net Promoter Score (NPS) paso hacer muy común para medir la lealtad de los clientes y su inclinación a recomendar la empresa, conforme a lo anterior se ha venido incrementando el uso de técnicas de análisis de datos y Machine Learning que permiten explorar los motivos de insatisfacción de los usuarios en las empresas, a continuación se detalla información relevante mencionada por diferentes autores luego de implementar algoritmos predictivos con problemáticas similares a las mencionabas en el actual trabajo.

Por ejemplo, Josefina, D. V. M. (2021) en su tesis de maestría “Modelo predictivo de detractores en casos de customer service” hace uso de los modelos Aprendizaje supervisado, Modelo de regresión logística LOGIT, Modelo de Naïve Bayes, Modelo de Random Forest donde tiene como resultado dar respuesta a ¿Cómo hacer que las decisiones de negocio que toman los managers de customer service tengan un enfoque proactivo y no reactivo? Indicando que los modelos predictivos pueden anticiparse a comportamientos futuros, por ejemplo, dentro de las oportunidades encontradas se dio cuenta que falta de conocimiento de los productos en los representantes que atienden las llamadas.

Otro dato importante, mencionado por Aitor, Z. G. (2021, Julio 1) en su tesis de pregrado “Clustering y Analítica de clientes de SEMIC mediante Machine Learning” fue que por medio de la implementación de cluster, que permiten agrupar la información logró agrupar los clientes según su similitud con las condiciones que utiliza para medir la satisfacción del cliente y adiciona visualizar las tendencias que ayudan a detectar los grupos de clientes que están muy

contentos y muy descontentos, con gráficos y filtros interesantes que le permiten explorar los datos de diferentes maneras.

Similar a lo que mencionan los autores Mandrai, Rahul;Sharma, Parth;Borkakaty, Bidisha en su artículo “Customer Risk Prediction: A Machine Learning Ensemble Approach” donde implementaron varios modelos entre ellos Clustering y Random Forest Classifier, teniendo como resultado identificar la importancia relativa de cada característica en la clasificación de riesgos de insatisfacción, la creación de clusters mediante K-Means++ facilitó la segmentación de datos, permitiendo la asignación de clases de riesgo y anticipar el riesgo de deserción de los clientes, adicional, identificar oportunidades de ventas adicionales y optimizar las recomendaciones.

Marco Conceptual

- KPI’s: viene de la sigla en inglés para Key Performance Indicator, o sea, Indicador Clave de Actuación. Es una forma de medir si una acción o un conjunto de iniciativas están efectivamente atendiendo a los objetivos sugeridos por la organización. Coutinho, V. (2021).
- NPS: El NPS es un sistema y un indicador para medir la satisfacción del cliente y también medir la lealtad. el NPS tiene un único objetivo: descubrir la probabilidad de que una persona recomiende una marca, una empresa, un producto o un servicio a otra persona. Ferreira & Ferreira. (2022).
- Machine Learning: El machine learning (ML) es una rama de la inteligencia artificial (IA) y la informática que se centra en el uso de datos y algoritmos para permitir que la IA imite la forma en que los humanos aprenden, mejorando gradualmente su precisión. IBM. (s. f.).

Marco Teórico

El marco teórico de este proyecto se centra en la aplicación de técnicas de machine learning para la identificación de causas de insatisfacción del cliente en una empresa de ecommerce. Se presentan los fundamentos teóricos de las técnicas seleccionadas.

Regresión Logística

Definición: La regresión logística es una técnica de análisis de datos que utiliza las matemáticas para encontrar las relaciones entre dos factores de datos. Luego, utiliza esta relación para predecir el valor de uno de esos factores basándose en el otro. AWS. (s. f.).

Fundamentos Teóricos:

Función Logística: La probabilidad de la clase positiva $\big(P(Y=1)\big)$ se modela mediante la función sigmoide: $P(Y=1 \mid X) = \frac{1}{1+e^{(\beta 0+\beta 1X1+\cdots+\beta nXn)}}$. donde $\beta 0$ es el intercepto y β1,…,βn son los coeficientes del modelo.

Estimación de Parámetros: Los coeficientes del modelo se ajustan utilizando el método de Máxima Verosimilitud, que maximiza la probabilidad de observar los datos dados los parámetros del modelo.

Función de Pérdida: Se minimiza la función de pérdida logarítmica (Log-Loss), definida como:

$$L(\beta) = -i = 1\sum n[yilogy\text{^}i + (1-yi)log(1-y\text{^}i)]$$

Aplicación: Se utiliza para identificar los factores que influyen en la satisfacción o insatisfacción del cliente, evaluando su impacto mediante los coeficientes β.

Random Forest

Definición: Es un método de conjunto que utiliza árboles de decisión como sus aprendices base, este método utiliza el bagging para crear conjuntos de trenes para sus alumnos

base. En cada nodo, cada árbol considera solo un subconjunto de las entidades disponibles y calcula la combinación óptima de entidad/punto de división. George Kyriakides, & Konstantinos G. Margaritis. (2019).

Fundamentos Teóricos:

Random Forest: Cada árbol en el bosque es entrenado con una muestra diferente del conjunto de datos mediante el método de Random Forest.

Importancia de Variables: El modelo proporciona medidas de importancia de las características, ayudando a identificar las variables más relevantes.

Aplicación: Es eficaz para manejar grandes conjuntos de datos con muchas características y relaciones complejas.

Gradient Boosting Machines (GBM)

Definición: Es otro algoritmo de boosting, que utiliza árboles de decisión de diferentes profundidades. George Kyriakides, & Konstantinos G. Margaritis. (2019), produce un modelo predictivo en forma de un conjunto de modelos de predicción débiles, típicamente árboles de decisión.

Fundamentos Teóricos:

Boosting: Combina predicciones de modelos simples para formar un modelo fuerte.

Función de Pérdida: Optimiza una función de pérdida mediante el ajuste secuencial de modelos.

Aplicación: Es útil para problemas de clasificación y regresión con alta precisión.

Support Vector Machines (SVM)

Definición: es un algoritmo de machine learning que permite encontrar un hiperplano que separe de la mejor forma posible dos clases diferentes de puntos de datos. “De la mejor forma

posible" implica el hiperplano con el margen más amplio entre las dos clases, representado por los signos más y menos. (SVM). (s. f.).

Fundamentos Teóricos:

Máximo Margen: Encuentra el hiperplano que separa las clases con el mayor margen posible.

Funciones Kernel: Permiten transformar los datos en un espacio de mayor dimensión para hacerlos separables linealmente.

Aplicación: Eficaz en espacios de alta dimensión y problemas de clasificación binaria y multiclase.

SMOTE (Synthetic Minority Over sampling Technique)

Definición: es una técnica estadística de sobremuestreo de minorías sintéticas para aumentar el número de casos de un conjunto de datos de forma equilibrada. El componente funciona cuando genera nuevas instancias a partir de casos minoritarios existentes que se proporcionan como entrada. Likebupt (2024), es una técnica muy útil para manejo de datos desbalanceados, donde hay alguna clase con muchos más datos vs otra clase.

Fundamentos Teóricos:

Método matemático: Dado un conjunto de datos con una clase minoritaria X_m para cada punto $x_i \in X_m$:

Se seleccionan k vecinos más cercanos dentro de la misma clase usando la distancia euclidiana:

$$d(xi,xj)= \sqrt{\sum_{n=1}^{N}(x_{in} - x_{jn})^2}$$

Se elige aleatoriamente uno de estos vecinos x_j

Se genera un nuevo punto sintético x_{new} mediante interpolación:

$$x_{new} = x_i + \lambda(x_j - x_i), \lambda \sim U(0,1)$$

donde U(0,1) es una variable aleatoria uniforme en el intervalo [0,1]

Este proceso se repite hasta alcanzar el nivel deseado de sobremuestreo.

Aplicación: SMOTE es ampliamente usado en problemas de clasificación desbalanceada en áreas como medicina, fraude financiero y reconocimiento de patrones. Sin embargo, puede inducir sobreajuste si se combina con un sobremuestreo excesivo y no aborda completamente problemas de solapamiento entre clases.

Metodología

Tabla 1

Metodología por Fase

Fase	Objetivo	Metodología (actividades por cada objetivo)	Productos / Resultados
Fase de Comprensión del Negocio	Reunir y entender los objetivos del negocio y los requisitos específicos del proyecto.	Reuniones con stakeholders para recoger requisitos. Definición de los KPIs (Key Performance Indicators) que se utilizarán para evaluar el éxito del proyecto. Documentación de los requisitos y expectativas del proyecto.	Se identifico las necesidades a trabajar con los modelos de clasificación, priorizando el seguimiento y control de la métrica de NPS como insumo principal para entender las causas de insatisfacción.
Fase de Comprensión y Preparación de los Datos	Entender y preparar los datos disponibles con el fin de garantizar uso de información relevante	Recolección de datos históricos de NPS. Análisis exploratorio de datos (EDA) para entender la estructura, calidad y características de los datos. Limpieza y preprocesamiento de los datos (manejo de valores faltantes, normalización, etc.).	Se obtuvo la base de datos a utilizar durante el desarrollo del proyecto aplicado, con el EDA y limpieza de datos, se pasó de tener 23 columnas y 23000 filas, a tener 18 columnas y 10062 filas

Fase de Modelado	Desarrollar y entrenar modelos de clasificación.	División de los datos en conjuntos de entrenamiento y prueba. Selección de algoritmos de machine learning para este caso: regresión logística, Random Forest, GBM y SVM. Entrenamiento de modelos y ajuste de hiperparámetros. Evaluación de los modelos con métricas apropiadas (Accuracy, recall y F1 score). Una vez obtenidos los primeros resultados de las métricas se identificó que se presentaba desequilibrio en las clases, a partir de ello se implementó la herramienta SMOTE (Synthetic Minority Over-sampling Technique)	Matriz de métricas desempeño para cada modelo (regresión logística, Random Forest, GBM y SVM). Matriz de métricas desempeño para cada modelo (regresión logística, Random Forest, GBM y SVM) con SMOTE

Nota. Esta tabla muestra las actividades realizadas en función de cada fase y objetivo propuesto para el desarrollo del proyecto aplicado.

Resultados

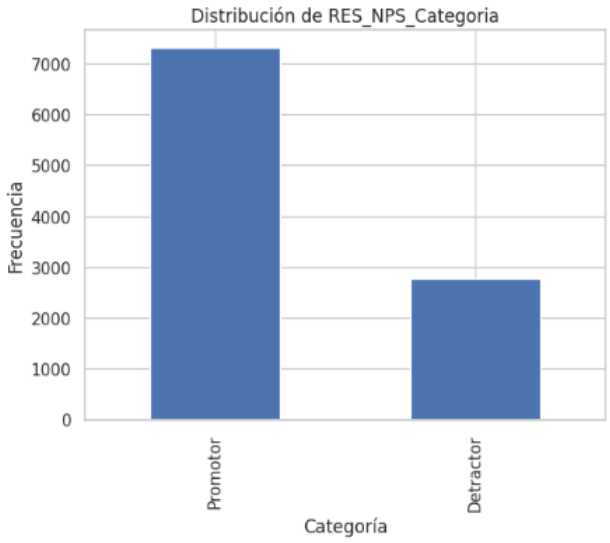
El análisis exploratorio de la información del NPS correspondiente al segundo trimestre de 2024 permitió examinar diversas variables que pueden influir en la satisfacción del usuario, tales como el medio de contacto (chat, llamada telefónica y correo electrónico), el proceso de atención, el país de origen del usuario, el representante que gestiono el caso y el equipo de trabajo que brindó la atención. A partir de este análisis, se identificó que las variables de país de origen y proceso de atención tenían mayor relevancia para comprender las razones de insatisfacción del cliente, ya que presentan patrones más marcados en la variabilidad del NPS. Con base en estos hallazgos, se decidió priorizar estas variables en la implementación de modelos de clasificación, con el objetivo de identificar factores clave asociados a la experiencia del cliente y mejorar la calidad del servicio.

Resultado 1

Dentro del análisis exploratorio se encontró la siguiente información.

Figura 1

Distribución de Promotores y Detractores



Se puede observar en la figura 1 que, en el conjunto de datos utilizado, la mayoría de las encuestas están en la categoría de promotor, esto quiere decir que, en términos generales, que la experiencia de los usuarios al momento de comunicarse con el servicio al cliente de la empresa en su mayoría es positiva.

Tabla 2

Datos de NPS por País del Usuario

País del usuario	% NPS	Numero de encuestas
MLA - Argentina	44.45%	7633
MLM - México	41.07%	1439
MLC - Chile	13.25%	513
MCO - Colombia	11.11%	477

*Nota***.** Esta tabla muestra el país desde donde se comunica el usuario, el porcentaje promedio de NPS y la cantidad de encuestas obtenidas.

En la empresa la variable más usada para entender como está siendo la experiencia del usuario es a nivel país, a partir de ello se tomó la decisión de revisar el NPS de esta categoría, en donde se evidencia que hay mayor oportunidad en mejorar la experiencia es en MLC y MCO, que hace referencia a los países de Chile y Colombia respectivamente; sin embargo, algo no menor por mencionar es que a nivel participación, los países que más se contactan es MLA y MLM, que hacen referencia a Argentina y México, y si bien son los países donde hay un NPS más alto (promedio de 43%) vs los demás, también hay ventana a la mejora de la experiencia en los usuarios, ya que el objetivo es tener un NPS cercano a 100%.

Tabla 3

Datos de NPS por Proceso

Proceso	% NPS	Numero de encuestas
Denuncia de fraude	57.65%	3770
Seguridad de cuenta	38.70%	3630
Sospecha ATO	31.52%	1525
Chargeback Comprador	18.81%	809
Restricciones	3.35%	328

Nota. Esta tabla muestra el proceso por el cual el usuario se contacta, el porcentaje promedio de NPS y la cantidad de encuestas obtenidas.

De acuerdo con los resultados obtenidos de NPS a nivel proceso, se logra identificar que los procesos con mayor participación son Denuncia de fraude y Seguridad de cuenta, sin embargo, son los proceso con mejor % de satisfacción según los usuarios, en cuanto a los procesos Chargeback Comprador y Restricciones si bien es donde hay mayor insatisfacción de os usuario, la participación en cuanto a contactos es pequeña.

Resultado 2

Luego de la implementación y entrenamiento de los modelos de clasificación se obtuvieron los siguientes resultados:

Figura 2

Matriz de Métricas de Desempeño para Regresión logística

Regresión Logística con SMOTE:

	precision	recall	f1-score	support
Detractor	0.38	0.48	0.42	844
Promotor	0.77	0.69	0.73	2175
accuracy			0.63	3019
macro avg	0.58	0.59	0.58	3019
weighted avg	0.66	0.63	0.64	3019

Nota. Esta tabla muestra los resultados de las métricas de desempeño para el modelo de regresión logística.

Teniendo en cuenta la información de métricas de desempeño disponible en tabla 4, se puede inferir que:

Detractor

- Accuracy (0,38): Indica que solo el 38% de las predicciones como "Detractor" fueron correctas, a partir de ello se puede decir que hay una alta tasa de falsos positivos para esta clase.
- Recall (0,48): Solo se identificaron correctamente el 48% de los "Detractores" reales. siendo una baja capacidad del modelo para capturar a todos los detractores.
- F1 Score (0,42): Refleja un equilibrio bajo entre la precisión y el recuerdo. El modelo no está logrando clasificar bien a los detractores.

Promotor

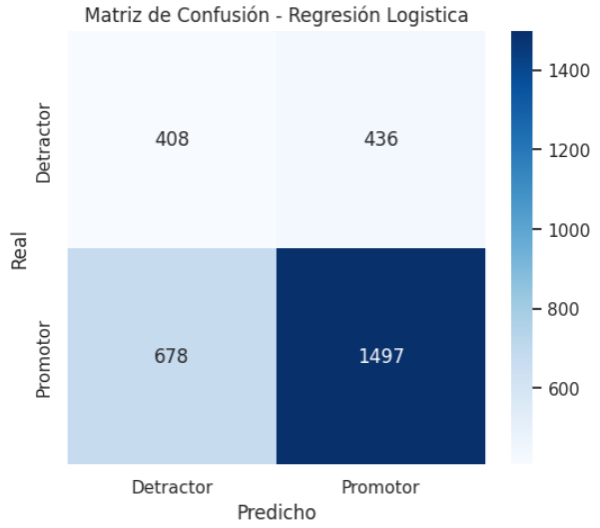
- Accuracy (0,77): El 77% de las predicciones como "Promotor" fueron correctas, siendo una buena precisión para esta clase.
- Recall (0,69): El modelo captura el 69% de los "Promotores" reales, aunque aceptable, aún hay margen para mejorar.
- F1 Score (0,73): Indica un buen equilibrio entre precisión y recuperación

Adicional, en el modelo de regresión logística se obtuvo los siguientes coeficientes: [-0.4527082, 0.38689236, -0.62752576, 0.27115695, -0.12982828, 0.12420174, -0.69670765, 0.70493789, -0.42478835], teniendo en cuenta que con este modelo se evalúa como el país del usuario y el proceso por el cual se comunica influyen en si un cliente es un promotor o no en función del NPS, lo que se estaría viendo en los coeficientes positivos obtenidos es que hay países y procesos que tienden a tener mayor probabilidad de que la experiencia del usuario sea buena, por ejemplo a nivel país Argentina o México y en cuanto a proceso “Seguridad en cuenta” o “Denuncia de fraude”, y en cuanto a los coeficientes negativos lo que se puede influir es que hay países y procesos donde la probabilidad de que la experiencia sea mala aumenta por ejemplo Chile y Colombia en cuanto países y del lado de procesos “Restricciones” y “Sospecha de ATO”

En cuanto al intercepto del modelo fue -0.43157757, haciendo referencia a la probabilidad de que el usuario tenga una buena experiencia y el caso de que las variables de países y proceso sean cero.

Figura 3

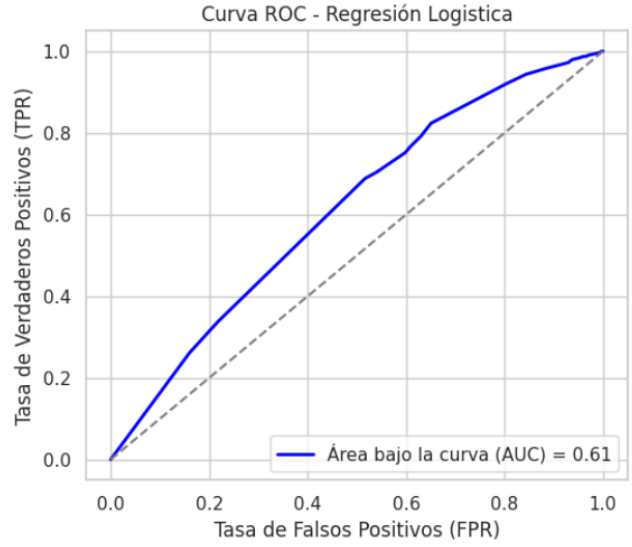
Matriz de Confusión para Modelo de Regresión Logística



En la Figura 2 se observa que el modelo tiene un buen desempeño al predecir la categoría "Promotor", con 1,497 verdaderos positivos. Sin embargo, la presencia de 678 falsos negativos sugiere que varios "Promotores" fueron clasificados erróneamente como "Detractores".

Figura 4

Curva ROC para Modelo de Regresión Logística



Teniendo en cuenta que las curvas ROC constituyen una herramienta importante para evaluar el rendimiento de un modelo de machine learning. (Curvas ROC, s. f.), para este caso se evidencia en la figura 3 que la curva de ROC del modelo de Regresión Logística presenta un AUC de 0.61, lo que indica un desempeño bajo en la clasificación. Dado que es un modelo lineal, es posible que no esté capturando adecuadamente las relaciones complejas entre las variables.

Figura 5

Matriz de Métricas de Desempeño para Random Forest

Random Forest con SMOTE:

	precision	recall	f1-score	support
Detractor	0.38	0.44	0.41	844
Promotor	0.77	0.72	0.74	2175
accuracy			0.64	3019
macro avg	0.58	0.58	0.58	3019
weighted avg	0.66	0.64	0.65	3019

Nota. Esta tabla muestra los resultados de las métricas de desempeño para el modelo de random forest.

Teniendo en cuenta la información de métricas de desempeño disponible en tabla 5, se puede inferir que:

Detractor

- Accuracy (0,38): Indica que solo el 38% de las predicciones como "Detractor" fueron correctas, a partir de ello se puede decir que hay una alta tasa de falsos positivos para esta clase.
- Recall (0,44): Solo se identificaron correctamente el 44% de los "Detractores" reales. siendo una baja capacidad del modelo para capturar a todos los detractores.

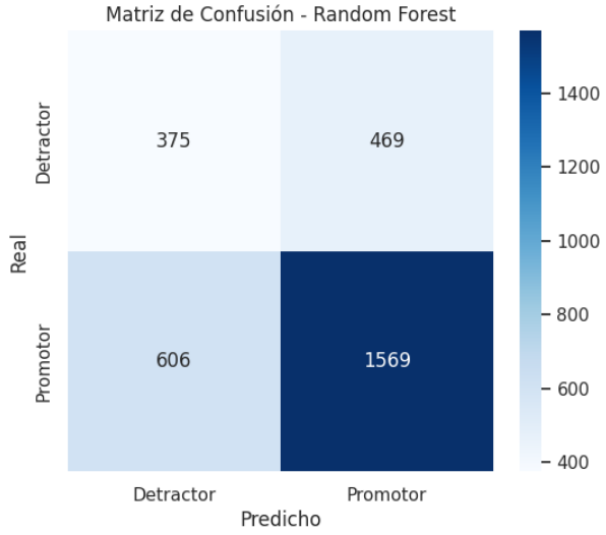
- F1 Score (0,41): Refleja un equilibrio bajo entre la precisión y el recuerdo. El modelo no está logrando clasificar bien a los detractores.

Promotor

- Accuracy (0,77): El 77% de las predicciones como "Promotor" fueron correctas, siendo una buena precisión para esta clase.
- Recall (0,72): El modelo captura el 72% de los "Promotores" reales, aunque aceptable, aún hay margen para mejorar, genera mejores resultados vs el modelo de regresión logística.
- F1 Score (0,74): Indica un buen equilibrio entre precisión y recuperación para esta clase.

Figura 6

Matriz de Confusión para Modelo de Random Forest

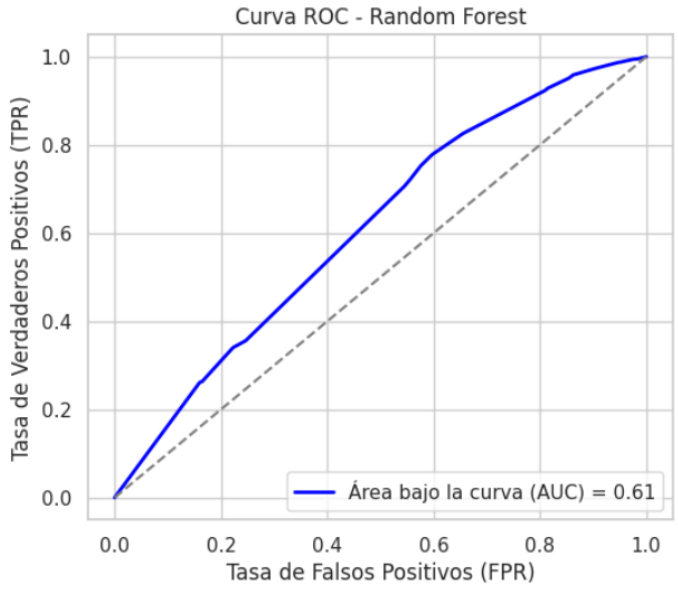


En la Figura 4 se observa que el modelo Random Forest tiene un desempeño similar al de regresión logística, se puede decir que también tiene un buen desempeño al predecir la categoría

"Promotor", con 1,569 verdaderos positivos. Sin embargo, la presencia de 606 falsos negativos sugiere que varios "Promotores" fueron clasificados erróneamente como "Detractores".

Figura 7

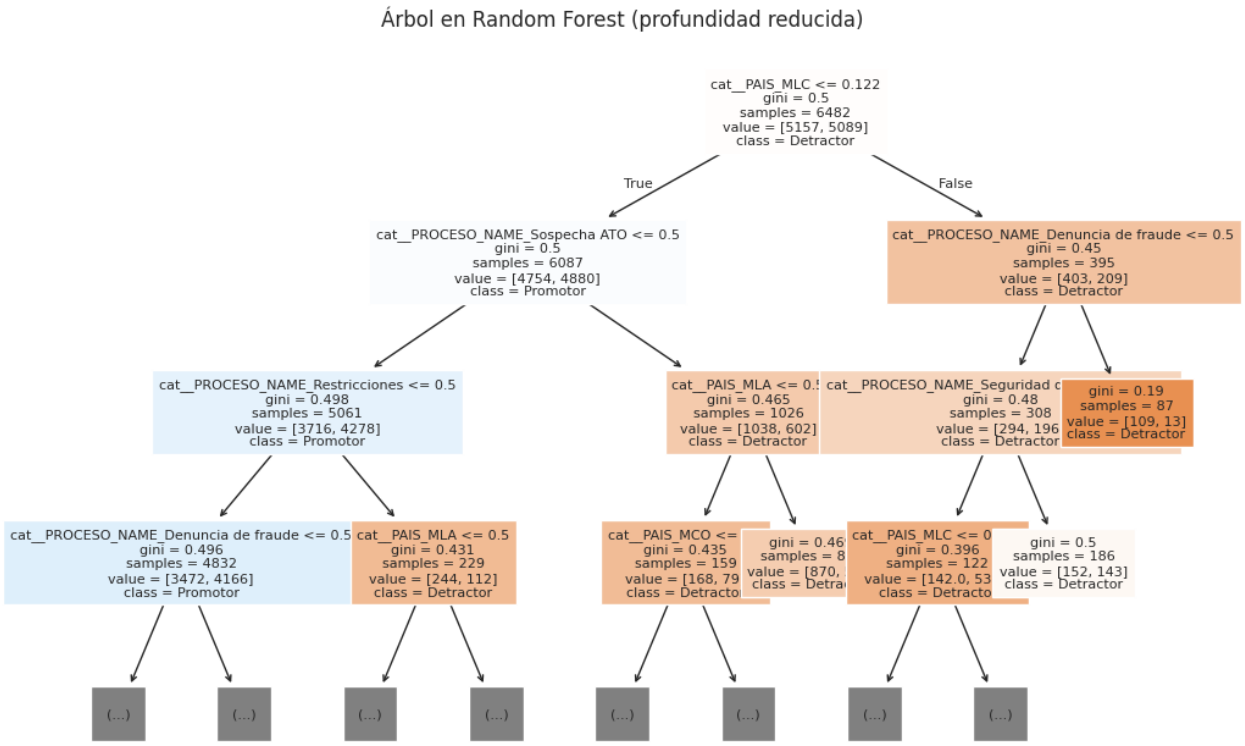
Curva ROC para Modelo de Random Forest



En la figura 5 se observa que la curva ROC para el modelo de Random Forest, muestra que a pesar de su capacidad para manejar relaciones no lineales y capturar patrones más complejos, no muestra una mejora significativa respecto a la regresión logística, con un AUC también de 0.61.

Figura 8

Árbol para Modelo de Random Forest



En la figura 6 se observa cómo el árbol representa el modo en que el modelo utiliza características específicas para clasificar entre "Promotor" y "Detractor", para este caso se muestra solo una parte del árbol debido a que es un árbol muy complejo y puede afectar la visualización de la información.

Adicional, las variables independientes cruciales para la clasificación son:

- cat_PROCESO_NAME_Sospecha ATO: Es el nodo raíz del árbol, por lo que es una variable clave para la división inicial.
- cat_PAIS_MLC: Aparece en una división importante después de la variable raíz, sugiriendo su relevancia.

- cat_PROCESO_NAME_Denuncia de fraude: Se utiliza en varias ramas del árbol, indicando su importancia en las decisiones del modelo.

Figura 9

Matriz de Métricas de Desempeño para Gradient Boosting Machines (GBM)

GBM:

	precision	recall	f1-score	support
Detractor	0.38	0.45	0.41	844
Promotor	0.77	0.71	0.74	2175
accuracy			0.64	3019
macro avg	0.57	0.58	0.57	3019
weighted avg	0.66	0.64	0.65	3019

Nota. Esta tabla muestra los resultados de las métricas de desempeño para el modelo de Gradient Boosting Machines (GBM).

Teniendo en cuenta la información de métricas de desempeño disponible en tabla 6, se puede inferir que.

Detractor

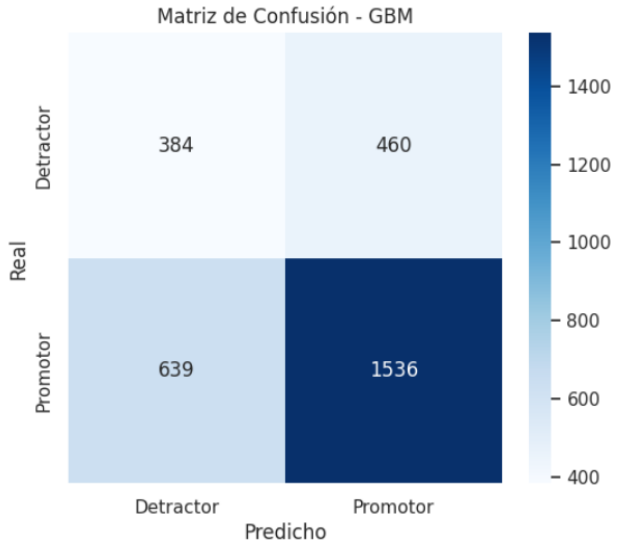
- Accuracy (0,38): Indica que solo el 38% de las predicciones como "Detractor" fueron correctas, a partir de ello se puede decir que hay una alta tasa de falsos positivos para esta clase.
- Recall (0,45): Solo se identificaron correctamente el 45% de los "Detractores" reales. siendo una baja capacidad del modelo para capturar a todos los detractores.
- F1 Score (0,41): Refleja un equilibrio bajo entre la precisión y el recuerdo. El modelo no está logrando clasificar bien a los detractores.

Promotor

- Accuracy (0,77): El 77% de las predicciones como "Promotor" fueron correctas, siendo una buena precisión para esta clase.
- Recall (0,71): El modelo captura el 71% de los "Promotores" reales, aunque aceptable, aún hay margen para mejorar, genera mejores resultados vs el modelo de regresión logística.
- F1 Score (0,74): Indica un buen equilibrio entre precisión y recuperación para esta clase.

Figura 10

Matriz de Confusión para Modelo de Gradient Boosting Machines (GBM)

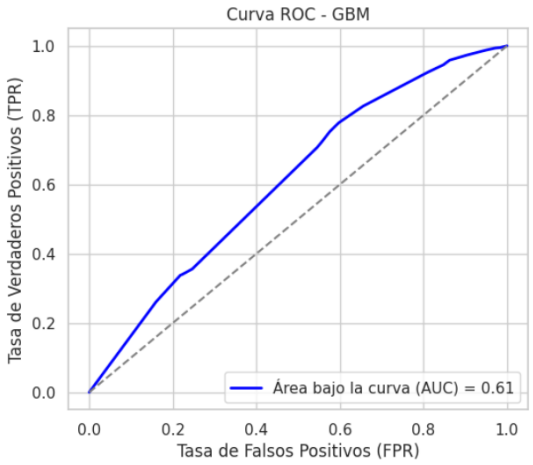


En la Figura 7 se mantiene comportamientos similares a los modelos pasados, donde GBM también tiene buen desempeño al predecir la categoría "Promotor", con 1,536 verdaderos positivos, pero se mantiene la oportunidad para predecir correctamente “Detractores” ya que hay

presencia de 639 falsos negativos sugiere que varios "Promotores" fueron clasificados erróneamente como "Detractores".

Figura 11

Curva ROC para Modelo de Gradient Boosting Machines (GBM)



En la Figura 8 se presenta la curva ROC del modelo Gradient Boosting Machines (GBM), la cual muestra un comportamiento similar al de los modelos previos. El área bajo la curva (AUC) indica que el modelo no logra una buena capacidad de discriminación entre clases, lo que sugiere que su desempeño es limitado.

Figura 12

Matriz de Métricas de Desempeño para Support Vector Machines (SVM)

SVM con SMOTE:

	precision	recall	f1-score	support
Detractor	0.38	0.44	0.41	844
Promotor	0.77	0.72	0.74	2175
accuracy			0.64	3019
macro avg	0.58	0.58	0.58	3019
weighted avg	0.66	0.64	0.65	3019

Nota. Esta tabla muestra los resultados de las métricas de desempeño para el modelo de Support Vector Machines (SVM).

Teniendo en cuenta la información de métricas de desempeño disponible en tabla 7, se puede inferir que:

Detractor

- Accuracy (0,38): Indica que solo el 38% de las predicciones como "Detractor" fueron correctas, a partir de ello se puede decir que hay una alta tasa de falsos positivos para esta clase.
- Recall (0,44): Solo se identificaron correctamente el 44% de los "Detractores" reales. siendo una baja capacidad del modelo para capturar a todos los detractores.
- F1 Score (0,41): Refleja un equilibrio bajo entre la precisión y el recuerdo. El modelo no está logrando clasificar bien a los detractores.

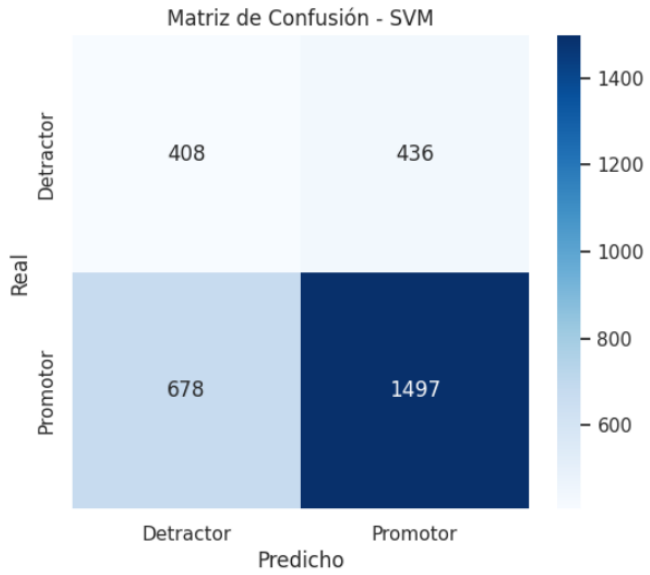
Promotor

- Accuracy (0,77): El 77% de las predicciones como "Promotor" fueron correctas, siendo una buena precisión para esta clase.

- Recall (0,72): El modelo captura el 72% de los "Promotores" reales, aunque aceptable, aún hay margen para mejorar, genera mejores resultados vs el modelo de regresión logística.
- F1 Score (0,74): Indica un buen equilibrio entre precisión y recuperación para esta clase.

Figura 13

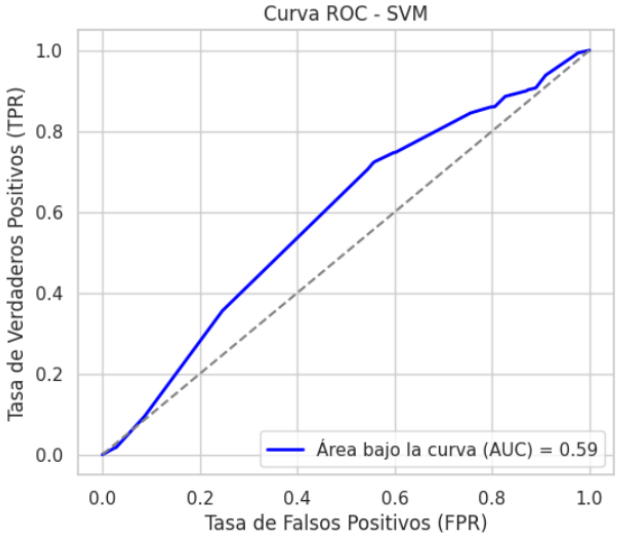
Matriz de Confusión para Modelo de Support Vector Machines (SVM)



Al comparar las matrices de confusión de modelo a modelo se puede ver que la variación entre ellos es mínima, por ejemplo, en la figura 9, para el modelo de SVM se mantiene comportamiento de predecir correctamente los casos de "Promotor" con 1497 verdaderos positivos, pero tiene muchos falsos negativos (678), lo cual sugiere que algunos "Promotores" fueron clasificados incorrectamente como "Detractores", incluso los resultados de SVM y regresión logística fueron prácticamente iguales.

Figura 14

Curva ROC para Modelo de Support Vector Machines (SVM)



En la Figura 10 se presenta la curva ROC del modelo Support Vector Machines (SVM), el cual obtuvo un AUC de 0.59, evidenciando un desempeño inferior al de los modelos previos. Este resultado sugiere que el modelo no logra una separación efectiva entre las clases, lo que puede deberse a una selección inadecuada del kernel o la falta de datos representativos.

Conclusiones

El análisis exploratorio reveló patrones claros en la satisfacción de los clientes según las variables de país, proceso, equipo y medio de contacto. Uno de los hallazgos más importantes fue la variación significativa del NPS entre diferentes nacionalidades de los usuarios. En los usuarios argentinos se registra el NPS más alto (44,4%); el servicio al cliente parece alinearse con las expectativas de los usuarios de este país. Sin embargo, en los usuarios de Chile y Colombia, se observa un NPS mucho más bajo (13%). Este resultado podría estar relacionado con el hecho de que los representantes de servicio al cliente son de Argentina y las diferencias culturales, incluyendo el tono, el lenguaje y el estilo de comunicación, pueden estar afectando la experiencia y gestión de casos.

La implementación de modelos de machine learning para el análisis de datos de NPS ha proporcionado a la empresa una herramienta importante para identificar y comprender las causas detrás de la insatisfacción de los clientes. A través del uso de diferentes modelos de clasificación es posible evaluar cómo variables específicas, como el proceso por el cual se comunica el usuario, el país de donde es el usuario o el equipo que atiende a su consulta, influyen en la experiencia y las expectativas del usuario. Este conocimiento detallado y específico sobre las fuentes de insatisfacción ofrece a la empresa la capacidad de adoptar medidas proactivas y personalizadas para mejorar la experiencia del cliente.

El desarrollo de este proyecto aplicado ha permitido comprender cómo el desequilibrio de clases afecta el rendimiento de los modelos de clasificación. En casos donde no se emplean herramientas como SMOTE (Synthetic Minority Over-sampling Technique) para gestionar datasets desbalanceados, los modelos enfrentan dificultades al clasificar las clases con menor

cantidad de información. Esto sucede porque el modelo tiene menos oportunidades para aprender y reconocer las características distintivas de estas clases.

La evaluación de los modelos de machine learning utilizados para clasificar la experiencia del cliente mostró un buen desempeño al identificar promotores, destacándose Random Forest y GBM con precisiones y F1 Scores superiores al 70%. Sin embargo, todos los modelos tuvieron dificultades significativas al clasificar a los detractores, con precisiones inferiores al 40% y un bajo F1 Score (~41%), lo que evidencia una alta tasa de falsos positivos y una capacidad limitada para identificar clientes insatisfechos. El no haber alcanzado el umbral del 80% en las métricas de precisión, sensibilidad y balance se debe principalmente al desbalance en las clases dentro del conjunto de datos, con una mayor proporción de promotores en comparación con detractores. Este desbalance, si bien refleja un aspecto positivo al indicar que la mayoría de los usuarios están satisfechos con la atención recibida, afectó la capacidad de los modelos para aprender patrones representativos de los clientes insatisfechos. Además, la limitación en la cantidad y diversidad de variables consideradas en la modelación pudo haber restringido la capacidad predictiva, reduciendo la efectividad de los modelos para clasificar correctamente todas las categorías de usuarios.

Recomendaciones

Para la elaboración de trabajos futuros se recomienda hacer uso de la variable de comentarios, donde registra la información dada por el usuario mencionando el porqué de su calificación en la encuesta de NPS, ya que es una variable que puede agregar valor en el análisis y entendimiento de las principales causas de insatisfacción de los usuarios y si esto va ligado a las variables específicas para la aplicación de los modelos de machine learning.

Se recomienda incorporar el análisis de series de tiempo en futuros estudios, ya que este enfoque puede aportar valor en identificar tendencias, patrones estacionales y variación en distintos periodos de tiempo, que justifiquen el porqué de la insatisfacción de los usuarios, presentes en contactos de servicio al cliente.

Se recomienda a la organización implementar dinámicas de seguimiento del NPS, apoyadas por modelos de machine learning, para clasificar las causas de satisfacción e insatisfacción relacionada a variables puntuales. Esta metodología permitirá diseñar planes de acción orientados a mejorar la experiencia del usuario. Por ejemplo, a partir de la información disponible sobre el NPS por proceso, se podrán identificar aquellos procesos con mayores oportunidades de mejora, analizar qué factores están afectando estos procesos, destacar las buenas prácticas que puedan ser replicadas y reconocer acciones que deben evitarse para no perjudicar al usuario.

Referencias Bibliográficas

Aitor, Z. G. (2021, 1 de julio). *Clustering y Analítica de clientes de SEMIC mediante Machine Learning*. https://repositori.udl.cat/items/2c792094-5479-4b10-b5be-f5c6af3b6d0c

Amazon Web Services, Inc. (s. f.). ¿Qué es la regresión logística? - Explicación del modelo de regresión logística - AWS. https://aws.amazon.com/es/what-is/logistic-regression/#:~:text=La%20regresi%C3%B3n%20log%C3%ADstica%20es%20una,factores%20bas%C3%A1ndose%20en%20el%20otro.

Bonaccorso, G. (2018). *Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning* (2nd ed.). Packt Publishing.

Boschetti, A., & Massaron, L. (2016). *Python Data Science Essentials - Second Edition* (Vol. 2). Packt Publishing.

Coutinho, V. (2021, 12 de febrero). *KPIs: descubre qué son los indicadores clave de rendimiento y cómo usarlos para orientar tus estrategias*. Rock Content.

Ferreira, A. C. (2022, 7 de marzo). *Net Promoter Score (NPS): ¿qué es y cómo se calcula?* https://www.inboundcycle.com/blog-de-inbound-marketing/net-promoter-score-nps-que-es-y-como-se-calcula

Galea, A. (2018). *Applied Data Science with Python and Jupyter: Use Powerful Industry-standard Tools to Unlock New, Actionable Insights From Your Data* (1st ed.). Packt Publishing.

IBM. (s. f.). ¿Qué es machine learning (ML)? https://www.ibm.com/mx-es/topics/machine-learning

Josefina, D. V. M. (2021). *Modelo predictivo de detractores en casos de customer service*. Universidad Torcuato Di Tella. https://repositorio.utdt.edu/handle/20.500.13098/11863

Kane, F. (2017). *Hands-On Data Science and Python Machine Learning*. Packt Publishing.

Kyriakides, G., & Margaritis, K. G. (2019). *Hands-On Ensemble Learning with Python: Build Highly Optimized Ensemble Machine Learning Models Using Scikit-learn and Keras*. Packt Publishing.

Likebupt. (2024, 1 de septiembre). *SMOTE - Azure Machine Learning*. Microsoft Learn. https://learn.microsoft.com/es-es/azure/machine-learning/component-reference/smote?view=azureml-api-2

Madhavan, S. (2015). *Mastering Python for Data Science: Explore the World of Data Science Through Python and Learn How to Make Sense of Data*. Packt Publishing.

Mandrai, R., Sharma, P., & Borkakaty, B. (2023). *Customer Risk Prediction: A Machine Learning Ensemble Approach*. *2023 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, 1–6. https://doi.org/10.1109/ICECET58911.2023.10389208

MATLAB & Simulink. (s. f.). *Curvas ROC*. https://la.mathworks.com/discovery/roc-curve.html

MATLAB & Simulink. (s. f.). *Support Vector Machine (SVM)*. https://la.mathworks.com/discovery/support-vector-machine.html

Pineda, R. (2022). *Aplicación de la Ciencia de Datos - Líneas de Investigación (ECBTI) [OVI]*. https://repository.unad.edu.co/handle/10596/50443

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning - Second Edition* (2nd ed.). Packt Publishing.

Thakur, A. (2016). *Python: Real-World Data Science*. Packt Publishing.