Proyecto de investigación

Aplicación de modelos de aprendizaje supervisado para predicción del tipo de contacto de clientes asignados a un BPO de cobranza.

Presentado por:
Karen Andrea Carvajal Jaramillo

Especialización en estadística aplicada

Bogotá D.C, Junio 2022

Facultad de ingeniería y ciencias básicas

Proyecto de investigación para optar por el título de

Especialista en estadística aplicada

Aplicación de modelos de aprendizaje supervisado para predicción del tipo de contacto de clientes asignados a un BPO de cobranza

Karen Andrea Carvajal Jaramillo

Director (a)

Jesus Antonio Villarraga Palomino, M.Sc

Colombia, Bogotá D.C

2022

RESUMEN

El artículo presenta un ejercicio de investigación sobre la aplicación y comparación de 3 modelos de aprendizaje supervisado que corresponden a regresión logística, árboles de decisión y redes neuronales, para predecir el tipo de contacto de clientes asignados a un business process outsourcing de cobranza; la metodología empleada, está basada en la metodología CRISP-DM de la cual se ejecutan las fases de comprensión del negocio, comprensión de los datos, preparación de los datos, modelado y evaluación de resultados. Para lo cual se emplea análisis exploratorio de datos, por medio de estadística descriptiva, data mining con FAMD (análisis factorial de datos mixtos), en el cual se concluye que el modelo de árbol de decisión y red neuronal con función de activación logística, obtienen resultados muy similares en métricas de evaluación. Por otra parte la regresión logística tiene los resultados de evaluación más bajos en comparación de los otros modelos, sin embargo, las redes neuronales presentan menor estabilidad en validación cruzada frente a la regresión logística y el árbol de decisión.

Palabras clave: *aprendizaje supervisado, FAMD, data mining, árboles de decisión, regresión logística, redes neuronales.*

ABSTRACT

The paper presents a research exercise on the application and comparison of 3 supervised learning models that correspond to logistic regression, decision trees and neural networks, to predict the type of contact of clients assigned to a collection business process outsourcing; the methodology used is based on the CRISP-DM methodology from which the

business understanding, data understanding, data preparation, modeling and results evaluation phases are executed. For which exploratory data analysis is used, through descriptive statistics, data mining with FAMD (factorial analysis of mixed data), in which it is concluded that the decision tree model and neural network with logistic activation function, obtain very similar results in evaluation metrics. On the other hand, logistic regression has the lowest evaluation results compared to the other models, however, the neural networks present less stability in cross-validation compared to logistic regression and the decision tree.

Keywords: *Supervised learning, machine learning, FAMD, data mining, decision trees, logistic regression, neural networks.*

OBJETIVO GENERAL

Seleccionar el mejor modelo de predicción para la fase de despliegue según la metodología CRISP-DM de acuerdo a métricas de evaluación de 3 modelos de aprendizaje supervisado para predecir el tipo de contacto de clientes asignados a un BPO de cobranza.

OBJETIVOS ESPECÍFICOS

- Realizar análisis exploratorio de datos por medio de estadística descriptiva de los datos de clientes asignados a un BPO de cobranza para compresión de los datos y el negocio.

- Realizar data mining por medio de técnicas estadísticas multivariadas de los datos de los clientes asignados para gestión de cobro a BPO, con el fin de identificar patrones y variables con mayor relevancia para el resultado de recuperación y contacto.
- Entrenar los modelos seleccionados para la obtención y comparación de métricas de evaluación de cada modelo.

INTRODUCCIÓN

El aprendizaje supervisado es una rama del aprendizaje automático perteneciente a la inteligencia artificial, cuyo objetivo principal se basa en que la máquina programada aprenda la posible respuesta con base a la experiencia, para lo cual se asignan datos de entrenamiento y datos de prueba, sin que sea necesario programar explícitamente a la máquina para la generación de los outputs, ya que él algoritmo aprende de acuerdo a los datos proporcionados de entrada y salida.

Resulta entonces interesante resolver problemas de predicción de una clase cuando no se tiene claro cómo obtener el resultado esperado, el artículo presenta un ejercicio de aplicación y comparación de 3 modelos de aprendizaje supervisado, correspondientes a regresión logística, árbol de decisión y red neuronal, con fin de predecir si un cliente va a ser contactado o no, de acuerdo a variables cuantitativas y cualitativas dentro de las cuales se incluyen datos demográficos, proporcionados al BPO de cobranza, pues esta clasificación permite definir estrategias para los clientes que se predice el no contacto, el artículo presenta resultados de las métricas de evaluación de cada uno de los modelos, donde se evidencia que el árbol de decisión obtiene mejores resultados de accuracy,

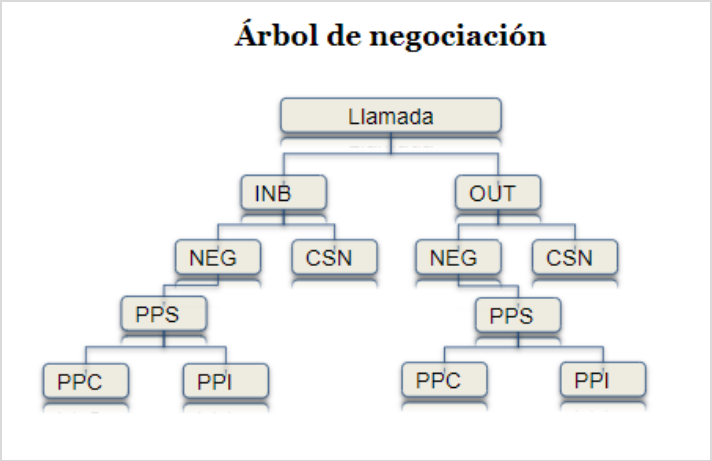
métricas por clase y validación cruzada, sin embargo la red neuronal obtiene resultados similares de desempeño en todas las métricas de evaluación a excepción de la validación cruzada donde se evidencia inestabilidad de la exactitud del modelo, por otra parte la regresión logística obtiene menores resultados de evaluación respecto a los demás modelos.

REFERENTES TEÓRICOS

De acuerdo a (BPO en la cobranza de una pyme | Microformas, s. f.) los BPO o business process outsourcing son compañías que se subcontratan para aliviar la carga administrativa de las empresas que utilizan esta alternativa, con el fin de centrarse por completo en su negocio, los BPO de cobranza se encargan específicamente de prestar el servicio de recuperación de cartera vencida de una compañía principalmente por medio de llamadas telefónicas para la obtención de una negociación con los clientes y con estos los pagos requeridos para que los clientes puedan salir del estado de morosidad.

El flujo de proceso del BPO está determinado en principio, por el árbol de negociación que inicia con la comunicación de una llamada telefónica, ésta puede ser tipo entrante (inbound) o saliente (outbound), de la cual se puede derivar un contacto con negociación o un contacto sin negociación, el efectivo se clasifica en el primer tipo , después esta acción puede evolucionar como; el cumplimiento o incumplimiento de la negociación realizada, conocida como promesas de pago, clasificándose como cumplida o incumplida como se muestra en la figura 1.

Figura 1. Árbol de negociación de un BPO de cobranza



Fuente: elaboración propia

La selección de un modelo de aprendizaje supervisado para despliegue, no solo depende de la métricas de evaluación de este, si no del objetivo de su aplicación, para lo cual resulta necesario un previo y correcto análisis sobre el entendimiento del negocio y los datos. (Vallalta ,2022)

Data Mining

El ejercicio de data mining basado en técnicas estadísticas multivariadas, es una forma eficiente de análisis exploratorio de datos para identificar patrones y relaciones entre individuos y variables, en grandes conjuntos de datos, adicionalmente permite reducir la dimensionalidad de estos para una mejor comprensión y aplicación de modelos de machine learning. (Pardo, 2015)

Aprendizaje automático

El aprendizaje automático (Machine learning) según Mahesh (2020), es el estudio científico de algoritmos y modelos estadísticos que utilizan los sistemas informáticos para realizar una tarea específica sin estar explícitamente programado.

Este cobra relevancia en la cuarta revolución industrial concepto acuñado por Klaus Schwab fundador del Foro Económico Mundial en el contexto de la edición del Foro Económico Mundial (2016), donde se habló sobre los avances tecnológicos que permiten la captura, recolección y procesamiento de datos, convirtiéndolos en un activo de relevancia para las compañías y cambiando el mundo que conocemos para dar paso a nuevos campos de estudio como la ciencia de datos, automatización de procesos, aprendizaje automático, entre otros.

Según Carbonell, Michalski & Mitchell, (1983) en la actualidad, instruir a una computadora o a un robot controlado por computadora para que realice una tarea requiere que se defina un algoritmo completo y correcto para esa tarea. y luego programar laboriosamente el algoritmo en una computadora.

Hastie, Tibshirani, Friedman, & Friedman, J. H. (2009). mencionan que la ciencia del aprendizaje juega un papel clave en los campos de la estadística, la minería de datos y la inteligencia artificial, intersectando áreas de ingeniería y otras disciplinas.

Resulta entonces interesante en investigación poder comprender los procesos de aprendizaje de los seres humanos, ya que estos pueden ser soluciones para la optimización de tareas repetitivas, como lo describen Hastie, Tibshirani, Friedman, & Friedman, (2009) que también sugieren que el enfoque de aprendizaje de la acción o tarea específica, ayudan a determinar la rentabilidad, compensaciones y limitaciones del aprendizaje.

El sistema de aprendizaje puede atribuir normas y descripciones de comportamiento de un objeto con base en la experiencia como los problemas de clasificación y muchos otros tipos de conocimiento útiles para el desempeño de tareas.

Aprendizaje Supervisado

El aprendizaje supervisado, es una tarea de aprendizaje automático en la cual la máquina aprende una función que asigna una entrada a una salida basada en ejemplos de pares de entrada-salida, para inferir el sistema de clasificación. Los datos incluyen un conjunto de ejemplos de entrenamiento. Los algoritmos de aprendizaje automático son aquellos que necesitan ayuda externa, ya que separan un conjunto de datos de entrada para entrenamiento y prueba que contendrán los datos de una variable resultante a predecir o categorizar. (Mahesh, 2020)

Árboles de decisión

Hastie, Tibshirani, Friedman, & Friedman, (2009) describen los árboles de decisión como sistemas que sirven para la diferenciación entre las categorías de características, ya que los nodos del árbol de decisión corresponden a las propiedades del objeto especificado y sus características corresponden a valores predeterminados que reemplazan estas propiedades.

Por otra parte, Quinlan, (1996), describe un árbol de decisión como un nodo hoja etiquetado con una clase o una estructura que consta de un nodo de prueba vinculado a dos o más subárboles. Una prueba nodo calcula algún resultado basado en los valores de atributo de una instancia, donde cada resultado posible está asociado con uno de los subárboles. Una instancia se clasifica comenzando por el nodo raíz del árbol. Si este nodo es una prueba, el

resultado para la instancia es determinado y el proceso continúa usando el subárbol apropiado. Cuando una eventualmente se encuentra una hoja, su etiqueta de la clase predicha de la instancia.

Regresión logística

La Biblioteca Nacional de Medicina (2020), define los modelos de regresión logística como "modelos estadísticos que describen la relación entre una variable dependiente cualitativa (es decir, una que puede tomar sólo ciertos valores discretos, como la presencia o ausencia de enfermedad) y una variable independiente".

La regresión logística se utiliza para obtener la razón de posibilidades cuando hay más de una variable explicativa. El procedimiento es bastante similar a la regresión lineal múltiple, excepto que la variable de respuesta es binomial. El resultado es el efecto de cada variable sobre la razón de posibilidades para el evento de interés observado. La principal ventaja es que se evitan los efectos de ruido cuando se analizan grupos de todas las variables juntas. (Sperandei, 2014).

Redes neuronales

Mahesh (2020), define las redes neuronales como "una serie de algoritmos que intentan aprender relaciones básicas en un conjunto de datos a través de un proceso que imita el funcionamiento del cerebro humano. En este sentido, las redes neuronales se refieren a sistemas neuronales, que son de naturaleza orgánica o artificial. Las redes neuronales pueden adaptarse a los cambios en las entradas; así que comprueba la mejor red y los posibles resultados sin tener que redefinir los criterios de salida."

Métricas de evaluación de modelos de aprendizaje supervisado de clasificación.

Accuracy: corresponde a "la precisión de la predicción del modelo, se puede definir como la relación entre la predicción correcta y el número total de instancias de entrada" Dridi, (2021).

$$Accuracy = \frac{\#\ de\ predicciones\ correctas}{Total\ de\ predicciones}$$

Precisión: La precisión se define como el número de resultados correctos, divididos por el número de acierto de las clases predichas por el modelo de predicción. Dridi, (2021).

$$Precisión = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ positivos}$$

Recall: "Se define como el número de resultados correctos divididos por todos muestras relevantes". Dridi, (2021).

$$Recall = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ negativos}$$

F1-Score: Es la media armónica entre la precisión y recall. Muestra la robustez del modelo de predicción. Dridi, (2021).

$$F1 - Score = \frac{2}{\frac{1}{Precisión} + \frac{1}{recall}}$$

METODOLOGÍA

La metodología empleada para el desarrollo del proyecto de investigación está basada en la metodología CRISP-DM representada en la figura 2, son las siglas de Cross-Industry

Standard Process for Data Mining. Consiste en un ciclo que comprende seis etapas descritas a continuación:

Fase 1. Comprensión del negocio

En esta fase inicial se desarrolló análisis exploratorio de los datos obtenidos con el fin de comprender las necesidades de aplicación de aprendizaje supervisado para la solución de uno de los problemas con mayor relevancia en el negocio del BPO de cobranza.

Fase 2. Comprensión de los datos

En esta fase se realizó análisis estadístico descriptivo multivariante para comprender y describir la relación entre las variables y los individuos con el fin de reducir la dimensionalidad de los datos para la fase de modelamiento.

Fase 3. Preparación de los datos

En esta fase se realiza la corrección de calidad y las transformaciones necesarias de los datos para la obtención del dataset final como son dumizar variables categóricas seleccionadas para poder realizar la etapa de modelamiento donde se obtiene un conjunto final de 43074 rows × 22 columns.

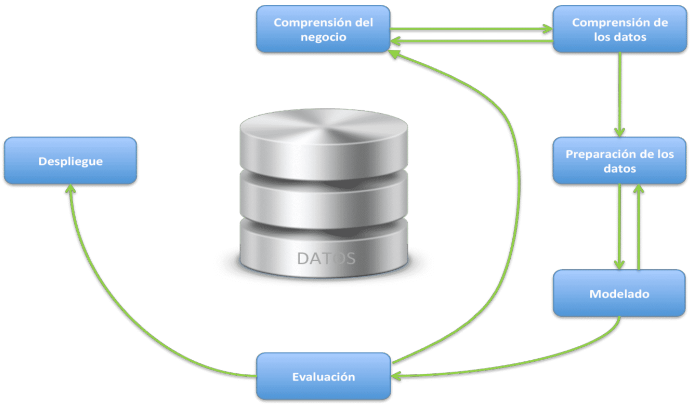
Fase 4. Modelado

En la fase de modelamiento se aplicaron 3 modelos de aprendizaje supervisado correspondientes a; árboles de decisión, regresión logística y redes neuronales con los datos finales de la etapa anterior y una segunda aplicación con variables escaladas.

Fase. 5 Evaluación

En la esta etapa se calcularon la métricas de evaluación de modelos clasificadores como son: accuracy, precision, recall, F1 score y la construcción de las matrices de confusión de cada uno de los modelos aplicados.

Figura 2. Metodología CRISP-DM



Fuente: (Vallalta Rueda, 2022)

RESULTADOS

Al realizar el análisis exploratorio de los datos se pueden describir algunas características relevantes correspondiente a las variables de género, edad, actividad laboral y saldo a capital, de los clientes asignados al BPO que permiten crear un perfilamiento de la población objetivo de contacto de acuerdo a la figura 3:

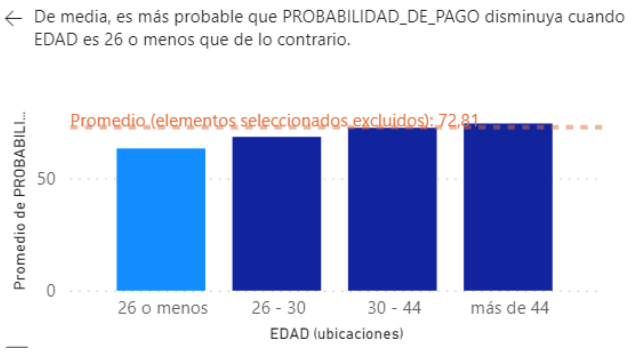
Figura 3. Distribución de de saldo a capital por profesión. Fuente: elaboración propia

donde el 57,61 % del capital asignado está concentrado principalmente en personas no profesionales, personas con carreras en administración de empresas, comerciantes, otras profesiones, profesionales en derecho y auxiliares de archivo.

Por otra parte, al analizar la variable de edades de los clientes asignado al BPO se encuentra que la participación de la generación millennials-Y es del 46,01% sobre la población, generación que corresponde a personas en rango de edad de 41-29 años.

Asimismo dentro del análisis exploratorio de los datos se evidencia que la media de la probabilidad de pago disminuye para personas menores de 26 años, representado en la figura 4.

Figura 4. distribución probabilidad de pago por edad. Fuente: elaboración propia



y las mujeres tienen una efectividad de recuperación más alta independiente de la actividad laboral que desempeñen.

Tabla 1. Efectividad de recuperación de capital por género y actividad laboral

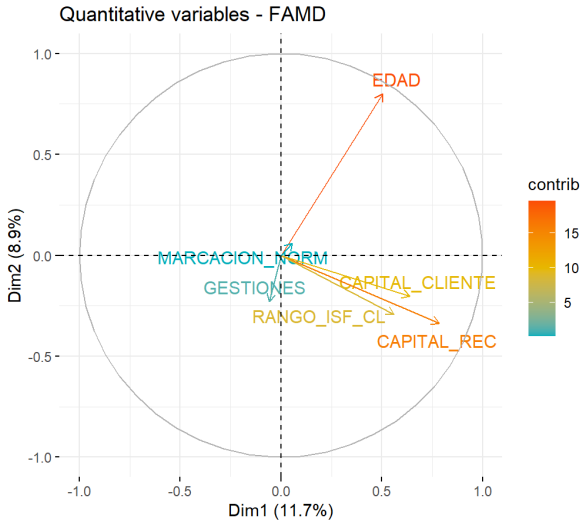
	Empleados	Independientes	Otros
F	70,10%	72,00%	72,50%
M	65,80%	68,70%	69,00%

Fuente: elaboración propia.

Cuando se realiza el análisis multivariado empleando análisis factorial de datos mixtos, se encuentra que las variables más representativas dentro de las variables cuantitativos corresponden a edad, el capital recuperado, y se evidencian correlación directa entre la variable Rango ISF que corresponde a la probabilidad de pago de de los clientes con el saldo a capital es decir que mientras mayor sea la el saldo a capital o deuda del cliente mayor es su probabilidad de pago.

Se evidencia que la variable de Marcación Nor que corresponde a la marcación que tiene un cliente que ha tenido arreglo de cartera anterior, no es una variable representativa ni tiene correlación con la variable de capital recuperado de acuerdo a la figura 5.

Figura 5. gráfica de análisis factorial de datos mixtos para variables cuantitativas

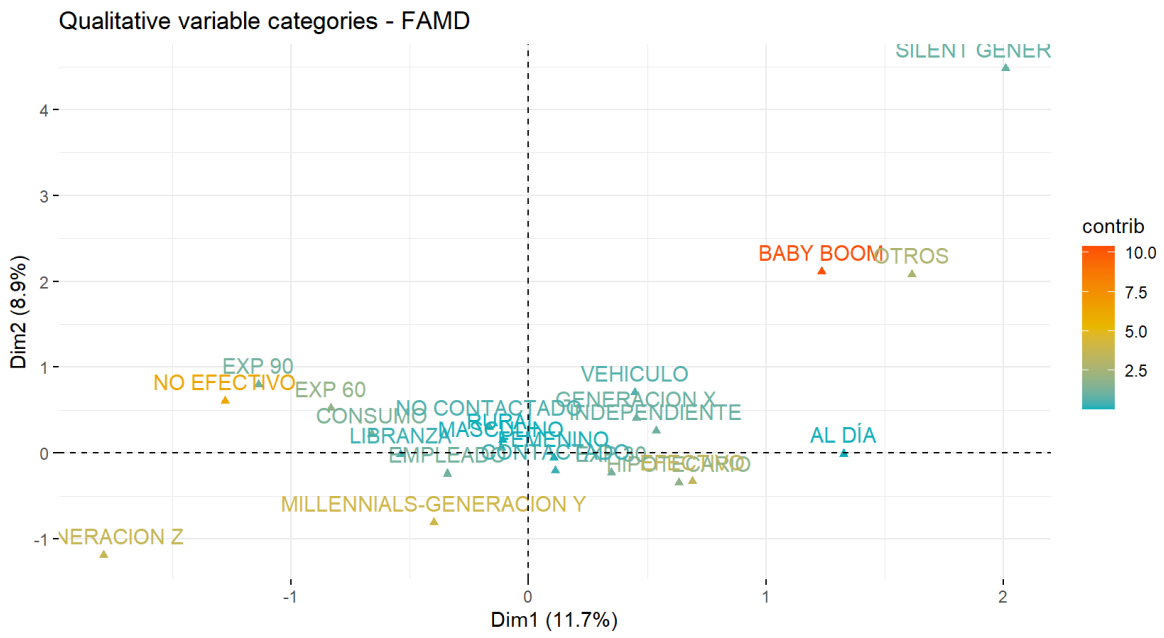


Fuente: elaboración propia

continuando con el análisis de la variables categóricas y de acuerdo a la observación de la figura 6, se identifican patrones y relaciones con los individuos de relevancia que permiten caracterizar mejor a la población objeto de estudio tales como;

- La generación de baby boom con rango de edades de 54 a 73 años tienen una actividad laboral diferente a empleados e independientes caracterizada principalmente por pensionados.
- Por otra parte se puede concluir que los clientes no efectivos, están concentrados en rangos de mora de 60 y 90, donde el principal segmento corresponde a productos de consumo, libranza y rural con contacto no efectivo.
- La generación x correspondiente a personas con rango de edades entre 53-42 años se encuentran en mora principalmente con créditos de vehículo y se caracterizan por tener una actividad laboral como independientes.
- La distancia entre la generación silent y la generación Z a pesar de no poderse caracterizar claramente con el resto de variables.
- Se observa que el producto con mejor efectividad corresponde a crédito hipotecario y el rango de mora 30.
- Posiblemente el producto con mayor dificultad de contacto está representado por los créditos de segmento rural.

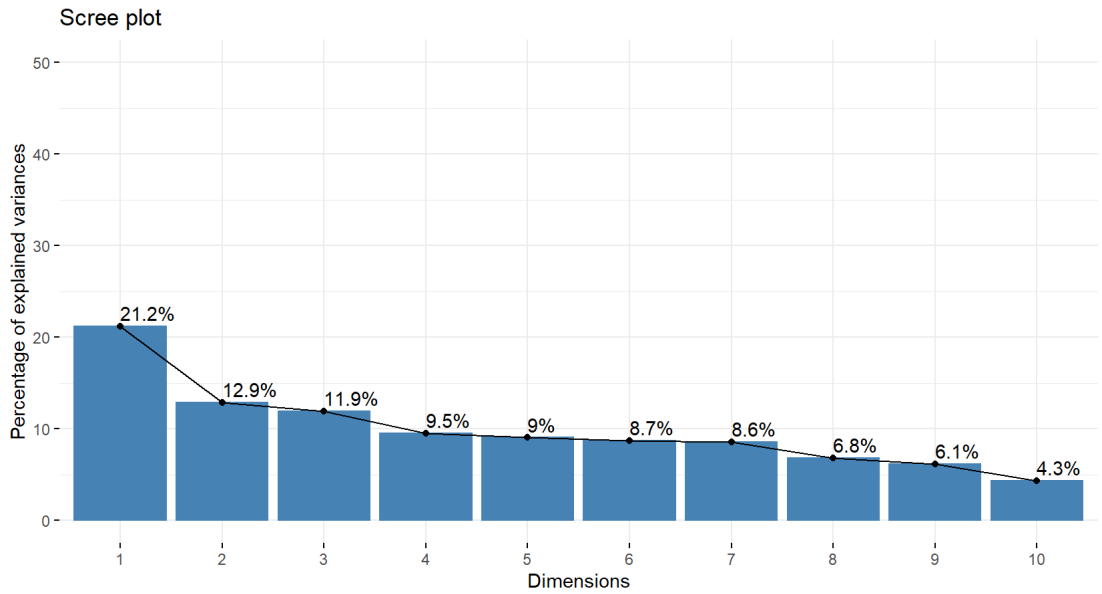
Figura 6. gráfica de análisis factorial de datos mixtos para variables cualitativas



Fuente: elaboración propia

Adicionalmente se valida la distribución de la varianza explicada dentro del conjunto de datos, donde el punto de inflexión se encuentra entre la dimensión 1 y 2 que contienen el 34,1% de varianza explicada de los datos representada en la figura 7.

Gráfica 7. gráfica de distribución de la varianza explicada en el conjunto de datos.



Fuente: elaboración propia.

Finalmente se seleccionan las variables para la aplicación de los modelos las cuales se presentan a continuación en tabla 2:

Tabla 2. Variables seleccionadas para aplicación de modelos

Variable	Tipo de variable
CAPITAL_CLIENTE	Cuantitativa
PRODUCTO	Categórica
RXM_MAX_INICIAL	Categórica
GÉNERO	Categórica
ACTIVIDAD_LABORAL	Categórica
PROBABILIDAD_DE_PAGO	Cuantitativa
EDAD	Cuantitativa
MARCACION_NORM	Cuantitativa
GESTIONES	Cuantitativa
CONTACTO_EFECTIVO	Cuantitativa
EFECTIVIDAD_REC	Cuantitativa

Fuente: elaboración propia

Los modelos seleccionados para aplicación corresponden a árboles de decisión, regresión logística y redes neuronales, la función de activación de la red neuronal es logística debido a que es la función de activación más utilizada para problemas de clasificación binaria, adicionalmente al entrenar la red con diferentes funciones tales como; identity, logistic, tanh y relu, se comprueba que la función que tiene mejor accuracy, es la función logística, seguida de identity y por último la Tanh y ReLu que comparten el mismo valor de exactitud, los resultados descritos se muestran en la tabla 3.

Tabla 3. tabla de resumen de accuracy del entrenamiento de la red neuronal con diferentes funciones de activación.

Función	Accuracy
Logistic	0,72978411
Identity	0,68614099
Tanh	0,59475354
ReLu	0,59475354

Fuente: elaboración propia

Al realizar la aplicación de los modelos seleccionados se obtuvieron los siguientes resultados de evaluación de exactitud, representado en la tabla 4:

Tabla 4. tabla de resumen de accuracy por modelo

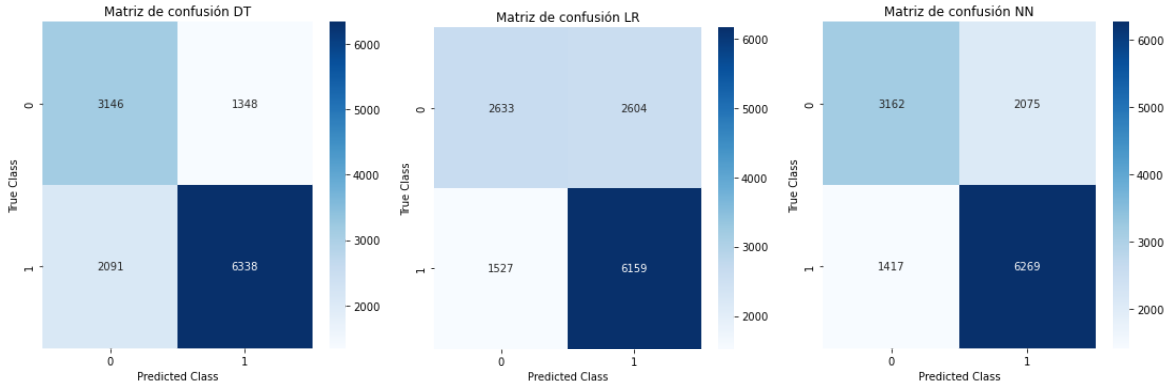
Model	Accuracy
Decision Tree	0,7339
Logistic regression	0,6803
Neural Network	0,7298

Fuente: elaboración propia

Donde el modelo con mayor exactitud corresponde al árbol de decisión, sin embargo se evidencia que no hay una diferencia significativa entre este y la red neuronal, por el contrario la regresión logística disminuye el rendimiento frente a los otros modelos en 4 puntos.

Al realizar la matriz de confusión de los 3 modelos se puede evidenciar que nuevamente el modelo que tiene mayor porcentaje y valores de aciertos sobre contacto efectivo, corresponde al árbol de decisión, seguido de la red neuronal y por último la regresión logística, sin embargo el árbol de decisión también cuenta con el mayor porcentaje y valor de falsos positivos donde clasificó el contacto no efectivo como efectivo, representadas en las figuras 8, 9 y 10 y la tabla 5.

Figura 8. **Figura 9.** **Figura 10.**



Nota: matrices de confusión por modelo. Fuente: elaboración propia.

Tabla 5. Tabla de % de resultados matriz de confusión por clase.

Clase	% Matriz DT		% Matriz LR		% Matriz NN	
0 (No_ Contacto)	24,30%	10,40%	20,40%	20,20%	24,50%	16,10%
1 (Contacto_efe)	16,20%	49,00%	11,80%	47,70%	11,00%	48,50%

Fuente: elaboración propia

Después se realizó la validación de las métricas por clasificación de clase, donde se observa que los valores de evaluación por clase del árbol decisión y la red neuronal son muy similares encontrando solo diferencia de 1 punto para la precisión de la clase 0 correspondiente a no contacto y en el f1-score.

Las métricas de evaluación de la regresión logística tienen valores más bajos en comparación de los otros modelos, teniendo el menor resultado de evaluación por clase presentados en la tabla 6.

Tabla 6. Tabla de resumen de métricas de evaluación por clase de los modelos aplicados

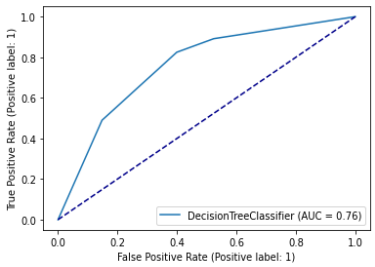
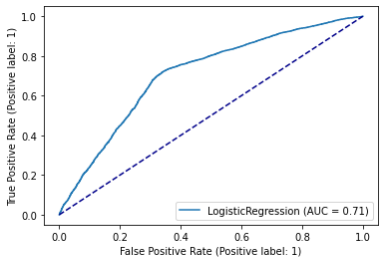
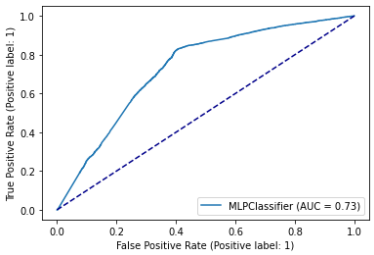
Decision Tree					Logistic Regression					Neural Network				
	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.70	0.60	0.65	5237	0	0.63	0.50	0.56	5237	0	0.69	0.60	0.64	5237
1	0.75	0.82	0.79	7686	1	0.70	0.80	0.75	7686	1	0.75	0.82	0.78	7686
accuracy			0.73	12923	accuracy			0.68	12923	accuracy			0.73	12923
macro avg	0.73	0.71	0.72	12923	macro avg	0.67	0.65	0.65	12923	macro avg	0.72	0.71	0.71	12923
weighted avg	0.73	0.73	0.73	12923	weighted avg	0.67	0.68	0.67	12923	weighted avg	0.73	0.73	0.73	12923

Nota: comparación de resultados de métricas de evaluación por clase de árbol de decisión, regresión logística y red neuronal.

Fuente: elaboración propia

La curva ROC más alta corresponde al árbol de decisión con 0.76, seguido de la red neuronal con 0.73 y por último la regresión logística con 0.71

Tabla 7. Tabla resumen de curvas ROC de los modelos aplicados

Árbol de decisión	Regresión Logística	Red Neuronal
		

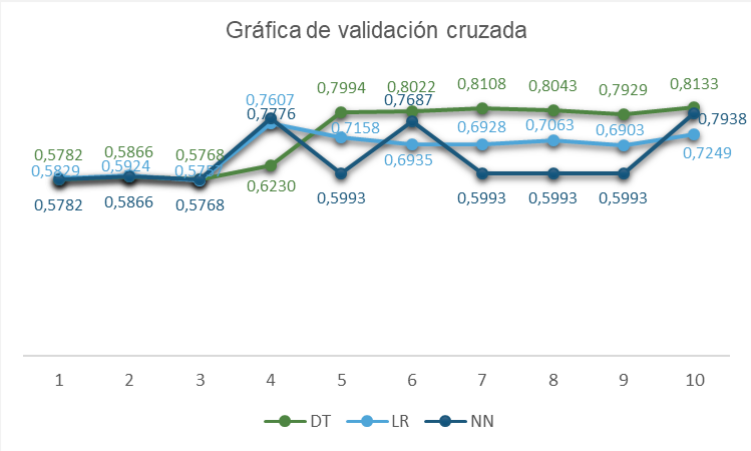
Nota: tabla comparativa de curva ROC de árbol de decisión, regresión logística y red neuronal.

Fuente: elaboración propia

Al realizar la validación cruzada de los modelos con 10 iteraciones, se evidencia que el modelo con mayor estabilidad en la métrica de exactitud, corresponde al árbol de decisión, por otra parte se observa que la red neuronal a pesar de obtener resultados muy cercanos a los del árbol de decisión presenta menor estabilidad respecto a los otros 2 modelos y la regresión logística que había obtenido los valores más bajos respecto a las métricas de

evaluación anteriores presenta un accuracy promedio más alto respecto a la red neuronal con 0,6735 y 0,6479 respectivamente, con valores más bajos en la iteraciones, como se muestra en la figura 11.

Figura 11. Gráfica de validación cruzada, accuracy de modelos con 10 iteraciones.



Fuente: elaboración propia

Por último, se realiza proceso de escalación de variables y se realiza nuevamente entrenamiento de los modelos de regresión logística y red neuronal, con el fin de validar si genera un aumento en la métrica de exactitud del modelo, pero al obtener los resultados se evidencia una disminución de esta para la regresión logística, mientras que no se evidencia ningún cambio en la medida de evaluación para la red neuronal.

Modelo	Accuracy	Accuracy Scal
LR	0,68033738	0,59475354
NN	0,72978411	0,72978411

CONCLUSIONES

1. De acuerdo a los resultados obtenidos, se puede concluir que los árboles de decisión son modelos bastante acertados y robustos para solucionar problemas de clasificación binaria en comparación a modelos de regresión logística y redes neuronales, sin embargo pueden tener mayor número de falsos positivos para la clase con mayor número de registros dentro de la matriz de confusión que la red neuronal y la regresión logística, requiriendo para el modelo de árbol de decisión en la etapa de preprocesamiento de datos, se realicen aplicaciones de técnicas de balanceo previas a la modelación.
2. Por último se concluye que las redes neuronales son clasificadores binarios eficientes que presentan menor sensibilidad frente a transformaciones como el escalamiento de los datos en comparación a la regresión logística, se propone para próximas investigaciones la optimización de modelos de aprendizaje supervisado con técnicas de aprendizaje por refuerzo, ya que es un campo de estudio poco explorado que puede contener soluciones para mejorar las métricas de evaluación de los modelos supervisados.

REFERENCIAS BIBLIOGRÁFICAS

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*, 381-386.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*(1), 71-72.

Praveena, M., & Jaiganesh, V. (2017). A literature review on supervised machine learning algorithms and boosting process. *International Journal of Computer Applications*, *169*(8), 32-35.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica*, *24*(1), 12-18.

Glonek, G. F., & McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(3), 533-546.

Pardo, C. E. (2015). Estadística descriptiva multivariada.

Menacho Chiok, C. H. (2014). Modelos de regresión lineal con redes neuronales. *Anales Científicos,* 75(2), 253. https://doi.org/10.21704/ac.v75i2.961

Vallalta Rueda, J. F. (2022). CRISP-DM: *una metodología para minería de datos en salud* [Image]. Recuperado de https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

BPO en la cobranza de una pyme | Microformas. (s. f.). Microformas | Soluciones para la Productividad. https://microformas.mx/blog/bpo-en-la-cobranza-de-una-pyme/

Sánchez, C. (08 de febrero de 2019). *Cita Textual o Directa*. Normas APA (7ma edición). https://normas-apa.org/citas/cita-textual/

Dridi, S. (2021). Supervised Learning-A Systematic Literature Review.

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1, pp. 29-39).

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology.

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168).