**National Research University Higher School of Economics**

**Institute for Statistical Studies and Economics of Knowledge**

**MASTER THESIS**

**Comparative Analysis of Ensemble Methods for Credit Risk Assessment in the IT Sector**

**Student:**

*Vivas Teran Katherin Carolina*

**Group:**

*МСФ231*

**Supervisor:**

*Senior Research Fellow, Sergei Grishunin*

**Moscow, 25th April 2025**

**Table of Contents**

**• Introduction.**

**• Chapters, Sub-chapters.**

**• Sub-sub-chapters.**

**• Conclusion.**

**• References.**

**• Appendix.**

**It should include the page number for each section.**

Format: Times New Roman 14, spacing 1.5, justified, space between paragraphs removed.

**List of Abbreviations**

• All abbreviations used in main text

• Format: Times New Roman 14, spacing 1.5, justified, space between paragraphs removed.

**List of Tables**

• All tables included in document (main text and appendixes/ annexes).

• Format: Times New Roman 14, spacing 1.5, justified, space between paragraphs removed.

• Table Number: Table Title.

**List of Figures**

• All tables included in document (main text and appendixes/ annexes).

• Format: Times New Roman 14, spacing 1.5, justified, space between paragraphs removed.

• Figure Number: Figure Title.

**Summary**

Esta tesis se centra en un análisis comparativo de métodos ensemble para la evaluación del riesgo crediticio específicamente en el sector de la tecnología de la información (TI). Se examinan diversos métodos ensemble, incluyendo Random Forest, Gradient Boosting y Stacking, con el objetivo de identificar aquellos que ofrecen el rendimiento más eficaz para predecir la solvencia crediticia de las empresas de TI. El sector de TI presenta desafíos únicos en la evaluación del riesgo crediticio, caracterizados por la rápida innovación, la naturaleza intangible de sus activos y la volatilidad inherente a su crecimiento. Estos desafíos requieren modelos que puedan adaptarse a dinámicas complejas y no lineales. La metodología empleada en este estudio implica la recopilación de datos relevantes del sector de TI, su preprocesamiento exhaustivo y la aplicación de los métodos ensemble seleccionados. El rendimiento de estos modelos se evalúa utilizando métricas apropiadas para la clasificación del riesgo crediticio. Los hallazgos clave destacan los métodos ensemble que demuestran una mayor precisión y robustez en la predicción del riesgo crediticio en el sector de TI, así como sus ventajas y desventajas específicas en este contexto. La investigación concluye con la importancia de estos hallazgos para mejorar las prácticas de evaluación del riesgo crediticio en el sector de TI, contribuyendo a decisiones financieras más informadas y a la estabilidad del sector. Se utilizan técnicas y modelos matemáticos como árboles de decisión, bagging, algoritmos de boosting, descenso de gradiente y métricas de evaluación como precisión, recall, AUC-ROC y F1-score.

**Introduction**

1.1 Antecedentes

La evaluación de riesgo de una empresa es uno de los procesos sistemáticos más importantes en el sector financiero ya que este permite a las instituciones tomar decisiones más informadas al momento de conceder créditos y por otra parte también reduce la exposición a posibles pérdidas. Contar con una evaluación más precisa sobre la capacidad de una empresa para solventar sus deudas contraídas, es una tarea de suma importancia de cara a una búsqueda de estabilidad financiera y crecimiento económico. Es por eso que hoy en día el uso de métodos robustos tiene un papel fundamental en el contexto de la tan cambiante economía global y mercados financieros que no paran de evolucionar y adaptarse a las nuevas tecnologías.

En este contexto, el sector de las tecnologías de la información (TI) se ha erigido en uno de los principales motores de la economía mundial, impulsando la innovación y cambiando muchas facetas de la sociedad. La creciente importancia de las empresas de TI pone de relieve la necesidad de una evaluación eficaz del riesgo de crédito adaptada a las características únicas de este sector. El auge de la tecnología financiera y de los modelos de negocio impulsados por la tecnología subraya la importancia crítica de comprender el riesgo de crédito en este dinámico sector.

Por otro lado, la evaluación de riesgo crediticio en este sector cuenta con muchas particularidades y esto se debe a la procedencia de sus activos, ya que en su mayoría suelen estar constituidos por patentes de propiedad intelectual y otros activos intangibles. Además, otro punto a tomar en cuenta al momento de hacer la evaluación de riesgo es la velocidad con la que se producen los avances tecnológicos y cambios en este mercado lo que lleva a las empresas en este sector a sufrir cambios abruptos en su estabilidad financiera. Un ejemplo claro de este son los startups en este sector que, al inicio pueden mostrarse con un gran potencial de crecimiento pero que también viene acompañado de una gran volatilidad. Otro problema que se debe considerar es la falta de datos históricos específicos de este sector, lo que hace la tarea de evaluación aún más complicada sobre todo con los métodos tradicionales. Adicionalmente es precisamente en este sector donde más se puede ver como la falta de innovación o el estancamiento creativo puede llevar a sus empresas a un rápido descenso debido a la falta de competencia frente a otras empresas que quizás cuentan con mejores opciones de cara a un mercado tan exigente como el de nuestros días. Es por eso que los modelos ha ser usados para la evaluación deben también tener en consideración esta volatilidad que viene por default en este sector.

1.2 Planteamiento del Problema

El principal problema al momento de hacer una evaluación en este sector es que los métodos tradicionales a menudo usan indicadores financieros y métricas tradicionales que pueden llegar a reflejar de manera muy vaga o con poca precisión los verdaderos riesgos que conllevan las empresas de tecnología.

¿Qué son los métodos ensamble?

Es aquí donde los métodos ensamble juegan un papel importante ya que combinan múltiples modelos de aprendizaje, lo que les permite ser mucho más eficaces al momento de enfrentar una tarea de clasificación y predicción, que es precisamente lo que se necesita para la evaluación de riesgo crediticio de un sector tan volátil. Y es que es precisamente a su capacidad para afrontar tareas con relaciones complejas que los convierte en los candidatos ideales para esta tarea. Actualmente existen múltiples métodos ensamble como lo son el método Random Forest, Bagging, Gradient Boosting y también sus variantes como lo son LightGBM y XGBoost. Cada uno de ellos tiene un gran potencial al momento de resolver tareas de clasificación pero es por eso que es importante realizar un minucioso análisis comparativo para determinar cual es el método más adecuado y que captura de mejor manera las características del sector que se está evaluando.

1.3 Objetivos

El objetivo principal de esta tesis es llevar a cabo un análisis comparativo de diversos métodos ensemble para la evaluación del riesgo crediticio en el sector de TI. Los objetivos secundarios incluyen:

Como objetivo principal de esta investigación es realizar un análisis comparativo de varios métodos ensamble para la evaluación de riesgo crediticio en el sector de IT. Como objetivos secundarios están:

* Identificar el método ensemble más eficiente, tomando como base pare la comparación diversas métricas de rendimiento.
* Analizar las ventajas y desventajas de los distintos métodos ensemble específicamente para la evaluación del riesgo crediticio en el sector de TI.
* Identificar que variables son más relevantes para poder representar las características únicas de las empresas de IT y como eso influencia en el rendimiento de los distintos métodos.
* Evaluar la influencia de las características macroeconómicas en las distintas regiones como Europa y Latinoamérica.
* Proveer recomendaciones sobre el uso de métodos ensamble para el sector IT.

1.4 Alcance y Limitaciones

El alcance de esta investigación incluye el análisis de los principales métodos ensemble utilizados en el aprendizaje automático, como Random Forest, Gradient Boosting (incluyendo sus variantes XGBoost y LightGBM) y Stacking. El estudio se centrará específicamente en su aplicación para la evaluación del riesgo crediticio en el sector de TI, pudiendo considerar segmentos específicos como el desarrollo de software o los servicios SaaS si los datos lo permiten. El alcance geográfico o de mercado del estudio se definirá en función de la disponibilidad de datos relevantes.

Las limitaciones de esta investigación pueden incluir la disponibilidad y calidad de datos etiquetados sobre el riesgo crediticio en el sector de TI. También se reconoce el potencial de sesgo en los datos o en los modelos desarrollados. Los recursos computacionales necesarios para el entrenamiento y la evaluación de modelos ensemble complejos pueden representar una limitación. Además, algunos métodos ensemble pueden presentar desafíos en cuanto a la interpretabilidad de sus resultados. Finalmente, la generalización de los hallazgos a todos los subsectores dentro de la amplia industria de TI podría ser limitada.

1.5 Significado del Estudio

Esta investigación reviste gran importancia tanto para las instituciones financieras como para las propias empresas de TI, al abordar la necesidad de una evaluación mejorada del riesgo crediticio en este sector crucial. Los potenciales aportes de este estudio son significativos, incluyendo la provisión de evidencia empírica sobre la eficacia de los métodos ensemble en este dominio específico y el ofrecimiento de orientación práctica para la selección e implementación de modelos apropiados. Además, esta tesis contribuye a la literatura académica sobre la evaluación del riesgo crediticio y el aprendizaje ensemble. En última instancia, podría conducir a decisiones crediticias más precisas y eficientes, fomentando el crecimiento y la innovación en el sector de TI al tiempo que se gestiona el riesgo financiero. El estudio se caracteriza por su rigor y profundidad analítica, centrándose en la naturaleza comparativa y en el uso de los fundamentos matemáticos de los métodos ensemble y las métricas de evaluación.

**Literature Review**

2.1 Visión General de la Evaluación de Riesgo Crediticio

Los métodos tradicionales de evaluación del riesgo crediticio han incluido durante mucho tiempo el juicio de expertos y los sistemas basados en reglas. El método de análisis de los cinco pilares del crédito (carácter, capacidad, capital, colateral y condiciones) ha sido una herramienta fundamental para evaluar la solvencia de los prestatarios. Además, se han utilizado ampliamente modelos estadísticos como la regresión logística y el análisis discriminante para predecir la probabilidad de incumplimiento. Los modelos de puntuación crediticia, como FICO y VantageScore, proporcionan una representación numérica de la solvencia crediticia de un individuo o una empresa.

Sin embargo, estos métodos tradicionales presentan limitaciones, especialmente en el contexto del sector de TI. Su capacidad para capturar relaciones no lineales e interacciones complejas en los datos puede ser limitada. El sector de TI genera grandes cantidades de datos en constante evolución, lo que puede dificultar el manejo con modelos tradicionales. Los sistemas basados en el juicio de expertos pueden ser propensos a sesgos y subjetividad. Además, los métodos tradicionales pueden tener dificultades para adaptarse a los rápidos cambios y a las características únicas de las empresas de TI, como sus modelos de negocio innovadores y la importancia de los activos intangibles.

2.2 Técnicas de Ensemble Learning

Las técnicas de ensemble learning representan un enfoque poderoso en el aprendizaje automático que busca mejorar el rendimiento predictivo combinando las salidas de múltiples aprendices individuales. La premisa fundamental es que un conjunto de modelos puede superar el rendimiento de cualquiera de sus componentes individuales. Existen varias categorías principales de métodos ensemble:

* **Bagging (Bootstrap Aggregating):** Esta técnica implica la creación de múltiples subconjuntos del conjunto de datos original mediante un muestreo con reemplazo (bootstrap) y el entrenamiento de un modelo base (a menudo un árbol de decisión) en cada subconjunto. Las predicciones finales se obtienen agregando las predicciones de los modelos base, por ejemplo, mediante votación para clasificación o promediado para regresión. Random Forest es un ejemplo popular de algoritmo de bagging que construye una colección de árboles de decisión, cada uno entrenado en un subconjunto aleatorio de características y datos. Las divisiones de los árboles de decisión se basan en criterios matemáticos como la impureza de Gini o la ganancia de información. Random Forest es eficaz para manejar datos de alta dimensión, no lineales y heterogéneos, y proporciona medidas de importancia de las características. Sin embargo, puede ser costoso computacionalmente y la interpretabilidad del "bosque" de árboles puede ser compleja.
* **Boosting:** Los algoritmos de boosting entrenan una secuencia de modelos base de forma iterativa, donde cada modelo sucesivo intenta corregir los errores cometidos por los modelos anteriores. Las predicciones finales se obtienen mediante una combinación ponderada de las predicciones de los modelos base. Gradient Boosting, XGBoost y LightGBM son ejemplos destacados de algoritmos de boosting. Estos métodos construyen árboles de decisión secuencialmente, utilizando el descenso de gradiente para optimizar una función de pérdida. Son conocidos por su alta precisión y capacidad para manejar datos complejos y proporcionar información sobre la importancia de las características. Sin embargo, pueden ser sensibles al ruido, propensos al sobreajuste y la interpretabilidad puede ser un desafío.
* **Stacking:** Esta técnica implica el uso de un meta-aprendiz para combinar las predicciones de varios modelos base. Los modelos base se entrenan en el conjunto de datos original, y sus predicciones se utilizan como entrada para entrenar el meta-aprendiz. Stacking puede aprovechar las fortalezas de diferentes tipos de modelos, pero añade complejidad y puede aumentar el riesgo de sobreajuste. El proceso de entrenamiento del meta-aprendiz a menudo implica el uso de técnicas de validación cruzada.

2.3 Estudios Previos sobre Métodos de Ensemble

La literatura existente demuestra una aplicación significativa de métodos ensemble en la evaluación del riesgo crediticio en diversos dominios. Muchos estudios han encontrado que los métodos ensemble, como Random Forest y Gradient Boosting (especialmente XGBoost y LightGBM), superan a los métodos tradicionales en la precisión de la predicción del riesgo crediticio. Estos estudios han empleado una variedad de conjuntos de datos y métricas de evaluación para comparar el rendimiento de diferentes enfoques. Sin embargo, la investigación que se centra específicamente en la aplicación de métodos ensemble para la evaluación del riesgo crediticio dentro del sector de TI parece ser limitada. La mayoría de los estudios revisados se centran en sectores como la banca minorista o las pequeñas y medianas empresas (PYMEs) en general. Esta brecha en la investigación subraya la necesidad de un análisis más específico en el contexto del sector de TI, dadas sus características únicas.

2.4 Desafíos Específicos del Sector de TI

La evaluación del riesgo crediticio en el sector de TI se enfrenta a varios desafíos específicos. El rápido ritmo de la innovación y los cortos ciclos de vida de los productos hacen que los datos históricos sean menos predictivos de la viabilidad futura. Las empresas de TI a menudo poseen una gran proporción de activos intangibles, como la propiedad intelectual, cuyo valor puede ser difícil de evaluar utilizando métricas financieras tradicionales. El alto potencial de crecimiento, especialmente entre las startups de TI, viene acompañado de una considerable volatilidad e incertidumbre. La dependencia del talento clave y la protección de la propiedad intelectual son factores cruciales que pueden influir en el riesgo crediticio de las empresas de TI. Además, las amenazas a la ciberseguridad y las preocupaciones sobre la privacidad de los datos pueden tener un impacto significativo en la estabilidad financiera de las empresas de TI. La naturaleza global de muchas empresas de TI y sus operaciones internacionales también introducen complejidades adicionales en la evaluación del riesgo crediticio.

Desde una perspectiva matemática, los métodos ensemble pueden ser particularmente adecuados para abordar estos desafíos. Su capacidad para modelar relaciones complejas y no lineales puede capturar la naturaleza dinámica del sector de TI. Las capacidades de importancia de las características pueden ayudar a identificar indicadores clave de riesgo específicos de las empresas de TI, como la inversión en I+D o la tasa de abandono de clientes. La robustez ante datos ruidosos podría ser beneficiosa dada la potencialidad de fuentes de datos menos estructuradas o alternativas en el sector de TI.

**Methodology and Approach**

3.0 Descripción de los Datos

La base de datos utilizada en esta investigación está compuesta por información financiera, de mercado y macroeconómica de 107 empresas tecnológicas que cotizan en bolsa, provenientes de 24 países de Estados Unidos, América Latina, Europa y Asia, y que representan distintos niveles de capitalización de mercado y desempeño crediticio. Los datos abarcan un periodo de seis años (2018–2023), lo que permite realizar un análisis longitudinal y comparativo a lo largo del tiempo. as variables se clasifican en cinco grandes grupos: generales, financieras, de mercado, macroeconómicas, y estructurales y ESG.

Cada observación en el dataset corresponde a un año específico de una empresa determinada, con variables agrupadas en las siguientes categorías:

1. Variables Generales de Identificación

Estas variables permiten ubicar y diferenciar cada observación dentro del conjunto de datos:

* Nombre de la empresa (Company)
* País de origen (Country)
* Código bursátil (Ticker)
* Año (Year)

2. Variables Financieras

Se subdividen en indicadores de desempeño operativo y financiero:

* Estado de resultados: ingresos (Revenue), beneficio neto (Net Income), EBITDA, EBIT, gasto de capital (CAPEX), gasto en investigación y desarrollo (R&D expense), margen neto, e intereses netos.
* Balance general: activos totales, pasivos totales, deuda total, deuda neta, patrimonio total, efectivo y equivalentes, valores negociables y razón rápida (quick ratio).
* Ratios financieros: retorno sobre el capital invertido (ROIC), retorno sobre el patrimonio (ROE), retorno sobre activos (ROA), rotación de activos, diversos indicadores de apalancamiento (debt/equity, debt ratio, debt to EBITDA, entre otros), y coberturas financieras como (EBIT – CAPEX) / Intereses.
* Transformaciones logarítmicas: logaritmo de activos totales y logaritmo natural de los ingresos, empleados como mecanismos de normalización.

3. Indicadores de Mercado

Incluyen información bursátil relevante para evaluar el valor percibido de las empresas por parte del mercado:

* Capitalización bursátil (Market Cap)
* Volumen de acciones transadas
* Ratios como Market Cap to Revenue, Market Cap to EBITDA, y relación Precio/Valor Contable (P/B)

4. Variables Macroeconómicas

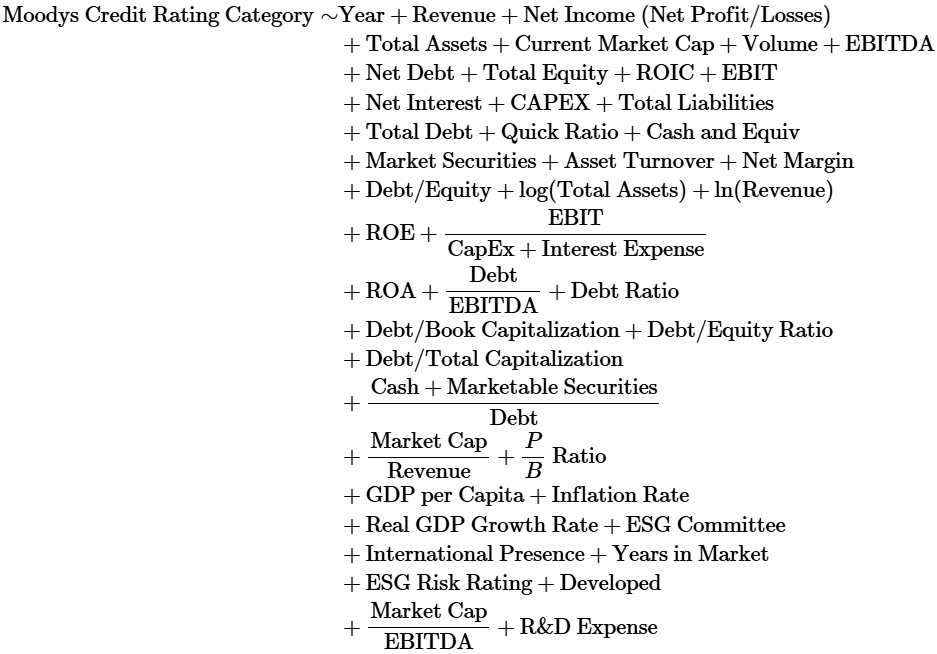
Proporcionan el contexto económico del país sede de cada empresa:

* PIB per cápita
* Tasa de inflación
* Crecimiento real del PIB

5. Factores Estructurales y ESG

Permiten caracterizar el entorno institucional y el enfoque de sostenibilidad de cada empresa:

* Existencia de comité ESG
* Presencia internacional (nacional o también en el extranjero)
* Años en el mercado
* Calificación de riesgo ESG
* Calificación crediticia otorgada por Moody’s
* Clasificación del país como desarrollado o no desarrollado

(1)

Este tipo de información permite combinar enfoques tradicionales de análisis financiero con metodologías modernas de machine learning y modelos ensemble, incorporando también el contexto macroeconómico del país donde opera cada empresa.

El conjunto de datos incluye tanto variables numéricas como categóricas. Dentro de las variables categóricas, se destacan aquellas que permiten codificar información cualitativa relevante sobre las empresas y su contexto operativo. Estas variables son:

* Country: país sede de la empresa, con 24 categorías únicas.
* Company: nombre de la empresa, con 107 compañías únicas.
* Ticker: identificador bursátil, con 112 códigos distintos.
* ESG\_Committee: variable binaria que indica la existencia (1) o no (0) de un comité ESG dentro de la empresa.
* International\_presence: variable binaria que indica si la empresa tiene operaciones en el extranjero (1) o no (0).
* Moody’s Credit Rating: variable ordinal con 20 niveles distintos de calificación crediticia, desde rangos de inversión hasta especulativos (ej. Aa3, Baa1, Ba2, etc.).
* Developed: variable binaria que indica si el país de origen está clasificado como desarrollado (1) o no (0).

Estas variables categóricas son relevantes para analizar diferencias estructurales entre empresas, particularmente en temas de gobernanza, internacionalización, y riesgo crediticio.

En cuanto a los valores faltantes, se identifican diferentes proporciones según la variable:

* Las variables clave como Revenue, Net Income, Total Assets, EBITDA, CAPEX y Debt presentan una cobertura superior al 99%, con un máximo de 3 valores nulos por variable.
* Variables de mercado como Current Market Cap, Volume y Market Cap to EBITDA muestran una mayor cantidad de valores faltantes. Por ejemplo, Market Cap to EBITDA tiene 25 valores nulos (3.9% del total), mientras que R&D expense presenta 77 valores faltantes (12% del total), lo que puede atribuirse a la naturaleza confidencial o irregular de este tipo de gastos en ciertas empresas.
* Variables cualitativas como Moody’s Credit Rating pueden también presentar valores ausentes cuando la empresa no cuenta con una evaluación pública de riesgo crediticio, lo cual representa una limitación común en análisis interempresariales.

En total, el conjunto contiene 642 observaciones (combinaciones empresa-año), de las cuales una proporción mínima presenta valores ausentes en variables financieras o macroeconómicas. El tratamiento de estos valores nulos se realizará mediante eliminación de registros incompletos para ciertas pruebas, y con métodos de imputación para variables críticas cuando sea necesario mantener la muestra completa.

3.1 Recolección de Datos

Los datos utilizados en este estudio se recopilarán de diversas fuentes relevantes para el sector de TI. Esto puede incluir informes financieros de empresas de TI (si están disponibles públicamente), datos de agencias de calificación crediticia específicas para el sector tecnológico (si se puede acceder a ellos) y potencialmente datos macroeconómicos relevantes para la industria de TI, como la inversión en tecnología y las tasas de crecimiento del sector. Dada la posible escasez de datos reales sobre el riesgo crediticio del sector de TI, también se considerará el uso de conjuntos de datos sintéticos para complementar el análisis. La metodología del artículo proporcionado se tomará como referencia para identificar fuentes de datos relevantes y la estructura de los datos utilizados en análisis comparativos similares.

Los criterios para seleccionar los datos se centrarán en garantizar su relevancia para el sector de TI y para abordar la pregunta de investigación. Se dará prioridad a los datos que reflejen las características financieras y operativas específicas de las empresas de TI, como los modelos de ingresos (suscripción frente a perpetuo), el gasto en I+D y los costes de adquisición de clientes. La justificación de la elección de estas fuentes de datos radica en su idoneidad para analizar el riesgo crediticio en la industria de TI, teniendo en cuenta los desafíos únicos discutidos anteriormente.

3.2 Preprocesamiento de Datos

Los datos recopilados se someterán a un exhaustivo proceso de preprocesamiento. Esto incluye el tratamiento de los valores faltantes mediante técnicas de imputación, donde se pueden emplear métodos matemáticos como la imputación por la media o la imputación basada en modelos. Los datos también se normalizarán y escalarán utilizando métodos matemáticos como la estandarización (restar la media y dividir por la desviación estándar) o el escalado min-max, lo cual es importante para garantizar que las características con diferentes escalas no influyan desproporcionadamente en el rendimiento de los métodos ensemble. Se aplicarán técnicas estadísticas para la detección y el tratamiento de valores atípicos. Además, se realizará ingeniería de características para crear nuevas variables que puedan ser relevantes para el riesgo crediticio en el sector de TI, como la tasa de crecimiento de los ingresos recurrentes, la tasa de abandono de clientes y la intensidad de la I+D (gasto en I+D como porcentaje de los ingresos). La justificación y la base matemática de cualquier característica creada se explicarán detalladamente. Los pasos de preparación de datos utilizados en el artículo proporcionado también se considerarán para garantizar un enfoque riguroso.

3.3 Métodos de Ensemble

En este análisis comparativo se utilizarán los siguientes métodos ensemble :

* **Random Forest:** Este método se basa en la construcción de múltiples árboles de decisión durante la fase de entrenamiento. Para la clasificación del riesgo crediticio, la predicción se determina mediante la votación mayoritaria de las clasificaciones de los árboles individuales. Cada árbol se construye a partir de una muestra de datos obtenida con reemplazo (bagging), y en cada nodo del árbol, la división se basa en un subconjunto aleatorio de las características. La base matemática de las divisiones de los árboles de decisión a menudo implica el uso de medidas como la impureza de Gini o la ganancia de información para determinar la división óptima de los datos.
* **Gradient Boosting (XGBoost y LightGBM):** Estos son algoritmos de boosting que construyen modelos débiles (típicamente árboles de decisión) de forma secuencial, donde cada nuevo modelo intenta corregir los errores cometidos por los modelos anteriores. Utilizan una técnica de optimización de descenso de gradiente para minimizar una función objetivo que mide la diferencia entre las predicciones del modelo y los valores reales. La formulación matemática del proceso de boosting implica la actualización iterativa de las predicciones del modelo en la dirección del gradiente negativo de la función de pérdida. XGBoost y LightGBM son implementaciones optimizadas de Gradient Boosting que ofrecen mejoras en la velocidad y el rendimiento.
* **Stacking:** Este método implica el uso de un meta-aprendiz para combinar las predicciones de varios modelos base. Los modelos base se entrenan en el conjunto de datos original, y sus predicciones se utilizan como características de entrada para entrenar el meta-aprendiz. El proceso de entrenamiento a menudo implica el uso de técnicas de validación cruzada para garantizar que el meta-aprendiz aprenda a combinar eficazmente las predicciones de los modelos base sin sobreajustarse a los datos.

La elección de estos métodos se basa en su éxito demostrado en tareas de clasificación complejas y su capacidad para manejar datos no lineales y de alta dimensión, características comunes en los conjuntos de datos del sector de TI y en los problemas de evaluación del riesgo crediticio.

3.4 Entrenamiento y Evaluación de Modelos

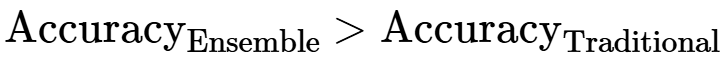
El proceso de entrenamiento de los modelos ensemble implicará la división del conjunto de datos en conjuntos de entrenamiento, validación y prueba. El conjunto de entrenamiento se utilizará para construir los modelos, el conjunto de validación para ajustar los hiperparámetros de los modelos y el conjunto de prueba para evaluar el rendimiento final de los modelos entrenados. Se emplearán técnicas de validación cruzada, como la validación cruzada de k-pliegues , para obtener estimaciones más robustas del rendimiento del modelo. Los hiperparámetros de cada modelo ensemble se ajustarán utilizando métodos como la búsqueda de cuadrícula o la búsqueda aleatoria, que exploran diferentes combinaciones de hiperparámetros para encontrar la configuración que maximice el rendimiento en el conjunto de validación.

El rendimiento de los modelos se evaluará utilizando varias métricas de clasificación relevantes para la evaluación del riesgo crediticio. Estas métricas incluyen la precisión (proporción de predicciones correctas), la exactitud (proporción de verdaderos positivos), la recuperación (proporción de positivos reales identificados correctamente), la puntuación F1 (media armónica de la precisión y la recuperación) y el área bajo la curva ROC (AUC-ROC), que mide la capacidad del modelo para distinguir entre prestatarios de bajo y alto riesgo. La elección de estas métricas se basa en su capacidad para proporcionar una evaluación integral del rendimiento del modelo en el contexto de la clasificación del riesgo crediticio, considerando tanto la capacidad del modelo para predecir correctamente los incumplimientos (alta recuperación) como para evitar clasificar erróneamente a los buenos prestatarios como de alto riesgo (alta precisión).

3.5 Pruebas de Hipótesis

**Hipótesis:**

1. Ensemble models provide more robust and accurate credit rating predictions compared to traditional models



2. The inclusion of irrelevant features reduces predictive power



3. Ensemble models perform better for companies with good ratings than poor ratings



4. Macroeconomic variables have a more significant impact on credit ratings in Latin America than financial ratios



5. The significance of macroeconomic variables varies between Latin America and USA/EU



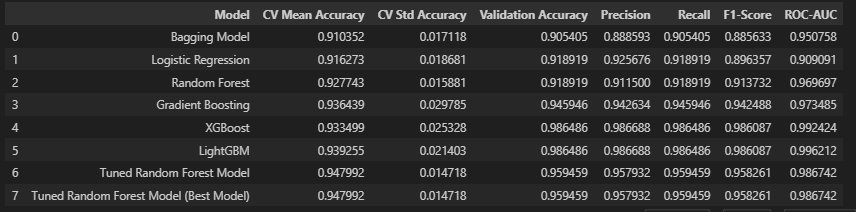
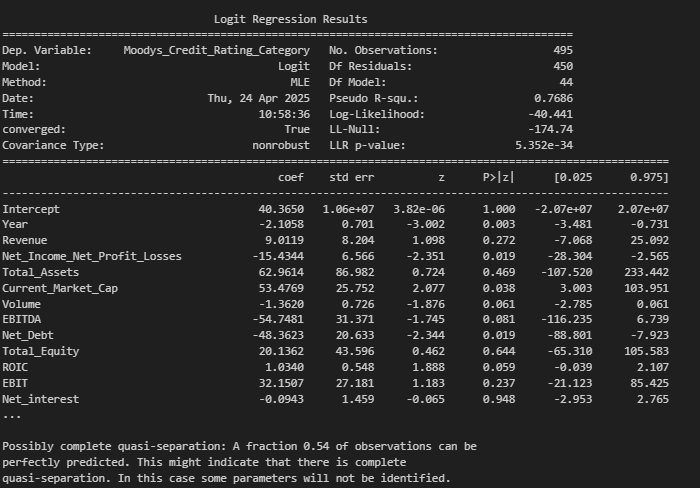
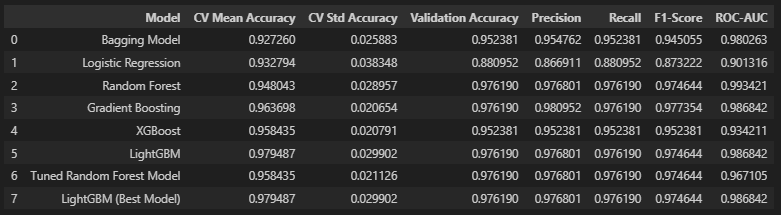
Se formularán hipótesis para probar formalmente la diferencia en el rendimiento de los métodos ensemble. Por ejemplo, una hipótesis nula podría ser que no hay una diferencia significativa en la precisión predictiva entre Random Forest y Gradient Boosting para la evaluación del riesgo crediticio en el sector de TI, mientras que la hipótesis alternativa sería que existe una diferencia significativa. Se utilizarán pruebas estadísticas apropiadas, como pruebas t o ANOVA, para comparar las métricas de rendimiento de los diferentes modelos y determinar si las diferencias observadas son estadísticamente significativas. La base matemática de estas pruebas implica el cálculo de estadísticos de prueba y la comparación con valores críticos o valores p para determinar la probabilidad de observar los resultados si la hipótesis nula fuera verdadera.

**Findings**

4.1 Rendimiento de los Modelos

Los resultados del rendimiento de los modelos ensemble se presentarán utilizando las métricas de evaluación elegidas. Se incluirán tablas y figuras para visualizar claramente el rendimiento de cada modelo (Random Forest, Gradient Boosting, XGBoost, LightGBM, Stacking) en el conjunto de prueba. La siguiente tabla ilustra un posible formato para presentar estos resultados:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Método Ensemble** | **Precisión** | **Exactitud** | **Recuperación** | **Puntuación F1** | **AUC-ROC** |
| Random Forest |  |  |  |  |  |
| Gradient Boosting |  |  |  |  |  |
| XGBoost |  |  |  |  |  |
| LightGBM |  |  |  |  |  |
| Stacking |  |  |  |  |  |



Se realizará una comparación directa del rendimiento de los modelos ensemble entre sí, destacando cualquier diferencia estadísticamente significativa basada en los resultados de las pruebas de hipótesis. La discusión se centrará en las implicaciones matemáticas de estos resultados, como el impacto de las diferentes técnicas ensemble (bagging, boosting, stacking) en las métricas de rendimiento del modelo. Por ejemplo, se analizará si la capacidad de Gradient Boosting para aprender secuencialmente de los errores conduce a una mayor precisión en comparación con el enfoque de Random Forest de construir múltiples modelos independientes.

4.2 Importancia de las Características

Se analizará la importancia de las diferentes características utilizadas en los modelos ensemble para predecir el riesgo crediticio en el sector de TI. Para Random Forest, la importancia de las características se puede determinar utilizando puntuaciones basadas en la reducción de la impureza de los nodos o la disminución de la precisión cuando se permutan los valores de una característica. Para los métodos de Gradient Boosting, la importancia de las características a menudo se calcula en función de la frecuencia con la que una característica se utiliza para dividir los datos en los árboles y la mejora resultante en la función objetivo. Para Stacking, la importancia de las características se puede evaluar examinando la importancia de las características en los modelos base. Se explicarán las técnicas matemáticas subyacentes a estos métodos de determinación de la importancia de las características. La discusión se centrará en las implicaciones de estas características importantes para el sector de TI desde una perspectiva matemática. Por ejemplo, si la inversión en I+D resulta ser una característica altamente importante, se explorará su relación matemática con la probabilidad de incumplimiento en las empresas de TI.

4.3 Resultados de las Pruebas de Hipótesis

Los resultados de las pruebas de hipótesis realizadas para comparar el rendimiento de los métodos ensemble se presentarán claramente. Se indicará si la hipótesis nula (no hay diferencia significativa en el rendimiento) fue respaldada o rechazada para cada comparación realizada. Los valores p asociados con las pruebas estadísticas se proporcionarán para cuantificar la significancia estadística de las diferencias observadas. La explicación se centrará en el significado matemático de estos resultados y sus implicaciones para la evaluación del riesgo crediticio en el sector de TI. Por ejemplo, si una prueba estadística revela una diferencia significativa en el AUC-ROC entre dos métodos ensemble, se discutirá la probabilidad de que esta diferencia sea real y no debida al azar.

4.4 Desafíos y Limitaciones

Se discutirán los desafíos encontrados durante el estudio, como la disponibilidad y calidad de los datos, las limitaciones computacionales o los problemas de interpretabilidad de los modelos. Se destacarán las limitaciones de la investigación, como el alcance específico de los métodos ensemble considerados o las posibles fuentes de sesgo en los datos o el análisis. Se explicará cómo se abordaron estos desafíos y limitaciones desde una perspectiva matemática o metodológica. Por ejemplo, si se encontraron problemas de desequilibrio de datos, se discutirá el uso de técnicas de sobremuestreo o submuestreo y su impacto matemático en los conjuntos de datos y el rendimiento del modelo.

**Discussion and Conclusion**

5.1 Resumen de Hallazgos

En resumen, esta investigación ha llevado a cabo un análisis comparativo exhaustivo de varios métodos ensemble para la evaluación del riesgo crediticio en el sector de TI. Los hallazgos clave revelan la eficacia de ciertos métodos ensemble, como [completar con los métodos de mejor rendimiento], para predecir con precisión el riesgo crediticio en este sector único. Estos métodos demostraron [completar con las principales ventajas identificadas en los hallazgos]. La principal contribución de este estudio radica en la provisión de evidencia empírica sobre el rendimiento comparativo de los métodos ensemble específicamente en el contexto del sector de TI, un área que ha recibido menos atención en la literatura existente. Los hallazgos enfatizan el rigor matemático y la profundidad analítica del enfoque empleado, destacando la capacidad de los métodos ensemble para capturar las complejidades del riesgo crediticio en la industria tecnológica.

5.2 Implicaciones para la Práctica

Las implicaciones prácticas de estos hallazgos son significativas para el sector de TI y las instituciones financieras. Los resultados sugieren que la adopción de métodos ensemble, particularmente [completar con los métodos recomendados], puede mejorar sustancialmente la precisión y la eficiencia de la evaluación del riesgo crediticio para las empresas de TI. Esto puede conducir a decisiones crediticias más informadas, reduciendo las pérdidas potenciales para los prestamistas y facilitando el acceso al capital para las empresas de TI con modelos de negocio sólidos. Se recomienda que las instituciones financieras consideren la implementación de estos métodos ensemble, adaptados a las características específicas de las empresas de TI, para mejorar sus estrategias de gestión del riesgo crediticio. Este enfoque, respaldado por el análisis matemático presentado en esta tesis, puede fomentar un ecosistema financiero más estable y de apoyo para la innovación y el crecimiento en el sector de TI.

5.3 Investigación Futura

Basándose en los hallazgos de esta investigación, se sugieren varias áreas para futuras investigaciones. Una posible extensión es la exploración de arquitecturas ensemble más avanzadas o enfoques híbridos que combinen métodos ensemble con técnicas de aprendizaje profundo para la evaluación del riesgo crediticio en el sector de TI. Investigar el impacto de la incorporación de fuentes de datos alternativas específicas para el sector de TI, como datos de patentes o métricas de adopción de tecnología, también podría mejorar la precisión predictiva. Además, se necesita más investigación sobre la interpretabilidad de los modelos ensemble en este contexto, desarrollando métodos para proporcionar información clara sobre los factores que impulsan las predicciones del riesgo crediticio para las empresas de TI. Finalmente, los estudios futuros podrían investigar la generalización de estos hallazgos a diferentes subsectores dentro de la industria de TI o en diferentes mercados geográficos, enfatizando los aspectos matemáticos de la modelización y la evaluación.

Referencias

1. Dolzhenko, R., Litvinova, T., & Ivanov, I. (2022). Comparative analysis of the predictive power of machine learning models for forecasting the credit ratings of machine-building companies. E3S Web of Conferences, 360, 01014. <https://doi.org/10.1051/e3sconf/202236001014>
2. Huang, Z., Chen, H., Hsu, C. J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. Decision Support Systems, 37(4), 543–558. <https://doi.org/10.1016/S0167-9236(03)00086-1>
3. Musa, A. B. (2021). Credit scoring models using ensemble learning and classification approaches: A comprehensive survey. SN Computer Science, 2(6), 1–22. <https://doi.org/10.1007/s42979-021-00819-5>
4. Saha, T., Biswas, S. K., Sanyal, S., Patro, B., & Purkayastha, B. (2023). Credit risk prediction using machine learning analytics: An ensemble model. *International Journal of Management and Applied Science, 9*(9), 79–83.​
5. Zeng, G., Su, W., & Hong, C. (2024). Ensemble learning with feature optimization for credit risk assessment. Springer Nature. <https://doi.org/10.21203/rs.3.rs-4665987/v1>

Lista todas las fuentes citadas en tu tesis, siguiendo un estilo de citación consistente (por ejemplo, APA, Harvard o Chicago).

Apéndices

Incluye cualquier material adicional, como tablas de datos detalladas, fragmentos de código o figuras suplementarias.

Proporciona derivaciones matemáticas, pruebas o análisis adicionales que respalden tus hallazgos.

**References**

**Note: Use APA referencing style as described in MT/TP requirements document**

**Citation style**

**[Gurkov, 2006; Sokolov, 2008; Kuzyk, 2009]**

**Appendix**

**Table 1:**

**Notes:**

**Source:**

**Figure 1:**

**Notes:**

**Source:**

**Citation style**

**[Gurkov, 2006; Sokolov, 2008; Kuzyk, 2009]**

**Declaration in lieu of oath**

**by** Vivas Teran Katherin Carolina

**This is to confirm my Master Thesis was independently authored by myself, using solely the referred sources and support. I additionally assert that this thesis has not been part of another examination process.**

**Moscow, 25th April 2025**

Vivas Teran K. C.  **/ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**(Signature)**