CODING DOJO

PREDICCIÓN DE INCUMPLIMINENTO DE PAGOS CREDITICIOS

Información general para sacar más partido en la gestión del riesgo de un préstamo bancario.

Alex Daniel Huarancca Moriano

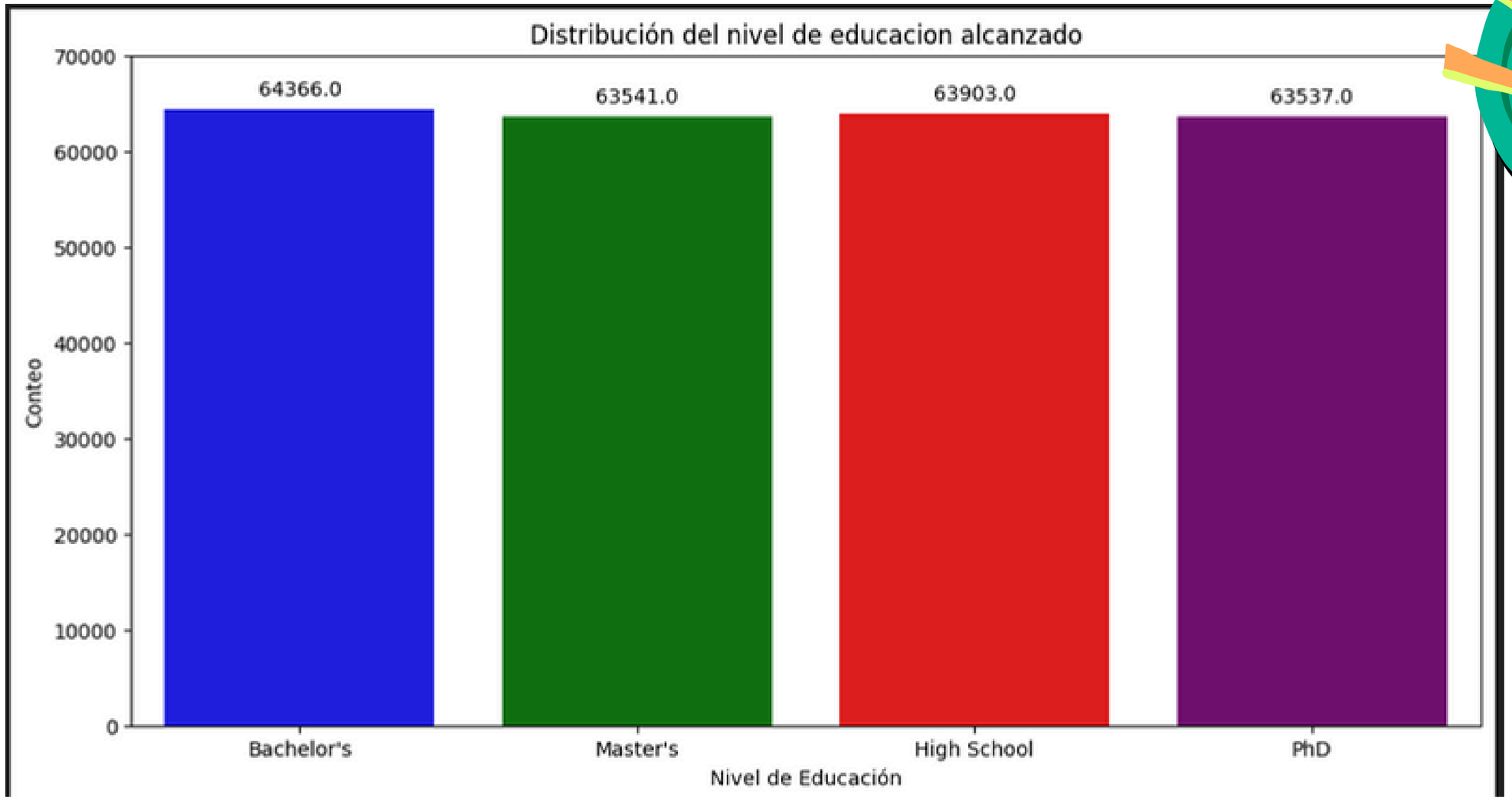
Introducción:

¿Se puede predecir quienes dejaran de pagar su prestamo bancario y la respuesta es...**Sí!!!**

¿Cómo?

Usando modelos de machine learning.

POBLACIÓN DE ESTUDIO DISTRIBUIDA POR EL NIVEL DE EDUCACIÓN

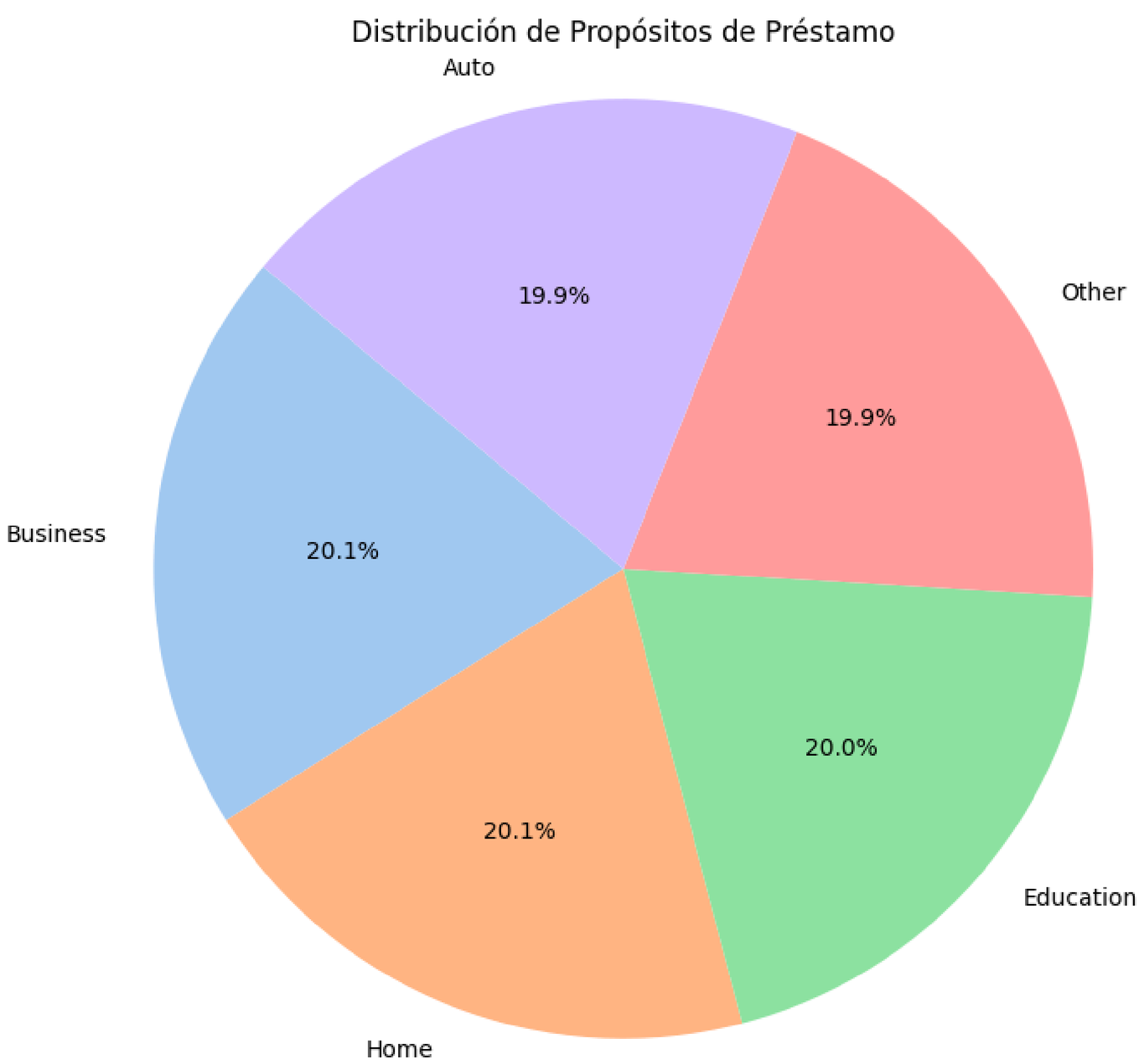


INFERENCIA : ***En base al grafico superior inferimos que la poblacion tiene 4 grupos de personas distribuidas por el nivel de estudio alcanzados, los cuales son: Bachelor´s, Master´s, High School, PhD. Los cuales se encuentran distribuidos de forma casi uniforme no presentando una gran diferencia en la cantidad de cada grupo. Por otro lado el grupo de personas que tienen educacion de Bachelor´s es el que presenta ligeramente mayor cantidad en comparación a los demas, por otro lado observando el Dataset no presenta sesgo hacia una categoria en especifico, asi mismo tambien concluimos que los clientes que solicitan los prestamos superaron el nivel de educación desde High School hacia adelante, y tambien puedo inferir que hay menor desigualdad en el acceso a la educación en los clientes de nuestro Dataset.***

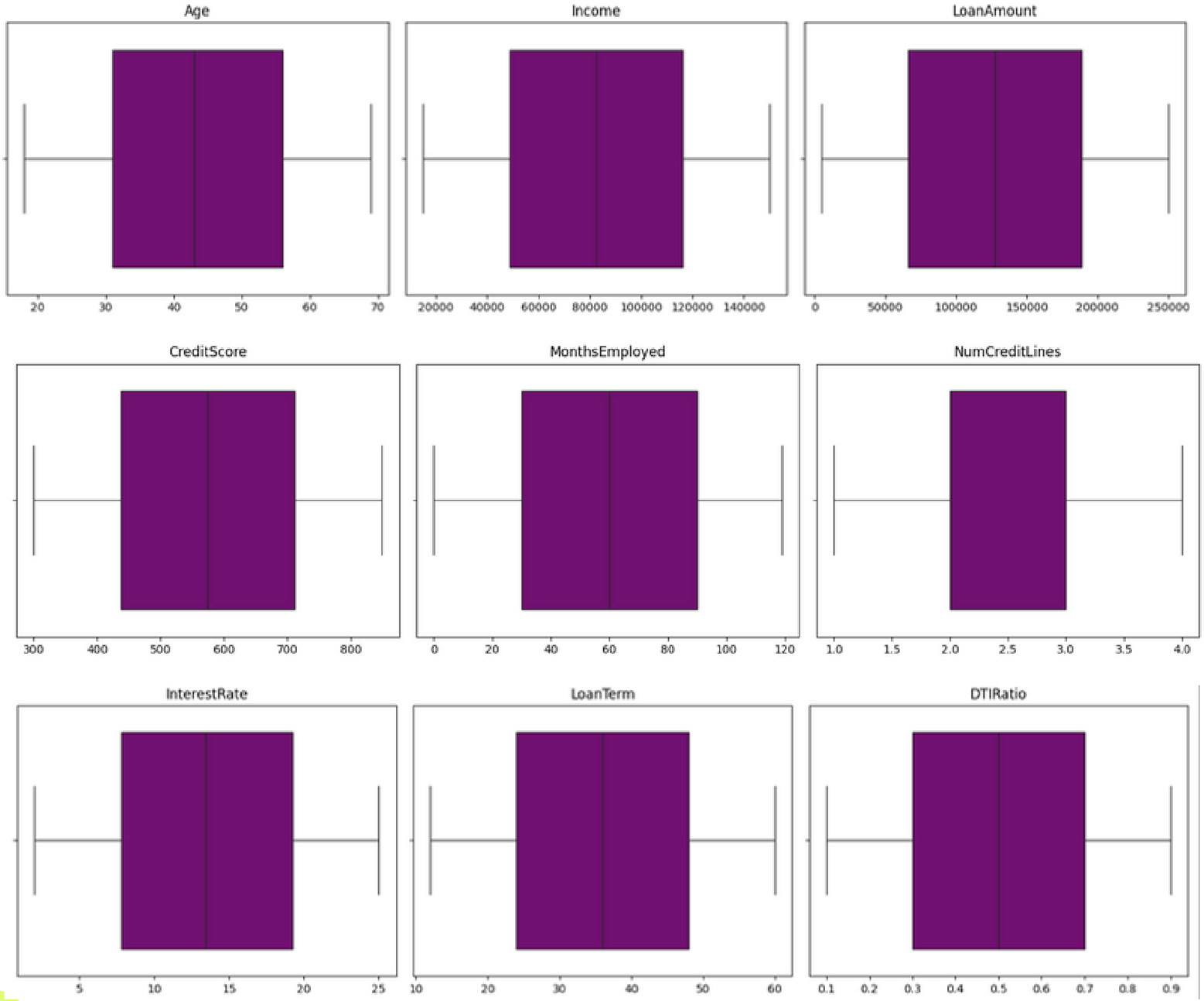
DISTRIBUCIÓN DEL PORCENTAJE DE PREFERENCIA DEL PROPOSITO DEL PRESTAMO

Inferencia:

El porcentaje de preferencia del tipo de prestamo se encuentra distribuidos de forma uniform en la población indicando que no existe un proposito superior en especifico del tipo de prestamo que se adquiere.



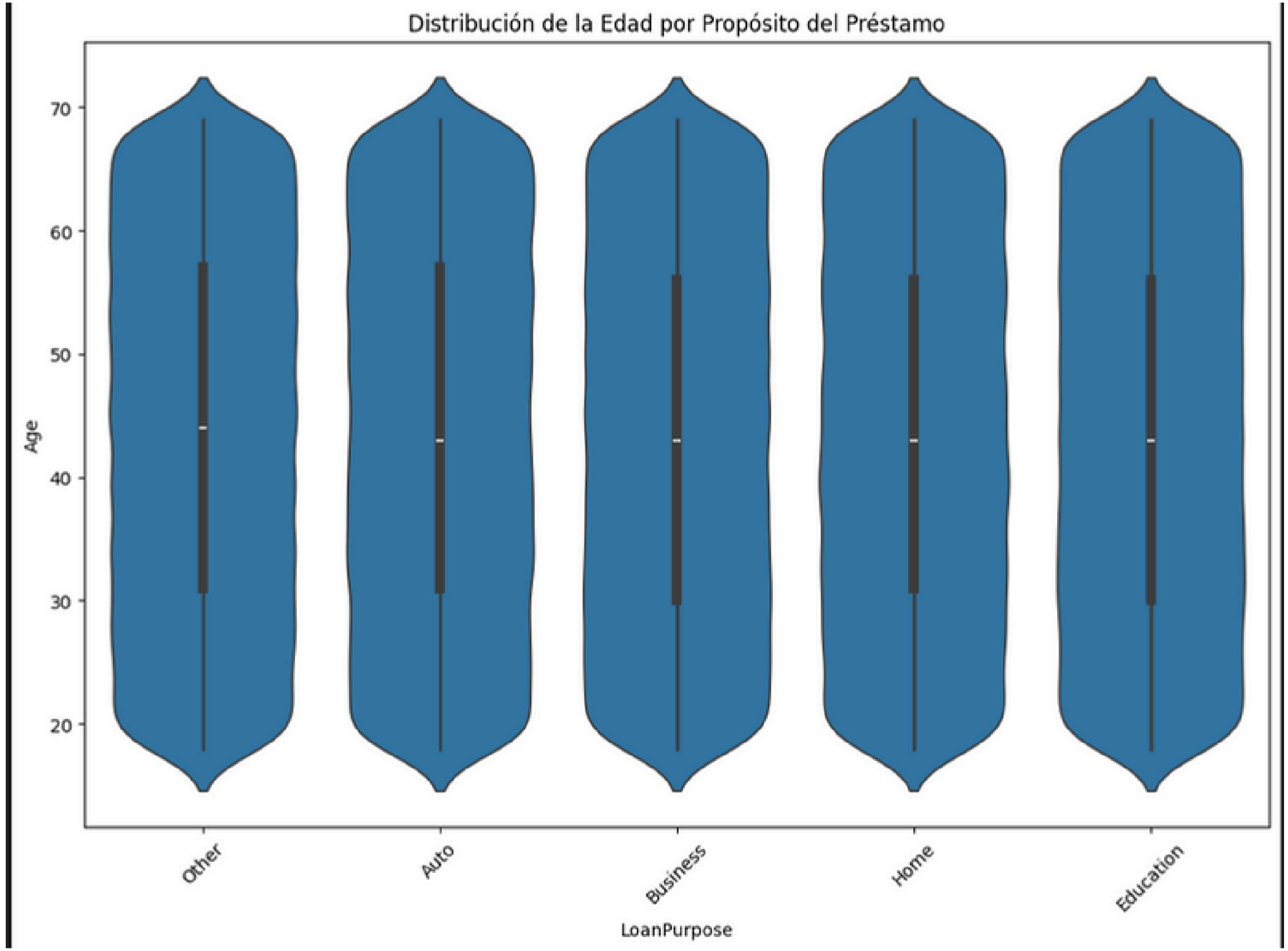
ESTUDIO DE VALORES ATIPICOS EN EL DATASET



Inferencia:

Las variables numericas se encuentran distribuidas de forma uniforme dentro de los rangos intercuartilicos, lo cual es optimo para proceder al escalamiento.

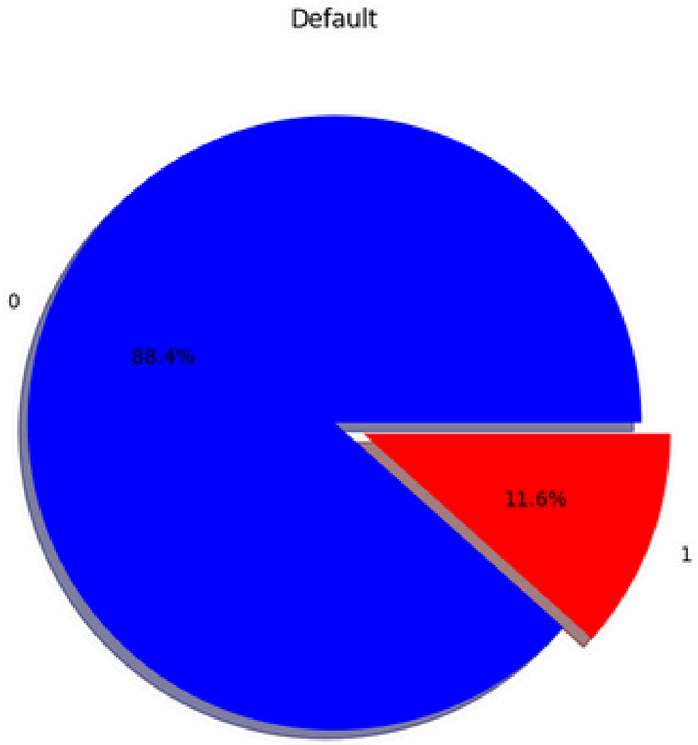
DISTRIBUCIÓN DE LA EDAD POR PROPOSITO DE L PRESTAMO

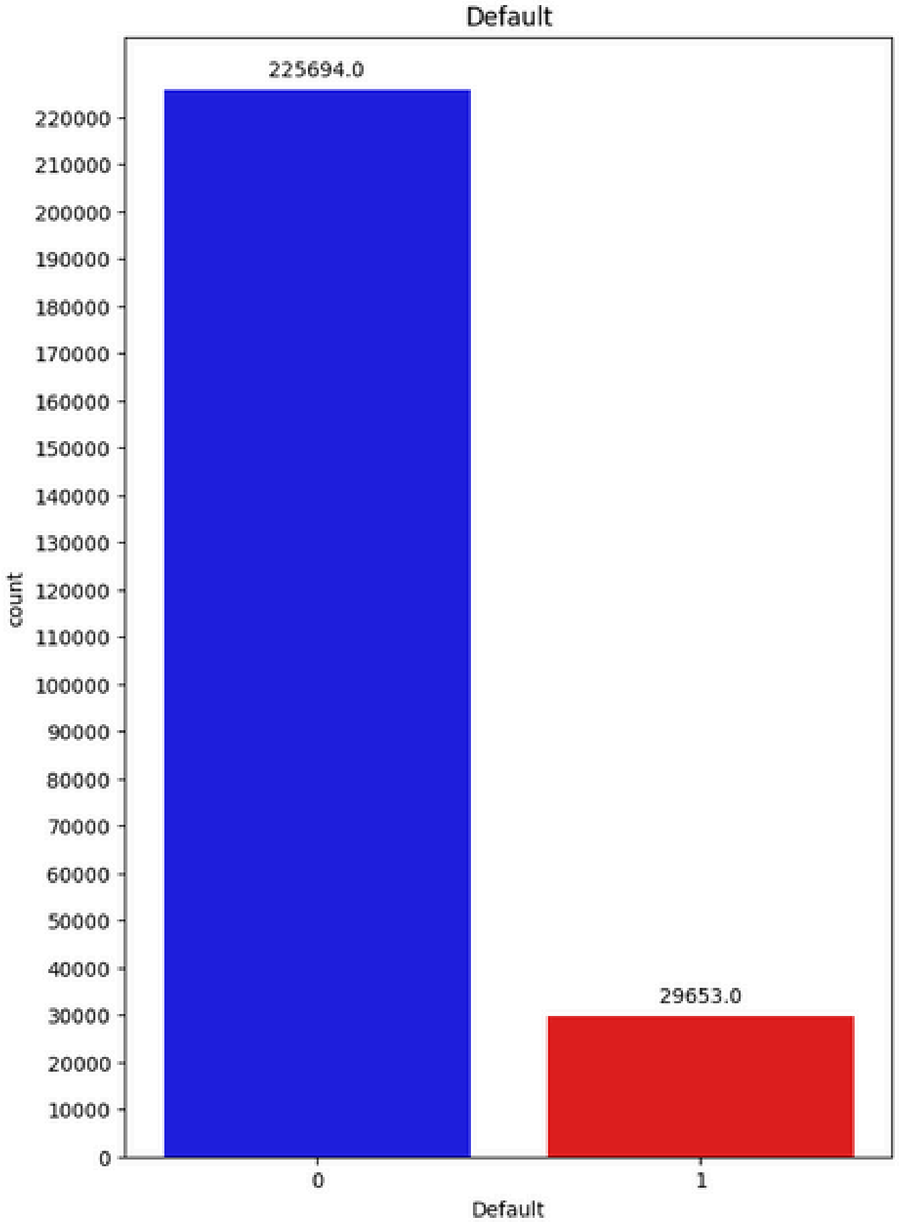


Inferencia:

No existe una preferencia sesgada por la edad del cliente, es decir que en relación a las edades el promedio oscila en 43 años y la preferencia de cualquier tipo de préstamo es uniforme en las distribuciones.

DISTRIBUCIÓN DEL TOTAL DE LA POBLACION EN RELACIÓN AL DEFAULT





Sí paga

No paga

Inferencia:

Las personas identificadas por el color azul son aquellas que cumplen con el pago de su prestamo bancario, los cuales representan el 88.4% del total y por otro lado las personas que presentan riesgo al momento de pagar su prestamo son solo el 11.6% del total de la población.

CONCLUSIONES FINALES DEL EDA

* La población de estudio se encuentra uniformemente distribuida en cuanto al nivel de estudio alcanzado suprando la Educacion secundaria en adelante.
* Asi mismo también que la distribución de la preferencai de prestamo no esta sesgada en especifico a una categoria particular de prestamo, ya que estas se encuentran ligeramente distribuidas de forma uniforme, a lo que podemos concluir que seria recomendable poder distribuir adecuadamente el presupuesto de marketing para cada categoria de prestaño.
* Observamos que la tasa de pagos del credito financiero es positiva y que se cumplen con los pagos de forma adecuada, esta se ve representada por el 88.4% sin embargo la tasa de clientes que abandona el pago de sus cuotas es minima teniendo solo el 11.6%, asi mismo es de principal estudio esta parte de la poblacioón para poder gestionar de mejor manera el otorgamiento del credito.
* Por ultimo observamos que la preferencia del tipo de prestamo en base a la edad promedio tambien resulta ser uniforme, obtniendo como pórmedio de la edad para cada tipo de prestamo el valor de 43 años.
.

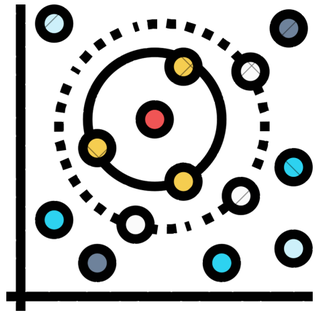
MODELOS DE MACHINE LEARNING A UTILIZAR

DECISION TREE CLASSIFIER

El modelo Decision Tree Classifier clasifica datos creando un árbol de decisiones basado en características, útil para tareas de clasificación.

RANDOM FOREST

El modelo Random Forest clasifica datos usando múltiples árboles de decisión, mejorando precisión y robustez mediante el voto mayoritario.

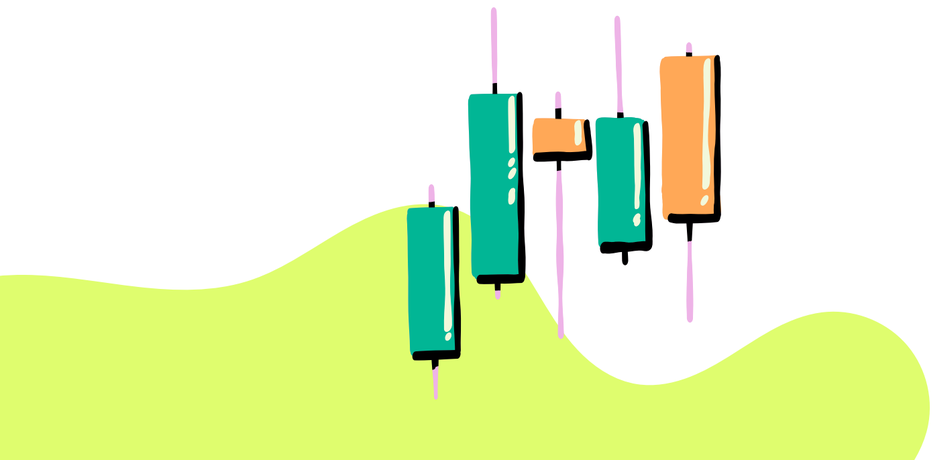


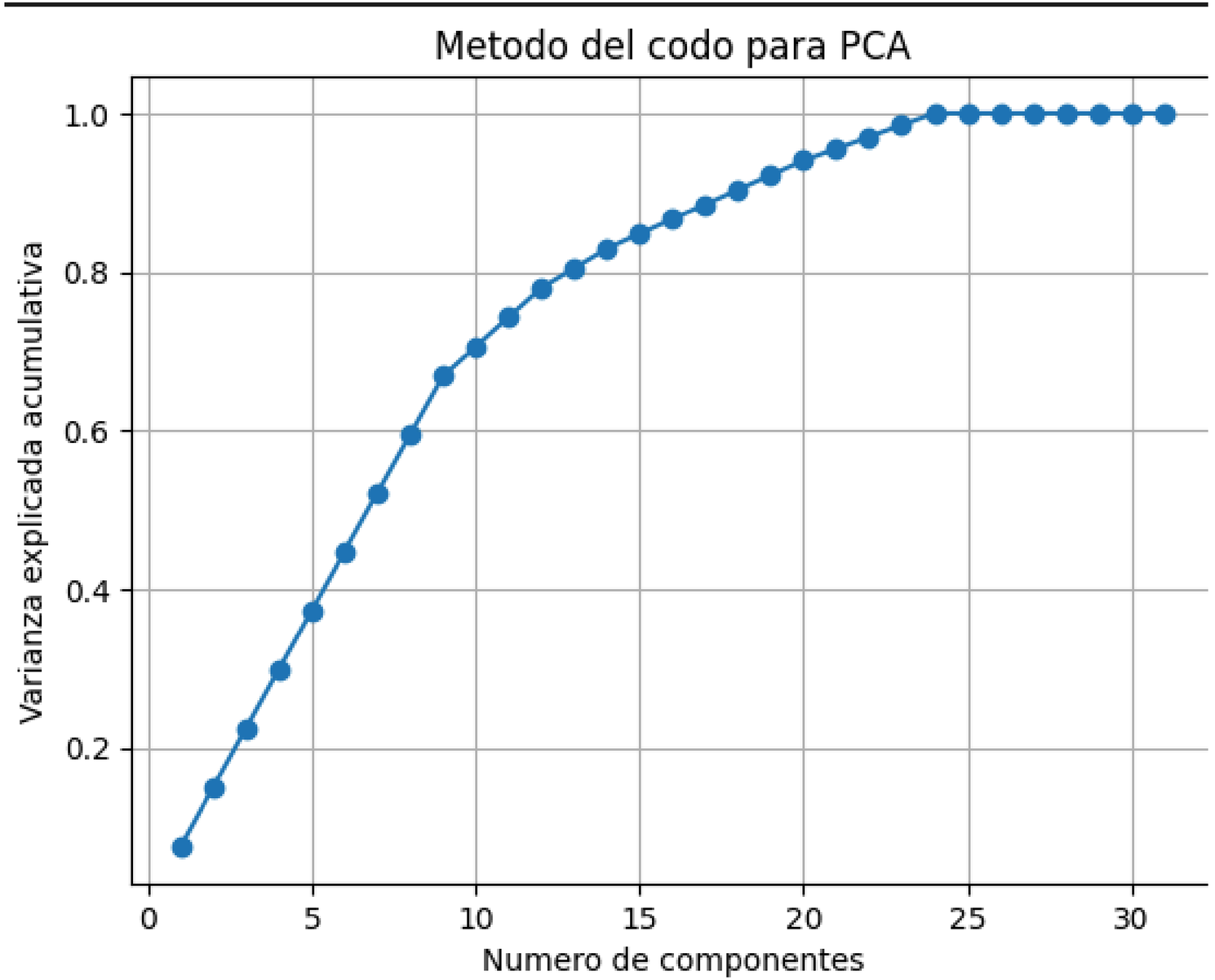
KNNEIGHBOURS

El modelo K-Nearest Neighbors clasifica datos basándose en la mayoría de los vecinos más cercanos, útil para clasificación y regresión.

ANALISIS DE COMPONENTES PRINCIPALES

PCA (Principal Component Analysis) reduce la dimensionalidad de datos, conservando variabilidad máxima, útil para preprocesamiento y visualización en machine learning





RENDIMIENTO DE LOS MODELOS DE MACHINE LEARNING SIN PCA Y CON PCA

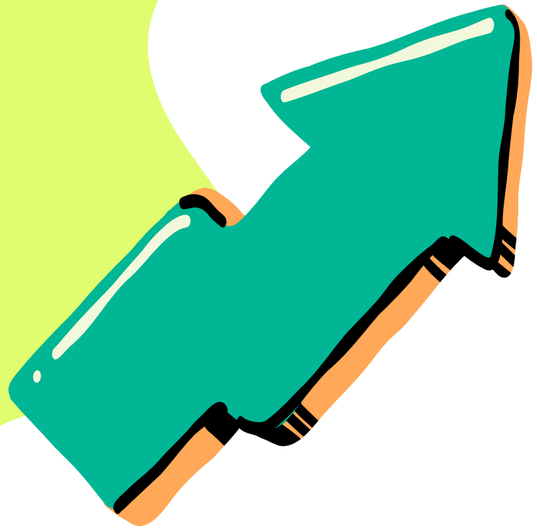
	SCORE	SCORE
MODELOS	ENTRENAMIENTO	TESTEO
Decission Tree Classifier sin PCA	88.47%	
Random Forest sin PCA	88.32%	
K-Nearest Neighbours sin PCA	88.12%	
Decission Tree Classifier con PCA	88.41%	88.32%
Random Forest con PCA	99.99%	88.40%

CONCLUSIONES FINALES

- **Concluimos que el mejor modelo para aplicar a este DataSet dentro del Machine Learning es el de Arbol de decision con PCA, sin embargo el accuracy es de 0.88, y el accuraccy tambien es de 0.88 observando un equilibrio de rendimiento tanto en los datos de entrenamiento y en los datos de testeo.**

- **Observamos tambein que el modelo de Random forest funciona casi a la perfeccion con los datos de entrenamiento, pero sin embargo el accuracy obtenido para los datos de testeo no se asemejan a los de entrenamiento, obteniendo uan diferencia del 0.11 entre los datos de entrenamiento y los de testeo.**

- **Por otro lado se sugiere en base a este avance del proyecto evaluar el Dataset con diferente modelos pero por el momento el algoritmo que funciona mejor es el de: **Arboles de Decisión**



Gracias

¿Tienes alguna pregunta? No dudes en consultarme:

EMAIL

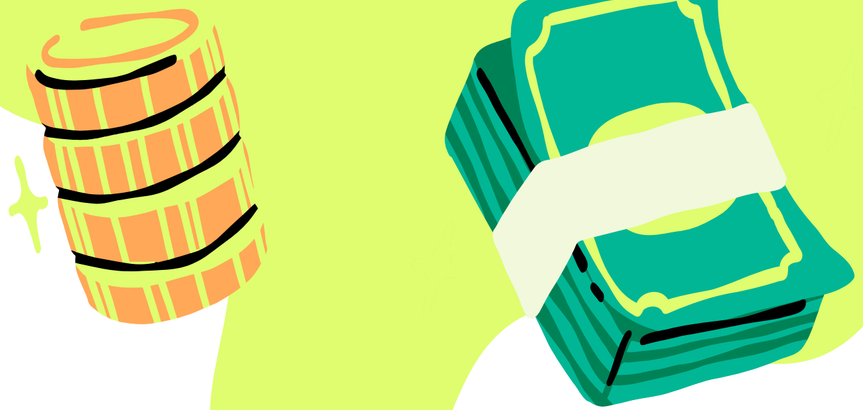
alexdaniel.huarancca@gmail.com

GITHUB:

https://github.com/alexdaniel-23

LINKEDIN

linkedin.com/in/alex-daniel-huarancca-moriano-50a5bb270



ENLACE AL REPOSITORIO

https://github.com/alexdaniel-23/Loan-Default-Prediction-Dataset.git