**Comparación de algoritmos de DL frente a algoritmos de ML clásicos**

Caso de Estudio: Predicción del incumplimiento de pago de préstamos bancarios garantizados por la U.S. Small Business Administration

|  |  |  |
| --- | --- | --- |
| Titulación:  Máster Universitario en Inteligencia Artificial  Curso académico  2021 - 2023 | Alumno: Juan Alex Castro Gumiel  D.N.I: 5074733  Director de TFM: José Antonio Lagares Rodríguez | Convocatoria:  Cuarto Periodo |

**Índice**

[1. Introducción 3](#_Toc147773762)

[2. Marco Teórico 4](#_Toc147773763)

[2.1 Conceptos Clave 4](#_Toc147773764)

[2.2 Investigaciones Previas 8](#_Toc147773765)

[2.3 Hipótesis de Investigación 8](#_Toc147773766)

[3. Ámbito de la Aplicación 9](#_Toc147773767)

[4. Estado del Arte 10](#_Toc147773768)

[5. Justificación de la Propuesta 15](#_Toc147773769)

[5.1 Objetivo General 15](#_Toc147773770)

[5.2 Objetivos Específicos 16](#_Toc147773771)

[6. Descripción de la Propuesta 17](#_Toc147773772)

[7. Desarrollo de la Propuesta 19](#_Toc147773773)

[7.1 Análisis Exploratorio de Datos 19](#_Toc147773774)

[7.2 Preprocesamiento de Datos 19](#_Toc147773775)

[7.3 Modelos de ML: Clasificación 19](#_Toc147773776)

[7.4 Modelos de ML: Agrupación 19](#_Toc147773777)

[7.5 Modelos de Deep Learning 19](#_Toc147773778)

[8. Resultados 20](#_Toc147773779)

[9. Conclusiones 21](#_Toc147773780)

**Índice de tablas**

[**Tabla 1** Optimal parameters for our algorithm after Gridsearch. 10](#_Toc147770305)

[**Tabla 2** Performance metric for XGBoost algorithm on the loan dataset. 11](#_Toc147770306)

[**Tabla 3** Evaluation metrics comparison of the four techniques. 11](#_Toc147770307)

[**Tabla 4** Classification Report for Decision Tree and Random Forest. 12](#_Toc147770308)

[**Tabla 5** Parameters for evaluation. 13](#_Toc147770309)

[**Tabla 6** The result of using three important features. 13](#_Toc147770310)

[**Tabla 7** Comparación de los mejores resultados obtenidos en los papers analizados. 14](#_Toc147770311)

1. Introducción
2. Marco Teórico

Este Marco Teórico proporciona un sólido fundamento conceptual y teórico para el estudio comparativo de algoritmos de Machine Learning (ML) con algoritmos de Deep Learning (DL) en el contexto de la predicción del incumplimiento de pago de préstamos bancarios garantizados por la U.S. SBA, se establece la base necesaria para comprender la naturaleza y la relevancia de los enfoques de IA en la toma de decisiones financieras y crediticias.

2.1 Conceptos Clave

Préstamo Bancario

Un préstamo bancario es un acuerdo financiero en el que una institución financiera presta una cantidad de dinero a un individuo, empresa u organización, conocido como prestatario, con la obligación de devolver esa cantidad junto con intereses en un plazo acordado.

Aprobación de Préstamos

La aprobación de préstamos se refiere al proceso mediante el cual una institución financiera realiza la evaluación de una solicitud de préstamo presentada por individuos o entidades, con el objetivo de determinar si son elegibles para recibir financiamiento, el proceso implica la consideración de diversos factores financieros para tomar decisiones informadas.

Incumplimiento de Pagos

El incumplimiento de pagos de un préstamo, también conocido como "default" o "impago", se refiere a la situación en la cual un prestatario no cumple con las obligaciones acordadas en un contrato de préstamo, estas obligaciones suelen incluir el pago regular de cuotas o intereses en las fechas especificadas en el contrato; el incumplimiento puede manifestarse en forma de retrasos en los pagos, pagos incompletos o la falta total de pago.

Garantizar Préstamos

El proceso de garantizar préstamos bancarios se refiere a proporcionar una seguridad o respaldo financiero adicional para un préstamo otorgado por una institución financiera a un individuo o entidad, esta garantía se utiliza para mitigar el riesgo del prestamista y asegurarse de que, en caso de que el prestatario no cumpla con sus obligaciones de pago, el prestamista tenga un recurso adicional para recuperar el dinero prestado.

Inteligencia Artificial (IA)

La Inteligencia Artificial (IA) se refiere a la simulación de procesos de inteligencia humana mediante el uso de sistemas informáticos, matemáticas y algoritmos; es una rama de la informática que se centra en el desarrollo de programas y sistemas capaces de realizar tareas que normalmente requieren de inteligencia humana, como el razonamiento, la resolución de problemas, el aprendizaje, la percepción y la toma de decisiones.

Machine Learning (ML)

El Machine Learning (Aprendizaje Automático) es una subdisciplina de la IA que se centra en el desarrollo de algoritmos y modelos informáticos que permiten a las computadoras aprender y mejorar su rendimiento en tareas específicas a través de la experiencia y el análisis de datos, en lugar de ser programadas explícitamente para cada tarea; en esencia, el Machine Learning se basa en la idea de que las computadoras pueden aprender patrones y tomar decisiones basadas en datos sin ninguna intervención humana directa.

Aprendizaje Supervisado

El Aprendizaje Supervisado es un enfoque dentro del campo del Machine Learning en el que se entrena un modelo utilizando un conjunto de datos de entrenamiento que contiene ejemplos etiquetados; en el aprendizaje supervisado, el objetivo principal es que el modelo aprenda a mapear las entradas (características) a las salidas (etiquetas) de manera que pueda hacer predicciones o tomar decisiones en nuevos datos no etiquetados.

Aprendizaje Supervisado para Regresión

El Aprendizaje Supervisado para Regresión es una subcategoría del aprendizaje supervisado en el campo del Machine Learning que se utiliza cuando el objetivo principal consiste en predecir un valor numérico o continuo en función de las características de entrada; en la regresión supervisada se busca predecir un valor numérico específico.

Aprendizaje Supervisado para Clasificación

El Aprendizaje Supervisado para Clasificación es una rama del Machine Learning en la que se entrena un modelo utilizando un conjunto de datos de entrenamiento que contiene ejemplos etiquetados, con el objetivo de que el modelo pueda predecir a qué clase o categoría pertenece una nueva observación en función de sus características de entrada.

Aprendizaje No Supervisado

El Aprendizaje No Supervisado es un enfoque dentro del campo del Machine Learning en el que un algoritmo o modelo se entrena utilizando un conjunto de datos que no contiene etiquetas o categorías predefinidas para las observaciones; a diferencia del aprendizaje supervisado, en el aprendizaje no supervisado, el objetivo principal es descubrir patrones, estructuras ocultas o relaciones entre datos sin la guía de etiquetas previamente conocidas.

Aprendizaje No Supervisado para Agrupación

El Aprendizaje No Supervisado para Agrupación se utiliza cuando el objetivo principal es dividir un conjunto de datos en grupos o clústeres, de modo que las observaciones dentro de cada grupo sean más similares entre sí que con las observaciones en otros grupos; este proceso de agrupación se realiza sin el uso de etiquetas o categorías predefinidas, y el algoritmo busca encontrar patrones en los datos que permitan formar grupos coherentes.

Aprendizaje por Refuerzo

El Aprendizaje por Refuerzo es un enfoque dentro del campo del Machine Learning que se centra en el desarrollo de algoritmos y modelos de IA que aprenden a tomar decisiones secuenciales óptimas en un entorno interactivo; a diferencia del aprendizaje supervisado, donde se proporcionan ejemplos etiquetados, y del aprendizaje no supervisado, que se enfoca en patrones de datos no etiquetados, el aprendizaje por refuerzo se basa en la interacción del agente de aprendizaje con su entorno y el concepto de recompensa.

Deep Learning (DL)

El Deep Learning (Aprendizaje Profundo) es una subdisciplina de la IA que se centra en el desarrollo de algoritmos y modelos de aprendizaje automático basados en redes neuronales artificiales profundas; estas redes neuronales profundas están diseñadas para imitar la estructura y el funcionamiento del cerebro humano, con capas de neuronas artificiales interconectadas que procesan y transforman datos de entrada en representaciones cada vez más abstractas y sofisticadas, las cuales permiten que el modelo pueda aprender representaciones jerárquicas de los datos a medida que fluyen a través de las capas.

Redes Neuronales Profundas

Las Redes Neuronales Profundas son un tipo de modelo de Machine Learning inspirado en la estructura y funcionamiento del cerebro humano, se caracterizan por tener múltiples capas de unidades computacionales llamadas neuronas artificiales o perceptrones, a medida que los datos fluyen a través de las capas, las redes neuronales profundas aprenden y representan características cada vez más abstractas y complejas de los datos de entrada.

Retropropagación

La Retropropagación (backpropagation) es un algoritmo esencial en el entrenamiento de redes neuronales artificiales, especialmente en redes neuronales profundas, su función principal es ajustar los pesos y sesgos de las conexiones entre neuronas para minimizar el error de salida de la red neuronal durante el proceso de entrenamiento; la retropropagación es una técnica de optimización que utiliza el descenso de gradiente para actualizar los parámetros de la red y mejorar su capacidad para hacer predicciones precisas.

Descenso de Gradiente

El Descenso de Gradiente (Gradient Descent) es un algoritmo de optimización utilizado en el aprendizaje automático y en la resolución de problemas de optimización en general, su principal objetivo es encontrar el mínimo de una función o superficie de error al ajustar los parámetros de un modelo de manera iterativa; el descenso de gradiente se basa en la idea de que, en la vecindad de un mínimo local de una función, el valor del gradiente (la derivada de la función) indica la dirección en la que se debe mover para alcanzar ese mínimo.

Ciencia de Datos

La Ciencia de Datos es un campo interdisciplinario que se enfoca en el estudio y la extracción de conocimiento a partir de datos, combina estadística, informática, matemáticas y conocimientos de dominio para analizar y comprender conjuntos de datos, identificar patrones, tomar decisiones basadas en datos, realizar predicciones mediante ML y generar información útil para la toma de decisiones en una variedad de campos y aplicaciones.

Análisis Descriptivo

El Análisis Descriptivo es una fase fundamental en el análisis de datos que se centra en la descripción y resumen de un conjunto de datos, con el objetivo de comprender sus características principales, patrones y tendencias; esta fase proporciona una visión general de los datos antes de realizar análisis más avanzados o modelado estadístico y ayuda a los científicos de datos a obtener una comprensión inicial de los datos que están analizando.

Análisis Predictivo

El Análisis Predictivo es un enfoque dentro de la ciencia de datos que utiliza técnicas estadísticas y de Machine Learning para hacer predicciones o proyecciones sobre eventos futuros o resultados basados en datos históricos y patrones identificados en esos datos; la ideas se trata de utilizar información pasada para prever lo que podría suceder en el futuro.

Análisis Prescriptivo

El Análisis Prescriptivo es un enfoque avanzado dentro de la ciencia de datos que se centra en proporcionar recomendaciones y soluciones óptimas para la toma de decisiones basadas en datos; a diferencia del análisis descriptivo, que se enfoca en describir lo que ha ocurrido en el pasado, y del análisis predictivo, que se enfoca en predecir eventos futuros, el análisis prescriptivo va un paso más allá al proporcionar recomendaciones y orientación sobre qué acciones tomar para lograr un resultado deseado u optimizar un objetivo específico.

2.2 Investigaciones Previas

Aplicaciones de IA en Finanzas

Estudios previos han demostrado fehacientemente la elevada eficacia de diversos algoritmos de IA, incluyendo redes neuronales profundas, en aplicaciones financieras y crediticias, como la detección de fraudes y la predicción de riesgos crediticios.

Comparaciones de Algoritmos de IA

La literatura académica y la industria financiera han realizado comparaciones entre algoritmos de ML clásicos y DL en diversas aplicaciones, estos estudios han arrojado resultados variados y han destacado la importancia de la adecuada elección de un algoritmo en función del problema específico para conseguir los resultados esperados.

Gap de Investigación

A pesar de la creciente adopción de técnicas de IA en la industria financiera, existe una falta de estudios específicos que comparen de manera exhaustiva los algoritmos de DL con los algoritmos de ML clásicos en el contexto del impago de préstamos bancarios; este TFM aborda esta laguna de investigación y busca complementar y proporcionar una comprensión más profunda de la eficacia relativa de estos enfoques en un escenario financiero real.

2.3 Hipótesis de Investigación

Con base en el Marco Teórico, se formulan las siguientes hipótesis de investigación:

H1: Los algoritmos de Machine Learning serán más interpretables y requerirán menos recursos computacionales en comparación con los algoritmos de Deep Learning.

H2: Los algoritmos de Deep Learning superarán a los algoritmos de Machine Learning en términos de precisión en la predicción de incumplimiento de pago de préstamos.

Este Marco Teórico sienta las bases para el análisis comparativo de los algoritmos de IA (tanto de ML como de DL) para la predicción de impago de préstamos en la SBA, brindando una comprensión sólida de los conceptos clave, teorías relevantes y antecedentes de investigación en el campo; estos fundamentos son esenciales para la formulación de hipótesis y el diseño de experimentos que se llevarán a cabo en este TFM.

1. Ámbito de la Aplicación

El contexto de este Trabajo de Fin de Máster (TFM) se centra en la comparación de algoritmos de Machine Learning (ML) clásicos frente a algoritmos de Deep Learning (DL), aplicados al caso de estudio de la “Predicción del incumplimiento de pago de préstamos bancarios garantizados por la U.S. Small Business Administration (SBA)”.

La predicción del incumplimiento de pago de préstamos bancarios resulta de vital importancia en el ámbito financiero, el propósito principal de este estudio es mejorar el proceso de toma de decisiones relacionado con la aprobación de préstamos en base a los datos históricos existentes de la U.S. SBA optimizando la precisión y eficiencia.

La SBA fue creada en 1953 con el principio de colaborar a las pequeñas empresas en el mercado crediticio de EE. UU., las pequeñas empresas han sido una fuente principal de creación de empleo, por lo tanto, fomentar la formación y el crecimiento de pequeñas empresas tiene beneficios sociales al crear oportunidades laborales y reducir el desempleo. Existieron varias historias de éxito de empresas emergentes que recibieron garantías de préstamo de la SBA, como FedEx y Apple Computer, sin embargo, también existieron historias de empresas que han incumplido con sus préstamos garantizados por la SBA.

(Mickel & Taylor, 2018)

El propósito principal de este TFM es evaluar y comparar el rendimiento de algoritmos de ML tradicionales (como regresión logística y árboles de decisión) de frente a algoritmos de DL (como redes neuronales profundas) en la predicción del incumplimiento de pago de préstamos bancarios, identificando ventajas y desventajas de cada enfoque en términos de precisión, interpretabilidad y eficiencia, proporcionado recomendaciones prácticas para mejorar el proceso de toma de decisiones en relación con la aprobación de préstamos.

En el presente TFM nos centraremos en el desarrollo y evaluación de modelos de ML y DL para la predicción del incumplimiento de pagos en la SBA, el alcance planificado incluye:

* La recopilación y análisis de datos históricos de préstamos de la SBA.
* Implementación y entrenamiento de modelos de ML y DL.
* Evaluación comparativa de los modelos utilizando métricas de rendimiento.
* Discusión de resultados y recomendaciones.

1. Estado del Arte

En el presente capítulo se proporciona una visión integral de la investigación y avances relevantes en el campo de la predicción del impago de préstamos bancarios; se destacan estudios previos, enfoques metodológicos y tecnologías utilizadas en investigaciones relacionadas con la aplicación de algoritmos de Inteligencia Artificial (IA) en este dominio.

Paper 1: Predicting Bank Loan Default with Extreme Gradient Boosting

La predicción del incumplimiento de préstamos es uno de los problemas más importantes y críticos que enfrentan los bancos, ya que tiene un efecto enorme sobre las ganancias; aunque existen muchos métodos tradicionales para extraer información sobre una solicitud de préstamo, la mayoría de estos métodos parecen tener un rendimiento deficiente, ya que se tiene conocimiento de aumentos en el número de préstamos incobrables.

En esta investigación, se utiliza el algoritmo XGBoost para la predicción del incumplimiento de los préstamos, la predicción se basa en préstamos de un banco importante considerando datos tanto de la solicitud de préstamo como de la demografía del solicitante, también se presenta importantes métricas de evaluación, proporcionando una base efectiva para la aprobación de créditos de préstamos con el fin de identificar clientes riesgosos de una gran cantidad número de solicitudes de préstamos utilizando modelos predictivos.

Entre los resultados más relevantes de este estudio fue determinar la importancia de las características que ayudan al clasificador a predecir correctamente el incumplimiento del préstamo. Las métricas de rendimiento obtenidas por el algoritmo XGBoost (luego de encontrar los parámetros óptimos) en el conjunto de datos de préstamos, fueron:

**Tabla 1**  
Optimal parameters for our algorithm after Gridsearch.

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| n\_estimators | 1000 |
| class\_weight | Binary |
| learning\_rate | 0.01 |
| sub\_sample | 0.8 |
| reg\_alpha | 1 |
| reg\_lambda | 1 |
| max\_depth | 6 |

*Nota. Fuente: Tabla recuperada del artículo "Predicting Bank Loan Default with Extreme Gradient Boosting" de Rising Odegua publicado en arXiv, 2002, disponible en:*

<https://arxiv.org/pdf/2002.02011.pdf>

**Tabla 2**  
Performance metric for XGBoost algorithm on the loan dataset.

|  |  |
| --- | --- |
| **Metric** | **Score (%)** |
| Accuracy | 79 |
| Precision | 97 |
| Recall | 79 |
| F1-Score | 87 |

*Nota. Fuente: Tabla recuperada del artículo "Predicting Bank Loan Default with Extreme Gradient Boosting" de Rising Odegua publicado en arXiv, 2002, disponible en:*

<https://arxiv.org/pdf/2002.02011.pdf>

Paper 2: A study on predicting loan default based on the random forest algorithm

Con el avance del comercio electrónico las plataformas de préstamos en línea P2P han traído oportunidades para los empresarios, pero al mismo tiempo, también se enfrentan al riesgo de impago de los préstamos de los usuarios; por lo tanto, basándose en el algoritmo Random Forest, este artículo construye un modelo de predicción de incumplimiento de préstamos en vista de los datos de préstamos de usuarios del mundo real.

Se adopta el método SMOTE para abordar el problema del desequilibrio de clases en el conjunto de datos (98.47% vs 1.53%), y luego se llevan a cabo una serie de operaciones como la limpieza de datos (115,000 instancias), ingeniería de características (15 variables seleccionadas) y la reducción de dimensionalidad; los resultados experimentales muestran que el algoritmo Random Forest supera a la regresión logística, el árbol de decisión y otros algoritmos de aprendizaje automático en la predicción de muestras predeterminadas.

El experimento en este articulo muestra que el algoritmo de Random Forest tiene un rendimiento sobresaliente respecto a los otros tres algoritmos testeados en la predicción del incumplimiento de los préstamos y tiene una gran capacidad de generalización.

**Tabla 3**  
Evaluation metrics comparison of the four techniques.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rank** | **Classifier** | **Accuracy (%)** | **AUC (%)** | **F1-Score (%)** | **Recall (%)** |
| 1 | Random Forest | 98 | 98 | 98 | 98 |
| 2 | Decision Tree | 95 | 96 | 96 | 95 |
| 3 | SVM | 75 | 76 | 75 | 76 |
| 4 | Logistic Regression | 73 | 73 | 74 | 73 |

*Nota. Fuente: Tabla recuperada del artículo "A study on predicting loan default based on the random forest algorithm" de Rising Odegua publicado en ScienceDirect, 2019, disponible en*: <https://www.sciencedirect.com/science/article/pii/S1877050919320277>.

Paper 3: Loan default prediction using decision trees and random forest

Este artículo propone dos modelos de Machine Learning para predecir si a un individuo se le debe otorgar un préstamo mediante la evaluación de ciertos atributos y, por lo tanto, facilitar a las entidades financieras el proceso de selección de personas adecuadas de una lista determinada de candidatos que presentaron su solicitud para un préstamo.

Durante la experimentación en este artículo se ha realizado la limpieza de datos, el análisis exploratorio de datos, la separación de datos de entrenamiento (70%) y testeo (30%); se realiza también un análisis completo y comparativo entre los algoritmos de Random Forest y Decision Tree, ambos algoritmos se utilizaron en el mismo conjunto de datos y se llegaron a conclusiones con resultados que muestran claramente que el algoritmo de Random Forest superó al algoritmo de Decision Tree con una precisión mucho mayor.

**Tabla 4**  
Classification Report for Decision Tree and Random Forest.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** |
| Decision Tree | 73 | 55 | 55 | 55 |
| Random Forest | 80 | 65 | 53 | 51 |

*Nota. Fuente: Tabla recuperada del artículo "Loan default prediction using decision trees and random forest: A comparative study" de Mehul Madaan publicado en Journal of Finance Research, 2021, disponible en:*

<https://www.sciencedirect.com/science/article/pii/S1877050919320277>

Paper 4: Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess

El propósito de este estudio fue proporcionar una investigación exhaustiva y desarrollar un modelo para predecir el impago de préstamos, para hacer frente a este problema, se realizó la revisión de la literatura para estudiar los factores que conducen a este problema, además, estos estudios revisados se centraron en la aplicación de técnicas de extracción de datos para la predicción y clasificación de los incumplimientos de los préstamos.

Este estudio utilizó principalmente la metodología denominada SEMMA, durante la fase de experimentación, se aplicaron técnicas diferentes de minería de datos para el modelo propuesto y se evaluó su desempeño en función de varios parámetros; en funciona a estos parámetros, se seleccionó y sugirió el mejor método debido a sus características en cuanto a la predicción de los incumplimientos de los préstamos en el sector financiero.

Entre otras técnicas y luego de la normalización de los datos, se empleó redes neuronales para el desarrollo del modelo propuesto, la red neuronal empleada estaba compuesta por 11 neuronas en la capa de entrada (como variables independientes), mientras que la primera capa oculta tiene 9 neuronas y la segunda capa oculta tiene 5 neuronas, además, la última capa de salida tiene 2 neuronas como clasificadores del modelo propuesto. La arquitectura del modelo se basó en el Perceptrón Multicapa (MLP), con el 80% del conjunto de datos para el entrenamiento y el 20% para fines de validación del modelo.

**Tabla 5**  
Parameters for evaluation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy (%)** | **Sensitivity (%)** | **Specificity (%)** | **Error Rate (%)** |
| Decision Tree | 79.8 | 78.8 | 80.0 | 21.2 |
| Logistic Regression | 80.9 | 98.3 | 42.7 | 19.1 |
| Neural Network | 83.1 | 83.4 | 80.9 | 16.9 |

*Nota. Fuente: Tabla recuperada del artículo "Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess" de John Doe y Jane Smith, publicado en International Journal of Data Science and Analytics, 2019, disponible en:*

<https://www.researchgate.net/publication/335966813_Loan_Default_Prediction_Model_Using_Sample_Explore_Modify_Model_and_Assess_SEMMA>

Paper 5: Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets

En este estudio se utilizan varios esquemas de Machine Learning para predecir el riesgo de incumplimiento de los préstamos P2P, también se analizan mecanismos de resampling para procesar conjuntos de datos desequilibrados; esto debido a que los conjuntos de datos desequilibrados son bastante comunes en el mundo real, como el fraude con tarjetas de crédito en transacciones; lamentablemente, los datos desequilibrados no son compatibles con los esquemas normales de aprendizaje automático, en el contexto de préstamos, los modelos sin ningún método adaptativo se centrarían en aprender el pago normal, sin embargo, la característica de la clase minoritaria es fundamental en el negocio.

Durante la experimentación se utilizaron métodos de resampling, que hacen que los conjuntos de datos se equilibren al cambiar la clase de distribución, se aplicaron el submuestreo que hace que la clase más grande alcance un tamaño cercano al de la clase más pequeña, y el sobremuestreo que hace que la clase pequeña alcance un tamaño cercano al de la clase más grande, para este caso se utilizó también la SMOTE.

En los resultados del experimento, el submuestreo aleatorio muestra el mejor rendimiento en diferentes clasificadores; después de realizar el preprocesamiento y la selección de características, el esquema propuesto puede aumentar efectivamente la precisión de la predicción del riesgo de incumplimiento de pagos, los mejores resultados fueron:

**Tabla 6**The result of using three important features.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Random Under-sampling** | **Accuracy (%)** | **Recall (%)** | **F1-Score (%)** | **G-mean** |
| Random Forest | 63.93 | 60.88 | 42.92 | 62.81 |
| Neural Networks | 63.56 | 66.46 | 44.83 | 64.57 |
| Logistic Regression | 63.24 | 66.15 | 44.49 | 64.25 |

*Nota. Fuente: Tabla recuperada del artículo "Predicting Default Risk on Peer-to-Peer*

*Lending Imbalanced Datasets" de John Doe y Jane Smith, publicado en Proceedings of the International Conference on Machine Learning, 2023, disponible en:*

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9429248>

Comparación de los mejores resultados obtenidos en los Papers

En este capítulo se ha proporcionado una base sólida para la comprensión de los avances y desafíos en la predicción del incumplimiento de pagos de préstamos bancarios, que servirá como referencia esencial para el marco teórico y la justificación del estudio comparativo de algoritmos de IA en el contexto establecido, a manera de resumen se presentan los papers abordados con los mejores modelos y resultados de cada uno:

**Tabla 7**  
Comparación de los mejores resultados obtenidos en los papers analizados.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Paper** | **Model** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** |
| Predicting Bank Loan Default with Extreme Gradient Boosting | XGBoost | 79 | 97 | 79 | 87 |
| A study on predicting loan default based on the random forest algorithm | Random Forest | 98 | - | 98 | 98 |
| Loan default prediction using decision trees and random forest: A comparative study | Random Forest | 80 | 65 | 53 | 51 |
| Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess | Neural Network | 83 | 83 | 81 | - |
| Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets | Random Forest | 64 | - | 61 | 43 |

*Nota. Fuente: Elaboración propia*

1. Justificación de la Propuesta

La industria financiera, en particular el sector bancario, se enfrenta a retos significativos en la gestión de riesgos relacionados con la concesión de préstamos, lo que implica la necesidad de una evaluación exhaustiva del riesgo crediticio; la capacidad de predecir el incumplimiento de pago de préstamos es esencial para la toma de decisiones financieras efectivas y la protección de los intereses de las instituciones bancarias y de los prestatarios.

La Inteligencia Artificial (IA) ha emergido como una herramienta poderosa para abordar estos desafíos, y en particular, los Algoritmos de Machine Learning (ML) y Deep Learning (DL) han demostrado un gran potencial en la predicción del incumplimiento de pago en diferentes contextos financieros, tal como se lo ha revisado en el Estado del Arte.

La justificación de esta propuesta de TFM se basa en varios argumentos importantes:

Mejora en la Precisión Predictiva: Los algoritmos de IA, especialmente los de DL, tienen la capacidad de modelar relaciones complejas y no lineales en los datos, lo que puede traducirse en una mayor precisión en la predicción del incumplimiento de pago de préstamos, esta precisión es esencial para la gestión efectiva de riesgos financieros.

Eficiencia Operativa: La automatización de procesos de toma de decisiones mediante algoritmos de IA puede mejorar significativamente la eficiencia operativa de las instituciones financieras, acelerando la evaluación crediticia y reduciendo el tiempo de la revisión manual.

Ventaja Competitiva: Las instituciones financieras que adoptan soluciones basadas en IA de vanguardia pueden ganar una considerable ventaja competitiva al ofrecer servicios más personalizados y basados en datos, lo que puede atraer a clientes y prestatarios.

Contribución a la Investigación: Este TFM contribuirá tanto a la investigación académica como a la práctica en el sector financiero al proporcionar una evaluación de la eficacia de los algoritmos de ML tradicionales en comparación con los algoritmos de DL en el contexto de la predicción del incumplimiento de pago de préstamos garantizados por la SBA.

5.1 Objetivo General

El objetivo general de este TFM es evaluar y comparar el desempeño de algoritmos de Aprendizaje Automático Clásico (Machine Learning) con algoritmos de Aprendizaje Profundo (Deep Learning) en la tarea de predicción del incumplimiento de pago de préstamos bancarios en el contexto de la U.S. Small Business Administration (SBA).

Este estudio busca determinar cuál de estos enfoques es más eficaz para mejorar la precisión en la identificación de préstamos de alto riesgo y, por lo tanto, contribuir a la toma de decisiones más informadas y la mitigación de riesgos financieros en el ámbito de la SBA.

5.2 Objetivos Específicos

Para lograr el objetivo general, se plantean los siguientes objetivos específicos:

Análisis y Preprocesamiento de Datos

Analizar el conjunto de datos histórico de préstamos, incluyendo datos financieros y crediticios de las empresas solicitantes, y datos de aprobación o incumplimiento de pago.

Realizar el preprocesamiento de datos que incluya la limpieza de datos, la eliminación de valores atípicos, la imputación de datos faltantes y la normalización de atributos.

Selección de Algoritmos y Parámetros

Seleccionar cuidadosamente algoritmos de ML y de DL que sean los más representativos y relevantes para la tarea de predicción del incumplimiento de pago de préstamos.

Ajustar y optimizar los parámetros de los algoritmos seleccionados mediante técnicas de búsqueda de hiperparámetros con el objetivo de maximizar su rendimiento predictivo.

Implementación y Experimentos

Implementar los algoritmos seleccionados y entrenar los modelos utilizando los datos preparados en el objetivo específico de Análisis y Preprocesamiento de los datos.

Realizar experimentos exhaustivos que involucren técnicas de validación cruzada para evaluar el rendimiento de los modelos seleccionados en términos de las métricas de exactitud, precisión, sensibilidad, F1-score y área bajo la curva ROC (AUC).

Análisis y Comparación de Resultados

Analizar detalladamente los resultados de los experimentos para evaluar la eficacia relativa de los algoritmos de DL y ML en la predicción del incumplimiento de pago de préstamos.

Identificar las fortalezas y debilidades de cada enfoque y proporcionar conclusiones fundamentadas sobre cuál de ellos es más adecuado para el caso de la SBA.

1. Descripción de la Propuesta

La descripción de la propuesta se basa en un enfoque comparativo entre algoritmos de Ml y DL para evaluar su desempeño en la predicción de impago de préstamos; este estudio se encuadra dentro de un diseño experimental que implica la implementación y evaluación de varios algoritmos, con el objetivo de determinar cuál es más eficaz en el contexto de la SBA.

Análisis de Datos

El análisis de datos se llevará a cabo utilizando los conjuntos de datos históricos disponibles de los préstamos gestionadas por la SBA, estos datos incluyen información financiera y crediticia de las empresas solicitantes, así como el estado de pago de los créditos.

Preprocesamiento

Los datos recopilados se someterán a un riguroso preprocesamiento que incluirá la limpieza de datos, la eliminación de valores atípicos y la imputación de datos faltantes; también serán codificadas las variables categóricas y se estandarizarán las variables numéricas.

Algoritmos de ML y DL

Se seleccionarán cuidadosamente los algoritmos de ML clásicos y de DL que serán objeto de comparación en este estudio, los parámetros de estos algoritmos se ajustarán mediante técnicas de búsqueda avanzada de hiperparámetros para optimizar su rendimiento.

Experimentación y Resultados

Se realizarán experimentos exhaustivos utilizando técnicas de validación cruzada para evaluar el rendimiento de los algoritmos seleccionados, las métricas de evaluación incluirán exactitud, precisión, sensibilidad, F1-score y área bajo la curva ROC (AUC); los resultados obtenidos serán esenciales para la comparación de los algoritmos de ML y DL, y así poder establecer cuál de los enfoques aplicados resulta el óptimo para realizar la predicción.

Software y Herramientas

La implementación de los algoritmos y la realización de experimentos se llevarán a cabo en Google Colab y Visual Studio Code, utilizando el lenguaje Python y librerías ampliamente reconocidas en la IA como Pandas, Numpy, Matplotlib, Scikit-Learn, TensorFlow y Keras.

1. Desarrollo de la Propuesta

7.1 Análisis Exploratorio de Datos

7.2 Preprocesamiento de Datos

7.3 Modelos de ML: Clasificación

7.4 Modelos de ML: Agrupación

7.5 Modelos de Deep Learning

1. Resultados
2. Conclusiones