APROBACIÓN DE PRÉSTAMOS

Misael Redrejo Fernández, Yerson Caleb Yarhui Sarate

Introducción

La toma de decisiones en la **aprobación de préstamos** constituye un desafío crítico en el ámbito financiero. La capacidad de evaluar con precisión la viabilidad de un solicitante para reembolsar un préstamo se ha vuelto cada vez más crucial para mitigar riesgos y garantizar la estabilidad financiera.

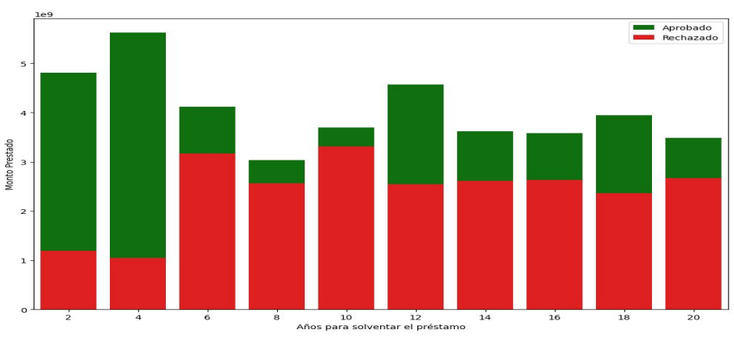
El objetivo central de este trabajo es explorar y aplicar diversas técnicas de aprendizaje automático para desarrollar modelos que predigan con precisión la aprobación o rechazo de préstamos.

Conjunto de Datos

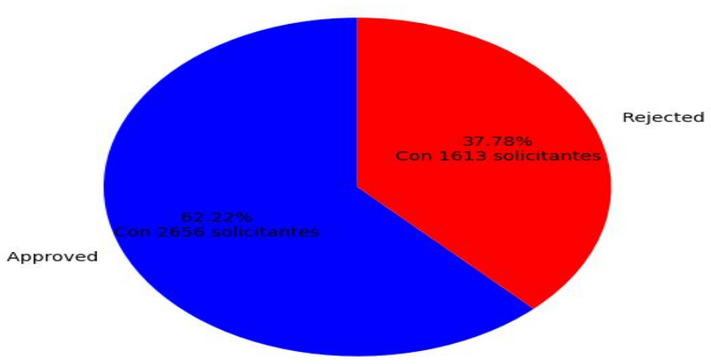
El conjunto se divide en 11 atributos(2 nominales binarias y 9 numéricos) y una salida nominal binaria.

- Atributos numéricos = no_of_depends, income_annum, loan_amount, loan_term, cibil_score, residential_assets_value, commercial_assets_value, luxury_assets_value, bank_assets_value
- Atributos categóricos binarios = loan_status(salida), education, self_employed.

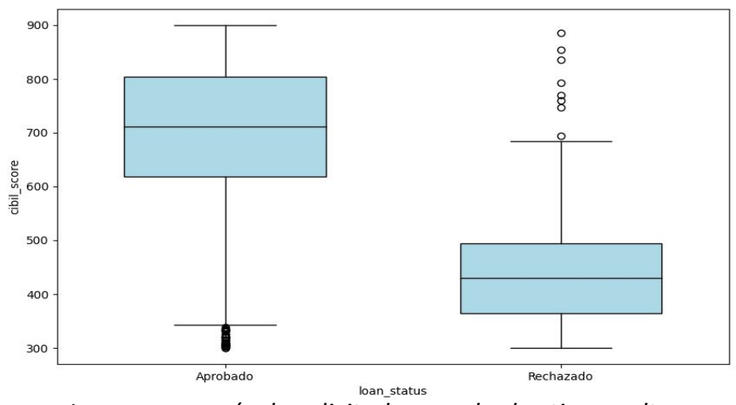
Se han dividido los datos en train, val y test con stratified.



Los montos grandes solicitados, que buscan solventarse en un corto período se conceden con mayor frecuencia que los montos pequeños que desean solventarse en un largo período.



Distribución del Estado de las Solicitudes de Préstamo de los Solicitantes.



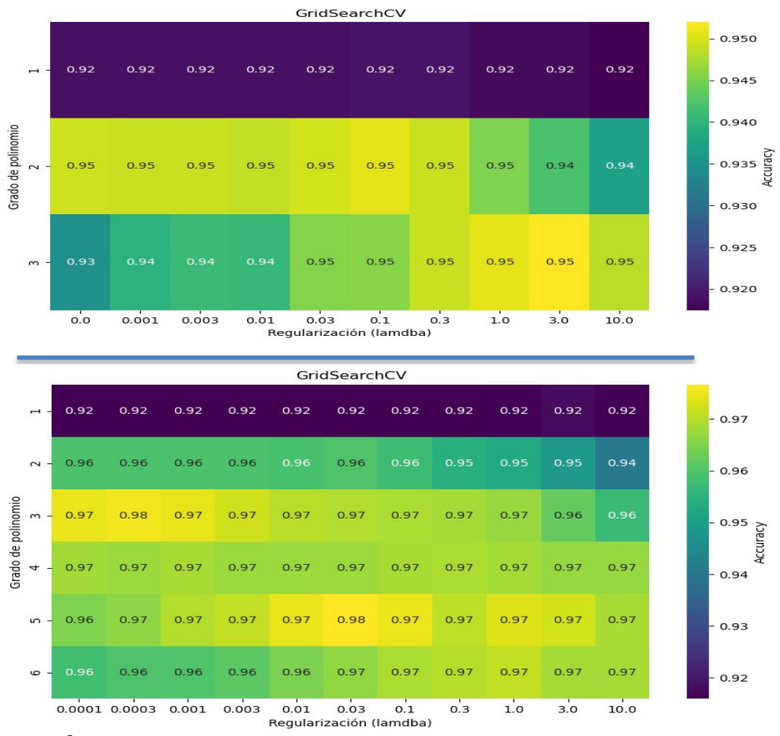
La gran mayoría de solicitudes aprobadas tienen alto cibil score mientras que en las solicitudes rechazadas ocurre lo contrario.

Marco Experimental

La investigación se enfoca en comparar técnicas de aprendizaje automático, incluyendo regresión logística, RFE, Naive Bayes, redes neuronales, K-means, ensambles (Bagging, Decision Tree, Random Forest, Boosting), OVA y OVO. Se evalúan en dimensiones críticas del conjunto de datos, considerando la cantidad y relevancia de las características, así como el contexto socioeconómico de los solicitantes de préstamos.

MÉTODO 1: Regresión logística

Para la regresión logística se probaron utilizando diferentes grados regularización y variables polinomiales, utilizando la regresión normal y la regresión con RFE (5 atributos).



MÉTODO 2: Naive Bayes

Con Naive Bayes se probaron diferentes técnicas de cálculo de probabilidad para variables categóricas y numéricas.

MÉTODO 3: Redes Neuronales

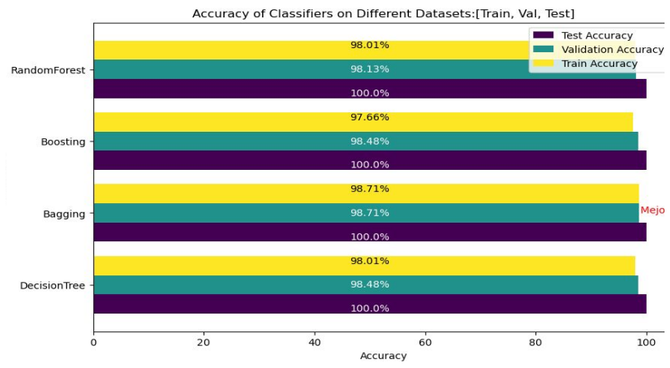
Con Redes Neuronales las siguientes combinaciones de capas ocultas: (11,), (50,), (11, 50), (50, 50); activación: logistic, relu, tanh; alpha: 0.001, 0.01, 0.1. Consiguiendo más éxito con capas ocultas (11, 50),activación relu y alpha 0.001.

MÉTODO 4: K-Means

K-Means, se realizó con un número fijo clusters de K=2 puesto que solo habían 2 posibles clases.

MÉTODO 5: Ensembles

Con los Ensembles se evaluaron los clasificadores Bagging, Decision Tree, Random Forest, Boosting



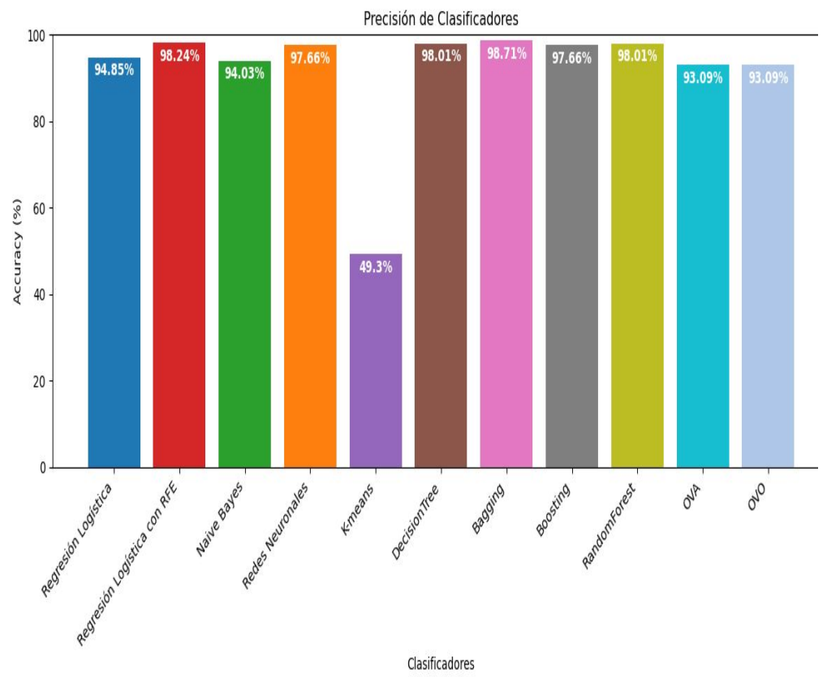
MÉTODO 6: OVO y OVA

Con OVA y con OVO se evaluaron los diferentes tipos de Kernel posibles linear, poly, rbf y sigmoid.

Estudio Experimental

Nuestro análisis concluye que el modelo basado en la técnica de Bagging logra el mayor accuracy en los datos de prueba con un 98.71%. Esto significa que ha demostrado ser el más robusto y efectivo para generalizar, es decir, en realizar predicciones precisas en datos no vistos, lo cual es idóneo para aplicaciones del mundo real.

En la siguiente gráfica se muestran el accuracy obtenido en el conjunto de pruebas para los diferentes modelos entrenados.

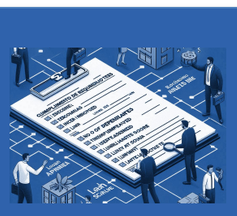


Conclusiones

En la selección de los mejores hiperparámetros para los diferentes modelos de aprendizaje automático se ha tenido en cuenta el accuracy en los datos de validación.
Al seleccionar el mejor modelo, observamos el rendimiento en el conjunto de prueba para obtener el modelo que mejor clasifique los datos.

Trabajo futuro

- Exploración de técnicas de normalización.
- Mitigación de falsos positivos o falsos negativos: depende de los objetivos y riesgos específicos del negocio.
- Optimización de hiper parámetros específicos: utilizar Google Colab para una búsqueda más exhaustiva.
- Exploración de nuevas técnicas: es posible que existan técnicas no estudiadas que permitan ajustarse a los datos mejorando la generalización del modelo.



Ingeniería informática - Aprendizaje automático

UPNA

Yerson Caleb Yarhui Sarate
Misael Redrejo Fernandez