Centro Oficial FP
Tecnología y Digitalización

Máster en Inteligencia Artificial y Big Data

Desarrollo de un Modelo de Aprendizaje Automático para la Perfilación del Riesgo del Inversor

Antonio Esquinas Fernández

Acceso al Proyecto

El contenido completo del proyecto, incluyendo notebooks, datos y documentación, puede consultarse online en el siguiente enlace de Google Drive:

https://drive.google.com/drive/folders/1b4C87MXNKw8QfIsOsPqzZpI4zqZyCU2_?usp=sharing

Resumen

El presente Trabajo de Fin de Master (TFM) aborda el desarrollo y la evaluación de un modelo de aprendizaje automático (ML) para la perfilación del riesgo del inversor, con el objetivo de superar las limitaciones inherentes a los sistemas tradicionales basados en reglas. El contexto es una aplicación financiera que actualmente utiliza un enfoque de reglas predefinidas para determinar la tolerancia al riesgo del usuario en una escala del 1 al 10. Este trabajo se centra en reemplazar o mejorar sustancialmente dicho sistema. La metodología comprende una exploración exhaustiva de fuentes de datos, seleccionando el *FINRA National Financial Capability Study (NFCS) 2021* como base principal debido a su alineación con los objetivos del modelo y la riqueza de sus variables. Se detalla un proceso iterativo de ingeniería de características, que incluye selección automatizada, curación manual guiada por el conocimiento del dominio y la creación de variables sintéticas y avanzadas. Se experimentó con la reducción de la dimensionalidad de la variable objetivo para mitigar el desbalance de clases, optando por un marco de clasificación binaria (riesgo bajo/medio vs. alto). Se entrenaron y compararon diversos modelos de ML, incluyendo Regresión Logística, Random Forest y LightGBM. Los resultados demuestran que la ingeniería avanzada de características, especialmente la guiada por el dominio, fue crucial para mejorar el rendimiento predictivo. Se desarrollaron modelos especializados: un "Modelo Académico" para optimizar métricas globales (AUC de 0.7752 y F1-macro de ≈0.70 en test para 2 clases), un "Modelo de Coste" para minimizar el impacto económico de errores de clasificación, y un "Modelo Detector" enfocado en maximizar la identificación de perfiles de alto riesgo (Recall del 81% para la clase de alto riesgo en test). Las conclusiones subrayan la sinergia entre los datos y el conocimiento del dominio, el valor de la ingeniería de características contextualizada y los beneficios de la optimización especializada de modelos. Se identifican limitaciones, como la dependencia de datos declarados, y se proponen líneas futuras de investigación, incluyendo la incorporación de datos de preferencias reveladas y la implementación de técnicas de explicabilidad (XAI).

Abstract

This thesis addresses the development and evaluation of a machine learning (ML) model for investor risk profiling, aiming to overcome the inherent limitations of traditional rule-based systems. The context is a financial application that currently uses a predefined rule-based approach to determine user risk tolerance on a 1-to-10 scale. This work focuses on substantially replacing or improving said system. The methodology includes an exhaustive exploration of data sources, selecting the FINRA National Financial Capability Study (NFCS) 2021 as the primary basis due to its alignment with the model's objectives and the richness of its variables. An iterative feature engineering process is detailed, encompassing automated selection, domain-knowledge-guided manual curation, and the creation of synthetic and advanced features. Dimensionality reduction of the target variable was experimented with to mitigate class imbalance, opting for a binary classification framework (low/medium vs. high risk). Various ML models were trained and compared, including Logistic Regression, Random Forest, and LightGBM. The results demonstrate that advanced feature engineering, particularly when guided by domain expertise, was crucial for enhancing predictive performance. Specialized models were developed: an "Academic Model" to optimize global metrics (AUC of 0.7752 and F1-macro of ≈0.70 on test for 2 classes), a "Cost Model" to minimize the economic impact of misclassifications, and a "Detector Model" focused on maximizing the identification of high-risk profiles (Recall of 81% for the high-risk class on test). The conclusions underscore the synergy between data and domain knowledge, the value of contextualized feature engineering, and the benefits of specialized model optimization. Limitations, such as reliance on self-reported data, are identified, and future research lines are proposed, including the

incorporation of revealed preference data and the implementation of eXplainable AI (XAI) techniques.

Glosario de siglas y abreviaturas

- **ACC** – *Accuracy* (exactitud global de un clasificador)
- **ANOVA** – *Analysis of Variance* (análisis de varianza)
- **AUC** – *Area Under the Curve* (área bajo la curva ROC)
- **CV** – *Cross-Validation* (validación cruzada)
- **ECF** – *Encuesta de Competencias Financieras*
- **EDA** – *Exploratory Data Analysis* (análisis exploratorio de datos)
- **ESMA** – *European Securities and Markets Authority* (Autoridad Europea de Valores y Mercados)
- **F1** – *F1-score* (media armónica de precisión y recall)
- **FL_SCORE** – *Financial Literacy Score* (puntuación de literacidad financiera)
- **FN** – *False Negative* (falso negativo)
- **FP** – *False Positive* (falso positivo)
- **GB** – *Gradient Boosting* (método de ensamble por gradiente)
- **IA** – *Inteligencia Artificial*
- **ILI** – *Financial Literacy Index* (índice de literacidad financiera)
- **JSON** – *JavaScript Object Notation* (formato ligero de intercambio de datos)
- **K-Means** – *K-Means Clustering* (agrupamiento por medias *k*)
- **KYC** – *Know Your Customer* (conoce a tu cliente)
- **LIME** – *Local Interpretable Model-agnostic Explanations* (explicaciones locales de modelos)
- **LightGBM** – *Light Gradient Boosting Machine* (implementación eficiente de GB)
- **LR** – *Logistic Regression* (regresión logística)
- **ML** – *Machine Learning* (aprendizaje automático)
- **NaN** – *Not a Number* (valor numérico inválido o ausente)
- **NFCS** – *National Financial Capability Study*
- **PCA** – *Principal Component Analysis* (análisis de componentes principales)

- **RF** – *Random Forest* (bosque aleatorio)
- **RFECV** – *Recursive Feature Elimination with Cross-Validation* (eliminación recursiva de variables con validación cruzada)
- **ROC** – *Receiver Operating Characteristic* (curva ROC)
- **SCF** – *Survey of Consumer Finances*
- **SHAP** – *SHapley Additive exPlanations* (valores de Shapley para interpretabilidad)
- **SMOTE** – *Synthetic Minority Over-sampling Technique* (sobremuestreo sintético de la clase minoritaria)
- **SVM** – *Support Vector Machine* (máquina de vectores de soporte)
- **TN** – *True Negative* (verdadero negativo)
- **TP** – *True Positive* (verdadero positivo)
- **VaR** – *Value at Risk* (valor en riesgo)
- **XAI** – *Explainable Artificial Intelligence* (inteligencia artificial explicable)

Índice General

Capítulo 1: Introducción

1.1. Contexto de la Perfilación del Riesgo del Inversor y la Gestión Patrimonial

La perfilación del riesgo del inversor constituye un pilar fundamental en la gestión patrimonial y el asesoramiento financiero moderno. Determinar adecuadamente la tolerancia y la capacidad de un individuo para asumir riesgos es esencial para la construcción de carteras de inversión que no solo busquen optimizar los rendimientos, sino que también se alineen con las expectativas y el confort emocional del inversor. El presente Trabajo de Fin de Máster (TFM) se enmarca en el desarrollo de una aplicación destinada a inversores. Esta herramienta tiene como objetivo principal facilitar la toma de decisiones financieras ofreciendo, tras la cumplimentación de un breve formulario, una evaluación del nivel de tolerancia al riesgo del usuario (en una escala del 1 al 10), una cartera de inversión diversificada recomendada automáticamente y la posibilidad de personalizar dicha cartera.

En la era digital, la proliferación de herramientas tecnológicas y el acceso a ingentes volúmenes de datos han transformado las expectativas de los inversores, quienes demandan servicios financieros cada vez más personalizados y adaptados a sus circunstancias individuales. La perfilación precisa del riesgo emerge, en este contexto, como el primer paso crítico hacia una genuina personalización de las estrategias de inversión, distanciándose de los enfoques genéricos que a menudo resultan inadecuados. Los servicios financieros, incluyendo la gestión de patrimonios y el asesoramiento automatizado (robo-advisors), están experimentando una profunda transformación impulsada por la inteligencia artificial (IA) y el aprendizaje automático, tecnologías que prometen ofrecer soluciones más eficientes, accesibles y, sobre todo, personalizadas.[4] La mejora continua en la precisión y la granularidad de la perfilación del riesgo no solo redunda en un mayor beneficio para el inversor individual, al permitirle tomar decisiones más informadas y coherentes con sus verdaderas preferencias y capacidades, sino que también puede contribuir a una asignación de capital más eficiente a nivel agregado en los mercados financieros.

1.2. Limitaciones de los Enfoques Tradicionales Basados en Reglas

El sistema de perfilación de riesgo que este TFM se propone mejorar se basa, actualmente, en un conjunto de reglas predefinidas. Este enfoque, si bien fundamentado en principios como la Teoría Moderna de Carteras de Markowitz y un sistema de puntuación propio que distingue entre la capacidad para asumir riesgo y la tolerancia psicométrica, presenta limitaciones significativas. La capacidad se evalúa mediante variables objetivas (edad, ingresos, patrimonio, horizonte temporal) y una fórmula con una función sigmoide, mientras que la tolerancia se mide con un cuestionario psicométrico, cuyas inconsistencias en las respuestas penalizan la puntuación. El perfil final resulta de una ponderación asimétrica fija (80/20) que favorece al componente más conservador entre capacidad y tolerancia.

Estos sistemas basados en reglas, aunque transparentes en su lógica para casos simples, a menudo exhiben una "fragilidad" inherente: pequeñas variaciones en los datos de entrada o matices no contemplados en la compleja psicología del inversor pueden conducir a evaluaciones de riesgo desproporcionadas o imprecisas.[3] La penalización por inconsistencias, por ejemplo, puede ser un indicativo de la dificultad del sistema para capturar la verdadera actitud del inversor de forma integral. Esta fragilidad no es meramente un problema técnico, sino un reflejo de la incapacidad de las reglas simples para encapsular la naturaleza multifacética y frecuentemente no lineal de la tolerancia al riesgo. Factores como las emociones, los sesgos cognitivos documentados extensamente por las finanzas conductuales [6], y las interacciones complejas entre variables demográficas y actitudinales son intrínsecamente difíciles de codificar en reglas explícitas y estáticas. La ponderación fija 80/20 es un claro ejemplo de una heurística que, si bien prudente, ignora la variabilidad individual y podría no ajustarse de manera óptima al apetito de riesgo real y matizado de todos los inversores.

La literatura académica ha señalado consistentemente las deficiencias de los cuestionarios tradicionales para evaluar la tolerancia al riesgo. Se ha observado que estos instrumentos pueden ser poco fiables, explicando en algunos casos menos del 15% de la variación en la asignación a activos de riesgo entre inversores.[1] Las preguntas formuladas pueden ser de baja calidad, ambiguas o malinterpretadas por los usuarios, y raramente están diseñadas para controlar los sesgos cognitivos que influyen en las respuestas.[8] Esto puede llevar a la recopilación de información incompleta o distorsionada, resultando en perfiles de riesgo inadecuados y, consecuentemente, en recomendaciones de inversión subóptimas.[9]

Aunque el sistema actual se apoya en la Teoría Moderna de Carteras de Markowitz, que asume un comportamiento racional del inversor, las limitaciones observadas (como la necesidad de penalizar inconsistencias) sugieren una desconexión con la realidad del comportamiento inversor. Este último está fuertemente influenciado por factores conductuales que los sistemas basados en reglas luchan por incorporar de manera efectiva.[6] La rigidez inherente a estos sistemas [3] y su dificultad para manejar datos incompletos o inherentemente sesgados pueden conducir directamente a perfiles de riesgo inexactos, lo que a su vez puede resultar en una asignación inadecuada de recursos, insatisfacción del cliente o, en el peor de los casos, pérdidas financieras evitables. La Tabla 1.1 resume las diferencias fundamentales entre el enfoque actual y la solución propuesta mediante aprendizaje automático.

Tabla 1.1: Comparación del Sistema Existente Basado en Reglas vs. Enfoque Propuesto de Aprendizaje Automático

Característica	Sistema Basado en Reglas	Sistema Propuesto de Aprendizaje Automático (ML)
Adaptabilidad	Baja, basado en lógica predefinida	Alta, aprende de datos y puede adaptarse a nuevos patrones
Manejo de Matices	Limitada capacidad para capturar interacciones complejas	Capacidad para aprender patrones sutiles y no lineales
Escalabilidad	Mantenimiento complejo a medida que aumentan las reglas	Escalable con nuevos datos y reentrenamiento
Explicabilidad	Alta para reglas simples, baja para lógica compleja	Potencialmente menor (caja negra), requiere técnicas XAI
Dependencia de Datos	Baja (basado en conocimiento experto codificado)	Alta (requiere grandes volúmenes de datos de calidad)
Personalización	Limitada por la generalidad de las reglas	Alto potencial para perfiles individualizados

Esta tabla es crucial al inicio para justificar la necesidad del TFM, estableciendo claramente las deficiencias del enfoque actual y las ventajas potenciales del ML, motivando así el resto del estudio. Al presentar esta comparación, se comprende inmediatamente el "por qué" del TFM, destacando cómo el ML puede abordar las limitaciones inherentes de los sistemas de reglas [1] y cómo el ML [11] puede ofrecer una solución superior.

1.3. El Potencial del Aprendizaje Automático para la Mejora de la Perfilación

Frente a las limitaciones expuestas, el aprendizaje automático (ML) emerge como una alternativa prometedora y poderosa. Los modelos de ML poseen la capacidad intrínseca de aprender a partir de grandes volúmenes de datos, descubriendo relaciones complejas y patrones sutiles, a menudo no lineales, que escaparían a un análisis basado en reglas predefinidas. Esta capacidad es particularmente relevante en el ámbito de la perfilación del riesgo del inversor, donde las interacciones entre múltiples factores demográficos, financieros, psicológicos y conductuales determinan la propensión individual al riesgo.

La principal ventaja del ML radica en su alta adaptabilidad y escalabilidad. A diferencia de los sistemas de reglas, que requieren una reconfiguración manual para incorporar nuevos conocimientos o adaptarse a cambios en el comportamiento del mercado o de los inversores, los modelos de ML pueden reentrenarse con nuevos datos, ajustando sus parámetros internos para reflejar patrones emergentes. Esto no solo busca una mayor precisión en la evaluación del riesgo, sino también una perfilación más robusta y dinámica, crucial en el contexto de mercados financieros en constante evolución. La capacidad de los modelos de ML para procesar grandes conjuntos de datos, identificar relaciones no lineales [14] y ofrecer evaluaciones de riesgo potencialmente en tiempo real representa un avance significativo sobre los métodos tradicionales.[11]

Además, el comportamiento financiero humano, como se ha mencionado, no siempre sigue los cánones de la racionalidad estricta. Las finanzas conductuales han demostrado la influencia de sesgos cognitivos en la toma de decisiones. Un modelo de ML, al aprender de datos que reflejan comportamientos reales o declarados de miles de inversores, tiene el potencial de identificar indirectamente la manifestación de estos sesgos o sus efectos sobre la tolerancia al riesgo, una tarea considerablemente más ardua para un sistema basado en reglas explícitas. En lugar de predefinir cómo interactúan la edad, los ingresos, la aversión a la volatilidad y otros factores, un modelo ML puede

aprender estas interacciones directamente de los datos, incluyendo posibles efectos no monótonos o la existencia de umbrales a partir de los cuales la relación entre variables cambia, como se ha observado en otros contextos financieros donde, por ejemplo, una tolerancia al riesgo excesiva puede ser perjudicial.[16]

1.4. Objetivos y Contribución del TFM

El objetivo primordial de este Trabajo de Fin de Máster es reemplazar o mejorar sustancialmente el sistema de reglas existente para la determinación del perfil de riesgo del inversor mediante la concepción, desarrollo, entrenamiento y evaluación de un modelo de Aprendizaje Automático (ML). Se busca que este nuevo modelo supere las limitaciones inherentes al enfoque actual, aprovechando la capacidad del ML para aprender de grandes volúmenes de datos y descubrir relaciones complejas que permitan una perfilación más precisa y fidedigna.

Las características deseables del modelo de ML a desarrollar son:

- **Entrenamiento con datos reales:** El modelo se entrenará utilizando datos de miles de casos de inversores reales, lo que permitirá capturar una amplia gama de perfiles y comportamientos.
- **Aprendizaje de patrones complejos:** Se espera que el modelo sea capaz de identificar y aprender patrones no triviales entre las variables de entrada, como, por ejemplo, la relación entre el nivel de educación financiera y la tolerancia real al riesgo manifestada, interacciones que a menudo escapan a la modelización mediante reglas explícitas.
- **Generación de perfiles más precisos y adaptativos:** El objetivo es lograr una mayor precisión en la asignación del perfil de riesgo, adaptándose de manera más fidedigna a las características individuales de cada inversor.
- **Explicabilidad:** Un aspecto crucial, especialmente en el dominio financiero, es la capacidad de generar perfiles que no solo sean precisos, sino también explicables. Esto implica una consideración temprana hacia modelos que, por su naturaleza, sean más interpretables o la necesidad de implementar técnicas de explicabilidad post-hoc (como SHAP o LIME) para justificar las predicciones del modelo.

Si bien el TFM se enfoca en mejorar una aplicación particular, la metodología desarrollada –que abarca la selección y el preprocesamiento de datos, la ingeniería de características iterativa, el manejo del desbalance de clases, la experimentación con diversos algoritmos y la especialización de modelos para objetivos concretos– ofrece un *framework* que podría ser replicable para abordar problemas similares de perfilación en otros contextos financieros o, incluso, en dominios no financieros donde se busque modelar preferencias o riesgos individuales complejos. El TFM no solo busca construir un modelo predictivo, sino también documentar un *proceso de investigación y desarrollo* que es inherentemente iterativo y que demuestra la importancia de combinar el rigor técnico con un profundo conocimiento del dominio para alcanzar soluciones efectivas y significativas.

1.5. Estructura del Documento

El presente documento se organiza de la siguiente manera:

El Capítulo 2 establece el marco teórico, revisando los fundamentos de la perfilación del riesgo del inversor, los principios de las finanzas conductuales, las aplicaciones del aprendizaje automático en la gestión de riesgos financieros y la importancia de la explicabilidad en modelos financieros.

El Capítulo 3 detalla el diseño metodológico concerniente a los datos y las características, incluyendo la exploración y selección de fuentes de datos, la justificación y el análisis exploratorio del conjunto de datos NFCS, el proceso de limpieza y preprocesamiento, la estrategia iterativa de ingeniería de características y la definición de la variable objetivo.

El Capítulo 4 describe el desarrollo y la experimentación con los modelos de aprendizaje automático, detallando la configuración experimental, los modelos base y el benchmarking inicial, el impacto de la ingeniería avanzada de características y la optimización especializada de modelos.

El Capítulo 5 presenta y analiza los resultados obtenidos, incluyendo una comparación del rendimiento de los modelos globales y una evaluación de los modelos especializados con sus métricas clave, además de una discusión sobre la importancia de las características identificadas.

Finalmente, el Capítulo 6 ofrece una discusión de los hallazgos, las conclusiones principales del estudio, una reflexión sobre la consecución de los objetivos del TFM, sus limitaciones, las implicaciones prácticas y las posibles líneas de investigación futuras.

Capítulo 2: Marco Teórico y Trabajos Relacionados

2.1. Fundamentos de la Perfilación del Riesgo del Inversor

La perfilación del riesgo del inversor es un proceso multifacético que busca comprender y cuantificar la actitud de un individuo hacia el riesgo financiero. La **tolerancia al riesgo financiero** se define como el grado máximo de incertidumbre que una persona está dispuesta a aceptar al tomar una decisión financiera que conlleva la posibilidad de una pérdida.[10] Es crucial distinguir este concepto de la **capacidad de riesgo**, que se refiere a la habilidad financiera objetiva del inversor para absorber pérdidas sin que ello comprometa sus objetivos financieros esenciales.[2] Mientras la tolerancia es una característica psicológica y subjetiva, la capacidad es una medida más tangible, dependiente de factores como los ingresos, el patrimonio neto, las obligaciones financieras y el horizonte temporal de la inversión.

El perfil de riesgo de un inversor no es un constructo monolítico, sino que se compone de diversas dimensiones. El sistema actual de la aplicación en la que se enmarca este TFM ya intenta capturar esta dualidad, evaluando tanto la capacidad como la tolerancia.[3] Sin embargo, la combinación de estas dimensiones mediante una ponderación fija (80% al valor más conservador) es una simplificación que un modelo de aprendizaje automático podría superar, aprendiendo una combinación más matizada o incluso identificando diferentes arquetipos de inversores donde la relación entre capacidad y tolerancia varía significativamente. Otros conceptos relevantes que contribuyen a un perfil de riesgo completo incluyen la preferencia de riesgo (una sensación general de que una situación es mejor que otra), la percepción de riesgo (una evaluación subjetiva de la peligrosidad de una decisión), la necesidad de riesgo (el nivel de riesgo que se debe asumir para alcanzar un objetivo financiero) y la compostura ante el riesgo (la propensión a comportarse de manera consistente ante situaciones de riesgo).[10]

Los factores que tradicionalmente se consideran influyentes en la determinación del perfil de riesgo incluyen, para la capacidad, la edad, los ingresos, el patrimonio y el horizonte temporal; y para la tolerancia psicométrica, la actitud del inversor hacia la volatilidad, el retorno esperado y su definición personal de riesgo. Adicionalmente, la situación financiera general, los objetivos de inversión específicos, el nivel de conocimiento y la experiencia previa en inversiones son determinantes clave.[17] Los cuestionarios son la herramienta más comúnmente utilizada para recabar esta información y evaluar la tolerancia al riesgo.[17] No obstante, la dificultad de los cuestionarios tradicionales para medir fiablemente la tolerancia al riesgo [1] se debe, en parte, a esta multidimensionalidad y a la influencia de factores psicológicos que no se capturan fácilmente con preguntas directas y estáticas, un aspecto que las finanzas conductuales intentan abordar.

2.2. Principios de Finanzas Conductuales y su Impacto en las Decisiones de Inversión

Las finanzas conductuales representan un campo de estudio que integra la psicología con la teoría financiera tradicional para explicar cómo los factores cognitivos y emocionales afectan la toma de decisiones de los inversores.[6] Este enfoque surge como una respuesta a las anomalías y comportamientos observados en los mercados financieros que no pueden ser explicados satisfactoriamente por los modelos basados en la premisa de la racionalidad del inversor y la eficiencia del mercado.[8] El TFM reconoce explícitamente que el comportamiento financiero humano no siempre es racional y que las reglas simples pueden omitir indicadores sutiles pero importantes de la tolerancia al riesgo.[3]

Una de las contribuciones fundamentales de las finanzas conductuales es la identificación de numerosos **sesgos cognitivos** que desvían sistemáticamente las decisiones de los inversores de la senda de la racionalidad. Entre estos se encuentran el exceso de confianza (sobreestimar la propia habilidad o la precisión de la información que se posee), la aversión a la pérdida (sentir el dolor de una pérdida de forma más intensa que el placer de una ganancia equivalente), el anclaje (depender excesivamente de una pieza inicial de información al tomar decisiones), el sesgo de confirmación (buscar información que confirme creencias preexistentes) y el comportamiento de rebaño (seguir las acciones de un grupo más grande). Estos sesgos pueden llevar a decisiones de inversión subóptimas, como asumir demasiado o muy poco riesgo, realizar operaciones excesivas o mantener posiciones perdedoras durante demasiado tiempo.

La percepción del riesgo, un componente clave de la tolerancia, también es altamente subjetiva y puede estar más influenciada por las emociones y sentimientos inmediatos ("risk-as-feelings") que por una evaluación analítica y objetiva de las probabilidades y los resultados potenciales.[8] Esto puede explicar por qué los inversores reaccionan de forma exagerada a las noticias del mercado o por qué su tolerancia al riesgo declarada puede fluctuar con el sentimiento del mercado.

La priorización del conjunto de datos NFCS (que contiene datos de preferencias declaradas) sobre el SCF (que ofrece información sobre la composición real de las carteras, o preferencias reveladas) en la metodología de este TFM introduce la relevancia del conocido

"say-do gap" (la brecha entre lo que se dice y lo que se hace). Las finanzas conductuales ofrecen explicaciones para esta discrepancia: los individuos pueden declarar una tolerancia al riesgo que no se alinea con sus acciones reales debido a sesgos como el optimismo, la deseabilidad social, o una autopercepción inexacta de su verdadera reacción emocional ante las pérdidas potenciales.[8] Un modelo de ML entrenado con datos declarados, como el que se desarrolla en este TFM, predice la

percepción que el usuario tiene de su riesgo o lo que *declara* estar dispuesto a tolerar. Esto es intrínsecamente útil para una aplicación que interactúa con el usuario *antes* de que este tome decisiones de inversión concretas. No obstante, para predecir el comportamiento *real* en el mercado, se requerirían datos de preferencias reveladas o un modelo más complejo que aprenda a identificar y, potencialmente, corregir el impacto de estos sesgos declarativos. Esta consideración se alinea directamente con las limitaciones identificadas y las futuras líneas de investigación propuestas en este TFM.

2.3. Aplicaciones del Aprendizaje Automático en la Gestión de Riesgos Financieros y Perfilación

El aprendizaje automático (ML) ha emergido como una tecnología transformadora en la industria financiera, con aplicaciones que abarcan desde la detección de fraude y la calificación crediticia hasta la optimización de carteras y la gestión de riesgos.[11] En el ámbito de la gestión de riesgos financieros, el ML ofrece la capacidad de analizar grandes volúmenes de datos históricos y en tiempo real para identificar patrones, predecir eventos futuros y, en última instancia, tomar decisiones más informadas y proactivas.[11] Los algoritmos comúnmente empleados incluyen la Regresión Logística, los bosques aleatorios (Random Forest), las redes neuronales, las máquinas de vectores de soporte (SVM) y los modelos de Gradient Boosting, muchos de los cuales son explorados en este TFM.[11]

Los beneficios de aplicar ML en este contexto son múltiples: mayor precisión predictiva en comparación con los modelos estadísticos tradicionales, capacidad para evaluar riesgos en tiempo real, adaptabilidad a condiciones de mercado cambiantes, y la habilidad para manejar conjuntos de datos de alta dimensionalidad y descubrir relaciones no lineales complejas entre variables.[11] En la perfilación de inversores específicamente, el ML puede utilizarse para segmentar clientes en grupos con perfiles de riesgo y comportamiento similares mediante técnicas de clustering, o para predecir el perfil de riesgo de un inversor basándose en sus respuestas a cuestionarios, datos demográficos, e incluso su comportamiento real en plataformas de inversión simuladas o reales.[25]

La perfilación del riesgo del inversor implica considerar una multitud de factores (demográficos, financieros, actitudinales, conductuales). Los enfoques tradicionales, a menudo basados en sistemas de puntuación lineales o árboles de decisión simples, luchan por manejar eficazmente esta alta dimensionalidad y las interacciones intrincadas entre estas variables. Los modelos de ML, especialmente aquellos basados en ensambles de árboles (como Random Forest y Gradient

Boosting, utilizados extensivamente en este TFM) o las redes neuronales, son inherentemente más aptos para manejar un gran número de características e identificar cuáles son verdaderamente predictivas, sin necesidad de que el analista especifique todas las interacciones posibles a priori.[15] Esto permite superar, en cierta medida, la "maldición de la dimensionalidad" que afecta a los modelos más simples cuando se enfrentan a espacios de características muy amplios. Existe una tendencia creciente en la industria y en la investigación académica hacia el uso de datos comportamentales —más allá de las simples respuestas a cuestionarios— para la perfilación del riesgo, y el ML se presenta como la herramienta natural para extraer señales predictivas valiosas de estos datos, que suelen ser más ricos y menos estructurados.[25] Aunque el presente TFM se basa principalmente en datos de encuestas, la metodología desarrollada y los modelos construidos sientan las bases para la futura incorporación de fuentes de datos más diversas y dinámicas.

2.4. Importancia de la Explicabilidad (XAI) en Modelos Financieros

A medida que los modelos de aprendizaje automático se vuelven más complejos y, a menudo, más precisos que sus contrapartes tradicionales, también pueden volverse más opacos, funcionando como "cajas negras" cuyas decisiones internas son difíciles de comprender. En el sector financiero, donde las decisiones algorítmicas pueden tener consecuencias significativas para los individuos (por ejemplo, la aprobación o denegación de un crédito, o la recomendación de una estrategia de inversión), esta opacidad es cada vez más inaceptable.[26] La explicabilidad de la inteligencia artificial (XAI, por sus siglas en inglés) emerge como un campo crucial que busca desarrollar métodos y técnicas para hacer que las decisiones de los modelos de ML sean transparentes e interpretables por humanos.[3]

La necesidad de XAI en finanzas está impulsada por varios factores interconectados:

- **Confianza del Usuario y Adopción:** Los clientes e inversores son más propensos a confiar y adoptar recomendaciones de sistemas de IA si comprenden la lógica subyacente a dichas recomendaciones.[26]

- **Cumplimiento Normativo:** Regulaciones como MiFID II en Europa exigen que las empresas financieras puedan justificar la idoneidad de sus recomendaciones de inversión.[20] Los modelos opacos dificultan la demostración de este cumplimiento.[27]
- **Gestión de Riesgos y Robustez del Modelo:** Comprender por qué un modelo toma ciertas decisiones ayuda a identificar posibles sesgos, errores o vulnerabilidades en el propio modelo, permitiendo su mejora y asegurando su robustez.
- **Responsabilidad y Equidad:** La explicabilidad es fundamental para asegurar que los modelos de ML no discriminen injustamente a ciertos grupos de individuos y para asignar responsabilidad cuando se producen errores.[28]

Técnicas populares de XAI incluyen LIME (Local Interpretable Model-agnostic Explanations) y SHAP (SHapley Additive exPlanations).[3] LIME funciona aproximando el comportamiento de un modelo complejo de forma local (es decir, para una predicción individual) mediante un modelo interpretable más simple, como una regresión lineal.[28] SHAP, por otro lado, se basa en la teoría de juegos cooperativos y calcula los valores de Shapley para cada característica, representando la contribución marginal de dicha característica a la predicción final, considerando todas las posibles combinaciones de características.[28] Estas técnicas se aplican en finanzas para explicar decisiones de modelos de calificación crediticia, sistemas de detección de fraude y estrategias de trading algorítmico.[26] La XAI también está ganando tracción en el ámbito de los robo-advisors para promover la transparencia algorítmica y mejorar la confianza del cliente.[5]

La mención temprana de la explicabilidad como una característica deseable del modelo en los objetivos de este TFM [3] y su reiteración en las conclusiones y líneas futuras [3] indican una conciencia clara de este imperativo. Aunque la implementación y evaluación exhaustiva de técnicas como SHAP o LIME no constituyen el foco central de los resultados presentados en este TFM, la consideración de la explicabilidad influye en la selección de los algoritmos de modelado (por ejemplo, los modelos basados en árboles como Random Forest y Gradient Boosting son, en general, más susceptibles de ser explicados mediante SHAP que algunas arquitecturas de redes neuronales muy profundas) y en la discusión sobre la interpretabilidad de los resultados. La XAI, por lo tanto, no es solo un complemento deseable, sino un componente cada vez más necesario para la adopción responsable y efectiva del ML en el sector financiero.

Capítulo 3: Diseño Metodológico: Datos y Características

3.1. Exploración y Selección de Fuentes de Datos

La selección de una fuente de datos adecuada es un paso fundamental en cualquier proyecto de aprendizaje automático, y su impacto se propaga a todas las fases subsiguientes del desarrollo del modelo. Para este TFM, se llevó a cabo un análisis preliminar de varios conjuntos de datos públicos de alta calidad que ofrecían información relevante sobre el comportamiento, las actitudes y las características financieras de los inversores. Se identificaron tres fuentes principales [3]:

1. **FINRA National Financial Capability Study (NFCS) (2021, EE. UU.):** Este estudio, con una muestra de aproximadamente 25,000 inversores estadounidenses, se destacó por incluir una pregunta directa sobre la disposición a asumir riesgo financiero para obtener retornos financieros, con una escala de respuesta de 0 a 10. Esta característica lo hacía particularmente atractivo, ya que la variable objetivo potencial era directamente comparable a la salida deseada de la aplicación (perfil de riesgo 1-10). Además, el NFCS contiene una rica variedad de variables psicológicas, demográficas y de comportamiento financiero que podrían actuar como predictores valiosos.[3]
2. **Survey of Consumer Finances (SCF) (2022, EE. UU.):** Con una muestra de alrededor de 11,000 hogares estadounidenses, el SCF ofrece información

sobre el nivel de riesgo auto-reportado (en 4 categorías) y, de manera crucial, sobre la composición real de las carteras de inversión (por ejemplo, el porcentaje invertido en acciones). Esto permitiría estimar la tolerancia al riesgo de forma indirecta, a partir de las decisiones de inversión efectivas (preferencias reveladas), lo que podría complementar la información obtenida de las preferencias declaradas.[3]

3. **Encuesta de Competencias Financieras (ECF) (España, 2021):** Esta encuesta, con una muestra de aproximadamente 21,000 personas en España, incluye preguntas sobre la preferencia por la seguridad en las inversiones y actitudes generales ante decisiones financieras. Su principal valor reside en que representa al público objetivo de la aplicación en el contexto español, permitiendo una potencial adaptación o validación del modelo a las particularidades locales en futuras etapas.[3]

La consideración inicial de estas tres fuentes de datos refleja un enfoque estratégico y multidimensional. Se reconoce implícitamente que la "tolerancia al riesgo" no es una característica unidimensional, sino un constructo complejo que se manifiesta tanto en lo que los individuos declaran (capturado potencialmente por el NFCS), como en lo que efectivamente hacen con su dinero (reflejado en el SCF), y cómo estos factores pueden variar según el contexto demográfico y cultural (abordado por la ECF).[3] Esta triangulación de fuentes, aunque presenta el desafío potencial de reconciliar información diversa, demuestra una comprensión sofisticada del problema de perfilación. La Tabla 3.1 ofrece una visión general comparativa de estos conjuntos de datos.

Tabla 3.1: Panorama de los Conjuntos de Datos Investigados para la Perfilación de Riesgo

Nombre del Dataset (Año)	Tamaño de Muestra (aprox.)	Variable(s) Objetivo Primaria(s) para Riesgo	Variables de Soporte Clave	Enfoque Geográfico	Razón Inicial para Uso
FINRA NFCS (2021)	25,000 inversores	Pregunta directa sobre disposición a asumir riesgo financiero (escala 0-10)	Demográficas, psicológicas, comportamiento, literacidad financiera	EE. UU.	Ajuste directo con la salida del modelo; riqueza en variables psicológicas y de comportamiento.
SCF (2022)	11,000 hogares	Nivel de riesgo (4 categorías); composición real de inversiones (% en acciones)	Demográficas, ingresos, patrimonio, activos y pasivos detallados	EE. UU.	Estimación indirecta de tolerancia al riesgo a partir de decisiones reales (preferencias reveladas).
ECF (España, 2021)	21,000 personas	Preguntas sobre preferencia por seguridad; actitudes ante decisiones financieras	Demográficas, nivel de estudios, situación laboral, conocimientos financieros básicos	España	Representatividad del público objetivo en España; adaptación del modelo al contexto local.

Esta tabla justifica la selección final del NFCS al compararlo con alternativas viables, mostrando un proceso de decisión informado y transparente, esencial para la rigurosidad metodológica del estudio.

3.2. Justificación y Análisis Exploratorio Detallado del Conjunto de Datos NFCS

Tras la exploración inicial, se tomó la decisión estratégica de priorizar el conjunto de datos **FINRA NFCS 2021** para el desarrollo principal del modelo de perfilación de riesgo. La justificación fundamental residió en que este conjunto de datos era el que "mejor se ajusta al objetivo del modelo; incluye variables psicológicas y de comportamiento financiero".[3] La disponibilidad de una pregunta directa sobre la disposición a asumir riesgo, en una escala idéntica (0-10) a la que la aplicación busca generar, simplificaba considerablemente el planteamiento inicial del problema de aprendizaje supervisado y la interpretación de la variable objetivo.[3]

No obstante, esta decisión conllevaba una reflexión metodológica importante. Al optar por el NFCS y su medida de preferencia declarada, el modelo se enfocaría en predecir lo que los usuarios *dicen* sobre su tolerancia al riesgo. Esto difiere de modelar lo que *hacen* en la práctica, que es lo que el SCF podría haber reflejado con mayor fidelidad a través de la composición de sus carteras. La brecha entre lo que se dice y lo que se hace ("say-do gap") es un fenómeno bien conocido en la economía conductual.[3] Si bien esto no invalida el enfoque —la aplicación busca generar un perfil *antes* de la inversión, basándose en la información proporcionada por el usuario—, es un matiz relevante a considerar en la interpretación de los resultados del modelo y para futuras líneas de investigación.

Una vez seleccionado el NFCS 2021, se procedió a un Análisis Exploratorio de Datos (EDA) exhaustivo (reflejado en los notebooks 0_NFCS_2021_EDA y 01_NFCS_2021_Full_EDA). El EDA preliminar del NFCS reveló un predominio de aversión al riesgo entre los encuestados: el 49.3% se ubicó en la categoría de "riesgo bajo" y un 29.4% en "riesgo moderado", indicando una postura general conservadora. Se crearon índices compuestos, como el *Financial Literacy Index (ILI)*, con una media de aproximadamente 0.47, que mostró una correlación positiva con la propensión al riesgo (*Risk Propensity*), sugiriendo que una mayor alfabetización financiera tiende a asociarse con una mayor disposición a asumir riesgos. Una segmentación mediante el algoritmo K-Means (con k=3) permitió identificar tres arquetipos de inversores: un clúster "Conservador" (caracterizado por un ILI de 0.46, bajo compromiso digital y baja propensión al riesgo), un clúster "Digital-nativo" (con ILI más bajo pero muy alto compromiso digital y alta propensión al riesgo) y un clúster "Trader experimentado" (con el ILI más alto, compromiso digital moderado y propensión al riesgo también moderada-alta).[3]

El EDA exhaustivo posterior [3] profundizó en la estructura de los datos:

- **Limpieza Inicial y Estadísticos Descriptivos:** Se convirtieron códigos especiales (como 98 o 99, usualmente indicando "No sabe" o "No contesta") en valores NaN (Not a Number) para su correcto tratamiento. Se calcularon medidas de tendencia central (media, mediana) y dispersión (desviación estándar, rangos) para las variables clave.
- **Distribuciones Clave:** Se analizó en detalle la distribución de la variable objetivo B10 (tolerancia al riesgo declarada), confirmando un predominio de respuestas en la zona de "riesgo promedio". La aplicación de pesos muestrales (WGT1), diseñados para que la muestra refleje mejor a la población general, no modificó sustancialmente esta distribución. También se examinaron las distribuciones de otras variables relevantes como el nivel de ingresos (S_Income) y la percepción de riesgo asociada a las criptomonedas (B24).
- **Análisis Bivariado y Correlaciones:** Se exploraron relaciones entre pares de variables, como edad vs. propensión al riesgo e ingresos vs. propensión al riesgo. Se calculó una matriz de correlación utilizando el coeficiente de Spearman (adecuado para variables ordinales o relaciones no lineales) para identificar asociaciones entre la variable objetivo B10 y otras características. Este análisis reveló una asociación moderada de B10 con el valor de la cartera de inversión (B4) y la confianza en el propio juicio financiero (C25), pero una correlación nula o muy baja con el uso de herramientas de inversión más sofisticadas o arriesgadas como las operaciones de margen y las opciones. Se observó que, en general, a mayor edad y en rangos de ingresos moderados, la tolerancia al riesgo tendía a ser menor. Un hallazgo interesante fue que la confianza en el propio juicio (C25) no siempre se alineaba con el conocimiento financiero real del individuo, sugiriendo la posible presencia de sesgos de exceso de confianza.
- **Datos Faltantes y Outliers:** Se identificaron variables con una alta tasa de valores faltantes (por ejemplo, B6, B20, B34, con una cobertura inferior al 5% del total de la muestra). Se exploró la técnica de winsorización para tratar valores extremos en la variable de peso muestral (WGT1), con el fin de limitar

su influencia desproporcionada.

Estos hallazgos del EDA no son meramente descriptivos; sirvieron como una validación inicial de que las variables contenidas en el NFCS poseían poder predictivo potencial y, crucialmente, guiaron las etapas posteriores de limpieza de datos e ingeniería de características. Por ejemplo, la correlación observada entre el índice de literacidad financiera y la propensión al riesgo sugirió la utilidad de construir una característica como FL_SCORE.[3] La identificación de arquetipos mediante K-Means también podría inspirar la creación de variables categóricas o la exploración de interacciones específicas en el modelado. La decisión de priorizar NFCS, aunque pragmática, tuvo una consecuencia directa en el tipo de modelo que se pudo construir y en la interpretabilidad de sus predicciones, enfocándose en la tolerancia al riesgo declarada por el usuario.

3.3. Proceso de Limpieza y Preprocesamiento de Datos

Con base en los hallazgos del EDA exhaustivo, se procedió a una fase de limpieza y preprocesamiento de datos rigurosa (documentada en el notebook 02_NFCS_Data_Cleaning).[3] El objetivo de esta etapa era transformar el conjunto de datos crudo en un formato limpio, consistente y adecuado para el entrenamiento de los modelos de aprendizaje automático. Las principales tareas realizadas incluyeron [3]:

- **Manejo de Valores Faltantes:**
  - **Eliminación de Columnas:** Se tomaron decisiones basadas en la proporción de datos faltantes por variable. Específicamente, se eliminaron tres columnas (B20, B6 y B34) que superaban un umbral predefinido del 30% de valores ausentes. Esta es una práctica común para evitar la imputación masiva de datos que podría introducir ruido o sesgos significativos.
  - **Imputación de Valores:** Para las variables de escala (numéricas continuas o discretas con muchas categorías) que aún contenían valores faltantes después de la eliminación de columnas, se optó por la

imputación utilizando la **mediana** de cada variable. La mediana es preferible a la media cuando la distribución de la variable es asimétrica o cuando se sospecha la presencia de outliers, ya que es una medida de tendencia central más robusta. Para las variables categóricas restantes con valores faltantes, se utilizó la **moda** (el valor más frecuente) para la imputación.

- **Tratamiento de Outliers:** Las variables de escala que fueron objeto de imputación también se sometieron a un proceso de **winsorización al 1%**. Esta técnica consiste en reemplazar los valores extremos (el 1% más bajo y el 1% más alto) por los valores correspondientes al percentil 1 y al percentil 99, respectivamente. El propósito es limitar el impacto desproporcionado que los outliers pueden tener en el entrenamiento de algunos modelos de ML, sin eliminarlos por completo.
- **Auditoría y Trazabilidad:** Todas las decisiones tomadas durante el proceso de limpieza de datos (columnas eliminadas, métodos de imputación por variable, umbrales de winsorización) quedaron meticulosamente registradas en un archivo de formato JSON. Esta práctica es fundamental para garantizar la trazabilidad y la reproducibilidad de la investigación, permitiendo a otros investigadores (o al propio autor en el futuro) comprender exactamente cómo se transformó el conjunto de datos original.

Al finalizar esta etapa, se obtuvo un conjunto de datos depurado que constaba de 2824 observaciones (inversores) y 92 variables (características). Este dataset estaba libre de valores faltantes y de valores atípicos extremos que pudieran distorsionar el proceso de modelado, dejándolo listo para las fases subsiguientes de división, selección de características y entrenamiento de modelos.[3] Este proceso de limpieza, aunque pueda parecer rutinario, representa un equilibrio delicado entre la necesidad de preservar la mayor cantidad de información útil posible y el imperativo de asegurar la robustez y la fiabilidad de los modelos que se construirán a partir de los datos.

3.4. Ingeniería de Características Iterativa

La ingeniería de características es a menudo descrita como más un arte que una ciencia, y en este TFM, se abordó como un proceso iterativo y multifacético, combinando enfoques automatizados con el indispensable conocimiento del dominio. Este proceso fue crucial para transformar los datos brutos en representaciones más informativas que pudieran mejorar el rendimiento de los modelos de ML.

3.4.1. División de Datos y Selección Inicial Automatizada

El primer paso metodológico tras la limpieza fue la división del dataset del NFCS 2021 (con sus 2824 observaciones y 92 variables) en conjuntos de entrenamiento, validación y prueba (documentado en 03.1_Division_TrainValTest).[3] Esta división se realizó de forma estratificada según la variable objetivo original B10 (tolerancia al riesgo), para asegurar que la distribución de las clases de riesgo fuera similar en los tres conjuntos. Se asignaron proporciones aproximadas de 70% para entrenamiento (1976 observaciones), 15% para validación (424 observaciones) y 15% para prueba (392 observaciones). Durante esta fase, también se crearon algunas variables compuestas iniciales, como FL_SCORE (Índice de Literacidad Financiera) y PORTFOLIO_DIVERSITY (una medida de la diversificación de la cartera).[3]

A continuación, se exploraron diversas técnicas de selección automática de características sobre los datos de entrenamiento (notebook 03.2_Seleccion_Features_Automatica).[3] Se aplicaron métodos univariados como el test ANOVA F-value (para evaluar la relación entre cada característica y la variable objetivo numérica o categórica) y la Información Mutua (para medir la dependencia entre variables). También se utilizaron métodos más sofisticados como la Eliminación Recursiva de Características con Validación Cruzada (RFECV) utilizando un modelo de Regresión Logística, y la importancia de características derivada de un modelo Random Forest. Paralelamente, se evaluó un enfoque de reducción de dimensionalidad mediante Análisis de Componentes

Principales (PCA), buscando retener el 90% de la varianza explicada, lo que resultó en 64 componentes principales.[3]

La validación cruzada sobre diferentes tamaños de subconjuntos de características seleccionadas por estos métodos automáticos mostró un estancamiento (plateau) en la métrica F1-macro (alrededor de 0.43) a partir de aproximadamente k=20 características. La selección univariada con k=20 características, de hecho, redujo la exactitud (accuracy) de un modelo base de Regresión Logística (de 0.46 a 0.41), mientras que RFECV con las 20 mejores características alcanzó una exactitud de aproximadamente 0.43. El pipeline que combinaba PCA (64 componentes) con Regresión Logística obtuvo el mejor rendimiento en esta fase inicial, con una exactitud en el conjunto de prueba de 0.49 y de 0.46 en el de validación. Entre las características que recurrentemente fueron identificadas como importantes por estos métodos se encontraban TRADER_SCORE, S_Age (edad del inversor), B4_log (valor de la cartera transformado logarítmicamente) y variables pertenecientes a las series F31 (fuentes de información financiera) y G30 (actitudes financieras) del cuestionario NFCS.[3]

Sin embargo, esta aproximación puramente algorítmica reveló una limitación práctica fundamental: "algunas variables son de preguntas muy técnicas o aisladas y eso me va a dificultar luego hacer una selección de input mínimos que hay que pedir al usuario".[3] Esta observación subraya una tensión crítica en el desarrollo de modelos de ML para aplicaciones del mundo real: la significancia estadística no siempre se traduce directamente en utilidad práctica o viabilidad para la implementación en un cuestionario de usuario que debe ser conciso y comprensible. El conocimiento del dominio emerge aquí como un filtro indispensable para los resultados de los métodos automatizados.

3.4.2. Curación Manual de Características Guiada por Dominio

Ante las limitaciones de la selección puramente automatizada, se realizó una curación manual de características (notebook 03.3_Seleccion_Features_Manual), donde el conocimiento del dominio y los requisitos de la aplicación final jugaron un papel preponderante.[3] Se seleccionaron 42 variables, agrupadas en seis categorías conceptuales: demográficas, información del portafolio, comportamiento de trading, actitudes financieras, alfabetización financiera e fuentes de información consultadas. Este proceso se guio no solo por la potencial relevancia predictiva inferida del EDA, sino también por el imperativo de que las variables seleccionadas pudieran razonablemente obtenerse a través de un cuestionario conciso y amigable para el usuario final de la aplicación. Se verificó que todas las 42 variables seleccionadas fueran de tipo numérico (float64) y no contuvieran valores nulos tras el preprocesamiento.[3]

Un análisis de correlación sobre este conjunto de 42 variables reveló la presencia de multicolinealidad entre algunos ítems, por ejemplo, entre diferentes preguntas sobre fuentes de información consultadas (e.g., F31_5 y F31_2 con una correlación absoluta de aproximadamente 0.65) y entre preguntas relacionadas con la actividad de trading (B30 y B31 con una correlación absoluta de ≈0.56).[3] Con estas 42 características seleccionadas manualmente, se entrenaron modelos base (Regresión Logística, Random Forest, Gradient Boosting) utilizando validación cruzada de 5 folds. La Regresión Logística, configurada con

class_weight='balanced' para mitigar el desbalance de clases, mostró el mejor rendimiento en esta etapa, alcanzando un F1-macro medio en validación cruzada de aproximadamente 0.45. En la evaluación final sobre el conjunto de test, este modelo obtuvo un F1-macro de 0.444 y un área bajo la curva ROC (AUC) de 0.732. Las variables que mostraron los coeficientes (pesos) más altos en este modelo de Regresión Logística, indicando una mayor influencia en la predicción, incluyeron F31_4 (frecuencia de consulta de asesores financieros), B2_23 y B2_24 (preguntas sobre productos de inversión específicos), S_Age (edad) y F31_2 (frecuencia de consulta de noticias financieras).[3]

Este paso de curación manual subraya el valor insustituible de la experiencia y el conocimiento del dominio. La selección manual no es arbitraria; implica que el investigador posee hipótesis implícitas sobre qué factores son *realmente* importantes y medibles para la perfilación del riesgo en un contexto práctico. Las categorías de variables seleccionadas (demográficas, portafolio, comportamiento, etc.) representan constructos que la teoría y la práctica financiera sugieren como relevantes para entender la propensión al riesgo. La Tabla 3 conceptual de [3] ilustra bien esta transición, donde variables "técnicas o aisladas" identificadas automáticamente son descartadas manualmente por su dificultad de implementación, a pesar de su posible poder predictivo.

3.4.3. Optimización y Reducción Adicional (17 Características)

A partir de las 42 características seleccionadas manualmente, se aplicó una etapa adicional de reducción de dimensionalidad y optimización (notebook 03.4_Optimizacion_Features).[3] Se realizaron tres análisis de componentes principales (PCA) separados sobre grupos específicos de variables correlacionadas, correspondientes a las series B2 (tipos de productos de inversión), F30_ (objetivos financieros) y F31_* (fuentes de información). El objetivo era condensar la información de estos grupos de preguntas en un número menor de variables sintéticas (componentes principales), eliminando las variables originales correspondientes para reducir la redundancia y la multicolinealidad. Este proceso, junto con otras selecciones y eliminaciones no detalladas, condujo a un conjunto final de 17 características.[3]

El uso de PCA en este contexto representa una técnica de compresión y, potencialmente, de reducción de ruido. Al combinar linealmente variables correlacionadas en componentes ortogonales, se pueden crear nuevas características más densas en información. Sin embargo, esto conlleva un trade-off: las componentes principales, al ser combinaciones de las variables originales, pueden perder interpretabilidad directa en términos de las preguntas originales del cuestionario. Es decir, puede ser más difícil explicar el impacto de una componente principal en la predicción final en comparación con el impacto de una variable original como la "edad".

Con este conjunto de 17 características, se construyó un ColumnTransformer para estandarizar el preprocesamiento (imputación y escalado) y se definieron pipelines de modelado que incluían el uso de SMOTE (Synthetic Minority Over-sampling Technique) para abordar el desbalance de clases, seguido de tres algoritmos clasificadores: Regresión Logística, Random Forest y Gradient Boosting. Los hiperparámetros de cada modelo se optimizaron mediante GridSearchCV con validación cruzada de 5 folds.[3]

Los resultados en el conjunto de test para estos modelos, probablemente evaluados sobre el problema original de 4 clases de riesgo (antes de la decisión de reducir a 2 clases), fueron los siguientes [3]:

- **Regresión Logística:** Accuracy = 0.3892, F1-macro = 0.3731, AUC_ovr = 0.6968.
- **Random Forest:** Accuracy = 0.5165, F1-macro = 0.4204, AUC_ovr = 0.7055.
- **Gradient Boosting:** Accuracy = 0.5283, F1-macro = 0.4302, AUC_ovr = 0.7032.

Gradient Boosting ofreció el mejor compromiso en términos de rendimiento en esta etapa. Se observó que los métodos de ensamble (Random Forest, Gradient Boosting) superaron consistentemente a la Regresión Logística, y que se había alcanzado un cierto plateau en las métricas de Accuracy y F1-macro, sugiriendo que se había extraído el límite de la información predictiva contenida en estas 17 variables para el problema de clasificación original (probablemente de 4 clases).[3]

3.4.4. Ingeniería Avanzada Manual ("95_Original" y "95_Ultimate")

A pesar de los esfuerzos de selección y optimización que llevaron al conjunto de 17 características, el rendimiento de los modelos no se consideró suficientemente convincente.[3] Esto llevó a la conclusión de que para mejorar significativamente el poder predictivo era necesario enriquecer la información de entrada mediante una ingeniería de características más avanzada, partiendo de un conjunto más amplio de las 95 variables base disponibles en el dataset NFCS limpio.[3]

Primero, se exploró la generación automatizada de características (notebook 05.0_Ingenieria_Avanzada_Automatica) sobre las 95 variables base. Se utilizó la librería Featuretools para generar aproximadamente 120 características agregadas y transformaciones (e.g., sumas, medias, diferencias entre variables relacionadas), complementado con la creación de PolynomialFeatures de grado 2 (para capturar interacciones multiplicativas simples). Esto expandió la matriz de características a unas 250 columnas. Posteriormente, un pipeline de filtrado (eliminación de características con baja varianza, selección de las 50 mejores con SelectKBest basado en F-value, y selección mediante regularización Lasso) redujo este conjunto a 48 características finales. Un modelo LightGBM entrenado con estas 48 características (optimizando F1-macro con GridSearchCV de 5 folds) alcanzó en el conjunto de test una Accuracy de 0.65 y un F1-macro de 0.61. En validación, las métricas fueron Accuracy=0.67 y F1-macro=0.63. Esto representó una mejora de +0.05 en el F1-macro de validación y +0.03 en el F1-macro de test en comparación con el LightGBM entrenado con las 17 características.[3] Sin embargo, se concluyó que "esta ingeniería automática no ha mejorado mucho los resultados comparándolo con los datos de entrenamiento de base 95 variables", sugiriendo que, aunque hubo una mejora sobre el conjunto de 17 características, el rendimiento con las 48 características generadas automáticamente no fue drásticamente superior al que se podría obtener modelando directamente y de forma adecuada las 95 variables originales.[3]

Este resultado impulsó un cambio de enfoque hacia una **ingeniería avanzada manual de características** (notebook 05.1_Ingenieria_Avanzada_95), partiendo nuevamente de las 95 variables base del NFCS. En este proceso, guiado por el conocimiento del dominio financiero y la comprensión de los constructos psicológicos relacionados con el riesgo, se crearon 11 nuevas variables derivadas. Estas incluían:

- **Scores de dominio:** Combinación de varias preguntas para medir conceptos como la sofisticación financiera o la propensión al trading activo.

- **Gap de percepción de riesgo:** Diferencias entre la percepción de riesgo de ciertos activos y el conocimiento objetivo sobre ellos, o entre la tolerancia al riesgo declarada y otros indicadores.
- **Interacciones específicas:** Productos de variables que, según la teoría o la intuición, podrían tener un efecto combinado sobre la tolerancia al riesgo (e.g., interacción entre edad y uso de ciertos productos financieros).
- Binning (agrupación) del valor de la cartera: Transformación de la variable continua del valor de la cartera en categorías ordinales, lo que puede ayudar a capturar relaciones no lineales.
  Este proceso resultó en un nuevo conjunto de datos de 103 características, denominado "95_ultimate".3

La comparación formal entre el conjunto de datos original con 95 características (denominado **"95_Original"**) y el conjunto enriquecido "95_ultimate" (notebook 06.0_Modelado_Comparacion), utilizando un modelo Random Forest con SMOTE y GridSearchCV (3 folds) para un problema de clasificación binaria (2 clases), arrojó resultados interesantes. En términos de AUC en el conjunto de test, los valores fueron muy similares: 0.7752 para "95_Original" y 0.7730 para "95_Ultimate". Sin embargo, se observó que el conjunto "95_Ultimate" demostró ser mejor para "casos específicos". Por ejemplo, para la clasificación binaria, el conjunto "95_Ultimate" logró un recall "detector" (probablemente para la clase de alto riesgo) de 0.7810, mientras que el recall "protector" (probablemente para la clase de bajo riesgo) fue de 0.6533.[3]

Este hallazgo es significativo: la superioridad de la ingeniería de características informada por el dominio sobre la puramente automática (o incluso sobre el conjunto original sin enriquecer) para capturar matices cruciales es evidente. El hecho de que la ingeniería automática con 48 características no mejorara sustancialmente sobre las 95 originales, mientras que la ingeniería manual avanzada "95_ultimate" sí lo hizo para "casos específicos", sugiere que el conocimiento del dominio es clave para crear características que capturen interacciones o conceptos (e.g., el "gap de percepción de riesgo") que los algoritmos automáticos, sin ese contexto semántico, no pueden generar fácilmente. Estas características "hechas a mano" probablemente encapsulan hipótesis más sofisticadas sobre los factores determinantes del comportamiento del inversor frente al riesgo, permitiendo a los modelos capturar señales que eran menos evidentes con las variables originales o las generadas automáticamente. Esto llevó directamente a un mejor rendimiento en la detección de perfiles de riesgo específicos, como se verá en los modelos especializados.

3.5. Estrategia de Definición de la Variable Objetivo (Reducción a Clasificación Binaria)

Uno de los desafíos más significativos identificados tempranamente en el análisis de los datos del NFCS fue el "gran desbalanceo" de la variable objetivo original B10 (tolerancia al riesgo en una escala de 0 a 10).[3] Una variable objetivo con muchas clases, donde algunas de ellas tienen muy pocas instancias, dificulta el aprendizaje de los modelos de ML, ya que estos tienden a favorecer a las clases mayoritarias y pueden tener un rendimiento pobre en la predicción de las clases minoritarias.

Para mitigar este problema y mejorar la robustez de los modelos, se exploró la reducción del número de clases de salida. Estos experimentos se realizaron utilizando las 17 características sintéticas obtenidas en la etapa de optimización descrita en la sección 3.4.3.

1. Modelo con 3 Clases de Salida (notebook 03.5_Modelo_3outputs):
   En este experimento, las etiquetas originales de la variable B10 se remapearon a tres categorías de riesgo: "bajo", "medio" y "alto". Se construyeron pipelines de modelado que incluían SMOTE (para sobremuestrear las clases minoritarias) y StandardScaler (para escalar las características). Se utilizó GridSearchCV (con validación cruzada de 5 folds) para optimizar los hiperparámetros de tres algoritmos (Regresión Logística - LR, Random Forest - RF, Gradient Boosting - GB), buscando maximizar el F1-macro. Los mejores F1-macro obtenidos en validación cruzada fueron aproximadamente 0.4529 para LR, 0.4900 para RF y 0.4851 para GB. En el

conjunto de test, el modelo Random Forest, tras una optimización de los umbrales de decisión, alcanzó una Accuracy de 0.5401, un F1-macro de 0.4623 y un AUC_ovr (área bajo la curva ROC para el enfoque One-vs-Rest) de aproximadamente 0.6625. Se observó que la optimización de umbrales mejoró el F1-macro en validación en aproximadamente 0.05 puntos.[3]

2. Modelo con 2 Clases de Salida (notebook 03.5_Modelo_2outputs):
   En este segundo experimento, las etiquetas originales se remapearon a solo dos categorías: un grupo combinado de "riesgo bajo/medio" y un grupo de "riesgo alto". Se compararon los mismos tres algoritmos (Regresión Logística, Random Forest y Gradient Boosting), pero esta vez sin utilizar SMOTE y empleando en su lugar el parámetro class_weight='balanced' en los modelos para dar más peso a la clase minoritaria. Se exploraron también estrategias de ajuste del umbral de decisión, especialmente para el Random Forest, con el objetivo de optimizar para diferentes escenarios, como "Detectar Alto Riesgo" (maximizando el recall de la clase de alto riesgo) o "Proteger Bajo Riesgo" (minimizando los falsos positivos para la clase de bajo/medio riesgo). Se concluyó que un modelo Random Forest con un umbral ajustado para "Proteger Bajo Riesgo" lograba el mejor compromiso global, controlando los errores de clasificación considerados más críticos desde una perspectiva prudencial.[3]

Tras estos experimentos, se tomó una decisión metodológica crucial: **adoptar un marco de clasificación binaria (2 clases de salida)** para el resto del desarrollo del TFM. Esta decisión se basó en la evidencia de que esta configuración, según se indica en el material de referencia, "probablemente ofrecía un mejor balance de clases y/o métricas de rendimiento superiores en los modelos preliminares, simplificando la tarea de clasificación".[3] La reducción de clases es una concesión pragmática que busca mejorar la tratabilidad del problema. Intentar predecir una escala fina de 0-10 (11 clases) o incluso 4 clases con datos desbalanceados y potencialmente ruidosos es un desafío considerablemente mayor que una clasificación binaria. La simplificación facilita el manejo del desbalance (ya sea mediante SMOTE o class_weight) y puede conducir a modelos más robustos y con mejor generalización, aunque se pierda granularidad en la predicción directa.

La Tabla 3.2 resume el impacto de esta decisión estratégica en la métrica F1-macro en el conjunto de test, utilizando los mejores resultados obtenidos para cada configuración con modelos y características comparables en esas etapas exploratorias.

Tabla 3.2: Impacto de la Reducción de Clases de la Variable Objetivo en F1-macro (Test) (Adaptada de 3)

Configuración Objetivo	Modelo y Características Base	F1-Macro (Test)	Razón para Elección Final
Original (4 clases)	Random Forest (95_Original features, Notebook 06.1)	0.4735	Desbalance dificulta el aprendizaje; distinciones entre clases adyacentes pueden ser ruidosas.
3 Clases	Random Forest (95_Ultimate features, Notebook 06.0)	0.5477	Mejora sobre 4 clases, pero la complejidad persiste.
2 Clases (Elegida)	**LightGBM (17 features, Notebook 04.3)**	**0.58**	**Logra un mejor rendimiento en F1-macro con un modelo más simple; simplifica el problema de clasificación y facilita el manejo del desbalance.**

Esta tabla es vital porque una de las decisiones metodológicas más impactantes fue la redefinición de la variable objetivo. Mostrar el impacto en el F1-macro [3] proporciona una justificación basada en evidencia para esta simplificación del problema, orientando el esfuerzo de modelado posterior. Es importante notar que esta decisión tiene implicaciones para la aplicación final: la salida binaria del modelo necesitará ser mapeada de nuevo a la escala de 1-10 requerida por la interfaz de usuario de la aplicación [3], un paso que podría introducir sus propias heurísticas o reglas adicionales.

Capítulo 4: Desarrollo y Experimentación de Modelos de Aprendizaje Automático

4.1. Configuración Experimental

Una vez definido el conjunto de características y la variable objetivo binaria, se procedió al desarrollo y evaluación sistemática de los modelos de aprendizaje automático. Para asegurar la validez y reproducibilidad de los resultados, se estableció una configuración experimental robusta:

- **División de Datos:** Como se mencionó en el Capítulo 3 (sección 3.4.1), el conjunto de datos NFCS limpio se dividió de forma estratificada en entrenamiento (70%), validación (15%) y prueba (15%).[3] El conjunto de entrenamiento se utilizó para ajustar los parámetros de los modelos; el de validación, para la selección de hiperparámetros y la toma de decisiones intermedias (como la elección del mejor conjunto de características o la optimización de umbrales); y el de prueba, reservado exclusivamente para la evaluación final del rendimiento de los modelos seleccionados, proporcionando una estimación insesgada de su capacidad de generalización a datos no vistos.
- **Métricas de Evaluación:** Dada la naturaleza del problema de clasificación, especialmente con clases potencialmente desbalanceadas y con diferentes costes de error asociados a la mala clasificación, se utilizaron múltiples métricas para evaluar el rendimiento [3]:
    - **Accuracy (Exactitud):** Proporción de predicciones correctas. Aunque es una métrica intuitiva, puede ser engañosa en problemas con clases desbalanceadas.
    - **F1-macro:** Media armónica de la precisión y el recall, calculada para cada clase y luego promediada. Es una métrica más robusta para clases desbalanceadas que la exactitud simple, ya que considera el rendimiento en todas las clases por igual. También se menciona el uso de F1-ponderado (F1_weighted), que pondera el F1 de cada clase por su soporte (número de instancias verdaderas).
    - **AUC-ROC (Área Bajo la Curva Característica Operativa del Receptor):**

Representa la capacidad del modelo para distinguir entre las clases. Un valor de AUC de 0.5 indica un rendimiento aleatorio, mientras que un valor de 1.0 indica una clasificación perfecta. Es una métrica útil porque es insensible al umbral de clasificación.
    - **Recall (Sensibilidad o Tasa de Verdaderos Positivos):** Proporción de instancias positivas reales que fueron correctamente identificadas por el modelo (TP / (TP + FN)). Es particularmente importante cuando el coste de los falsos negativos es alto, como en la detección de perfiles de alto riesgo.
    - **Precision (Precisión):** Proporción de instancias clasificadas como positivas que realmente lo son (TP / (TP + FP)). Es importante cuando el coste de los falsos positivos es alto.
    - **Matriz de Confusión:** Tabla que visualiza el rendimiento de un clasificador, mostrando el número de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN).
- **Validación:** Para la optimización de hiperparámetros (e.g., mediante GridSearchCV o RandomizedSearchCV) y para obtener estimaciones más estables del rendimiento durante el desarrollo, se empleó la **validación cruzada** (generalmente de 5 o 3 folds, según la etapa del proyecto) sobre el conjunto de entrenamiento.[3]
- **Software y Librerías:** Todo el proceso de análisis de datos, ingeniería de características y modelado se llevó a cabo utilizando el lenguaje de programación Python, apoyándose en librerías estándar del ecosistema de ciencia de datos como Pandas para la manipulación de datos, NumPy para operaciones numéricas, Matplotlib y Seaborn para visualización, y Scikit-learn para la implementación de algoritmos de ML, preprocesamiento y evaluación. También se utilizaron librerías específicas como LightGBM para el algoritmo homónimo y Featuretools para la ingeniería automática de características.[3]

La elección de estas métricas, en particular el F1-macro y el enfoque posterior en el Recall de la clase de alto riesgo para el "Modelo Detector" [3], refleja una comprensión profunda de los desafíos inherentes al problema: el desbalance de clases y la importancia diferencial de los errores de clasificación. Clasificar erróneamente a un inversor de alto riesgo como de bajo riesgo (un falso negativo para la clase "alto riesgo") puede tener consecuencias más perjudiciales para el inversor y para la entidad financiera que el error opuesto. Una configuración experimental robusta, con una división adecuada de los datos, el uso de validación cruzada y la selección de métricas apropiadas, es fundamental para obtener resultados fiables, evitar el sobreajuste a los datos de entrenamiento y asegurar que las conclusiones extraídas sean válidas.

4.2. Modelos Base y Benchmarking Inicial (con 17 características optimizadas, objetivo binario)

Tras la decisión de adoptar un marco de clasificación binaria y la optimización que resultó en un conjunto de 17 características (ver secciones 3.4.3 y 3.5), se realizó una primera ronda formal de entrenamiento y evaluación comparativa de modelos. El objetivo era establecer una línea base de rendimiento y determinar si este conjunto reducido de características era suficiente o si se requería una ingeniería más profunda. Se evaluaron tres algoritmos para la clasificación binaria [3]:

1. **Regresión Logística (notebook 04.1_Modelo_RegresionLogistica):**
    - Se identificó la mejor configuración de hiperparámetros como C=0.01 (inverso de la fuerza de regularización) y solver='saga' (adecuado para datasets grandes y regularización L1 o L2).
    - En validación cruzada (CV), con un umbral de decisión de 0.5, el modelo alcanzó un F1-ponderado de aproximadamente 0.6889 y un AUC superior a 0.7.
    - Se experimentó con la optimización del umbral de decisión. Al ajustar el umbral a 0.77 con el objetivo de maximizar el recall de la clase 0 (Bajo/Medio Riesgo), este recall subió a un impresionante 0.99. Sin embargo, esto tuvo un coste significativo en la detección de la clase de Alto Riesgo, cuyo recall cayó drásticamente a 0.02. Este resultado ilustra el clásico trade-off entre sensibilidad y especificidad (o entre recalls de diferentes clases) al modificar el umbral.

2. **Random Forest (notebook 04.2_Modelo_RandomForest):**
    - La mejor configuración de hiperparámetros incluyó n_estimators=200 (número de árboles), max_depth=10 (profundidad máxima de los árboles), max_features='sqrt' (número de características a considerar en cada división) y class_weight='balanced' (para manejar el desbalance de clases).
    - En CV, el modelo mostró un F1-ponderado de aproximadamente 0.71 y un AUC ROC de alrededor de 0.77.
    - Evaluado en el conjunto de test (con umbral de 0.5), el modelo obtuvo una Accuracy de 0.58 (0.60 en validación), un F1-macro de 0.55 (0.57 en validación) y un AUC de 0.78 (0.80 en validación).
    - Al ajustar el umbral de decisión a aproximadamente 0.42 (probablemente para mejorar la detección de la clase minoritaria de Alto Riesgo), el recall para esta clase en el conjunto de validación pasó de 0.52 a 0.75.
3. **LightGBM (notebook 04.3_Modelo_LightGBM):**
    - La configuración óptima para LightGBM fue n_estimators=200, learning_rate=0.05, num_leaves=31, max_depth=6 y class_weight='balanced'.
    - En CV, este modelo alcanzó un F1-ponderado de aproximadamente 0.74 y un AUC ROC de alrededor de 0.80.
    - En el conjunto de test (con umbral de 0.5), LightGBM demostró ser el mejor de los tres, con una Accuracy de 0.62 (0.64 en validación), un F1-macro de 0.58 (0.61 en validación) y un AUC de 0.82 (0.84 en validación).
    - Similarmente, al ajustar el umbral a aproximadamente 0.47, el recall para la clase de Alto Riesgo en validación subió de 0.56 a 0.79.

La Tabla 4.1 resume estos resultados en el conjunto de test con el umbral de decisión estándar de 0.5.

Tabla 4.1: Comparación del Rendimiento en Test de Modelos Iniciales (Objetivo de 2 Salidas, 17 Características, Umbral 0.5)
(Adaptada de 3)

Modelo	Accuracy (Test)	F1-macro (Test)	AUC-ROC (Test)	Recall Alto Riesgo (Test, umbral 0.5, estimado de validación)
Regresión Logística	(No disponible en test para umbral 0.5)	(No disponible en test para umbral 0.5)	(CV:>0.70)	(No disponible en test para umbral 0.5)
Random Forest	0.58	0.55	0.78	0.52
LightGBM	0.62	0.58	0.82	0.56

Aunque LightGBM mostró el mejor desempeño global en esta ronda inicial (F1-macro en test de 0.58 y AUC de 0.82), se consideró que estos resultados no alcanzaban el nivel de rendimiento deseado para una mejora sustancial sobre el sistema existente.[3] La experimentación con los umbrales de decisión, incluso en esta etapa temprana, ya demostraba un intento de alinear el comportamiento del modelo con objetivos de negocio específicos, prefigurando la optimización de modelos especializados que se realizaría posteriormente. El rendimiento "no convincente" con este conjunto de 17 características, a pesar de la optimización de hiperparámetros y el ajuste de umbrales, fue un catalizador crucial. Demostró que la simple selección y reducción de dimensionalidad de las características originales no era suficiente, impulsando la investigación hacia una fase de ingeniería de características más profunda y guiada por el conocimiento del dominio.

4.3. Impacto de la Ingeniería Avanzada de Características en el Rendimiento

La conclusión de la fase de benchmarking inicial fue clara: para mejorar significativamente el rendimiento de los modelos, era necesario enriquecer la información de entrada mediante una **ingeniería avanzada de características**, partiendo de un conjunto más amplio de las 95 variables base del NFCS.[3] Este proceso se abordó en dos vertientes: una automática y otra manual, guiada por el dominio.

- Ingeniería Automática de Características (notebook 05.0):
  Se aplicaron técnicas de generación automática de características sobre los datos de entrenamiento con las 95 variables originales. Se utilizó Featuretools para crear aproximadamente 120 nuevas características basadas en transformaciones y agregaciones de las existentes, y se complementó con PolynomialFeatures (grado 2) para capturar interacciones simples, resultando en un espacio de unas 250 características. Un pipeline de filtrado posterior (baja varianza, SelectKBest con k=50, y regularización Lasso) redujo este conjunto a 48 características finales. Un modelo LightGBM entrenado con estas 48 características alcanzó en el conjunto de test una Accuracy de 0.65 y un F1-macro de 0.61. Si bien esto representó una mejora sobre los modelos entrenados con las 17 características (un incremento de +0.03 en F1-macro de test), se concluyó que esta mejora era modesta y no superaba drásticamente lo que se podría obtener modelando adecuadamente las 95 variables originales.3
- Ingeniería Manual Avanzada ("95_ultimate", notebook 05.1):
  El enfoque se desplazó entonces hacia una ingeniería de características manual, más intensiva en conocimiento del dominio. Partiendo de las 95 variables base, se crearon 11 nuevas variables derivadas, diseñadas para capturar conceptos financieros y comportamentales más complejos. Estas incluían "scores de dominio" (e.g., un índice de propensión al trading), "gap de percepción de riesgo" (e.g., la diferencia entre la aversión al riesgo declarada y la tenencia de activos volátiles), interacciones específicas entre variables y la categorización (binning) del valor de la cartera. Este proceso resultó en un dataset de 103 características, denominado "95_ultimate".3

La comparación del rendimiento de modelos entrenados con el conjunto de 95 características originales ("95_Original") y el conjunto enriquecido "95_ultimate" (notebook 06.0, utilizando Random Forest con SMOTE para clasificación binaria) reveló que, aunque los AUC generales en test eran muy similares (0.7752 para "95_Original" vs. 0.7730 para "95_Ultimate"), el conjunto "95_Ultimate" demostró ser superior para "casos específicos". Notablemente, el conjunto "95_Ultimate" logró un recall "detector" (para la clase de alto riesgo) del 78.10%, indicando una mejor capacidad para identificar a los inversores más propensos al riesgo.[3]

La Tabla 4.2 resume el impacto de las diferentes estrategias de ingeniería de características en el rendimiento del modelo para la clasificación binaria.

Tabla 4.2: Impacto de la Ingeniería de Características en el Rendimiento del Modelo (2 Clases, Test)
(Adaptada de 3)

Conjunto de Características Utilizado	Modelo Principal Usado	AUC-ROC (Test)	F1-Macro (Test)	Observaciones Clave (basadas en resúmenes)
Base ~17 Características (Optimizadas manualmente)	LightGBM (04.3)	0.82	0.58	Buen punto de partida, pero rendimiento general limitado.
Base 95 Características (NFCS Limpio, "95_Original")	Random Forest (06.1)	0.7752	≈0.70	Buen rendimiento general, sólida línea base.
Características Avanzadas Automáticas (48 feat. desde 95 base)	LightGBM (05.0)	(No disponible directamente)	0.61	Mejora modesta sobre las 17 feat., pero no supera a las 95_Original en F1-macro.
"95_ultimate" (Ingeniería Manual Avanzada, 103 feat.)	**Random Forest (06.0/06.3)**	**0.7730 (06.0) / 0.7711 (06.3)**	**≈0.69 (06.3)**	**AUC similar a 95_Original, pero mejor para "casos específicos", como lo demuestra el recall de 0.78-0.81 para la clase de alto riesgo (detector).**

Esta tabla es fundamental para narrar la historia de la mejora del modelo. Demuestra que, si bien la ingeniería automática proporcionó una mejora modesta, fue la ingeniería manual avanzada, informada por el conocimiento del dominio, la que permitió capturar matices cruciales para tareas específicas como la detección de perfiles de alto riesgo. Las 11 características creadas manualmente para el conjunto "95_ultimate" (como los scores de dominio o el gap de percepción de riesgo) [3], al encapsular conceptos financieros y conductuales relevantes, tuvieron un impacto más significativo que las transformaciones genéricas producidas automáticamente. Estas características "hechas a mano" permitieron a los modelos capturar señales predictivas que eran menos evidentes con los conjuntos de características anteriores, lo que condujo directamente a un mejor rendimiento en la identificación de perfiles de riesgo específicos.

4.4. Optimización Especializada de Modelos

El descubrimiento de que diferentes conjuntos de características (particularmente "95_Original" y "95_Ultimate") y, por extensión, diferentes configuraciones de modelos, sobresalían en distintos aspectos del problema de predicción, llevó a una fase de especialización y optimización de modelos. Se reconoció que un único modelo "óptimo" podría no ser la mejor solución para todos los posibles casos de uso o requisitos de la aplicación. Por lo tanto, se decidió explorar la optimización de modelos para tres objetivos distintos, utilizando el conjunto de características que se había mostrado más prometedor para cada contexto.[3]

Se tomó la decisión de "para que casos utilizar X_train base [95_Original] y para que casos utilizar los datos de 95_ultimate".[3] Esto implicó un análisis para definir cuándo cada conjunto de características sería más apropiado, sentando las bases para un sistema de perfilación potencialmente más adaptativo, donde diferentes modelos o umbrales podrían aplicarse según el contexto del usuario o los objetivos de la entidad financiera.

Los tres modelos especializados desarrollados fueron:

1. **Optimización del Modelo Académico (notebook 06.1.Optimizacion_Modelo_Academico):**
    - **Objetivo:** Maximizar el rendimiento predictivo general, utilizando métricas estándar como AUC y F1-macro. Este modelo serviría como un benchmark de la mejor capacidad predictiva global alcanzable.

- **Dataset:** Se utilizó el conjunto "95_Original" (95 características base del NFCS).
- **Modelo y Técnicas:** Se empleó un Random Forest, combinado con StandardScaler para el escalado de características y SMOTE para manejar el desbalance de clases. La optimización de hiperparámetros se realizó en dos etapas: primero con RandomizedSearchCV para una exploración amplia del espacio de hiperparámetros, seguida de GridSearchCV para un ajuste fino alrededor de las mejores regiones encontradas.
- **Resultados (Test, 2 clases):** El modelo alcanzó un **AUC de 0.7752**, una Precisión Global (Accuracy) de aproximadamente **0.73**, y un **F1-macro de aproximadamente 0.70**.
- **Resultados (Test, 4 clases):** Para tener una comparativa con el problema original más granular, este mismo enfoque se aplicó al problema de 4 clases, obteniendo un F1-macro de 0.4735 y una Exactitud de aproximadamente 0.58. Estos valores, inferiores a los de 2 clases, reafirman la dificultad del problema más granular y la ventaja de la simplificación a 2 clases para el rendimiento general.

2. **Optimización del Modelo de Coste (notebook 06.2.Optimizacion_Modelo_Coste):**
   - **Objetivo:** Minimizar el coste económico asociado a las predicciones erróneas. Esto implica definir una matriz de costes que asigne diferentes penalizaciones a los distintos tipos de errores de clasificación (e.g., un falso negativo para la clase "alto riesgo" podría tener un coste mayor que un falso positivo).
   - **Dataset:** Se utilizó el conjunto "95_Ultimate" (103 características, incluyendo las 11 creadas manualmente), dado su mejor rendimiento en "casos específicos".
   - **Modelo y Técnicas:** Se empleó un Random Forest con SMOTE. La optimización de hiperparámetros se realizó mediante RandomizedSearchCV, utilizando una función de coste personalizada

como métrica de scoring para guiar la búsqueda hacia modelos que minimicen el coste total definido.

- **Resultados (Test, 3 clases):** Para este escenario, se evaluó un modelo con 3 clases de salida. El coste medio por predicción fue de **0.5354**, con una Accuracy de 0.64 y un F1-score ponderado de 0.62. Es importante destacar que el recall para la clase de riesgo alto fue bajo (0.12), lo que sugiere que, aunque el coste medio general fuera aceptable según la matriz definida, la detección de perfiles de muy alto riesgo podría ser deficiente si la matriz de costes no penaliza suficientemente este tipo de error.
- **Resultados (Test, 4 clases):** Con 4 clases, el coste medio fue de 0.7358, con una Accuracy de 0.58 y un F1-score ponderado de 0.56.

3. **Optimización del Modelo Detector (notebook 06.3.Optimizacion_Modelo_Detector):**
   - **Objetivo:** Maximizar la identificación de perfiles de alto riesgo, es decir, maximizar el recall para la clase positiva (definida como "Alto Riesgo") en un problema de clasificación binaria. Este modelo es crucial desde una perspectiva de gestión de riesgos y protección al inversor.
   - **Dataset:** Se utilizó el conjunto "95_Ultimate", por su demostrada capacidad para mejorar la detección en casos específicos.
   - **Modelo y Técnicas:** Se empleó un RandomForest con SMOTE, donde el parámetro k_neighbors de SMOTE se ajustó a 9. La optimización de hiperparámetros también se realizó en dos fases (RandomizedSearchCV y GridSearchCV). Los hiperparámetros finales del RandomForest incluyeron n_estimators=200, max_depth=40, min_samples_split=5, min_samples_leaf=1 y max_features='log2'.
   - **Resultados (Test, 2 clases):** El modelo alcanzó un **AUC de 0.7711**. Para la Clase 1 (Alto Riesgo), se logró un **Recall del 81%** y una **Precisión del 77%**. La Accuracy global fue del 72%, y el F1-macro de aproximadamente 0.69. La matriz de confusión fue [,], indicando 223 verdaderos positivos para Alto Riesgo, 51 falsos negativos, 84 verdaderos negativos y 66 falsos

positivos.

Esta triple optimización resulta en una suite de modelos, cada uno adaptado a un propósito específico. El desarrollo de estos tres modelos distintos con diferentes funciones objetivo (rendimiento general, minimización de costes ponderados, detección específica de alto riesgo) es una conclusión sofisticada del proceso de modelado. Demuestra que no existe un único "mejor" modelo para todas las situaciones posibles. El "Modelo Académico" puede ser útil para benchmarking y comprensión general, el "Modelo de Coste" para una implementación que considere explícitamente las implicaciones económicas de los errores de clasificación (siempre que la matriz de costes esté bien definida y validada), y el "Modelo Detector" para priorizar la identificación proactiva de aquellos perfiles de inversor que requieren mayor atención o estrategias de mitigación de riesgos. Este enfoque de "suite de modelos" es considerablemente más flexible y adaptable a las necesidades de negocio que un enfoque monolítico de "un solo modelo para todo". Implica que la aplicación final podría incorporar una lógica para seleccionar qué modelo o qué umbral de decisión utilizar según el contexto del usuario, los objetivos de la empresa o incluso el nivel de confianza de la predicción de cada modelo. El bajo recall de la clase alta en el "Modelo de Coste" para 3 clases (0.12) es una advertencia importante: si la matriz de costes no penaliza suficientemente el fallo en detectar "alto riesgo", el modelo podría ser subóptimo desde una perspectiva prudencial, lo que refuerza la necesidad y el valor del "Modelo Detector" especializado.

Capítulo 5: Resultados y Análisis

Este capítulo presenta un análisis detallado de los resultados obtenidos tras el desarrollo y la optimización de los modelos de aprendizaje automático. Se examina el rendimiento comparativo de los modelos globales y se evalúan en profundidad los modelos especializados, destacando las métricas clave y discutiendo la importancia de las características identificadas.

5.1. Análisis Comparativo del Rendimiento de los Modelos Globales

Antes de la especialización de modelos para objetivos específicos (coste, detección), se evaluó el rendimiento de los modelos entrenados con los conjuntos de características más prometedores: "95_Original" y "95_Ultimate". El "Modelo Académico", optimizado para un rendimiento general en la tarea de clasificación binaria (2 clases) y utilizando el conjunto de características "95_Original", sirve como un buen representante de este rendimiento global. Este modelo logró un **F1-macro de aproximadamente 0.70** y un **AUC de 0.7752** en el conjunto de test.[3] Estas cifras representan una mejora sustancial respecto a los modelos iniciales entrenados con solo 17 características (que alcanzaron un F1-macro máximo de 0.58 con LightGBM, ver Tabla 4.1).

Es particularmente revelador comparar el rendimiento del Modelo Académico en la tarea de 2 clases con su rendimiento en la tarea original de 4 clases. Para el problema de 4 clases, el mismo enfoque de modelado (Random Forest con optimización similar) sobre el conjunto "95_Original" obtuvo un F1-macro de solo 0.4735.[3] La marcada diferencia en F1-macro (0.4735 para 4 clases vs. ≈0.70 para 2 clases) valida empíricamente la decisión estratégica de reducir la dimensionalidad de la variable objetivo. Esta simplificación no solo facilitó el proceso de modelado y el manejo del desbalance de clases, sino que condujo a un rendimiento predictivo significativamente superior en términos de la métrica F1-macro agregada. Aunque se pierde granularidad en la predicción directa (el modelo ya no distingue entre 4 niveles de riesgo, sino solo 2), la ganancia en robustez y en la capacidad general del modelo para discriminar entre perfiles de riesgo más amplios puede justificar esta simplificación, especialmente si la salida final requerida por la aplicación (una escala del 1 al 10) puede reconstruirse satisfactoriamente mediante un mapeo posterior de la salida binaria del modelo.

5.2. Evaluación de los Modelos Especializados y sus Métricas Clave

La verdadera fortaleza del enfoque metodológico desarrollado en este TFM se manifiesta en la creación y evaluación de modelos especializados, cada uno optimizado para un objetivo particular. La Tabla 5.1 resume los resultados finales de estos modelos en el conjunto de test.

Tabla 5.1: Rendimiento Final Optimizado del Modelo para Casos de Uso Especializados (Test)
(Adaptada de 3)

Modelo Especializado	Dataset Usado	Métrica Principal Optimizada	Valor Métrica Principal (Test)	Otras Métricas Clave (Test)	Hiperparámetros/Técnicas Clave
Modelo Académico (2 clases)	95_Original	AUC	0.7752	F1-Macro ≈0.70, ACC≈0.73	SMOTE, RF, Búsqueda 2 etapas
Modelo de Coste (3 clases)	95_Ultimate	Coste Medio (minimizar)	0.5354	Acc=0.64, F1−w=0.62, Recall Clase Alta =0.12	SMOTE, RF, Función de coste personalizada
Modelo Detector (2 clases)	95_Ultimate	Recall (Clase Alto Riesgo)	0.81	Precisión (Alto Riesgo) =0.77, AUC =0.7711, F1-Macro ≈0.69	SMOTE (k_neighbors=9), RF (max_features ='log2', etc.), Búsqueda 2 etapas

El **Modelo Académico** (con 2 clases, sobre "95_Original") sirve como la mejor representación del rendimiento global equilibrado, alcanzando un AUC de 0.7752 y un F1-macro de ≈0.70.[3] Este modelo es útil para comparaciones generales y para entender la capacidad predictiva base del sistema.

El **Modelo de Coste** (evaluado con 3 clases, sobre "95_Ultimate") intentó optimizar las decisiones minimizando un coste económico predefinido asociado a los errores. Logró un coste medio de 0.5354.[3] Sin embargo, es crucial notar el bajo recall para la clase de riesgo alto (0.12). Esto sugiere que la matriz de costes utilizada podría no haber penalizado suficientemente los errores de clasificación de perfiles de alto riesgo como de menor riesgo, lo cual es una consideración crítica desde una perspectiva prudencial. Si el objetivo principal es evitar clasificar erróneamente a los inversores de alto riesgo, este modelo, tal como está configurado, podría no ser el más adecuado a pesar de un coste medio general "bajo".

El **Modelo Detector** (con 2 clases, sobre "95_Ultimate") fue específicamente optimizado para maximizar la identificación de perfiles de Alto Riesgo. Logró un **Recall del 81%** para esta clase, con una Precisión del 77%.[3] Aunque su F1-macrogeneral (≈0.69) y AUC (0.7711) son ligeramente inferiores a los del Modelo Académico, su capacidad para capturar la gran mayoría de los casos de alto riesgo es notablemente superior. La matriz de confusión para este modelo fue [,] (Verdaderos Negativos, Falsos Positivos; Falsos Negativos, Verdaderos Positivos para la clase Alto Riesgo). Esto significa que de 274 (51+223) verdaderos casos de Alto Riesgo, el modelo identificó correctamente 223. Sin embargo, también clasificó erróneamente 51 casos de Alto Riesgo como de Bajo/Medio Riesgo (falsos negativos) y 66 casos de Bajo/Medio Riesgo como de Alto Riesgo (falsos positivos).[3] Estas cifras concretas son mucho más informativas para la toma de decisiones operativas que las métricas agregadas por sí solas, permitiendo una evaluación más matizada del comportamiento del modelo.

La comparación entre estos modelos especializados ilustra un punto fundamental: no existe una única métrica o modelo "mejor" para todos los propósitos. El "Modelo Académico" puede ser el más equilibrado en general, pero si la prioridad es la detección exhaustiva de perfiles de alto riesgo, el "Modelo Detector" es superior en esa tarea específica, aunque pueda incurrir en más falsos positivos. La elección del conjunto de características "95_Ultimate" para los modelos de Coste y Detector se basó en la observación previa de que este conjunto, enriquecido con características diseñadas manualmente, era mejor para "casos específicos".[3] Los resultados de estos modelos especializados validan esa elección, sugiriendo que las características que encapsulan conocimiento del dominio en "95_Ultimate" ayudan a discernir mejor los perfiles de riesgo más extremos o matizados.

5.3. Discusión sobre la Importancia de las Características Identificadas (XAI)

Aunque este TFM no presenta una implementación exhaustiva y visualización de resultados de técnicas de XAI como SHAP o LIME para los modelos finales (siendo esta una línea de trabajo futuro) [3], es posible inferir la importancia de ciertos tipos de características a partir del proceso de desarrollo y los resultados obtenidos.

En las etapas iniciales de selección automática de características, variables como TRADER_SCORE, S_Age (edad) y B4_log (valor de la cartera) emergieron consistentemente como predictoras.[3] Posteriormente, en el modelo de Regresión Logística entrenado con 42 características seleccionadas manualmente, variables relacionadas con la consulta de fuentes de información financiera (F31_4, F31_2), el uso de ciertos productos de inversión (B2_23, B2_24) y la edad (S_Age) mostraron coeficientes elevados, sugiriendo su influencia.[3]

Sin embargo, el salto más significativo en el rendimiento para tareas específicas (como la detección de alto riesgo en el Modelo Detector) se observó al utilizar el conjunto de características "95_Ultimate". Este conjunto incluía 11 nuevas variables creadas manualmente, tales como "scores de dominio", "gap de percepción de riesgo", interacciones específicas y el "binning" del valor de la cartera.[3] Se puede argumentar que estas características construidas son particularmente importantes porque el conjunto "95_Ultimate" demostró ser superior para el Modelo Detector. Estas variables, diseñadas con un entendimiento del comportamiento del inversor y de los constructos financieros y psicológicos relevantes, probablemente capturan dimensiones del riesgo que las variables originales por sí solas, o las transformaciones automáticas, no lograban representar adecuadamente. Por ejemplo:

- Un **"score de dominio"** que combine experiencia, frecuencia de trading y uso de productos sofisticados podría ser un fuerte indicador de una mayor (o menor, dependiendo de cómo se construya) tolerancia al riesgo.

- Un **"gap de percepción de riesgo"**, que podría medir la discrepancia entre la aversión al riesgo declarada por un inversor y su comportamiento de inversión real (inferido de otras preguntas) o su comprensión de los riesgos de ciertos activos, podría ser una proxy de sesgos como el exceso de confianza o una aversión al riesgo no realista.
- Las **interacciones** entre variables (e.g., cómo la edad modula el efecto del nivel educativo sobre la tolerancia al riesgo) pueden capturar relaciones no aditivas que los modelos lineales o los árboles simples podrían pasar por alto si no se especifican explícitamente.

Una discusión sobre qué *tipos* de características parecen ser más influyentes (demográficas, actitudinales, de comportamiento, o las construidas que combinan varias de estas) puede ofrecer insights valiosos para el diseño de futuros cuestionarios de perfilación de riesgo o para la priorización en la recopilación de datos. La necesidad de técnicas de XAI formales para justificar las predicciones de los modelos finales y para comprender en detalle la contribución de cada característica sigue siendo un aspecto crucial para la implementación práctica y la confianza en el sistema.[3]

Capítulo 6: Discusión, Conclusiones y Trabajo Futuro

6.1. Síntesis de los Hallazgos Principales

El recorrido metodológico y experimental de este Trabajo de Fin de Máster ha arrojado una serie de hallazgos clave que merecen ser destacados, los cuales se derivan directamente del proceso iterativo de análisis, desarrollo y evaluación [3]:

- **La Sinergia entre Datos y Conocimiento del Dominio es Fundamental:** Quizás el hallazgo más contundente es la demostración empírica de que, si bien los algoritmos de aprendizaje automático son herramientas poderosas, su aplicación efectiva en dominios complejos como la perfilación del riesgo del inversor requiere una profunda comprensión contextual. Las mayores ganancias de rendimiento, especialmente para identificar "casos específicos" como los perfiles de alto riesgo, se obtuvieron cuando el conocimiento del dominio guio la creación de características avanzadas (el conjunto "95_ultimate"). Los métodos automatizados de selección y generación de características sirvieron como herramientas exploratorias útiles, pero fue la intervención humana informada, basada en la teoría financiera y la psicología del inversor, la que condujo a los avances más significativos en la capacidad predictiva para nichos relevantes.

- **Valor de la Ingeniería de Características Específica del Contexto:** Se constató que las características diseñadas manualmente, aquellas que encapsulaban conceptos financieros y comportamentales complejos (como los "scores de dominio" o el "gap de percepción de riesgo"), fueron particularmente efectivas para mejorar las predicciones en segmentos de la población de inversores más difíciles de clasificar o en la detección de perfiles de riesgo más extremos. Esto sugiere que la información más matizada y conceptualmente rica es crucial para ir más allá de una perfilación superficial y capturar las sutilezas de la propensión individual al riesgo.
- **Evolución hacia un Sistema Adaptativo y Especializado:** El proyecto evolucionó desde la búsqueda inicial de un único modelo de reemplazo para el sistema basado en reglas hacia el desarrollo de un sistema de modelos más matizado y potencialmente adaptativo. La creación de modelos optimizados para diferentes objetivos (académico/general, coste económico, detección de alto riesgo) refleja un reconocimiento de que la "mejor" solución puede ser, en realidad, una combinación de componentes especializados que se activan o ponderan según el contexto o el propósito específico de la evaluación.
- **Beneficios de la Optimización Especializada de Modelos:** La optimización de modelos para diferentes funciones objetivo demostró la flexibilidad inherente del enfoque de aprendizaje automático y su capacidad para adaptarse a diversas necesidades operativas o analíticas. Este enfoque permite ir más allá de la optimización de una métrica de rendimiento única y

global, y considerar los trade-offs específicos que pueden ser más relevantes para la aplicación final (e.g., priorizar el recall de una clase crítica sobre la exactitud general).

El TFM, en su conjunto, puede ser visto como un caso de estudio sobre la interacción entre el "arte" y la "ciencia" en la aplicación del Machine Learning. El recorrido desde la selección automática de características (ciencia algorítmica), pasando por la curación manual y la ingeniería avanzada guiada por el dominio (una combinación de arte y ciencia), hasta la especialización final de modelos (ingeniería de soluciones), ilustra que el desarrollo de ML efectivo en dominios complejos no se reduce a la mera aplicación de algoritmos. Es, más bien, un proceso iterativo de formulación de hipótesis, experimentación rigurosa y refinamiento continuo, informado y validado constantemente por el contexto del problema. Un tema subyacente que emerge es el de la "inteligencia aumentada": los modelos de ML son herramientas potentes, pero su máximo potencial se alcanza cuando se combinan sinérgicamente con la experiencia y el juicio humano.

6.2. Consecución de los Objetivos del TFM

Al inicio de este trabajo [3], se plantearon varios objetivos clave para el desarrollo del nuevo modelo de perfilación de riesgo del inversor. A continuación, se evalúa en qué medida se han cumplido:

- **Reemplazar o mejorar sustancialmente el sistema de reglas existente:** Los modelos de ML desarrollados, particularmente el Modelo Académico (con un F1-macro de ≈0.70 y AUC de 0.7752 para 2 clases) y el Modelo Detector (con un recall del 81% para la clase de alto riesgo), demuestran un rendimiento predictivo que supera cualitativa y cuantitativamente lo que se esperaría de un sistema simple basado en reglas con ponderaciones fijas y sin capacidad de aprendizaje. La capacidad de los modelos para aprender de los datos y adaptarse a diferentes objetivos de optimización representa una mejora sustancial.

- **Entrenamiento con datos reales:** El modelo se entrenó utilizando el conjunto de datos FINRA NFCS 2021, que contiene miles de casos de inversores reales de EE. UU., cumpliendo con este requisito.
- **Aprendizaje de patrones complejos:** La mejora en el rendimiento al pasar de conjuntos de características simples a conjuntos más ricos y con ingeniería avanzada ("95_ultimate") sugiere que los modelos fueron capaces de aprender relaciones más complejas y no triviales entre las variables de entrada. La efectividad de las características diseñadas manualmente para capturar "casos específicos" apoya esta conclusión.
- **Generación de perfiles más precisos y adaptativos:** La precisión, medida por métricas como F1-macro y AUC, mejoró a lo largo del proceso iterativo de desarrollo. La adaptabilidad se demuestra a través de la especialización de modelos, que pueden ajustarse a diferentes necesidades o perfiles de usuario.
- **Explicabilidad:** Si bien se reconoció la importancia crucial de la explicabilidad desde el inicio y se consideró en la discusión, la implementación y evaluación rigurosa de técnicas de XAI (como SHAP o LIME) para los modelos finales se ha identificado como una línea de trabajo futuro. Por lo tanto, este objetivo se ha abordado a nivel conceptual y de planificación, pero no se ha completado en términos de implementación detallada en los resultados presentados.

En resumen, se puede considerar que los objetivos principales relacionados con la mejora del rendimiento predictivo y la adaptabilidad del sistema de perfilación se han cumplido satisfactoriamente. El objetivo de la explicabilidad ha sido reconocido y su camino delineado, aunque su plena realización práctica queda para etapas posteriores.

6.3. Limitaciones del Estudio Y Líneas de Investigación Futuras

A pesar de los avances logrados y los resultados prometedores, este trabajo presenta ciertas limitaciones inherentes al alcance de un TFM y a las decisiones metodológicas tomadas. Es importante reconocer estas limitaciones para contextualizar adecuadamente los hallazgos y guiar futuras investigaciones [3]:

- **Dependencia de Datos Declarados (NFCS):** El modelo principal se basa en la tolerancia al riesgo declarada por los encuestados en el NFCS. Como se discutió en el marco teórico (sección 2.2), existe un conocido "say-do gap" entre lo que las personas declaran y lo que realmente hacen.[3] La literatura sobre finanzas conductuales [6] y las críticas a los cuestionarios de riesgo [1] sugieren que las preferencias declaradas pueden estar sujetas a sesgos y no reflejar perfectamente el comportamiento de inversión real.
- **Exploración Limitada de Arquitecturas de ML Alternativas:** Aunque se probaron algoritmos robustos y populares (Regresión Logística, Random Forest, LightGBM), la exploración de otras arquitecturas de ML, como redes neuronales profundas (que podrían capturar interacciones aún más complejas) o modelos factoriales (que podrían ofrecer una mejor interpretabilidad de los constructos latentes de riesgo), fue limitada.
- **Refinamiento del Mapeo de Salida:** La decisión de reducir la variable objetivo a una clasificación binaria (riesgo bajo/medio vs. alto) simplificó el modelado y mejoró el rendimiento, pero la aplicación final requiere una salida en una escala del 1 al 10. El proceso de mapeo de la salida binaria del modelo a esta escala más granular no se ha desarrollado en detalle en este TFM y es un componente crítico que podría introducir sus propias heurísticas o requerir un segundo nivel de modelado.[3]
- **Implementación de XAI Pendiente:** Aunque la explicabilidad se identificó como un objetivo importante, la implementación y evaluación rigurosa de técnicas de XAI (como SHAP o LIME) para los modelos finales, especialmente para el Modelo Detector o un eventual Modelo de Coste desplegado, es una tarea pendiente y fundamental para la confianza del usuario y el cumplimiento normativo.[3]
- **Validación en el Contexto Español (ECF) No Realizada:** Si bien se identificó la Encuesta de Competencias Financieras (ECF) de España como una fuente de datos relevante para el público objetivo de la aplicación, la validación o adaptación del modelo utilizando estos datos no se llevó a cabo y sigue siendo una línea de trabajo futura importante.[3] Las actitudes hacia el

riesgo pueden tener componentes culturales, y un modelo entrenado predominantemente con datos de EE. UU. podría no generalizar perfectamente al contexto español.

- **Calidad Inherente a los Datos de Encuestas:** Los datos de encuestas, incluso de alta calidad como el NFCS, pueden estar sujetos a limitaciones como errores de respuesta, sesgos de deseabilidad social, o la incapacidad de las preguntas para capturar perfectamente constructos psicológicos complejos.[8]

Estas limitaciones son, en muchos casos, un reflejo realista de las fronteras de un proyecto de investigación con un alcance temporal y de recursos definido, como es un TFM. Reconocerlas no debilita el trabajo realizado, sino que demuestra una comprensión madura del proceso de investigación y, lo que es más importante, abre caminos claros y bien fundamentados para la continuidad y la mejora del sistema de perfilación desarrollado.

6.4. Implicaciones Prácticas y Contribuciones

Los resultados de este TFM tienen varias implicaciones prácticas y contribuciones potenciales para la aplicación financiera en la que se enmarca y para el campo de la perfilación del riesgo del inversor en general:

- **Hacia una Herramienta de Inversión más Inteligente y Personalizada:** El desarrollo de modelos de ML con un rendimiento predictivo superior al de los sistemas basados en reglas tiene el potencial de hacer que la aplicación financiera sea más "inteligente" en su capacidad para evaluar el riesgo del inversor. Esto, a su vez, puede conducir a recomendaciones de cartera más personalizadas y adecuadas al perfil individual de cada usuario.[3]
- **Mejora en la Precisión de la Perfilación del Riesgo:** Los modelos desarrollados, especialmente aquellos que utilizan el conjunto de características "95_Ultimate" y las optimizaciones especializadas, ofrecen una mejora tangible en la precisión de la perfilación en comparación con lo que se podría esperar de un sistema de reglas estático. Esto puede traducirse en una

mejor alineación entre el inversor y sus inversiones, potencialmente aumentando la satisfacción del cliente y la probabilidad de alcanzar sus objetivos financieros.

- **Posibilidad de Ofrecer Perfiles de Riesgo Dinámicos y Contextuales:** La suite de modelos desarrollados (Académico, Coste, Detector) y la discusión sobre el uso selectivo de diferentes conjuntos de características ("95_Original" vs. "95_Ultimate") abren la puerta a una perfilación de riesgo más dinámica y contextual. La aplicación podría, por ejemplo, presentar al usuario un perfil base generado por el Modelo Académico, pero luego refinarlo o emitir alertas específicas si el Modelo Detector lo marca como un perfil de alto riesgo. Incluso la comunicación y las explicaciones proporcionadas al usuario podrían adaptarse basándose en la confianza del modelo o en las características que más influyeron en su predicción.
- **Fundamento para Decisiones de Diseño de Producto:** El análisis de la importancia de las características (aunque sea inferido en esta etapa) puede informar futuras decisiones sobre el diseño del cuestionario de entrada de la aplicación. Se podrían priorizar las preguntas que se demuestren más predictivas y eliminar o reformular aquellas con bajo impacto, optimizando la experiencia del usuario sin sacrificar (e incluso mejorando) la calidad de la perfilación.

La contribución principal de este TFM no radica solo en la creación de modelos predictivos, sino en la documentación de un proceso metodológico riguroso y reflexivo para abordar un problema complejo en la intersección de las finanzas, la psicología y el aprendizaje automático. Este proceso, que enfatiza la iteración, la importancia del conocimiento del dominio y la adaptación a objetivos específicos, puede servir de guía para proyectos similares.

En conclusión, este TFM ha demostrado el potencial significativo del aprendizaje automático para avanzar en la perfilación del riesgo del inversor, superando muchas de las limitaciones de los enfoques tradicionales basados en reglas. El camino recorrido no solo ha producido modelos con una capacidad predictiva potencialmente superior, sino que también ha generado una comprensión más profunda de los datos subyacentes y de los complejos factores que configuran la tolerancia al riesgo. Los hallazgos y la metodología aquí presentados sientan una base robusta para futuras mejoras y para la eventual implementación de una herramienta de inversión más inteligente, personalizada y, en última instancia, más útil para el inversor.

7. Anexos

Obras citadas

1. INVESTOR RISK PROFILING: AN OVERVIEW - CFA Institute Research and Policy Center, fecha de acceso: junio 13, 2025, https://rpc.cfainstitute.org/sites/default/files/-/media/documents/article/rf-brief/rfbr-v1-n1-1-pdf.pdf
2. Risk Capacity vs. Risk Tolerance: What's the Difference? - SmartAsset, fecha de acceso: junio 13, 2025, https://smartasset.com/investing/risk-capacity-vs-risk-tolerance
3. Notebooks
4. La IA en la gestión de patrimonios: Casos de uso y herramientas - Botpress, fecha de acceso: junio 13, 2025, https://botpress.com/es/blog/ai-in-wealth-management
5. The Rise of AI and Robo-Advisors: Redefining Financial Strategies in the Digital Age - ijrpr, fecha de acceso: junio 13, 2025, https://ijrpr.com/uploads/V6ISSUE1/IJRPR38153.pdf
6. FINANZAS CONDUCTUALES: - Dialnet, fecha de acceso: junio 13, 2025, https://dialnet.unirioja.es/descarga/articulo/3202463.pdf
7. (PDF) THE IMPACT OF BEHAVIORAL BIASES ON FINANCIAL RISK TOLERANCE OF INVESTORS AND THEIR DECISION MAKING - ResearchGate, fecha de acceso: junio 13, 2025, https://www.researchgate.net/publication/366836734_THE_IMPACT_OF_BEHAVIORAL_BIASES_ON_FINANCIAL_RISK_TOLERANCE_OF_INVESTORS_AND_THEIR_DECISION_MAKING

8. (PDF) Assessing Investors' Risk Tolerance Through a Questionnaire - ResearchGate, fecha de acceso: junio 13, 2025, https://www.researchgate.net/publication/256045201_Assessing_Investors'_Risk_Tolerance_Through_a_Questionnaire
9. Limitaciones y Desafíos en la Gestión de Riesgos - Visure Solutions, fecha de acceso: junio 13, 2025, https://visuresolutions.com/es/gu%C3%ADa-fmea-de-gesti%C3%B3n-de-riesgos/limitaciones-y-desaf%C3%ADos/
10. FINANCIAL RISK TOLERANCE: A PSYCHOMETRIC REVIEW, fecha de acceso: junio 13, 2025, https://rpc.cfainstitute.org/sites/default/files/-/media/documents/article/rf-brief/rfbr-v4-n1-1.pdf
11. 160+ million publication pages organized by topic on ResearchGate, fecha de acceso: junio 13, 2025, https://www.researchgate.net/publication/387106436_MACHINE_LEARNING_IN_FINANCIAL_RISK_MANAGEMENT_TECHNIQUES_FOR_PREDICTING_EARLY_PAYMENT_AND_DEFAULT_RISKS
12. From rule-based to risk-based approach, fecha de acceso: junio 13, 2025, https://4639135.fs1.hubspotusercontent-na1.net/hubfs/4639135/2024%20Website/WP%20%7C%20From%20rule-based%20to%20risk-based%20approach.pdf
13. Fraud Detection Using Machine Learning vs. Rules-Based Systems - FraudNet, fecha de acceso: junio 13, 2025, https://www.fraud.net/resources/fraud-detection-using-machine-learning-vs-rules-based-systems
14. What Is a "Nonlinear" Exposure in Value at Risk (VaR)? - Investopedia, fecha de acceso: junio 13, 2025, https://www.investopedia.com/ask/answers/042215/what-non-linear-exposure-value-risk-var.asp
15. Risk Management in Investment Portfolios Using Machine Learning Models - ResearchGate, fecha de acceso: junio 13, 2025, https://www.researchgate.net/publication/387445252_Risk_Management_in_Investment_Portfolios_Using_Machine_Learning_Models

16. The non-linear impact of risk tolerance on entrepreneurial profit and business survival - EconStor, fecha de acceso: junio 13, 2025, https://www.econstor.eu/bitstream/10419/283245/1/1880609991.pdf
17. Cuestionario tolerancia de riesgo: descubre más sobre tu perfil de inversor - The Investor U, fecha de acceso: junio 13, 2025, https://theinvestoru.com/blog/cuestionario-tolerancia-de-riesgo/
18. Perfil de inversor - PELLEGRINI, fecha de acceso: junio 13, 2025, https://www.pellegrinifci.com.ar/perfil-del-inversor
19. Tolerancia al riesgo: ¿Qué es y cómo determinarla? - QuestionPro, fecha de acceso: junio 13, 2025, https://www.questionpro.com/blog/es/tolerancia-al-riesgo/
20. Directrices, fecha de acceso: junio 13, 2025, https://www.esma.europa.eu/sites/default/files/library/esma35-43-1163_guidelines_on_certain_aspects_of_mifid_ii_suitability_requirements_es.pdf
21. Las Finanzas Conductuales, el rol de las emociones en las decisiones financieras. - Biblioteca Digital UNCUYO, fecha de acceso: junio 13, 2025, https://bdigital.uncuyo.edu.ar/objetos_digitales/19422/las-finanzas-conductuales-el-rol-de-las-emociones.pdf
22. Determinants of Financial Risk Tolerance: An Analysis of Psychological Factors - MDPI, fecha de acceso: junio 13, 2025, https://www.mdpi.com/1911-8074/16/2/74
23. IA en las finanzas: aplicaciones, ejemplos y ventajas - Google Cloud, fecha de acceso: junio 13, 2025, https://cloud.google.com/discover/finance-ai?hl=es
24. Machine Learning For Risk Profiling: Enhancing AML And KYC Compliance Through Machine Learning In The Digital Age - Financial Crime Academy, fecha de acceso: junio 13, 2025, https://financialcrimeacademy.org/machine-learning-for-risk-profiling/
25. Machine Learning Methods in Investor Risk Profiling - ResearchGate, fecha de acceso: junio 13, 2025,

https://www.researchgate.net/publication/384732242_Machine_Learning_Methods_in_Investor_Risk_Profiling
26. Top Use Cases of Explainable AI: Real-World Applications for Transparency and Trust, fecha de acceso: junio 13, 2025, https://smythos.com/developers/agent-development/explainable-ai-use-cases/
27. www.lumenova.ai, fecha de acceso: junio 13, 2025, https://www.lumenova.ai/blog/ai-banking-finance-compliance/#:~:text=In%20the%20banking%20and%20finance,fairness%2C%20and%20maintaining%20stakeholder%20trust.
28. Techniques for Explainable AI: LIME and SHAP - Unnat Bak ..., fecha de acceso: junio 13, 2025, https://www.unnatbak.com/blog/techniques-for-explainable-ai-lime-and-shap
29. How to Interpret Machine Learning Models with LIME and SHAP - Svitla Systems, fecha de acceso: junio 13, 2025, https://svitla.com/blog/interpreting-machine-learning-models-lime-and-shap/
30. Robo-Advisors in Wealth Management: A Bibliometric Study of Research Evolution, fecha de acceso: junio 13, 2025, https://esj.eastasouth-institute.com/index.php/esaf/article/view/501