UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE ESTUDIOS ESTADÍSTICOS

TESIS DOCTORAL

Aplicación del Modelo Lingüístico Difuso 2-Tuplas para Mejora de las Técnicas XAI: Casos de uso en Fintech y Smart City

MEMORIA PARA OPTAR AL GRADO DE DOCTORA

PRESENTADO POR

Leticia Monje García

DIRECTORES

Ramón Alberto Carrasco González

Manuel Antonio Sánchez-Montañés Isla

© Leticia Monje García 2025

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE ESTUDIOS ESTADÍSTICOS

Programa de Doctorado en **Análisis de Datos (Data Science)**

TESIS DOCTORAL

Aplicación del modelo lingüístico difuso 2-tuplas para Mejora de las Técnicas XAI: Casos de uso en Fintech y Smart City

MEMORIA PARA OPTAR AL GRADO DE DOCTORA

PRESENTADO POR

Leticia Monje García

DIRECTORES

Ramón Alberto Carrasco González

Manuel Antonio Sánchez-Montañés Isla

Madrid, 2025

A mis padres, que siempre me han animado a superarme académicamente.

A Julio, por acompañarme en este camino.

A mis hijas Leticia y Leire por vuestra comprensión.

Agradecimientos

" Convierte tu muro en un peldaño" - Reiner María Rilke

En primer lugar, mi sincero agradecimiento a mis directores de tesis, Ramón Alberto Carrasco y Manuel Antonio Sánchez-Montañés, por motivarme a nunca darme por vencida y por su apoyo en los momentos de agobio. Su actitud positiva, su valiosa orientación y sus sabios consejos han sido cruciales para la conclusión exitosa de esta tesis.

Deseo expresar mi más profundo agradecimiento a mis padres, porque sin los valores que me inculcaron no sería la persona que soy hoy. Gracias por su contribución a mi educación y por enseñarme que todas las metas se logran con trabajo y esfuerzo. A mi padre, quien siempre predicó con el ejemplo; estoy segura de que hoy seguiría aprendiendo de él. Sé que se sentiría profundamente orgulloso de tener a su hija doctora, y aún más feliz me hubiese hecho a mí, que hubiese podido vivir este momento conmigo

En especial, a mi madre, mi mayor referente en la vida, una persona excepcional que hace fácil hasta lo más difícil. Con ella, todas mis preocupaciones desaparecen. Siempre me ha alentado a no rendirme y a intentarlo todo, recordándome con esa frase suya que llevo siempre presente: *"Inténtalo, porque el 'no' ya lo tienes."* Gracias a ella he llegado hasta donde estoy hoy.

Quiero expresar mi más profundo agradecimiento a mi marido Julio, por su amor, cariño y el inquebrantable apoyo emocional que me ha brindado durante este desafiante periodo. Su comprensión y fe incondicional en mí, han sido pilares fundamentales para que esta tesis sea una realidad. Gracias por estar siempre a mi lado, por asumir con generosidad el doble rol de padre y madre en tantos momentos y, por acompañarme en cada paso de este

camino, incluso cuando no era fácil, como en la transición de la ingeniería a la banca. Esta tesis es tanto mía como tuya, un reflejo de todo lo que hemos logrado juntos.

Agradezco a mis hijas Leticia y Leire, por entender que durante el desarrollo de esta tesis fue necesario sacrificar situaciones y momentos a su lado para así, poder completar exitosamente mi trabajo académico. A partir de ahora, recuperaremos ese tiempo.

Por último y no menos importante, agradecer a mi amiga Mar Dorado los momentos compartidos a lo largo de estos años, que me han permitido desconectar y recargar energías. Gracias por tu generosidad al ayudarme en la lectura de esta tesis. Aunque la distancia no nos permita vernos tanto como quisiéramos, como bien dice Manuel Carrasco: "Qué bonito es saber que siempre estás ahí... afortunada yo por contar con tu amistad".

Gracias a todos.

Resumen

Las empresas se hallan inmersas en un continuo proceso de toma de decisiones diarias, abarcando una diversidad en cuanto a su alcance y magnitud. La capacidad de utilizar la información asequible, apoyada por la implementación de modelos de toma de decisiones, emerge como un camino hacia la elección de la solución más idónea frente a sus respectivos desafíos.

El ámbito de la ciencia de datos ofrece la posibilidad de efectuar elecciones fundamentadas y precisas. Desde una perspectiva empresarial, las grandes bases de datos y el análisis de la clientela otorgan ventajas competitivas al proporcionar una comprensión detallada de sus patrones de comportamiento a través de diversas variables.

A lo largo de este proceso de investigación, se ha llevado a cabo una exhaustiva labor de investigación dedicada a abordar diversos problemas aplicados a conjuntos de datos reales. Se han formulado y evaluado distintas metodologías, todas ellas orientadas a un propósito común: la mejora de la toma de decisiones por parte de la empresa.

Para ello, hemos abordado dos casos de estudio concretos de diferente ámbito:

1. Tecnología financiera (en inglés Fintech): préstamos entre particulares a través de plataforma online.
2. Ciudades Inteligentes (en inglés Smart Cities).

El caso de uso de ámbito financiero cobra particular relevancia en el contexto actual, donde las instituciones financieras se ven sometidas a una creciente exigencia de un análisis de riesgo más exhaustivo y una gestión más eficaz. Por otro lado, las Ciudades Inteligentes representan entornos urbanos que hacen uso de la tecnología y los datos para elevar la calidad de vida de sus habitantes. El transporte eficiente de pasajeros en áreas urbanas es un problema cada vez

más importante en nuestra sociedad y el conocimiento de la demanda de pasajeros es crucial para la toma de decisiones.

En esta tesis se han utilizado, para ambos casos de uso, datos reales proporcionados por: Lending Club para el caso de uso de Fintech y, la Empresa Municipal de Transportes (EMT) para la línea de investigación de Ciudades Inteligentes. En ambos casos de uso se plantea una metodología que permite crear un modelo predictivo fácilmente interpretable desde un punto de vista lingüístico. Este modelo resulta beneficioso para la toma de decisiones en situaciones que involucran grandes cantidades de datos.

Como resultado de nuestras investigaciones, hemos desarrollado metodologías de Inteligencia Artificial Explicable (en ingles eXplainable Artificial Intelligence - XAI) que aplicamos al área de Fintech y Smart Cities, donde los modelos más precisos que hemos encontrado son de tipo "caja negra": en el caso de las Fintech, el mejor modelo es un clasificador eXtreme Gradient Boosting (XGBoost), y en el caso de las Ciudades Inteligentes, una red neuronal recurrente (en inglés Recurrent Neural Network - RNN) del tipo Long Short Term Memory (LSTM). La interpretabilidad de estos modelos se ha logrado mediante la implementación de modelos surrogados, los cuales son de tipo "caja blanca" y la inclusión del modelo difuso lingüístico 2-tuplas (Monje et al., 2022), que sin perder precisión, mejoró la interpretabilidad lingüística del modelo generado de XAI, permitiéndonos para cada caso de uso separar la predicción en "muy baja", "baja", "media", "alta" y "muy alta", cada uno de ellos con diferente traducción simbólica, de forma que pudimos distinguir por qué una probabilidad era menor que la otra.

Abstract

Companies are immersed in a continuous process of daily decision-making, encompassing a diversity of scope and magnitude. The ability to utilize accessible information, supported by the implementation of decision-making models, emerges as a pathway towards choosing the most suitable solution for their respective challenges.

The field of data science offers the possibility of making well-founded and precise choices. From a business perspective, large databases and customer analysis provide competitive advantages by offering a detailed understanding of behavioral patterns through various variables.

Throughout this research process, extensive research work has been dedicated to addressing various problems applied to real data sets. Different methodologies have been formulated and evaluated, all aimed at a common purpose: improving the company's decision-making.

To this end, we have addressed two specific case studies in different domains:

1. Financial technology (Fintech): peer-to-peer lending through an online platform.

2. Smart Cities.

The financial domain case is particularly relevant in the current context, where financial institutions are increasingly required to conduct more thorough risk analysis and more efficient management. On the other hand, Smart Cities represent urban environments that leverage technology and data to enhance the quality of life of their inhabitants. Efficient passenger transport in urban areas is an increasingly important problem in our society, and knowledge of passenger demand is crucial for decision-making.

In this thesis, real data provided by Lending Club for the Fintech use case and the Community of Madrid for the Smart Cities research line were used. In

both cases, a methodology was proposed to create a predictive model that is also easy to interpret from a linguistic perspective. This model is beneficial for decision-making in situations involving large amounts of data.

As a result of our research, we have developed Explainable Artificial Intelligence (XAI) methodologies applied to the Fintech and Smart Cities areas, where the most accurate models we found are "black box" models: in the case of Fintech, the best model is an eXtreme Gradient Boosting (XGBoost) classifier, and in the case of Smart Cities, a recurrent neural network (RNN) of the Long Short Term Memory (LSTM) type. The interpretability of these models has been achieved through the implementation of surrogate models, which are "white box" models, and the inclusion of the 2-tuple linguistic fuzzy model (Monje et al., 2022), which, without losing accuracy, improved the linguistic interpretability of the generated XAI model. This allowed us to separate the prediction in each use case into "very low," "low," "medium," "high," and "very high," each with different symbolic translations, enabling us to distinguish why one probability was lower than another.

Índice

Agradecimientos I

Resumen III

Abstract V

Índice VII

Índice de tablas IX

Índice de figuras X

Índice de acrónimos XI

Desarrollo de la memoria 2

1. Estructura de la memoria 4

2. Introducción 7

2.1. Objetivos 12

2.1.1. Objetivo general 12

2.1.2. Objetivos específicos 13

3. Estado del arte 14

3.1. Uso del modelado lingüístico difuso en XAI 14

3.2. Aplicaciones de XAI en predicción en el contexto de Fintech: técnicas de interpretabilidad en la predicción de la mora en préstamos P2P 28

3.3. Aplicaciones de XAI en predicción en el contexto de Smart Cities: técnicas de interpretabilidad en la predicción de pasajeros. 35

4. Marco teórico de la investigación 47

4.1. Técnicas de Machine Learning supervisadas 48

4.1.1. Clasificación 48

4.1.2. Regresión 62

4.2. Modelo lingüístico difuso 2-tuplas 72

4.3. Técnicas de XAI 75

4.4. Metodología KDD 88

5. Aplicación del modelo lingüístico difuso 2-tuplas en las técnicas XAI de árboles surrogados: 2-Tuplas, árboles de surrogación lingüística difusa y reglas 92

5.1. Propuesta de mejora de la interpretabilidad lingüísticas de un modelo XAI surrogado 92

5.2. Caso de uso en Fintech. Predicción del incumplimiento temprano (clasificación) 95

5.3. Caso de uso en ciudad inteligente. Predicción de demanda de pasajeros (regresión) 105

6. Conclusiones y trabajo futuro 116

7. Bibliografía 119

Anexo 137

a) Artículos publicados 138

Índice de tablas

Tabla 1. Trabajos relevantes sobre modelos Machine Learning, lingüísticos difusos y XAI. 24
Tabla 2. Trabajos relevantes sobre predicción del riesgo de impago y XAI. 32
Tabla 3. Trabajos relacionados con la predicción del flujo de pasajeros en autobuses. 41
Tabla 4. Principales características de las distintas técnicas de XAI 82
Tabla 5. AUC de los distintos modelos considerados 99
Tabla 6. Reglas surrogadas obtenidas para el modelo XGBoost. 101
Tabla 7. Reglas surrogadas obtenidas para el modelo XGBoost con representación lingüística de 2 tuplas. 102
Tabla 8. Comparación de modelos. 109
Tabla 9. Reglas surrogadas obtenidas para el modelo LSTM. 111
Tabla 10. Representación de 2 tuplas de la variable objetivo predicha. 112
Tabla 11. Reglas surrogadas obtenidas para el modelo LSTM con representación lingüística de 2 tuplas. 113

Índice de figuras

Figura 1. Esquema de la estructura de la memoria 5
Figura 2. Mapa estratégico de predicción de pasajeros. 36
Figura 3. Redes temáticas para cada tema Redes temáticas para cada tema. ... 40
Figura 4. Regresión logística. 49
Figura 5. Árboles de decisión. 51
Figura 6. Árboles aleatorios 53
Figura 7. SVM. 54
Figura 8. K-NN 56
Figura 9. eXtreme Gradient Boosting classifier (XGBoost). 61
Figura 10. Regresión lineal. 64
Figura 11. Regresión Polinómica. 65
Figura 12. Redes Neuronales Recurrentes. 68
Figura 13. LSTM. 71
Figura 14. Definición del conjunto S para la variable probabilidad de impago temprano 74
Figura 15. Proceso XAI Machine Learning. 84
Figura 16. Fases del proceso KDD 89
Figura 17. Porcentaje de impago temprano. 96
Figura 18. Frecuencia del impago temprano. 97
Figura 19. Distribución del estado del préstamo en función del grado. 97
Figura 20. Distribución del estado del préstamo en función del plazo del préstamo 98
Figura 21. Distribución del estado del préstamo en función de la antigüedad en el trabajo 99
Figura 22. Curva ROC. 100
Figura 23. Número diario de pasajeros por la tarde. 106
Figura 24. Histograma de pasajeros. 107
Figura 25. Ampliación de la predicción en el conjunto de pruebas. Verde discontinuo: valor real. Rojo sólido: predicción. 111

Índice de acrónimos

AE:	AutoEncoders
ALE:	Accumulated Local Effects
ANFIS:	Adaptive Neuro-Fuzzy Inference System
ANN:	Artificial Neural Networks
AUC:	Area Under the Curve
BERT:	Bidirectional Encoder Representations from Transformers
BiLSTM:	Bidirectional Long Short Term Memory
Bi-RNN:	Bidirectional Recurrent Neural Networks
BPNN:	Back Propagation Neural Network
BPTT:	BackPropagation Through Time
CART:	ClAssification and Regression Trees
CBR:	Case-Based Reasoning
CNN:	Convolutional Neural Networks
COVID-Net:	Fuzzy Multi-Input Transfer Learning -
DCNN:	Deep Convolutional Neural Network
DM - ACME:	Data Mining - Adaptative Clustering Methodology Enhancement
DNN:	Deep Neural Network
DTs:	Decision Trees
EBM:	Explainable Boosting Machine
FCMs:	Fuzzy Cognitive Maps
FI:	Feature Interaction
FMTL-COVID-Net:	Fuzzy Multi-Input Transfer Learning - COVID-Net
FNN:	Feed-Forward Neural Networks
FRF:	Fuzzy Random Forest
GAN:	Generative Antagonistic Neural Networks
GB:	Gradient Boosting
GBDT:	Gradient Boosting Decision Trees
GCN:	Graph Convolutional Network

Grad CAM++:	Gradient-weighted Class Activation Mapping
HDFS:	Hadoop Distributed File System
ICE:	Individual Condicional Expectation
IML:	Intepretable Machine Learning
IoT:	Internet of Things
KDD:	Knowledge Discovery in Databases
K-NN:	K-Nearest-Neighbor
LightGBM:	Light Gradient Boosting Machine
LIME:	Local Surrogate Model
LR:	Logistic Regression
LRP:	Layer-wise Relevance Propagation
LSTM:	Long Short Term Memory
LUS:	Lung UltraSound
LVQ:	Learning Vector Quantization
MCDM:	Multiple Criteria Decision Making
ML:	Machine Learning
MLP:	MultiLayer Perceptron model
P2P:	Peer to Peer
PDA:	Pixel Density Analysis -
PDP:	Partial Dependence Plot
PR:	Polynomial Regression
qLM:	cuasi-Levenberg-Marquardt
RBF:	Radial Basis Function
REMDD:	Resampling Ensemble Model based on Data Distribution
RF:	Random Forest
RNN:	Recurrent Neural Networks
RoBERTa:	Robustly Bidirectional Encoder Representations
SHAP:	Shapley Additive Explanation
SMOTE:	Synthetic Minority Oversampling Technique
SNN:	Siamese Neural Networks

SOFM:	Kohonen Self-Organising Maps
SP-LIME:	Submodular-Local Interpretable Model-agnostic Explanations
SSA-CatBoost:	Sparrow Search Algorithm - Category and Boosting
SVM:	Support Vector Model
SVR:	Support Vector Regression
UMCDD:	Undersampling based on Majority Class Data Distribution
XAI:	eXplainable Artificial Intelligence
XGBoost:	eXtreme Gradient Boosting

Desarrollo de la memoria

1. Estructura de la memoria

El contenido de esta memoria se estructura en dos secciones principales. En la primera sección, se presentan los fundamentos y objetivos teóricos esenciales necesarios para la formulación de las soluciones propuestas. Además, se realiza un resumen de los casos de uso que dan solución a la problemática de la explicabilidad relacionada con esta tesis. Esta sección inicial termina con las conclusiones derivadas del trabajo realizado y con la identificación de nuevas áreas de investigación que se han identificado para profundizar en las metodologías propuestas en este estudio.

La segunda sección, presenta los artículos científicos publicados/en revisión en revistas internacionales de alto impacto durante los años de trabajo dedicados a esta tesis.

La Figura 1 representa el esquema de la estructura de esta tesis, que ha servido de base para la generación de varias soluciones empresariales.

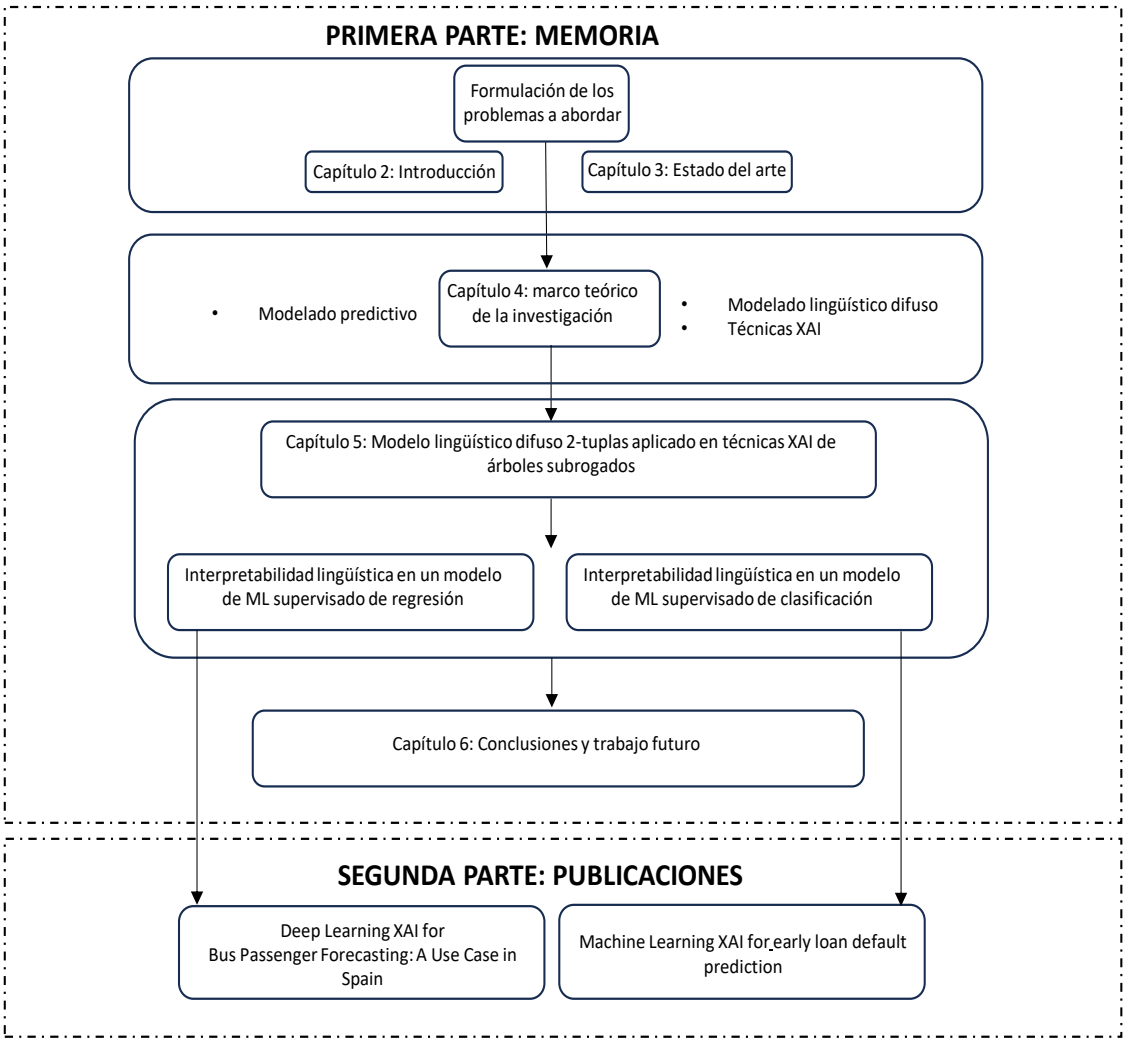


Figura 1. Esquema de la estructura de la memoria

La primera sección presenta los capítulos principales de la tesis, que se describen a continuación:

En el segundo capítulo de la tesis, se introducen los campos de aplicación de la investigación: Fintech y Ciudades Inteligentes. Se presentan las soluciones basadas en Ciencia de Datos que se han sugerido previamente para abordar los desafíos presentes en estos ámbitos y se detallan los objetivos que esta tesis busca alcanzar con el fin de contribuir a la resolución de problemas en los sectores anteriormente mencionados.

El tercer capítulo, presenta una revisión exhaustiva del estado del arte en relación con la explicabilidad lingüística en modelos de aprendizaje automático (en inglés Machine Learning - ML). Se inicia con un análisis exploratorio general y se procede

a llevar a cabo dos análisis exploratorios adicionales, uno para cada uno de los casos de uso abordados en esta investigación: Fintech y Ciudades Inteligentes.

El cuarto capítulo expone el marco teórico necesario para el desarrollo de las metodologías planteadas para la solución de los diferentes problemas incluidos en esta tesis. Principalmente, en este capítulo se exponen las diversas técnicas de ML supervisado empleadas para tareas de clasificación y regresión en los casos de estudio. Asimismo, se presentan técnicas de explicabilidad con un enfoque especial en las técnicas surrogadas, diseñadas para explicar las decisiones de los modelos de ML. Finalmente, se introduce el modelo lingüístico difuso 2-tuplas, que proporciona la capacidad de interpretación lingüística en el contexto de XAI.

En el quinto capítulo de esta investigación, se aborda inicialmente de forma general para después aplicarlo en los contextos de Fintech y ciudad inteligente, la implementación del modelo lingüístico difuso de 2-tuplas en el contexto de las técnicas de XAI mediante el uso de árboles surrogados. Este enfoque específico se conoce en inglés como "2-Tuple Fuzzy Linguistic Surrogate Trees and Rules".

Por último, el sexto capítulo muestra las conclusiones fundamentales derivadas de la investigación. También, se definen las posibles direcciones para futuras investigaciones.

La segunda parte de esta memoria expone dos artículos: uno de los cuales ya ha sido publicado y el otro se encuentra actualmente bajo revisión, ambos en destacadas revistas vinculadas a las áreas de Ciencias de la Computación y Economía.

2. Introducción

Como se ha indicado previamente, en esta primera parte se procede a la presentación de los casos de uso que han sido desarrollados en el marco de esta tesis. Cada uno de ellos muestra la aplicación de los modelos con el propósito de resolver la problemática de la explicabilidad tanto en el dominio de Fintech como en el contexto de las Ciudades Inteligentes.

I. Caso de uso de Fintech. Después de la devastadora crisis financiera de 2008 (Thakor, 2015), que dejó al público con una marcada desconfianza en la solidez de las instituciones bancarias tradicionales, surgió un modelo de préstamo que, aunque puede parecer innovador, se fundamenta en principios que tienen raíces históricas.

Un hito importante en esta evolución fue el lanzamiento de Lending Club en mayo de 2007 (Basha et al., 2021). Esta plataforma pionera introdujo el concepto de préstamos entre particulares o pares (en inglés Peer to Peer - P2P) e inversores, eliminando la intermediación de las entidades financieras convencionales. Lending Club se basó en la técnica del préstamo participativo (en inglés crowdlending), que permitía a los prestatarios obtener financiación directamente de inversores individuales a través de la plataforma en línea (Namvar, 2014), desempeñando así, un papel importante en la transformación del sector financiero después de la crisis de 2008, al proporcionar una alternativa moderna y basada en la confianza entre pares.

Este modelo ofrecía préstamos P2P con plazos flexibles que oscilaban entre tres y cinco años, y las cantidades solicitadas para los préstamos podían llegar hasta los 35.000 dólares (Namvar, 2014). Se trata de préstamos entre personas: personas con dinero (inversores), basándose en información crediticia, prestan dinero a través de una plataforma a personas anónimas que lo necesitan (prestatarios) (Wang et al., 2015). Los prestatarios tenían la oportunidad de obtener financiación sin tener que recurrir a bancos tradicionales, mientras que los inversores tenían la posibilidad de

diversificar sus inversiones al financiar múltiples préstamos a través de la plataforma.

Este enfoque de crowdlending no solo permitió a las personas acceder a crédito de manera más directa, sino que también ofreció un proceso de solicitud de préstamo más ágil y transparente. A lo largo del tiempo, este modelo inspiró el desarrollo de otras plataformas similares y contribuyó a la creación de un ecosistema financiero más diversificado y accesible para las personas y las empresas (Arnaut & Bećirović, 2023).

En la actualidad, el modelo de préstamos entre pares ha experimentado una notable evolución. Ahora, los inversores tienen la capacidad de canalizar sus fondos hacia porciones específicas de préstamos o distribuirlos entre varios, sin necesidad de comprometer su dinero en un único préstamo dirigido a un cliente en particular, (Chishti, 2016). Este avance ha dado lugar a un grupo de nuevos inversores impulsados por la innovación, quienes emplean enfoques analíticos revolucionarios que incorporan elementos como el Big Data, entendido como el procesamiento de grandes datos de información, datos no convencionales y técnicas de ML (Awotunde et al., 2021).

En este sentido, como resultado, estas tecnologías emergentes están transformando el paisaje financiero (Havrylchyk & Verdier 2018), impactando de manera significativa en la gestión del riesgo crediticio. Esto se debe a su capacidad para procesar grandes volúmenes de datos en tiempo récord, al tiempo que mejoran la precisión en la predicción de los posibles incumplimientos, un aspecto crítico en el sector financiero.

Hay que destacar que la morosidad representa uno de los desafíos más significativos para las entidades financieras, ya que puede comprometer su estabilidad y solvencia (Ghenimi et al., 2017). Por ello, el análisis del riesgo de crédito es fundamental y desempeña un papel crítico, ya que constituye la primera línea de defensa para salvaguardar la solvencia y garantizar una provisión de liquidez

eficiente. En este contexto, la detección temprana de señales de deterioro en los clientes adquiere una importancia relevante, ya que permite la adopción de medidas preventivas con el fin de evitar un empeoramiento de la situación del cliente. Un aspecto clave en este proceso es el seguimiento de la morosidad temprana, considerando mora temprana cuando el deudor no paga una cuota del préstamo en la fecha pactada, situándose el retraso del impago en un plazo inferior a 90 días.

La predicción de la morosidad temprana permite tomar medidas preventivas (Koyuncugil & Ozgulbas, 2012), como la reestructuración de la deuda antes de que ocurra el impago, evitando así los impagos contables, impago de más de 90 días desde la fecha pactada, y reduciendo la necesidad de realizar provisiones que afecten al capital disponible. En este sentido también permite a las instituciones financieras buscar soluciones mutuamente beneficiosas antes de que se materialice el impago, contribuyendo a una mejor relación con los clientes al abordar los problemas de pago de manera proactiva y ofrecer soluciones adecuadas.

Las entidades financieras están sujetas a regulaciones y normativas (Neuberger & Rissi, 2024), que requieren que mantengan reservas adecuadas para cubrir las pérdidas esperadas en sus carteras de préstamos. Para ello, efectúan provisiones para impagos como un componente fundamental de su responsabilidad financiera y su cumplimiento normativo. Esto les proporciona la capacidad de anticiparse a situaciones imprevistas, estar en consonancia con las regulaciones vigentes, mantener la confianza de los inversores y salvaguardar su estabilidad financiera en un contexto donde los impagos son una eventualidad.

Por todo ello, el enfoque proactivo de detección del impago temprano se utiliza para prevenir el impacto posterior en las provisiones y, por consiguiente, en el capital de la entidad financiera (Hidayat et al., 2021). Al identificar y abordar de manera anticipada los problemas de pago de los clientes, la entidad financiera puede reducir la necesidad de realizar provisiones adicionales y proteger así su posición financiera.

Actualmente, el uso de modelos y técnicas de ML en el sector financiero se ha extendido ampliamente para la detección temprana de impagos (Wang et al., 2021). En estos casos, las empresas pueden desarrollar modelos internos que no necesariamente tienen que ser transparentes en términos de su funcionamiento interno. Esto abre la puerta para la utilización de modelos analíticos más sofisticados, incluso los considerados "caja negra", como aprendizaje profundo (en inglés Deep Learning -DL) y XGBoost, (Adadi & Berrada, 2018), permiten predicciones más precisas al reducir su explicabilidad frente a modelos tradicionales como el probit (Abdou & Pointon, 2011) y la regresión logística (en inglés Logistic Regression - LR) (Bolton, 2009). Por ello, la interpretabilidad de estos algoritmos, a través de técnicas de XAI, desempeña un papel crucial (Heng & Subramanian, 2022).

En los últimos años, ha sido notable el crecimiento de los mercados de préstamos P2P, dada su naturaleza de préstamos entre particulares facilitados a través de plataformas en línea (Lenz, 2016). Sin embargo, en este contexto, los prestamistas se enfrentan a un riesgo de crédito significativo en caso de que los prestatarios incumplan con sus obligaciones de repago. Por ello, la capacidad de predecir el impago temprano reviste una importancia crucial. Los enfoques tradicionales para la predicción del incumplimiento, principalmente basados en modelos financieros probabilísticos, como las calificaciones crediticias convencionales, han hecho uso de modelos lineales, como la LR (Dreiseitl et al., 2002). Con el propósito de incrementar la capacidad predictiva del impago, se han incorporado modelos de ML. Resulta esencial que las predicciones generadas por estos modelos sean interpretables, lo que facilita su validación. En este contexto, la XAI desempeña un rol fundamental. La interpretabilidad de los modelos (Arrieta et al., 2019), adquiere una relevancia crítica en aras de garantizar la transparencia y la trazabilidad, especialmente en el ámbito de la predicción de impagos (Leo et al., 2019).

II. En el contexto de las Ciudades Inteligentes, la predicción del flujo de tráfico (Manibardo et al., 2021), los tiempos de viaje (Cristóbal et al., 2019), la prevención de congestiones y la demanda de pasajeros, entre otros, son aspectos críticos para la toma de decisiones. La comprensión del comportamiento de la demanda de pasajeros se convierte en un elemento crucial para tomar decisiones relacionadas con la planificación de servicios de transporte público y el diseño de rutas adaptadas a las necesidades de la población. A través de Internet de las Cosas (en inglés Internet of Things - IoT), los datos pueden recopilarse a lo largo de todo el viaje, como la recogida del pasajero, el tiempo de recorrido y la parada.

La demanda de pasajeros depende de muchos factores como la población, los ingresos y los establecimientos comerciales y de servicios. Otros factores dependen únicamente de la oferta de transporte, como el tiempo de viaje, las frecuencias ofrecidas y el confort. Para lograr la máxima eficiencia en el servicio, es esencial conocer la demanda de pasajeros a lo largo del tiempo, ya que generalmente se caracteriza por una alta estacionalidad con períodos pico anuales, semanales o diarios.

La mayoría de los sistemas de transporte público tienen frecuentes retrasos (Martin et al., 2021) debido a una mala planificación de las diferentes rutas, lo que conduce a un transporte abarrotado en ciertas de ellas y completamente vacío en otras. Esta mala gestión del transporte público también tiene otras consecuencias que afectan no solo a la vida de sus usuarios, sino también a las personas que no viajan en él.

Por ello, la recopilación de datos en tiempo real, así como su posterior análisis y estudio, proporciona toda la información necesaria para implementar y optimizar los procesos involucrados en la planificación de la demanda, además de ayudar en la toma de decisiones estratégicas y eficientes (Li et al., 2021). La predicción de la demanda de pasajeros se puede considerar como un aspecto fundamental en el contexto del transporte público, especialmente en las Ciudades Inteligentes (Liu et

al., 2019). La explicabilidad como ayuda para la toma de decisiones tiene un papel crucial en la ciencia de datos y, los casos de uso desarrollados en esta tesis ponen de manifiesto esa necesidad (Barredo-Arrieta et al., 2019; Heng & Subramanian, 2022).

El propósito de esta tesis es la creación de un método estadístico que pueda mejorar esa toma de decisiones, consiguiendo una interpretación lingüística sin pérdida de precisión del modelo generado por XAI. Los modelos presentados a lo largo de esta investigación buscan abordar diversos problemas mencionados previamente. A pesar de que las contribuciones se centran en los sectores de Fintech y Ciudades Inteligentes, es importante destacar que todos estos modelos son aplicables a una amplia gama de sectores que requieran de la interpretación lingüística de sus modelos de ML para la toma de decisiones.

2.1. Objetivos

En esta sección, se expondrán tanto el objetivo general como los objetivos específicos que se han formulado en el contexto de esta investigación. Estos objetivos serán objeto de revisión en la sección de Conclusiones y trabajo futuro (Sección 6), donde se evaluará su grado de cumplimiento.

2.1.1. Objetivo general

El objetivo principal de esta investigación es desarrollar un método estadístico eficaz que mejore la toma de decisiones, logrando una interpretación lingüística sin pérdida de precisión en el modelo generado por XAI. Con ello se mejora la interpretabilidad de los modelos XAI mediante las técnicas de modelos lingüísticos difusos, concretamente 2-tuplas y, a partir de esa mejora de interpretabilidad lo aplicamos al caso de uso de Fintech y Ciudades Inteligentes.

2.1.2. Objetivos específicos

Una vez realizada la mejora de la interpretabilidad, se han fijado diversos objetivos específicos con el propósito de complementar el objetivo principal de esta tesis. Estos objetivos se presentan a continuación:

- Implementar el modelo de explicabilidad mejorado con lenguaje de programación Python.
- Aplicación del modelo propuesto de explicabilidad mediante el uso del modelo de 2 -tuplas a la gestión efectiva de riesgos.
- Aplicación del modelo propuesto de explicabilidad a la gestión de Ciudades Inteligentes.
- Integración del modelo de explicabilidad mejorado con el modelo de 2-tuplas en la metodología de Descubrimiento de Conocimiento en Bases de Datos (en inglés Knowledge Discovery in Databases - KDD).

3. Estado del arte

La Inteligencia Artificial (IA) en el ámbito empresarial posibilita el análisis rápido de grandes volúmenes de datos, lo que capacita a las empresas para tomar decisiones más precisas e informadas. Esto se basa en un análisis detallado y en la mejora de la eficiencia operativa y de productividad, proporcionando una ventaja competitiva.

Las empresas recurren a modelos de caja negra en una variedad de aplicaciones, desde la predicción y análisis de datos, hasta la personalización de la experiencia del cliente y la optimización de procesos empresariales. La utilización de ML puede conferir a las empresas una ventaja competitiva al permitirles tomar decisiones más informadas y adaptarse ágilmente a los cambios en el entorno empresarial.

Es esencial destacar que la XAI busca garantizar que los sistemas de IA sean justos, seguros, transparentes y confiables, proporcionando una comprensión clara de cómo el modelo llega a sus predicciones y ofreciendo transparencia en el proceso de toma de decisiones. Sin embargo, es esencial alcanzar una interpretación sin comprometer la precisión del modelo generado por XAI. Con este fin, en esta investigación utilizaremos modelos lingüísticos difusos, los cuales son comprensibles por los humanos y permiten entender el razonamiento que se encuentra detrás de las decisiones del modelo.

3.1. Uso del modelado lingüístico difuso en XAI

De cara a perfilar el uso de modelos lingüísticos difusos en las técnicas XAI, el proceso de búsqueda de literatura se lleva a cabo mediante la combinación de un conjunto de palabras clave específicas, que incluyen términos como "XAI", "Fuzzy", "Machine Learning" y "Deep Learning". La consulta se ejecutó a primeros de diciembre de 2023 y se descargaron 25 registros bibliográficos de la principal colección de Web of Science que fueron revisados manualmente. Además, se realizó

una exhaustiva revisión de la literatura que nos permitió seleccionar los artículos más relevantes para nuestra investigación que a continuación comentamos:

En el artículo (Kamal et al., 2022) se abordó el hecho de que los modelos existentes de ML sólo pueden predecir el glaucoma, sin proporcionar explicaciones claras sobre cómo se toman las decisiones. Para abordar esta limitación, se propusieron modelos de XAI y aprendizaje automático interpretable (en inglés Intepretable Machine Learning - IML) que permitieran analizar y explicar las predicciones de glaucoma. Se utilizaron enfoques como sistema de inferencia neurodifusa adaptativo (en inglés Adaptive Neuro-Fuzzy Inference System - ANFIS), análisis de densidad de píxeles (en inglés Pixel Density Analysis - PDA) y explicaciones submodulares-locales interpretables y agnósticas del modelo (en inglés Submodular-Local Interpretable Model-agnostic Explanations - SP-LIME) que brindasen explicaciones coherentes y confiables sobre las decisiones del modelo.

Este estudio (Murray et al., 2020) se enfocó en la fusión de datos e información, específicamente abordando cuatro aspectos clave de la XAI relacionados con la integral de Choquet. Estas cuestiones incluyen la calidad de las entradas y sus interacciones, la combinación de información, la calidad de los datos de entrenamiento y la confianza en los resultados. A través de datos sintéticos y experimentos de teledetección, este enfoque mejoró la precisión de XAI, proporcionando una comprensión más profunda de lo aprendido y orientando hacia soluciones más efectivas en el futuro.

En el siguiente estudio (Tjoa et al., 2021) propuso una revisión de las interpretabilidades sugeridas por diferentes trabajos de investigación en relación a la naturaleza de la caja negra del DL categorizándolas para mostrar las distintas dimensiones de la investigación en este campo. Se destacó la necesidad de enfoques que proporcionaran información interpretable de manera clara, así como estudios sobre patrones complejos. La aplicabilidad de estas categorías a la investigación médica buscó fomentar su uso con cautela entre los profesionales, promover

conocimientos sobre interpretabilidad en prácticas médicas y respaldar iniciativas de investigación basadas en datos. En línea con la investigación médica, (Akpan et al., 2022), presentó una aplicación mejorada del algoritmo de C-means difuso y Red Neuronal Artificial para la segmentación de imágenes médicas, específicamente en la detección automática de tumores cerebrales mediante resonancia magnética. La explicabilidad de estos procesos buscaba generar seguridad en las decisiones médicas, así como mejorar la confianza, aceptación y la posibilidad de integración en el flujo de trabajo de segmentación de imágenes médicas. En última instancia, se destacó la relevancia de identificar y clasificar correctamente las lesiones patológicas en sus etapas iniciales para mejorar la probabilidad de supervivencia.

En respuesta a la pandemia de COVID-19, esta investigación (Hasan et al., 2023) propuso una metodología inteligente asistida por ordenador para clasificar imágenes pulmonares de ultrasonido. Utilizó una red neuronal convolucional (en inglés Convolutional Neural Networks – CNN) basada en agrupamiento difuso, FP-CNN, que destaca características representativas cruciales para la clasificación. El modelo clasificó las imágenes en tres clases: COVID, normal y neumonía. Para garantizar la equidad, se empleó la Explicación aditiva de Shapley (en inglés Shapley Additive Explanation - SHAP) para explicar las decisiones del modelo. Se probaron diversas arquitecturas de CNN, y el modelo Xception con agrupamiento difuso en todas las capas logró la mejor precisión del 97.2%. Este enfoque buscó ser útil en el diagnóstico clínico de COVID-19 a través de imágenes de ultrasonido pulmonar (en inglés Lung UltraSound - LUS). En este estudio (Hu et al., 2022), se presentó una CNN Transferencia difusa de múltiples entradas (en inglés Fuzzy Multi-Input Transfer Learning - COVID-Net - FMTL-COVID-Net) para detectar instancias de COVID-19 a partir de radiografías torácicas, motivados por la investigación existente y los esfuerzos de código abierto en esta área. Además, utilizaron un método de explicabilidad para investigar varias predicciones de la CNN COVID-Net no solo para obtener una comprensión más profunda de los factores críticos asociados con las instancias de COVID-19, sino también para ayudar a los médicos a mejorar la

detección. Mostraron que mediante el uso de transfer learning y modelos pre-entrenados, pudieron detectarlo con un alto grado de precisión. Utilizaron imágenes de rayos X y seleccionaron cuatro redes neuronales para predecir su probabilidad. Finalmente, consideraron varios métodos para verificar las técnicas propuestas y lograron crear un modelo con un área bajo la curva (en inglés Area Under the Curve - AUC) de 1,0 y una precisión y sensibilidad de 0,97. El modelo se cuantificó para su uso en dispositivos de IoT y mantuvo una precisión del 0,95%. Para el diagnóstico definitivo de la enfermedad COVID-19, la reacción en cadena de la polimerasa con transcriptasa inversa sigue siendo el estándar de oro. Sin embargo, las pruebas de laboratorio disponibles actualmente pueden no ser capaces de detectar a todos los individuos infectados, por lo que se requieren nuevos métodos de detección. Para ello se propuso en este artículo (Zeltner et al., 2021) un modelo neuronal basado en lógica continua que utiliza funciones de activación denominadas Squashing, que forman una familia paramétrica de funciones que satisfacen requisitos de invarianza natural y contienen unidades lineales rectificadas como un caso particular. Este trabajo presentó los primeros test de referencia que miden el rendimiento de las funciones Squashing en redes neuronales. Se llevaron a cabo tres experimentos para examinar su utilidad y se realizó una comparativa con las funciones de activación más populares en cinco tipos diferentes de redes. El rendimiento se determinó midiendo la precisión, la pérdida y el tiempo por época. Estos experimentos y las pruebas realizadas han demostrado que el uso de funciones Squashing es posible y similar en rendimiento a las funciones de activación convencionales.

Para mejorar la comprensibilidad de las DNN mediante la integración de lógica difusa (Yeganejou et al., 2019) propuso una CNN como extractor automático de características. Posteriormente, se aplicó un proceso de agrupamiento difuso en el espacio de características resultantes. Después de consolidar los grupos, empleó el algoritmo de Rocchio para la clasificación de los puntos de datos. Los experimentos realizados indicaron que la extracción automática de características conlleva a una mejora significativa en la precisión del clasificador difuso. Para entender la lógica

subyacente en el funcionamiento de estas redes, (Aghaeipoor et al. 2023) propuso sistemas explicativos basados en reglas difusas que mejorasen la comprensión de modelos basados en DNN. El algoritmo propuesto aprende un conjunto de reglas difusas compactas y precisas utilizando la importancia de las características obtenidas de las redes neuronales entrenadas. Estos sistemas explicativos pueden aplicarse tanto a nivel local como global. Los resultados de la evaluación en diversas aplicaciones demostraron que los explicadores difusos mantuvieron la fidelidad y precisión de las DNN originales, al tiempo que presentaron una menor complejidad y mayor comprensibilidad, mejorando la interpretabilidad de las DNN mediante la representación lingüística basada en reglas difusas, permitiendo pasar de un alto número de reglas específicas a un conjunto más reducido de reglas generales y confiables, mejorando la interpretabilidad y manteniendo la precisión del modelo. Los resultados experimentales indicaron que el algoritmo propuesto logró mejorar tanto la explicabilidad como el rendimiento predictivo de los clasificadores basados en reglas difusas. En la siguiente línea de investigación (Zaman & Hassan 2021) se propusieron técnicas de ML como redes neuronales artificiales (en inglés Artificial Neural Networks - ANN), árboles de clasificación y regresión (en inglés ClAssification and Regression Trees - CART) y sistemas de inferencia difusa para mejorar el gráfico de control de Shewhart en el diagnóstico de procesos. Aunque los clasificadores de ANN son opacos, la inferencia difusa y CART son más transparentes. Este artículo comparó heurísticas difusas y la técnica CART para el reconocimiento de patrones en gráficos de control. El clasificador CART superó ligeramente a la heurística difusa Mamdani en precisión de clasificación (98.58% frente a 95.76%). En línea con redes neuronales (Pradhan et al., 2022) utilizó una CNN y la SHAP para identificar posibles ubicaciones de mineralización de oro en la provincia de Rengali, India. Integró datos geológicos, geofísicos, mineralógicos y geoquímicos, y el modelo CNN logró una precisión del 90%. Los factores clave incluyen antimonio, arcilla, plomo, arsénico y anomalías magnéticas, con indicadores geoquímicos y geofísicos prominentes. Las áreas centrales, especialmente el valle del

río, mostraron perspectivas más altas de oro, asociadas posiblemente con metavolcánicos intermedios y zonas de cizallamiento. El enfoque busca la replicabilidad en áreas circundantes al modelar ocurrencias conocidas.

En este artículo, (Deepanshi et al., 2023) utilizaron para diagnósticos eficientes y explicables en el contexto del COVID-19, modelos de aprendizaje profundo, como Red Residual (en inglés Residual Network-50), Inception V3, Red Neuronal Densa (en inglés Dense neural network -121) y Red neuronal convolucional profunda (en inglés Deep Convolutional Neural Network – DCNN), con la integral de Choquet integrada como función de agregación para mejorar la clasificación. La explicabilidad se incrementa mediante la integración de un Mapa de activación de clases ponderado por gradiente (en inglés Gradient-weighted Class Activation Mapping - Grad CAM++). Los resultados mostraron un aumento del 2% en la precisión para Densenet-121 y DCNN, y del 1% para ResNet-50 e Inception V3. El uso de Grad CAM++ mejoró la aceptación entre los profesionales médicos, al proporcionar claridad sobre cómo se tomaron las decisiones. El modelo propuesto, Densenet 121 integrado con ChI, destacó como el más efectivo en comparación con otros métodos.

Este trabajo (Csiszar et al., 2023), presentó un modelo de aprendizaje profundo en el que se buscó ofrecer no solo predicciones precisas, sino también explicaciones comprensibles de cómo se relacionan las características de entrada en el proceso de toma de decisiones. Para ello el enfoque utilizado se basó en una función de activación parametrizada y diferenciable, inspirada en la lógica difusa nilpotente y la toma de decisiones multicriterio (en inglés Multiple Criteria Decision Making - MCDM). El parámetro aprendible en el modelo tiene un significado semántico que indica el nivel de compensación entre las diversas características de entrada. La red neuronal ajusta estos parámetros mediante el descenso de gradiente, buscando encontrar relaciones comprensibles por los humanos entre las características de entrada. En este sentido (Loh et al., 2022) realizaron la revisión de 99 artículos de categoría Q1 que se centraban en diversas técnicas de XAI, como SHAP, Explicación

agnóstica del modelo local interpretable (en inglés Agnostic explanation of the local interpretable model – LIME) , GradCAM, Propagación de Relevancias por Capas (en inglés Layer-wise Relevance Propagation – LRP), clasificadores difusos, Máquina de Refuerzo Explicable (en inglés Explainable Boosting Machine – EBM), Razonamiento basado en Casos Reales (en inglés Case-Based Reasoning – CBR) y sistemas basados en reglas. Los resultados destacan áreas específicas de la atención médica, como la detección de anomalías en biosignales unidimensionales y la identificación de texto clave en notas clínicas, que requieren más atención en la investigación de la explicabilidad. Este artículo (Islam et al., 2019) analizó los mapas cognitivos difusos (en inglés Fuzzy Cognitive Maps - FCMs), destacando su evolución y aplicación en la IA, incorporando avances como algoritmos de aprendizaje y CNN. Los FCMs, conocidos por su transparencia e interpretabilidad, se presentaron como una solución eficaz para abordar la explicabilidad. Se defienden las propiedades explicativas de los FCMs y destacan ejemplos de implementación exitosa en diversos dominios, tales como: apoyo a decisiones médicas, agricultura de precisión, ahorro de energía, monitoreo ambiental y formulación de políticas públicas. En el siguiente análisis (Apostolopoulos et al., 2023), demostraron que la integral difusa de Choquet (ChI), una potente función de agregación no lineal se puede representar como una red de múltiples capas, denominada en adelante ChIMP. También se presentó una mejora de ChIMP (iChIMP) que conduce a una optimización basada en el descenso de gradiente estocástico, teniendo en cuenta el número exponencial de restricciones de desigualdad de ChI. Un beneficio adicional de ChIMP/iChIMP es que permite la explicabilidad. Se aplicó iChIMP a la fusión de un conjunto de modelos profundos de arquitectura heterogénea en teledetección consiguiendo una mejora en la precisión del modelo. En esta investigación (Rohlfing et al., 2020) se revisó el concepto actual de XAI destacando tres limitaciones: el explicado pasivo, una visión estrecha del proceso social y una evaluación no diferenciada de la comprensión del explicado. Para superar estas limitaciones, los autores propusieron un enfoque donde ven la explicabilidad como una práctica social coconstruida a nivel micro, donde tanto el

que explica como el explicado contribuyen a la comprensión. Esta coconstrucción a nivel micro se incorporó a un nivel macro, generando expectativas sobre roles sociales y modelos de colaboración. La siguiente línea de investigación (Maaroof et al., 2023) presentó un método innovador denominado Multi-Class Fuzzy-LORE (mcFuzzy-LORE) diseñado para explicar las decisiones de clasificadores multi-clase basados en lógica difusa, como los Bosques Aleatorios Difusos (en inglés Fuzzy Random Forest - FRF). La técnica mcFuzzy-LORE es una adaptación del método Fuzzy-LORE que sustituye los DTs clásicos por DTs difusa. Esto proporciona reglas interpretables y comprensibles para los humanos, permitiendo entender el razonamiento que hay detrás de las decisiones del modelo para una entrada específica. Los resultados revelaron que mcFuzzy-LORE superó a los métodos anteriores basados en LORE clásicos en la generación de instancias contrafactuales.

En este artículo (Biedma-Rdguez et al., 2022) se propuso una extensión novedosa de la gramática lingüística difusa y una extensión lineal para problemas de regresión. Se utilizó un enfoque mejorado de aprendizaje evolutivo multiobjetivo basado en árboles lingüísticos para expresar tanto el comportamiento general como la variabilidad específica de los datos en una sola regla. El proceso de aprendizaje buscó maximizar dos métricas semánticas mientras minimizaba el número de reglas y el error cuadrático medio del modelo, garantizando transparencia y precisión.

En el siguiente documento (Fernández et al., 2019) se buscó proporcionar una visión integral y orientadora sobre el papel y el desarrollo futuro de los sistemas difusos evolutivos en la investigación actual.

En este artículo (Aghaeipoor et al., 2021), se propuso un clasificador difuso interpretable, denominado IFC-BD, basado en celdas, abarcando tres etapas: aprendizaje de reglas iniciales, generalización de reglas y selección heurística de reglas, para abordar el desafío de mantener la interpretabilidad en sistemas complejos basados en reglas difusas, especialmente cuando se enfrentan a grandes volúmenes de datos. (Salimi-Badr et al., 2023) presentaron un clasificador

interpretable para diagnosticar la Enfermedad de Parkinson (EP) basándose en el análisis del ciclo de la marcha. El método utiliza características clínicas extraídas de la Fuerza de Reacción del Suelo Vertical (vGRF) medida por sensores portátiles, y emplea la lógica difusa de Tipo-2 para aumentar la robustez frente a datos ruidosos del sensor. El proceso comenzó extrayendo reglas difusas iniciales mediante un enfoque de agrupamiento basado en Vecinos Más Próximos (en inglés K-Nearest-Neighbor – K-NN). Luego, se introdujo un enfoque de aprendizaje cuasi-Levenberg-Marquardt (qLM) para ajustar las reglas iniciales, minimizando la función de pérdida de entropía cruzada mediante un método de optimización de región de confianza. Además, se propuso un aprendizaje en línea complementario para mejorar las reglas mediante la incorporación de nuevas muestras etiquetadas. La evaluación del método abarcó la clasificación de pacientes sanos en diversas condiciones, incluyendo la presencia de ruido y la observación de nuevas muestras.

Esta propuesta (Fernández et al., 2022) definió explicaciones fácticas y contrafactuales en el contexto de árboles de decisión difusos, donde varias ramas podían activarse simultáneamente. La propuesta se centró en explicaciones fácticas que podían incluir más de una regla, en contraste con el estándar actual que se limitaba a considerar una sola regla como explicación fáctica. Además, introdujo la noción de explicación fáctica robusta. Finalmente, el trabajo presentó procedimientos para obtener explicaciones contrafactuales tanto a partir de la instancia como de una explicación fáctica. En conjunto, esta propuesta buscó mejorar la interpretación de los modelos de clasificación, particularmente en el contexto de árboles de decisión difusos. Este trabajo (Artelt et al., 2023) se enfocó en cómo calcular explicaciones contrafactuales de manera eficiente para clasificadores basados en prototipos, como los modelos de cuantización de vectores de aprendizaje. Propuso programas específicos, tanto convexos como no convexos, dependiendo de la métrica utilizada. El enfoque se centró en cambios mínimos en el modelo en sí, en lugar de en la entrada, explorando contrafactuales de la prescripción del modelo en lugar de los puntos de datos. La investigación se focalizó en clasificadores basados en distancias, como

modelos de cuantización de vectores de aprendizaje, y desarrolló técnicas de optimización eficientes para generar cambios métricos contrafactuales según el modelo seleccionado.

El siguiente trabajo (Yuksel et al., 2022) examinó el progreso y tendencias de investigación en los últimos 15 años, destacando métodos como el ML, algoritmos genéticos y lógica difusa. Revisó estudios de diseño impulsados por IA en diversas etapas, resaltando el aumento del interés en métodos basados en datos y XAI. El uso de IA en diseño de ingeniería ofrece resultados eficientes, rápidos y precisos, especialmente con métodos de aprendizaje profundo. La elección adecuada del método es crucial, y el trabajo proporcionó una perspectiva sobre cómo seleccionar el enfoque correcto basándose en la literatura para problemas de diseño.

Este estudio (Marques-Silva & Ignatiev, 2023) estudió vincular la interpretabilidad con la capacidad del modelo para explicar por qué realizaba una predicción en un punto específico del espacio de características. Al buscar explicaciones para las predicciones, el estudio reveló limitaciones adicionales en modelos interpretables. Se enfocó en dominios donde el propósito era ayudar a los tomadores de decisiones humanos a comprender el razonamiento que existe detrás de las predicciones, buscando información irreducible. En estos contextos, el estudio argumentó que las respuestas a preguntas del tipo "por qué" pueden presentar redundancia arbitraria, permitiendo simplificaciones mediante la inspección humana de la representación del modelo interpretable de ML.

El artículo (Bárcena et al., 2023) propuso un marco de trabajo de Aprendizaje Federado (AF) como servicio para permitir el entrenamiento colaborativo de modelos sin revelar datos privados. Se enfocó en la aplicabilidad de este marco al pronóstico de calidad de experiencia (QoE) en redes vehiculares, destacando cómo el AF de modelos de XAI mejora la precisión y la confiabilidad.

Finalmente, el artículo (Junior et al., 2023) examinó el estado del arte de los sistemas difusos profundos para regresión, que integran DL y los sistemas de lógica difusa para lograr precisión e interpretabilidad. En el contexto de la XAI, destacó la importancia de la interpretabilidad desde el desarrollo del modelo, no solo como un proceso posterior al aprendizaje.

La Tabla 1 muestra los trabajos relacionados específicamente con la predicción, modelado lingüístico difuso y explicabilidad de los modelos utilizados.

Tabla 1. Trabajos relevantes sobre modelos Machine Learning, lingüísticos difusos y XAI.

Referencia	**Fundamentos**	**Aplicación**	**Uso modelos difusos en XAI**
Islam et al., 2019	ChIMP/iChIMP	Optimización de ChIMP	No
Apostolopoulos et al., 2023	FCMs	Análisis de la naturaleza de los mapas cognitivos difusos	
Artelt & Hammer, 2022	Modelos de cuantización de vectores de aprendizaje	Calcular eficientemente explicaciones contrafactuales	
Bárcena et al., 2023	AF	Redes inalámbricas	
Deepanshi et al., 2023	ResNet-50, Inception V3, Densenet-121 y DCNN	Interpretación de imágenes de radiografías de tórax	
Hasan et al., 2023	FP-CNN	Diagnóstico médico	
Hu et al., 2022	modelo de CNN difusa llamado Multi-Input Transfer Learning	Detección COVID-19	
Marques-Silva et al., 2023	Revisión de literatura	Relación predicción-interpretabilidad	
Kamal et al., 2022	ANFIS, PDA y SP-LIME, IML	Predicciones de glaucoma	
Murray et al., 2020	integral de Choquet	Aumento del rendimiento y comprensión	

Referencia	Fundamentos	Aplicación	Uso modelos difusos en XAI
Pradhan et al., 2022	CNN	Pronosticar la mineralización de oro	No
Tjoa et al., 2021	Revisión de literatura	Revisión de las interpretabilidades sugeridas por diferentes trabajos de investigación y aplicación a la investigación médica	
Zaman & Hassan, 2021	CART, CNN	Procesos de fabricación	
Zeltner et al., 2021	CNN	Detección COVID-19	
Aghaeipoor et al., 2021	IFC-BD	Reducción del número de reglas siendo éstas más generales y confiables	Sí
Aghaeipoor et al., 2023	DNN	mejorar la comprensión de modelos basados en DNN	
Biedma-Rdguez et al., 2022	Regresión	Obesidad infantil	
Csiszar et al., 2023	MCDM	Relaciones comprensibles para el ser humano entre las características de entrada	

Referencia	**Fundamentos**	**Aplicación**	**Uso modelos difusos en XAI**
Fernández et al., 2019	Revisión de literatura. Responde a las 4W	Comprensión del contexto actual e los sistemas difusos evolutivos	Sí
Fernández et al., 2022	Revisión de literatura	Definición de explicaciones fácticas y contrafactuales en el marco de los árboles de decisión difusos y proporción de procedimientos para la obtención de explicaciones.	
Loh et al., 2022	Revisión de literatura	Identificar áreas de la atención médica que requieren más atención por parte de la comunidad de investigación en XAI	
Maaroof et al., 2023	mcFuzzy-LORE	Explicación de las decisiones tomadas por clasificadores multi-clase basados en lógica difusa	
Junior et al., 2023	DL	Estado del arte de las técnicas difusas profundas	
Salimi-Badr et al., 2023	K-NN, qLM	Diagnóstico de la Enfermedad de Parkinson	

Referencia	**Fundamentos**	**Aplicación**	**Uso modelos difusos en XAI**
Yeganejou et al., 2019	Redes profundas	Introducir un mecanismo de interpretación efectivo	Sí
Yuksel et al., 2022	Revisión de literatura	Diseño de ingeniería. Cómo elegir el método de Inteligencia Artificial adecuado basándonos en los resultados de la literatura para problemas de diseño	

Tras estudiar la literatura, hemos observado que no se han incorporado los modelos lingüísticos difusos que ayudan a interpretar mejor los modelos desde el punto de vista humano. La asignación de etiquetas lingüísticas es fruto de la agregación de datos. Cada nodo del árbol lleva agregado un conjunto de datos, y con el uso del modelo de 2-tuplas propuesto en este estudio se mejora la interpretabilidad y la precisión.

3.2. Aplicaciones de XAI en predicción en el contexto de Fintech: técnicas de interpretabilidad en la predicción de la mora en préstamos P2P

Alineada con los objetivos de la tesis, la investigación de la literatura científica se divide en dos vertientes: una enfocada en la predicción de incumplimientos a partir de modelos considerados caja negra y, la otra, en la interpretabilidad de los modelos que realizan estas predicciones. El proceso de búsqueda de literatura se llevó a cabo mediante la combinación de un conjunto de palabras clave específicas, incluyendo términos como "forecast", "prediction*", "credit", "loan", "mortgage", "lending", así como "Machine Learning", "Deep Learning", y conceptos relacionados con plataformas P2P como "Peer to Peer" y "P2P". Además, se incorporaron términos relevantes a la interpretabilidad de la IA, como "XAI" y "explainable artificial intelligence". La consulta se ejecutó a primeros de abril de 2023 y se descargaron 22 registros bibliográficos de la principal colección de Web of Science que fueron revisados manualmente.

En el estudio (Chen et al., 2019) se propuso un modelo de ML con selección de características y se demuestró que la predicción del riesgo crediticio para los préstamos P2P se podia mejorar utilizando la LR además de una selección de características adecuada. Se utilizó el submuestreo para equilibrar el conjunto de datos. Otro estudio similar (Zhou et al., 2019) propuso el uso de árboles de decisión de gradiente de refuerzo (en inglés Gradient Boosting Decision Trees - GBDT), XGBoost y la máquina de refuerzo de gradiente ligero (en inglés Light Gradient Boosting Machine - LightGBM). En la misma línea (Ma et al., 2018) y (Ko et al., 2022) propusieron LightGBM basado en múltiples conjuntos de datos que clasifican los resultados de predicción como un modelo más predictivo. En (Li et al., 2018) se diseñó un modelo de aprendizaje ensemble multirronda basado en conjuntos heterogéneos para la predicción del riesgo de impago. En este modelo, XGBoost se utilizó inicialmente para el aprendizaje de conjuntos, y XGBoost, la red neuronal

profunda (en inglés Deep Neural Network - DNN) y la LR se consideraron aprendices individuales heterogéneos para someterse a la ponderación lineal. En (Yang et al., 2023) se propuso un nuevo modelo Algoritmo de búsqueda Sparrow – Categoría Reforzamiento (en inglés Sparrow Search Algorithm - Category and Boosting - SSA-CatBoost), que combinó SSA y CatBoost para mejorar la predicción y la clasificación. Comparando otros modelos de ML conocidos y el modelo CatBoost, el modelo SSA-CatBoost tiene la mejor precisión de clasificación y predicción. En el siguiente trabajo (Song et al., 2020) se propuso un método de aprendizaje ensemble multivista basado en distancia a modelo y clustering adaptativo (en inglés Data Mining - Adaptative Clustering Methodology Enhancement - DM - ACME) que utiliza refuerzo de gradiente (en inglés: Gradient Boosting - GB). Este método produjo un conjunto de diversos árboles de decisión. En (Zhu et al., 2019) se propuso un modelo de predicción de impago de préstamos basado en un bosque aleatorio (en inglés Random Forest - RF) entrenado utilizando el método de remuestreo "técnica de Sobremuestreo Sintético de la Minoría" (en inglés Synthetic Minority Oversampling Technique - SMOTE) para hacer frente al problema de desequilibrio de clases en el conjunto de datos. Por otro lado, en (Chen, 2021) los autores utilizaron el modelo XGBoost con tres métodos de remuestreo SMOTE, Cercano al Menor (en inglés NearMiss) y selección aleatoria, para predecir el incumplimiento sin el problema del desequilibrio de clases. En (Niu et al., 2020), se realizó un remuestreo conjunto basado en la distribución de datos (en inglés Resampling Ensemble Model based on Data Distribution - REMDD), en el que el uso del método de submuestreo basado en la distribución de datos de clase mayoritaria (en inglés Undersampling based on Majority Class Data Distribution - UMCDD) resuelve el problema del desequilibrio de clases mejorando la clasificación del REMMD. En (Li et al., 2020) se propuso un modelo que añade valores perdidos al modelo para el autoentrenamiento mejorando el rendimiento y resuelve el problema de desequilibrio de clases proponiendo un nuevo modelo basado en el aprendizaje de conjuntos heterogéneos. En (Li, 2022) basandose en redes neuronales de aprendizaje profundo propuso un

modelo predictivo para la predicción de impago de préstamos online. Simula el modelo de evaluación Red Neuronal con Retropropagación (en inglés Back Propagation Neural Network - BPNN) y lo compara con la máquina de vectores soporte (en inglés Support Vector Model - SVM) y el modelo de regresión, siendo el modelo BPNN el de mayor precisión. En esta línea, (Kim & Cho, 2019), presentó otro modelo de aprendizaje profundo, una CNN para la predicción del reembolso en préstamos sociales P2P. Los datos de préstamos P2P se generan en tiempo real, pero muchos de los préstamos están pendientes de decisiones de reembolso porque el plazo aún no ha expirado. En (Kim & Cho, 2022) se consideraron los datos no vencidos etiquetados para mejorar el rendimiento de predicción del modelo. Para ello, se propuso un clasificador conjunto compuesto por varias CNN de GoogLeNet, ResNet (en inglés Residual Network) y DenseNet (en inglés Densely Connected Convolutional Networks) y los datos adicionales etiquetados por fusión Dempster-Shafer potenciando el entrenamiento de varios modelos CNN.

En (Kriebel & Stitz, 2021) los autores propusieron un método para mejorar la predicción del riesgo de impago utilizando la extracción de palabras clave y breves fragmentos de texto de los comentarios de los usuarios. De este modo, mostraron la mejora de la predictividad de métodos más sencillos, como las redes neuronales de incrustación media, frente a métodos más complejos, como las representaciones codificadoras bidireccionales a partir de transformadas (en inglés Bidirectional Encoder Representations from Transformers - BERT) y las BERT optimizadas de forma robusta (en inglés Robustly Bidirectional Encoder Representations - RoBERTa).

El trabajo (Fu et al., 2020) propuso extraer palabras clave de los comentarios de los inversores y, a continuación, utilizó un modelo basado en la memoria bidireccional a corto y largo plazo (en inglés Bidirectional Long Short Term Memory - BiLSTM). En (Lee et al., 2021), se propuso un modelo gráfico basado en redes convolucionales gráficas no lineales (en inglés Graph Convolutional Network - GCN) entre el riesgo

de impago y los atributos del prestatario. Para sus relaciones se utilizaron por separado tres tipos de información sobre los prestatarios: información sobre el préstamo, información sobre el historial crediticio e información blanda. Por otro lado, en (Xu et al., 2021) se analizaron los factores que afectan al reembolso. Se aplicaron cuatro métodos de ML (RF, XGBoost, GB y Neural Network) para predecir los factores importantes que afectan al reembolso. En el artículo (Arora & Kaur, 2020) se realizó una selección de características del conjunto que hacen que los modelos sean más predictivos. Las características preseleccionadas generadas por Bolasso se aplicaron a varios algoritmos de clasificación, como RF, SVM, Naïve Bayes (NB) y K-NN para probar su precisión predictiva. Se observó que el algoritmo RF habilitado por Bolasso (BS-RF) proporcionó los mejores resultados para la evaluación del riesgo crediticio. Por otro lado, en (Cho et al., 2019) se propuso un modelo de decisión de inversión en el mercado de préstamos P2P consistente en la clasificación de préstamos totalmente pagados utilizando el método de vectores de soporte difusos basados en instancias (IEFSVM), que es una versión modificada de la máquina de vectores de soporte difusos basados en entropía (EFSVM). Aplicando el modelo al conjunto de datos de préstamos, se determinaron los préstamos que se esperaba que fueran reembolsados en su totalidad.

Para predecir el impago de préstamos (Stevens et al., 2020) utilizaron el modelo XGBoost y la explicabilidad del modelo se comprobó mediante explicaciones post hoc utilizando el método SHAP. En (Moscato et al., 2021) se realizó un estudio comparativo de algunos de los modelos de scoring más utilizados para la predicción del riesgo crediticio. Utilizaron diferentes técnicas para resolver el problema del desequilibrio de clases y, finalmente, los tres mejores enfoques fueron evaluados en términos de su explicabilidad, utilizando diferentes herramientas XAI.

La Tabla 2 muestra los trabajos relacionados específicamente con la predicción de impagos y la explicabilidad de los modelos utilizados. La tabla también incluye

nuestra propuesta, en la que consideramos el modelo XGBoost, un modelo ampliamente utilizado en la literatura citada.

Tabla 2. Trabajos relevantes sobre predicción del riesgo de impago y XAI.

Referencia	**Conjunto de datos**	**Fundamentos**	**Aplicación**	**Interpretabilidad**
Arora & Kaur, 2020	Préstamos P2P	Several ML methods: RF, SVM, NB, K-NN BS-RF	Predicción del riesgo de impago	Modelos de caja negra, no interpretabilidad
Chen et al., 2019	Préstamos P2P	LR	Predicción del riesgo de impago	
Cho et al., 2019		IEFSVM	Predicción del riesgo de impago	
Fu et al., 2020	Préstamos P2P	BiLSTM	Extraer palabras clave de los comentarios de los inversores para predecir el riesgo de impago	
Kim & Cho, 2019	Préstamos P2P	CNN	Predicción del riesgo de impago	
Kim & Cho, 2022	Préstamos P2P	CNN	Clasificador conjunto para la predicción de reembolsos en préstamos sociales	
Ko et al., 2022	Préstamos P2P	LightGBM	Mitigar los riesgos de impago y de información asimétrica	
Kriebel & Stitz, 2021	Préstamos P2P	BERT, RoBERTa	Predicción del riesgo de impago mediante textos breves generados por los usuarios	

Referencia	Conjunto de datos	Fundamentos	Aplicación	Interpretabilidad
Lee et al., 2021	Préstamos P2P	GCN	Predicción del riesgo de impago	Modelos de caja negra, no interpretabilidad
Li et al., 2018	Préstamos P2P en China	Several ML methods: XGBoost, DNN, LR	Predicción del riesgo de impago. Comparación con modelos tradicionales de ML y modelos de aprendizaje por conjuntos.	
Li et al., 2020	Préstamos P2P en China	Several ML methods: XGBoost, DNN, LR	Predicción del riesgo de impago. Problema de desequilibrio	
Li, 2022	Préstamos P2P	NN	Predicción del riesgo de impago	
Ma et al., 2018	P2P loans in Asia	LightGBM	Predicción del riesgo de impago	
Niu et al., 2020	Préstamos P2P	REMDD, UMCDD	Predicción del riesgo de impago. Problema de desequilibrio	
Song et al., 2020	Préstamos P2P	DM-ACME	Predicción del riesgo de impago	
Chen, 2021	Préstamos P2P	SMOTE, NearMiss	Predicción del riesgo de impago. Problema de desequilibrio	
Xu et al., 2021	Préstamos P2P en China	Several ML methods: RF, XGBT, GBM, NN	Predicción del riesgo de impago usando datos de renrendai.com	
Yang et al., 2023	Credit rating	SSA-CatBoost	Predicción y clasificación del riesgo de impago	

Referencia	Conjunto de datos	Fundamentos	Aplicación	Interpretabilidad
Zhou et al., 2019	Préstamos P2P	Several ML methods, GBDT, XGBoost, LightGBM	Predicción del riesgo de impago	Modelos de caja negra, no interpretabilidad
Zhu et al., 2019	Préstamos P2P	RF-SMOTE	Predicción del riesgo de impago	
Arora & Kaur, 2020		Several ML methods: RF, SVM, NB, K-NN BS-RF	Predicción del riesgo de impago	
Cho et al., 2019	Préstamos P2P	IEFSVM	Predicción del riesgo de impago	
Moscato et al., 2021	Préstamos P2P	Several ML methods: LR, RF, MLP	Predicción y explicabilidad del riesgo de impago	SHAP
Stevens et al., 2020	Préstamos P2P	SHAP	Predicción y explicabilidad del riesgo de impago	SHAP, LIME, Anchors, BEEF and LORE
Nuestra propuesta	**P2P loans**	**XGBoost**	**Predicción y explicabilidad del riesgo de impago**	**2-Tuple Fuzzy Linguistic Surrogate Trees and Rules**

A partir de una revisión exhaustiva de la literatura relacionada, se concluye que la mayoría de los algoritmos de ML utilizados son considerados como "caja negra". Esto significa que, aunque son capaces de realizar predicciones, su funcionamiento interno no es comprensible para los seres humanos. En el contexto de la predicción del incumplimiento encontramos numerosos trabajos que abordan esta problemática,

pero son escasos los autores que han prestado atención a la predicción del incumplimiento temprano y a la interpretabilidad de los modelos, como se evidencia en la Tabla 2.

Cabe mencionar que existen diferentes enfoques en el problema de XAI como se detalla en la Sección 4.3. En nuestro trabajo, se emplea una técnica difusa denominada modelo de 2-tuplas, que mejora la interpretabilidad lingüística del modelo XAI resultante sin sacrificar la precisión. Es relevante destacar que no encontramos en la literatura ejemplos de la combinación de ambos enfoques.

3.3. Aplicaciones de XAI en predicción en el contexto de Smart Cities: técnicas de interpretabilidad en la predicción de pasajeros.

Nuestra propuesta se centra en la predicción de la demanda de pasajeros en el transporte público de autobuses a través del empleo de técnicas de aprendizaje profundo. En esta sección, procederemos a examinar el estado de la cuestión en la literatura respecto a este tipo de trabajo. Con este propósito, hemos llevado a cabo un estudio bibliométrico sobre esta temática. En concreto, hemos elaborado un mapa bibliométrico, un componente de la cartografía científica que busca establecer una estructura conceptual de un área particular previo a su interpretación. Hemos seguido la metodología descrita en (Cobo et al., 2012) y hemos empleado la herramienta SciMat para generar dicho mapa.

En la primera etapa, se deben recuperar los datos, es decir, los artículos que componen el análisis. Por lo tanto, se descargaron registros bibliográficos de la principal colección de Web of Science utilizando la consulta: TS = ((“forecast*” OR “prediction*”) NEAR/6 (“passenger” o “traveler” o “traveller” o “voyager”) AND (“Machine Learning” o “Deep Learning”)), donde el campo TS es una búsqueda basada en un tema dado (título, resumen y palabras clave), NEAR/n es un operador de proximidad que requiere que ambos términos aparezcan cerca el uno del otro,

hasta n palabras, en el texto. La consulta se ejecutó a finales de marzo de 2022 y resultaron 120 artículos que fueron revisados manualmente.

Posteriormente, en la etapa de preprocesamiento, llevamos a cabo la consolidación de términos duplicados y la corrección de errores tipográficos, entre otras actuaciones.

A continuación, procedimos con las siguientes fases del proceso que incluyeron la extracción de redes, la normalización, el mapeo, el análisis y la visualización.

Una ventaja inherente a la metodología seguida radica en su capacidad para generar un diagrama estratégico de los temas basado en dos medidas fundamentales: la centralidad, que refleja el grado de interacción de una red con otras redes, y la densidad, que indica la fortaleza interna de la red o las palabras clave que caracterizan el tema en el contexto de cualquier proceso de cartografía científica. Presentamos el diagrama estratégico obtenido para nuestro análisis en la Figura 2, que incluye la interpretación de los temas ubicados en cada cuadrante.

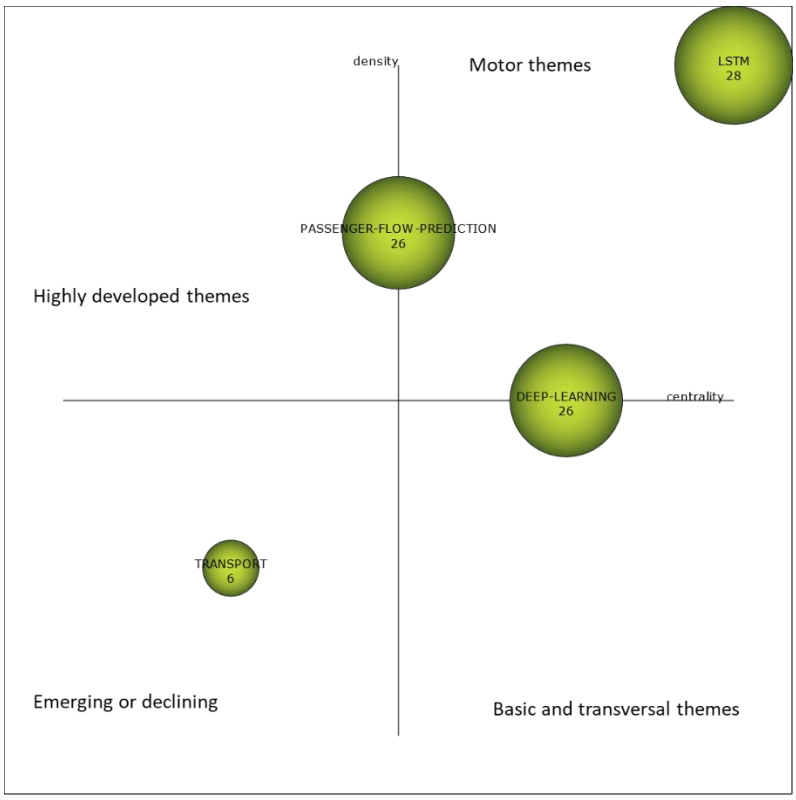
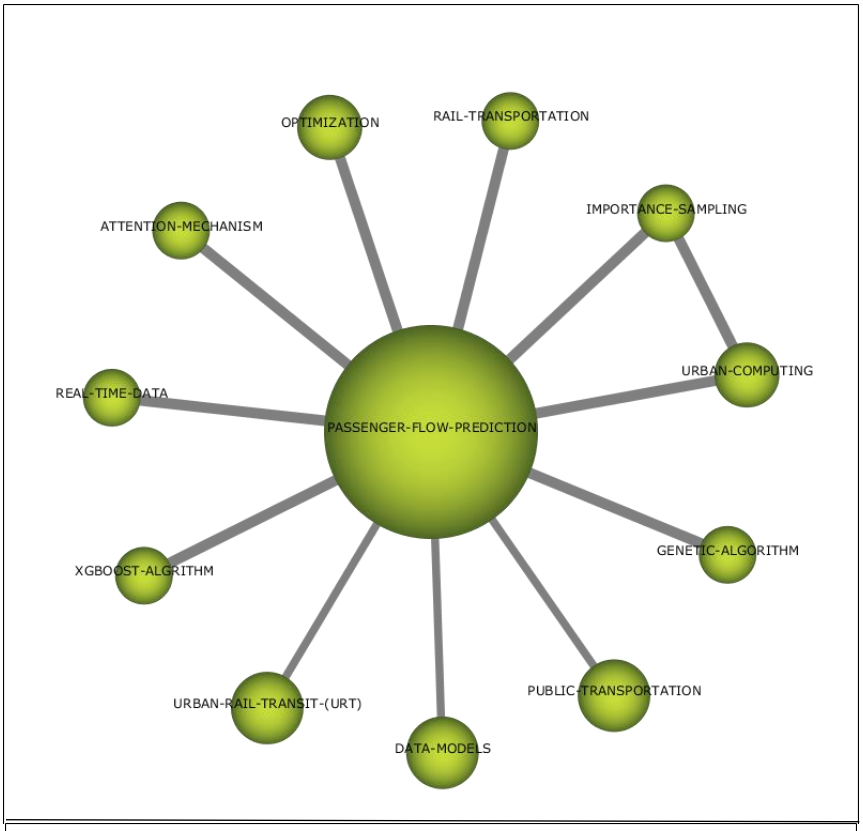
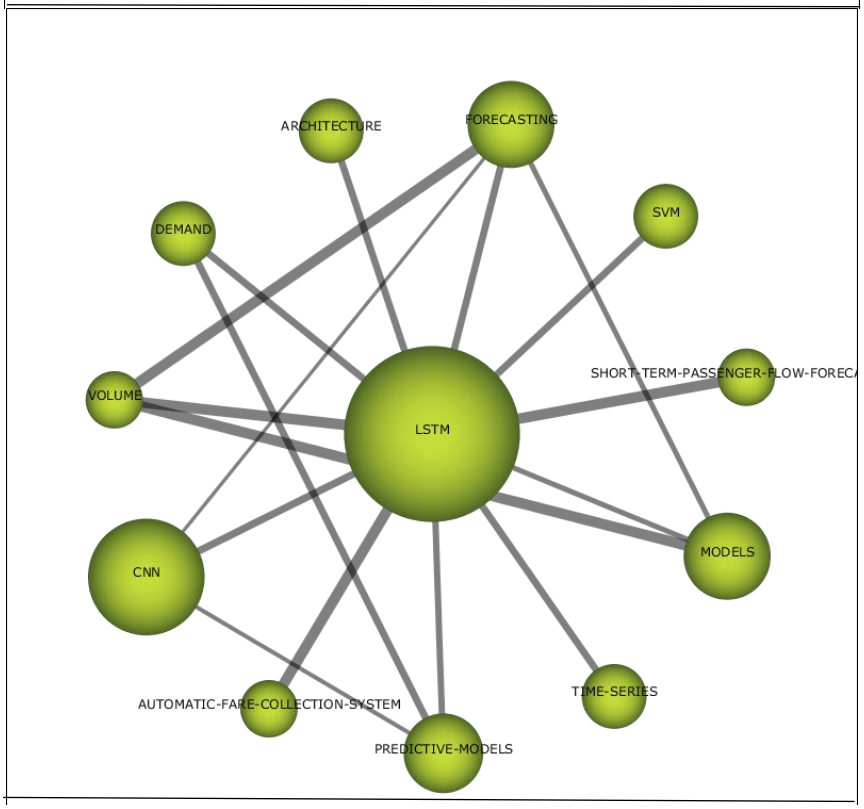


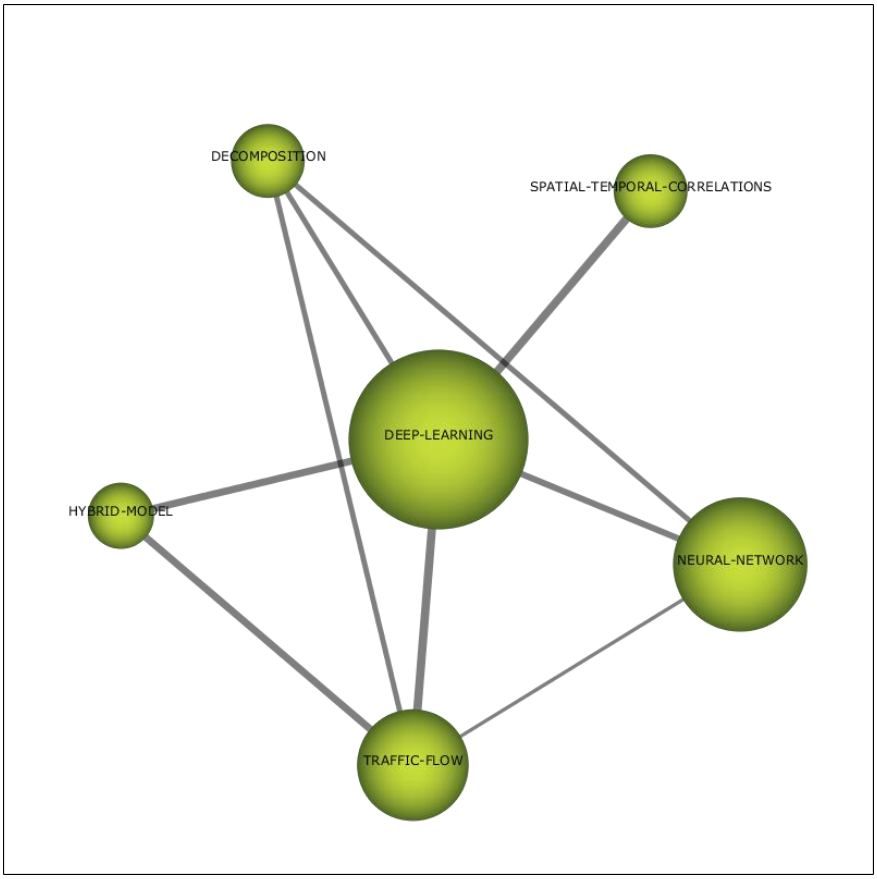
Figura 2. Mapa estratégico de predicción de pasajeros.

Podemos identificar cuatro temas principales, que se muestran en la Figura 3 a través de sus redes temáticas:

- Aprendizaje profundo. Basados en redes neuronales, este tipo de algoritmos se utilizan ampliamente para el problema de predicción del flujo de tráfico que se presenta aquí.
- LSTM. Este tema incluye algoritmos específicos de aprendizaje profundo/máquina que son importantes o fundamentales para el problema en cuestión. Así, los modelos LSTM son el tema central y predominante, seguidos en segundo lugar por las CNN y, finalmente, las SVM que tienen mucha menos relevancia. Este tema está relacionado con la predicción de modelos temporales, ya que en los problemas que estamos tratando, son muy comunes.
- Predicción del flujo de pasajeros. Se trata de un tema heterogéneo que engloba problemas y técnicas de optimización estudiadas, incluidos algoritmos genéticos. También abarca otras técnicas predictivas como XGBoost. Comprende otros términos interesantes que apuntan al uso de datos en tiempo real y aplicaciones predictivas en el transporte urbano y público.
- Transporte. Este es un tema muy pequeño y en declive que incluye el uso de análisis estadístico y aprendizaje automático genérico para el problema planteado.


OPTIMIZATION
RAIL-TRANSPORTATION
ATTENTION-MECHANISM
IMPORTANCE-SAMPLING
URBAN-COMPUTING
REAL-TIME-DATA
PASSENGER-FLOW-PREDICTION
GENETIC-ALGORITHM
XGBOOST-ALGRITHM
PUBLIC-TRANSPORTATION
URBAN-RAIL-TRANSIT-(URT)
DATA-MODELS


ARCHITECTURE
FORECASTING
SVM
DEMAND
SHORT-TERM-PASSENGER-FLOW-FORECA
VOLUME
LSTM
MODELS
CNN
TIME-SERIES
AUTOMATIC-FARE-COLLECTION-SYSTEM
PREDICTIVE-MODELS


DECOMPOSITION
SPATIAL-TEMPORAL-CORRELATIONS
DEEP-LEARNING
HYBRID-MODEL
NEURAL-NETWORK
TRAFFIC-FLOW

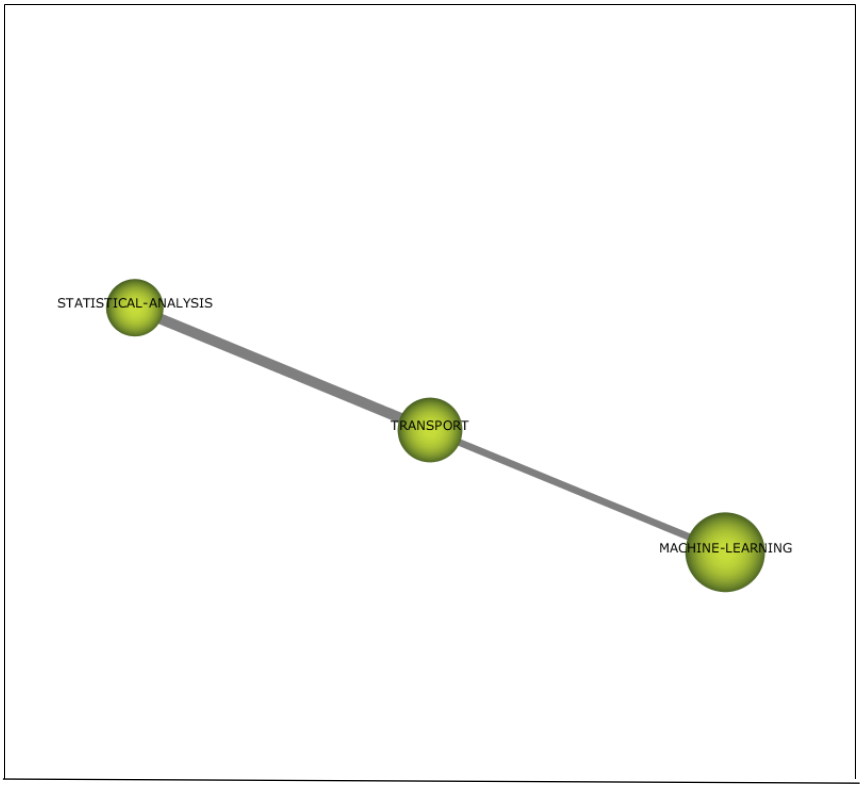


Figura 3. Redes temáticas para cada tema Redes temáticas para cada tema.

Nuestra investigación presenta un modelo ampliamente empleado en la literatura relacionada, conocido como LSTM. Por otro lado, la predicción de pasajeros involucra varios modos de transporte. La Tabla 3 que se presenta a continuación detalla los trabajos específicamente relacionados con la predicción del flujo de pasajeros en autobuses. Nuestra propuesta también se encuentra incluida en esta tabla.

Tabla 3. Trabajos relacionados con la predicción del flujo de pasajeros en autobuses.

Referencia	**Fundamentos**	**Conjunto de datos**	**Aplicación**	**Interpretación del modelo**
Bai et al., 2017	Fusión Profunda de Múltiples Patrones (MPDF) basada en Red de Creencias Profundas	Datos de flujo de pasajeros	Predicción del flujo de pasajeros de autobús en Guangzhou (China)	Contribuciones de variables a la predicción del flujo de pasajeros
Chen et al., 2022	Secuencia de Gráficos Espaciales-Temporales con Red de Atención (STGSAN)	Datos de flujo de pasajeros	Predicción del flujo de pasajeros de autobús en Urumqi (China	
Cristóbal et al., 2019	medoids clustering	Datos de flujo de pasajeros	Predicción del tiempo de viaje en Gran Canaria	
Gummadi & Edara, 2019	Red Neuronal Artificial	Datos de flujo de pasajeros	Predicción del flujo de pasajeros de autobús	
Han et al., 2019	LSTM	Datos de flujo de pasajeros	Predicción del flujo de pasajeros de autobús en Qingdao (China)	
Jiao et al., 2021	Modelo basado en STL y LSTM	Datos de pasajeros, datos meteorológicos, datos de vacaciones, etc.	Predicción del flujo de pasajeros de autobús en Beijing (China)	
Jin et al., 2019	Modelo Shepard	Datos de pasajeros y datos de múltiples fuentes	Predicción del flujo de pasajeros de autobús	

Referencia	Fundamentos	Conjunto de datos	Aplicación	Interpretación del modelo
Liu & Chen, 2017	Autoencoders Apilados (SAE) y Red Neuronal Profunda (DNN)	Datos de flujo de pasajeros recopilados por el conteo automático de pasajeros y datos de vacaciones	Predicción del flujo de pasajeros de autobús rápido en Xiamen (China)	Contribuciones de variables a la predicción del flujo de pasajeros
Liu et al., 2019	Revisión del mundo inteligente	Revisión de Ciudades Inteligentes	Revisión	
Liu et al., 2020	Red Neuronal Profunda y Red Neuronal Convolucional Modular	Datos de flujo de pasajeros	Predicción del flujo de pasajeros de autobús en Nanjing (China)	
Liu et al., 2020	SVM	Flujo de pasajeros de datos de estación de autobuses y datos de tarjetas de viaje inteligentes	Predicción del flujo de pasajeros de autobús en Shenzhen Tong (China	
Lv et al., 2022	XGBoost	Datos de pasajeros y puntos de interés	Predicción del flujo de pasajeros de autobús en Beijing (China)	
Manibardo et al., 2021	DL	Revisión del uso de DL en Ciudades Inteligentes	Revisión	

Referencia	Fundamentos	Conjunto de datos	Aplicación	Interpretación del modelo
Nagaraj et al., 2022	LSTM	Datos de flujo de pasajeros	Predicción del flujo de pasajeros de autobús en el transporte de Bus Rapid Transit (KSRTCBRT) en Karnataka, India	Contribuciones de variables a la predicción del flujo de pasajeros
Ouyang et al., 2020	Extracción de características basada en modelo XGBoost y predicción basada en LSTM	Datos históricos y en tiempo real de pasajeros y puntos de interés	Predicción del flujo de pasajeros de autobús en Beijing (China)	
Tan et al., 2019	Bosque Aleatorio	Datos de tarjetas de viaje inteligentes	Predicción del flujo de pasajeros de autobús en Shenzhen (China)	
Wu et al., 2020	Árbol de Decisión de Reforzamiento Apilado (SS-GBDT)	Datos de flujo de pasajeros y información de múltiples fuentes	Predicción del flujo de pasajeros de autobús en Guangzhou (China)	
Zhai et al., 2020	Modelo híbrido basado en modelo de series temporales (ARIMA), Redes de Creencias Profundas (DBNs), y Máquina de Aprendizaje Extremo Incremental Mejorada (Im-ELM)	Datos de flujo de pasajeros recopilados por el conteo automático de pasajeros	Predicción del flujo de pasajeros de autobús en Dalian (China)	

Referencia	Fundamentos	Conjunto de datos	Aplicación	Interpretación del modelo
Zhang et al., 2016	XGBoost	Datos de flujo de pasajeros, datos de puntos de interés y datos meteorológicos	Elecciones de abordaje de pasajeros de autobús y flujo de pasajeros del transporte público en Guangzhou (China)	Contribuciones de variables a la predicción del flujo de pasajeros
Zou et al., 2022	XGBoost	Datos de pasajeros de tarjetas de viaje	Predicción del flujo de pasajeros de autobús en Guangzhou (China)	
Luo et al., 2020	Aprendizaje Profundo Multitarea (MDL) que incluye LSTM, CNN, etc.	Datos de flujo espaciotemporal y heterogéneo	Predicción del flujo de pasajeros de autobús en Jinan (China)	
Toqué et al., 2017	LSTM, Bosque Aleatorio, Red Neuronal	Registros de emisión de boletos de pasajeros	Predicción del flujo de pasajeros de autobús en París (Francia)	
Wang et al., 2022	LSTM y CNN	Datos de flujo espaciotemporal	Predicción del flujo de pasajeros de autobús	

Referencia	Fundamentos	Conjunto de datos	Aplicación	Interpretación del modelo
Zhai et al., 2018	Varios métodos de aprendizaje automático: SVM, Redes Neuronales, ARIMA, etc.	Datos de flujo de pasajeros de contadores automáticos de pasajeros, sistemas automáticos de recaudación de tarifas, etc.	Predicción del flujo de pasajeros de autobús en Shanghai (China)	Contribuciones de variables a la predicción del flujo de pasajeros
Nuestra propuesta	**LSTM**	**Datos de flujo de pasajeros y datos de vacaciones**	**Predicción del flujo de pasajeros de autobús en Madrid (España)**	**Modelo Global Surrogado y Modelo Lingüístico Difuso de 2-tuplas".**

A partir de una revisión exhaustiva de la literatura relacionada, se desprende que la mayoría de los algoritmos de ML utilizados son considerados "caja negra". Esto significa que son capaces de realizar predicciones, pero su funcionamiento interno no es interpretable por seres humanos. En el contexto de la predicción del flujo de pasajeros en autobuses, encontramos numerosos trabajos que abordan esta problemática; sin embargo, son escasos los autores que han prestado atención a la interpretabilidad de los modelos, como se evidencia en la Tabla 3. Los pocos trabajos que se han enfocado en este aspecto se limitan a determinar la importancia de las variables involucradas en el modelo (Zou et al., 2022).

Por lo tanto, resulta apropiado recurrir a técnicas de XAI para interpretar los resultados de estos algoritmos. Es importante destacar que las técnicas de XAI aún son relativamente novedosas y su ámbito de aplicación principal sigue siendo bastante genérico, no estando tan enfocado en problemas específicos (Corchado et al., 2021, Kostopoulos et al., 2021). Si bien existen trabajos orientados a Ciudades Inteligentes (Barredo-Arrieta et al., 2019; Herrera et al., 2000), son escasos los que se

centran específicamente en el transporte público. En este sentido, se pueden hallar investigaciones relacionadas con la predicción y comprensión del flujo de tráfico (Marín et al., 2021), así como trabajos que buscan conseguir que las recomendaciones respecto a la flota de transporte sean interpretables (Bueno et al., 2022).

Hasta dónde llega nuestro conocimiento, no se ha realizado un trabajo específico sobre la previsibilidad e interpretabilidad en relación a los usuarios del autobús público, como el que presentamos en este estudio.

Cabe mencionar que existen diferentes enfoques en el problema de XAI (Elshawi et al., 2019), como se detalla en la Sección 4.3. En nuestro trabajo, se emplea una técnica difusa denominada modelo de 2-tuplas, que mejora la interpretabilidad lingüística del modelo XAI resultante sin sacrificar la precisión. Es relevante destacar que no encontramos en la literatura ejemplos que combinen ambos enfoques.

4. Marco teórico de la investigación

En este capítulo se exponen los conceptos teóricos fundamentales que sirven de base para la investigación presentada en esta memoria. Específicamente, se introduce el modelado mediante técnicas de ML supervisadas, las técnicas de XAI mediante modelos surrogados y el modelado lingüístico difuso 2-tuplas.

El modelo surrogado se utiliza para facilitar la interpretación, desde una perspectiva empresarial, del modelo de ML construido. Por otro lado, incorporaremos el modelo lingüístico difuso 2-tuplas a las técnicas XAI y se aplicará en los distintos casos de uso (sección 5) con el propósito de reducir la incertidumbre inherente a los datos utilizados, sin sacrificar la precisión. Esto supone que sin perder precisión podamos tener interpretabilidad lingüística del modelo de XAI. Estos modelos son especialmente relevantes en situaciones en las cuales los criterios de evaluación se relacionan con probabilidades, lo que es común en problemas de toma de decisiones.

Las investigaciones abordadas en este trabajo se centran en la necesidad de priorizar y seleccionar alternativas, lo que motiva la presentación de un modelado lingüístico difuso 2-tuplas como una herramienta para la toma de decisiones en contextos caracterizados por la incertidumbre y la necesidad de comprensión intuitiva.

4.1. Técnicas de Machine Learning supervisadas

Los modelos de ML supervisado utilizan datos etiquetados para el entrenamiento. Los podemos distinguir en función de si su proceso de toma de decisiones es entendible y explicable por los humanos por lo que son modelos de caja blanca o, modelos de caja negra, si no son interpretables por el ser humano.

Otra manera de dividir los modelos es en función de su variable objetiva. Cuando la variable objetivo que se quiere predecir es una etiqueta o una categoría discreta, son modelos de clasificación, como el caso de uso de Fintech. Si la variable objetivo que se quiere predecir es numérica o continua serán modelos de regresión como el caso de uso de ciudad inteligente.

4.1.1. Clasificación

Las técnicas de ML supervisado utilizadas para la clasificación se basan en la predicción de la categoría o etiqueta a la que pertenece un conjunto de datos, tomando como referencia un conjunto de ejemplos previamente etiquetados. Las técnicas más prominentes en el campo del ML supervisado aplicadas a la clasificación son las siguientes:

- LR: es un modelo de ML (Schober & Vetter, 2021) donde el objetivo es predecir la probabilidad de que una observación pertenezca a una de las categorías o clases. El modelo de LR se emplea para predecir probabilidades y asignar observaciones a categorías discretas. A continuación, mostramos cómo funciona:

En un problema de LR, se tiene una variable dependiente binaria o categórica, comúnmente etiquetada como 0 o 1. También se tiene un conjunto de variables independientes (predictoras) que se utilizan para predecir la variable dependiente.

RL es un modelo de caja blanca que utiliza una función logit o log-odds para modelar la probabilidad de que una observación pertenezca a la categoría 1. Esta función logit se calcula tomando el logaritmo natural (ln) de la razón de las probabilidades

$\left(\frac{probabilidad\ de\ éxito}{probabilidad\ de\ fracaso}\right)$ y se expresa como una combinación lineal de las variables independientes. La ecuación tiene la forma, y la representación gráfica de la Figura 4.

$$Ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_n X_n \quad (1)$$

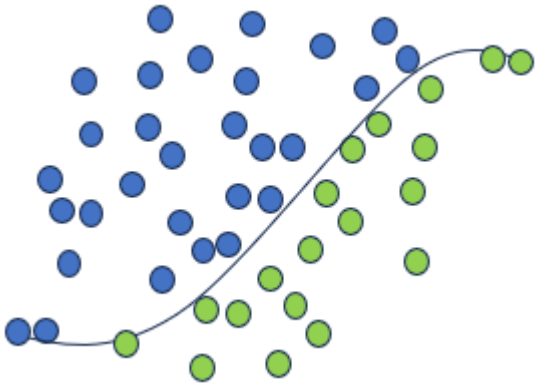


Figura 4. Regresión logística.

Donde $P(Y = 1)$ es la probabilidad de que la observación pertenezca a la categoría 1, X_i son las variables independientes y, β_i son los coeficientes del modelo.

La función logit se transforma mediante una función sigmoide (logística) que limita la salida entre 0 y 1. Los coeficientes β_i se estiman a partir de los datos de entrenamiento utilizando técnicas de optimización, como la máxima verosimilitud.

Una vez que el modelo ha sido entrenado, se puede usar para realizar predicciones. Dado un conjunto de valores para las variables independientes, el modelo calcula la probabilidad de que la observación pertenezca a la categoría 1. Si esta probabilidad supera un umbral (generalmente 0,5), se clasifica como categoría 1; de lo contrario, se clasifica como categoría 0.

- Árboles de Decisión (en inglés Decision Trees - DTs) de clasificación: es una herramienta de ML (Pérez-Rave & González Echavarría 2018), donde el objetivo es asignar observaciones a categorías predefinidas, se trata de un modelo de caja blanca.

La estructura de este modelo se asemeja a un árbol invertido, con un nodo raíz en la parte superior y nodos hoja en la parte inferior. La operativa de un modelo de DT de clasificación se detalla de la siguiente manera:

El proceso se inicia en un nodo raíz que representa la totalidad de la observación tomada del conjunto de entrenamiento. En este nodo, se selecciona una variable (característica) que será utilizada para la primera división del conjunto de datos en subconjuntos más pequeños. Cada subconjunto resultante se divide nuevamente en función de otra variable, y este proceso se repite hasta que se cumple un criterio de parada. Los criterios habituales de parada incluyen alcanzar una profundidad máxima del árbol, contar con un número mínimo de observaciones en un nodo o lograr una pureza mínima. Las divisiones se efectúan con el objetivo de maximizar la pureza o separación entre las categorías. Los nodos que no son hojas, representan divisiones y contienen reglas de decisión basadas en las características. Los nodos hoja representan las categorías finales a las que se asigna una observación. Para determinar la categoría de una observación, ésta sigue las reglas de decisión desde el nodo raíz hasta una hoja correspondiente. Una vez construido el árbol, se utiliza para predecir la categoría a la que pertenece una nueva observación. La observación sigue un camino en el árbol continuando las reglas de decisión en cada nodo hasta llegar a una hoja que dictamina su categoría.

Los DTs son altamente interpretables, ya que cada nodo representa una decisión basada en una característica específica. Esto facilita la comprensión de cómo se lleva a cabo una decisión de clasificación y qué características son las más influyentes en el proceso, pero tienden a sufrir de sobreajuste, por lo que se pueden podar o regularizar para mejorar su capacidad de generalización en datos no observados. En la Figura 5 se muestra el funcionamiento de este método de clasificación.

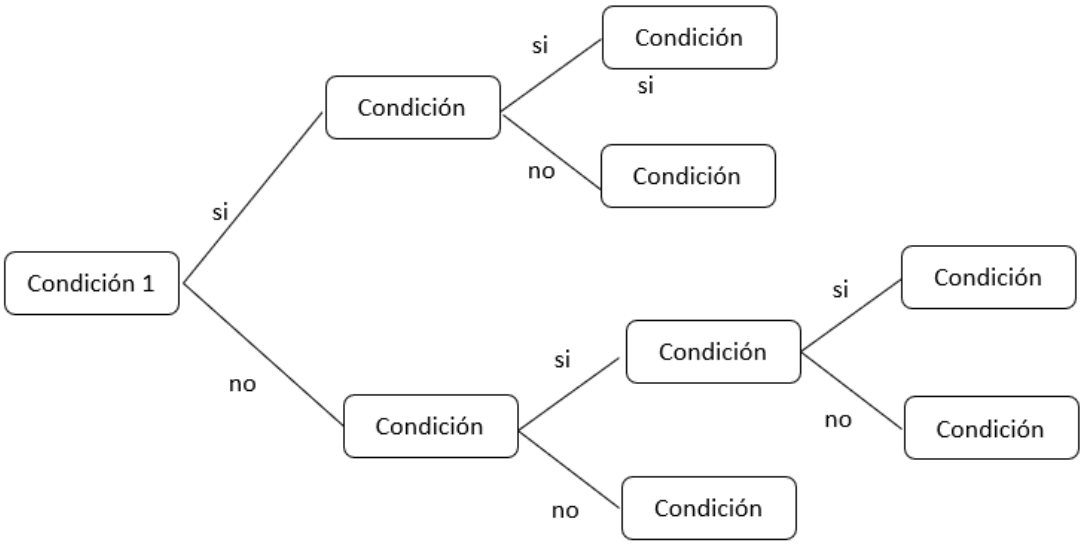


Figura 5. Árboles de decisión.

- El modelo de RF: es una técnica de ML (Liu et al., 2021) utilizada para problemas de clasificación y regresión. Su enfoque se basa en la construcción de múltiples DTs y la combinación de sus resultados para mejorar la precisión del modelo y mitigar el riesgo de sobreajuste, es un modelo de caja blanca. A continuación, se presenta una explicación detallada del funcionamiento de un modelo de RF para clasificación:

El proceso comienza con un conjunto de datos de entrenamiento que contiene observaciones con etiquetas. Cada observación está caracterizada por un conjunto de atributos que se utilizan para predecir su categoría o clase. Para la construcción del árbol, en lugar de crear un solo DT, el modelo RF genera una colección de DTs donde cada árbol se entrena utilizando una porción aleatoria del conjunto de datos de entrenamiento y una selección aleatoria de características. Esta aleatoriedad introduce diversidad y reduce la posibilidad de sobreajuste. Cuando se realiza una predicción para una nueva observación, cada árbol en el bosque emite su propia predicción. En el contexto de clasificación, cada árbol "vota" por una clase y la predicción final se determina mediante una votación mayoritaria, es decir, se selecciona, como la predicción final la clase que recibe la mayoría de votos.

El proceso de selección aleatoria de observaciones del conjunto de datos de entrenamiento se conoce como agregación de Bootstrap (en inglés Bootstrap Aggregating o Bagging), e implica la creación de múltiples subconjuntos de datos de entrenamiento y, en cada subconjunto se construye un DT. Posteriormente, se combinan las predicciones de todos los árboles para obtener una decisión final. Además de muestrear observaciones, se realiza una selección aleatoria de un subconjunto de características en cada árbol. Esto contribuye a garantizar que los árboles en el bosque sean diversos y no estén fuertemente correlacionados. La combinación de múltiples árboles a través del Bootstrap Aggregating y la selección aleatoria de características reduce la varianza del modelo, lo que disminuye el riesgo de sobreajuste. En consecuencia, el modelo se vuelve más resistente y generaliza mejor a datos no observados. Además, proporciona información sobre la importancia de las características. Esto se logra al evaluar qué características se utilizan con mayor frecuencia para dividir nodos en los árboles, lo que permite identificar que atributos son más influyentes en la clasificación.

Por todo ello, RF representa una técnica poderosa de clasificación que aprovecha la construcción de múltiples DTs con variabilidad controlada para mejorar la precisión y reducir el riesgo de sobreajuste. En la Figura 6 se muestra el funcionamiento de este método de clasificación.

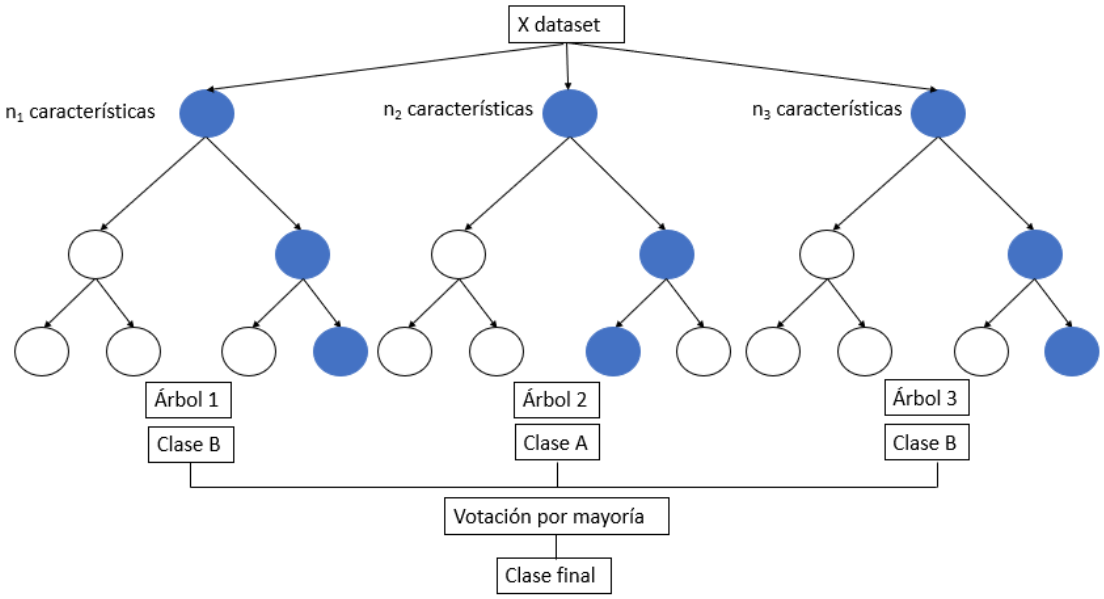


Figura 6. Árboles aleatorios.

- SVM: Es considerado un modelo de caja negra. Las SVM (Cortes & Vapnik, 1995) buscan encontrar un hiperplano óptimo para separar las clases en un espacio de características. Este hiperplano se selecciona de tal manera que maximiza la distancia entre los puntos más cercanos de ambas clases, los cuales se conocen como "vectores de soporte". La principal ventaja de las SVM radica en su capacidad para abordar problemas en espacios de características de alta dimensionalidad y para resolver situaciones de separación no lineal a través del uso de "kernels" que transforman los datos en espacios de características no lineales.

En la Figura 7 se muestra el funcionamiento de este método de clasificación

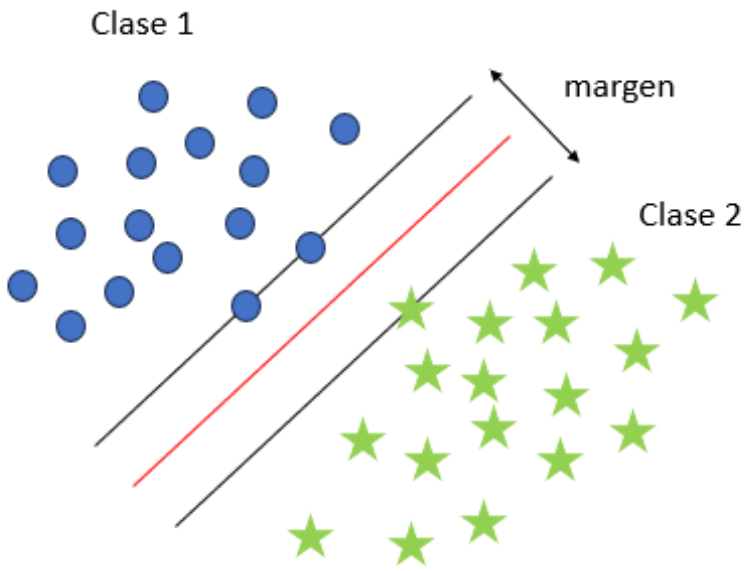


Figura 7. SVM.

Las ventajas de este modelo son:

- Eficacia en espacios de características de alta dimensionalidad.

- Capacidad para abordar problemas de separación no lineal mediante la aplicación de kernels.

- Robustez frente a datos ruidosos y la capacidad de manejar multicolinealidad.

- Flexibilidad en la selección de funciones de kernel, lo que permite adaptar el modelo a diferentes tipos de datos.

- Buena capacidad de generalización a partir de un número limitado de vectores de soporte.

Sin embargo, el SVM también presenta ciertos desafíos, como la necesidad de ajustar cuidadosamente los hiperparámetros y los requerimientos computacionales en conjuntos de datos extensos.

- K-NN: El método k-NN (Mora – Flórez et al., 2008), que fue propuesto por Fix y Hodges en 1951, es un enfoque de clasificación supervisada que se basa en un conjunto de datos de entrenamiento y prototipos para estimar la función de densidad $F\left(\frac{x}{Cj}\right)$ de las variables predictoras x para cada clase C_j.

Este método se caracteriza por ser no paramétrico, lo que significa que no realiza suposiciones sobre la distribución de las variables predictoras. Por el contrario, utiliza la información proporcionada por el conjunto de prototipos o ejemplos para estimar el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento x pertenezca a la clase C_j. Su fundamento radica en la premisa de que las observaciones que comparten similitudes en cuanto a características tienden a pertenecer a la misma categoría o a mostrar valores de salida semejantes. En esta exposición, se presenta una explicación detallada del modelo K-NN:

El proceso se inicia con un conjunto de datos de entrenamiento que comprende observaciones etiquetadas. Cada observación está definida por características que se emplean para anticipar su categoría (en el contexto de la clasificación) o un valor numérico (en la regresión). La denominación "K-Vecinos más Próximos" alude al hecho de que el algoritmo al efectuar una predicción busca los K puntos más próximos a una nueva observación. K constituye un valor entero positivo que se selecciona previamente. Para determinar qué observaciones se encuentran más cercanas a una nueva observación, se recurre a una medida de similitud, generalmente la distancia euclidiana. Esta medida cuantifica la distancia entre las características de las observaciones en un espacio multidimensional. En el caso de clasificar una nueva observación, se identifican los K-NN en el conjunto de datos de entrenamiento y se evalúa cuántos de ellos pertenecen a cada clase. La nueva observación se clasifica en la categoría que prevalezca entre sus K-NN.

La elección del valor K representa un aspecto crucial en K-NN, un valor K más reducido genera una mayor sensibilidad al ruido en los datos, mientras que un valor K más amplio puede suavizar la predicción, aunque también puede ocultar patrones significativos.

K-NN es un modelo considerado caja negra y puede acarrear costos computacionales considerables en conjuntos de datos extensos, ya que implica calcular la distancia entre la nueva observación y todas las observaciones de entrenamiento.

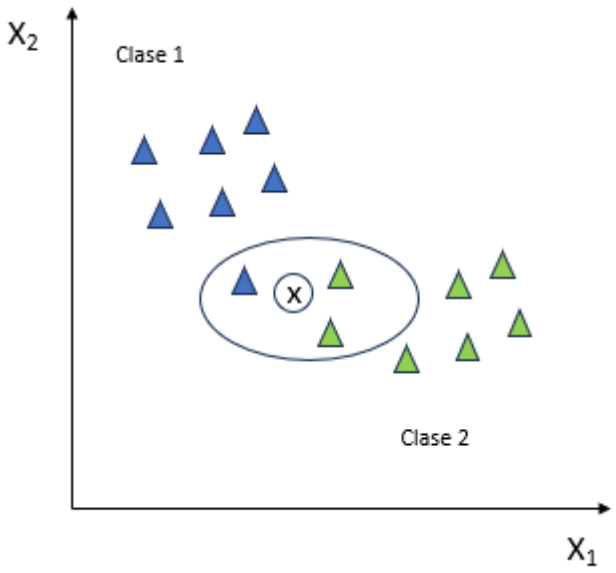


Figura 8. K-NN

En la Figura 8 se muestra el funcionamiento de este método de clasificación. En esta representación gráfica, se muestran un total de 14 muestras, divididas en dos clases distintas: la Clase 1, compuesta por 7 triángulos de color azul y, la Clase 2, formada por 7 triángulos de color verde. Para este ejemplo, se ha seleccionado un valor de (k=3), lo que significa que se consideran tres vecinos más próximos. De los tres vecinos más próximos a la muestra x, identificada en la figura como una cruz, uno de ellos pertenece a la Clase 1, mientras que los otros dos se encuentran en la Clase 2. En consecuencia, la regla 3-nn asignará la muestra a la Clase 2.

- El modelo NB: es una técnica de ML (Wickramasinghe & Kalutarage, 2021) de amplio uso en tareas de clasificación considerada caja negra. Se fundamenta en el teorema de Bayes (Berrar, 2019) y se caracteriza por sus suposiciones de independencia condicional entre las características, lo que le otorga el atributo "ingenuo". Su capacidad para manejar datos de alta dimensión y su eficiencia en el entrenamiento lo convierten en una elección destacada en muchas tareas de

clasificación. A continuación, se presenta una explicación detallada del modelo NB en el contexto de la clasificación:

El proceso se inicia con un conjunto de datos de entrenamiento que consta de observaciones etiquetadas. Cada observación se compone de características que se emplean para predecir su categoría o clase.

El modelo NB se basa en el teorema de Bayes, que establece que la probabilidad de una hipótesis (en este contexto, una clase) dado un conjunto de evidencia (las características de una observación) es proporcional al producto de la probabilidad de la evidencia dada la hipótesis y la probabilidad previa de la hipótesis. Esta relación se expresa mediante la siguiente ecuación:

$$P(\frac{clase}{evidencia}) \propto P(\frac{evidencia}{clase}) * P(Clase)$$

Donde:

- $P(\frac{clase}{evidencia})$ representa la probabilidad de que la observación pertenezca a una clase dada la evidencia.

- $P(\frac{evidencia}{clase})$ denota la probabilidad de que la evidencia se observe dado un valor de clase.

- $P(Clase)$ se refiere a la probabilidad previa de una clase.

La característica "ingenua" en NB se refiere a la suposición de que las características son condicionalmente independientes, dado un valor de clase. Esto implica que las características de una observación no están relacionadas entre sí, ya que conocemos la clase. A pesar de esta simplificación, NB demuestra ser eficaz en numerosos escenarios de la vida real.

Para aplicar NB, es necesario estimar las probabilidades $P(\frac{evidencia}{clase})$ y $P(Clase)$ utilizando los datos de entrenamiento. En el caso de la clasificación binaria, estas probabilidades se calculan como sigue:

- $P(Evidencia|Clase)$ se estima mediante la frecuencia relativa de cada valor de característica dado un valor de clase en el conjunto de entrenamiento.

- $P(Clase)$ se estima mediante la frecuencia relativa de cada clase en el conjunto de entrenamiento.

Una vez estimadas las probabilidades, se utiliza el teorema de Bayes para calcular la probabilidad posterior de cada clase para una nueva observación. La clase con la probabilidad más alta se asigna como la predicción final.

El modelo NB tiene distintas variantes, como el Naïve Bayes Multinomial (adecuado para la clasificación de texto) y el Naïve Bayes Gaussiano (idóneo para datos numéricos continuos).

- ANN: Las redes neuronales profundas (Moayedi, et al., 2020), reconocidas como DL y consideradas cajas negras, emplean capas de neuronas para aprender representaciones sofisticadas de los datos y se aplican en diversas tareas de clasificación. Las ANN representan un modelo matemático inspirado en la estructura y el funcionamiento de las neuronas biológicas y su organización en el cerebro. Estas redes buscan aprender mediante iteraciones con el objetivo de optimizar su propia estructura para maximizar la capacidad de predicción.

En un modelo de red neuronal, los nodos desempeñan funciones de entrada, salida o procesamiento intermedio, y se conectan entre sí mediante conexiones ponderadas. Siguiendo un paradigma de aprendizaje, el modelo comienza con el primer caso y realiza una evaluación inicial basada en las ponderaciones. Posteriormente, se calcula el error de predicción y se ajustan las ponderaciones para mejorar la precisión. Este proceso se repite para cada caso, en lo que se denomina fase de entrenamiento o evaluación. Una vez que el modelo está calibrado, se evalúan los resultados globales utilizando un conjunto de prueba.

Las redes neuronales se hicieron operativas por primera vez a finales de la década de 1950, cuando Frank Rosenblatt creó el perceptrón (Rosenblatt, 1958), un algoritmo de

reconocimiento de patrones basado en una red de dos capas de aprendizaje de computadora que utiliza operaciones simples de suma y resta. A pesar de un período de estancamiento en la investigación a finales de la década de 1960, el desarrollo teórico de las redes neuronales experimentó un renacimiento a partir de la década de 1980. Existen varios modelos de redes neuronales, siendo el algoritmo de retropropagación, creado por Paul Werbos en 1990 (Werbos, 1990), uno de los más ampliamente utilizados. En la actualidad, se han desarrollado más de 40 paradigmas de ANN. No obstante, se ha observado que solo las siguientes arquitecturas abarcan aproximadamente el 90% de las aplicaciones prácticas de redes neuronales. Entre ellas, el perceptrón multicapa destaca como el modelo más ampliamente utilizado, siendo empleado en alrededor del 70% de las aplicaciones.

1. El modelo del perceptrón multicapa (en inglés MultiLayer Perceptron model - MLP), ver por ejemplo (Taud et al., 2018).

2. Los mapas autoorganizados de Kohonen (en inglés Kohonen Self-Organising Maps - SOFM), un caso es (Kohonen, 1997).

3. El vector de cuantificación (en inglés Learning Vector Quantization - LVQ), muestra de ello es (Kohonen, 1995).

4. Las redes de base radial (en inglés Radial Basis Function - RBF), un ejemplo es (Ghosh & Nag, 2001).

- GB: (Bentéjac et al., 2021). Se basa en la construcción secuencial de múltiples modelos, donde cada nuevo modelo se crea con el propósito de corregir las deficiencias de los modelos anteriores. Se utiliza para abordar problemas de regresión y clasificación y es considerado caja negra. El proceso se puede desglosar de la siguiente manera:

La metodología comienza con la generación de un modelo base, generalmente un árbol de decisión, que se emplea para efectuar predicciones iniciales. Aunque este modelo inicial puede carecer de alta precisión, cumple la función de servir como

punto de partida. Posteriormente, se procede al cálculo de los residuos, que representan el error del modelo actual, y reflejan las discrepancias entre las predicciones del modelo inicial y los valores reales. A continuación, se genera un segundo modelo con el propósito de rectificar las deficiencias del primero. Este segundo modelo se adapta a los residuos calculados previamente, con el objetivo de disminuir las disparidades entre las predicciones y los datos reales. Las predicciones provenientes del primer y del segundo modelo se combinan para mejorar la precisión global, calculándose nuevamente los residuos entre las predicciones combinadas de los dos modelos y los valores reales.

El proceso de construcción y ajuste de modelos se repite en múltiples ocasiones, lo que implica la creación de nuevos modelos en cada iteración, cada uno de los cuales se enfoca en reducir las disparidades restantes entre las predicciones y los valores reales. Las predicciones finales se obtienen sumando las predicciones de todos los modelos construidos durante el proceso de GB. Esto da como resultado predicciones más precisas en comparación con cualquier modelo individual. Para prevenir el sobreajuste, se aplican técnicas de regularización, como limitar la profundidad de los árboles o ajustar la tasa de aprendizaje.

El modelo GB es conocido por su capacidad para producir modelos altamente precisos y es ampliamente utilizado en diversas aplicaciones de ML. Algunos de los algoritmos populares basados en GB incluyen GB Machines (GBM), LightGBM, CatBoost y XGBoost, el cual desarrollaremos a continuación.

En nuestro caso de uso de clasificación (Fintech) hemos utilizado el XGBoost, que es un algoritmo de ML supervisado (Chen & Guestrin, 2016) que consiste en un ensamblaje de árboles secuenciales como se muestra en la Figura 9, donde el árbol siguiente aprende el resultado de los árboles anteriores y corrige el error producido por ellos, hasta que el error ya no puede ser corregido. Esto es lo que se conoce como descenso

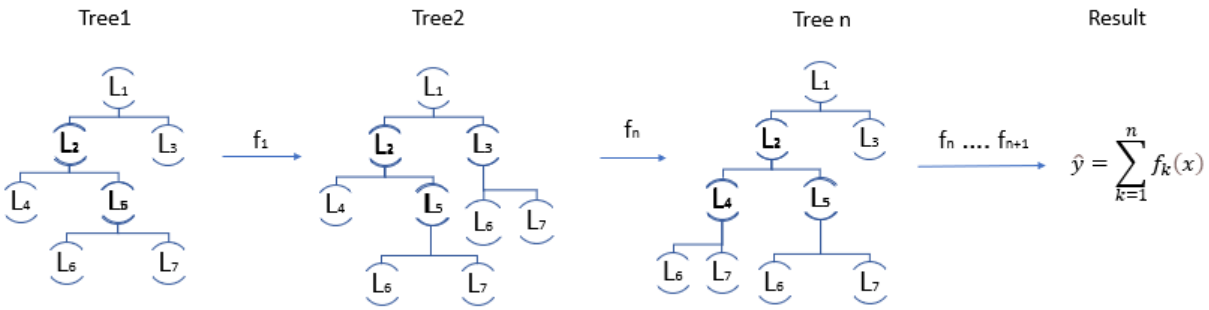


Figura 9. eXtreme Gradient Boosting classifier (XGBoost).

El XGBoost se vale de un proceso de boosting para mejorar el desempeño del modelo. Incorpora términos de regularización L1 y L2 con el propósito de mitigar el sobreajuste y potenciar la generalización del modelo. Emplea el gradiente descendente para minimizar una función de pérdida, lo que otorga una alta eficiencia en la convergencia hacia una solución óptima.

El XGBoost incorpora mecanismos para tratar los valores ausentes en los datos de entrada y es capaz de capturar interacciones no lineales entre las características mediante la creación de funciones de interacción. Además, se caracteriza por su eficiencia en la utilización de recursos y la capacidad de ejecutarse en paralelo, lo que lo hace escalable a conjuntos de datos extensos. Proporciona una estimación de la importancia de las características, lo que facilita la selección de variables y su rendimiento superior y su capacidad para manejar una diversidad de tipos de datos lo convierten en una herramienta poderosa en el campo del ML supervisado.

XGBoost es una función de funciones, una suma de árboles aditivos, donde el primer término es la función pérdida y, el segundo es el término de regularización, minimizando así la pérdida de lo que hemos aprendido previamente y añadiendo un nuevo árbol. No se puede optimizar a través de los métodos de optimización tradicionales, y para transformar la función inicial en una función en el espacio euclídeo (Chen & Guestrin, 2016).

Pérdida Regularización

$$L^{(t)} = \sum_{i=1}^{n} (y_i, \widehat{y_i}^{(t-1)} + f_t(x_i)) + \Omega\,(f_t) \quad (2)$$

También puede ser visto como $f(x+\Delta x)$ donde $x = \widehat{y_i}^{(t-1)}$ y Δx es el nuevo aprendiz que tenemos que añadir en el paso t.

Tratamos de medir lo bueno que es el árbol, y de optimizar un nivel del árbol a la vez, ya que no podemos optimizar directamente el árbol. Obtenemos así un resultado final:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_r^2}{H_R+\lambda} - \frac{(H_L+H_R)^2}{H_L+H_R+\lambda}\right] - \gamma \quad (3)$$

4.1.2. Regresión

El enfoque del ML supervisado de regresión es un proceso esencial en la ciencia de datos y la IA que se utiliza para abordar la tarea de predecir valores numéricos a partir de datos previamente etiquetados. Es una herramienta fundamental en una amplia variedad de aplicaciones, incluyendo pronósticos financieros, análisis de series temporales, así como otros contextos en los que la predicción precisa de valores numéricos es esencial. Para la evaluación del modelo se emplean métricas de evaluación, como el error cuadrático medio (MSE) o el coeficiente de determinación (R^2), midiendo así la calidad de las predicciones del modelo en comparación con los valores reales. Las técnicas más destacadas aplicadas a la regresión son las siguientes:

- El modelo de regresión lineal (en inglés Lineal Regression – LR) considerado caja blanca, es uno de los modelos (Krämer & Sonnberger, 2012) más simples y ampliamente utilizados en la regresión, y que busca establecer una relación lineal entre la variable de entrada y la variable de salida.

El modelo de RL se formula mediante la ecuación:

$$Y = \beta_0 + \beta_1 + \beta_2 + \ldots\ldots + \beta_n + \varepsilon \quad (4)$$

Donde:

- Y representa la variable dependiente que deseamos pronosticar.

- β_0 es el término de intercepción, reflejando el valor de Y cuando todas las variables independientes son nulas.

- $\beta_1 + \beta_2 + \ldots\ldots + \beta_n$ son los coeficientes asociados a las variables independientes (X_1, X_2, ……, $X_{n)}$ indicando la contribución de cada predictor a la predicción de Y.

- ε denota el término de error, capturando la variabilidad no explicada por el modelo.

El objetivo central de la LR radica en estimar los valores de los coeficientes β de manera que se minimice alguna función de costo, comúnmente el error cuadrático medio (MSE), que cuantifica la discrepancia entre las predicciones del modelo y los valores observados. Este proceso de ajuste se lleva a cabo a través de técnicas de optimización.

La LR se caracteriza por su interpretabilidad, ya que los coeficientes proporcionan información acerca de la relación y la magnitud de influencia de las variables independientes sobre la variable dependiente. No obstante, se debe tener presente que este modelo asume una relación estrictamente lineal como se ve en la Figura 10, lo cual puede constituir una limitación en casos en los que las relaciones son más complejas. En tales circunstancias, se recurre a modelos de regresión no lineal o enfoques avanzados de ML.

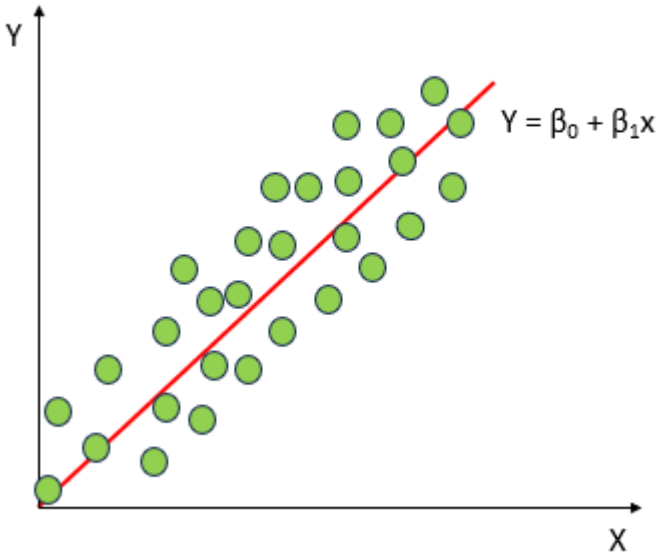


Figura 10. Regresión lineal.

- Regresión Ridge y Lasso (Bedoui & Lazar, 2020): Estas son extensiones de la LR que introducen regularización para evitar el sobreajuste. Ridge agrega una penalización L2 a los coeficientes, mientras que Lasso agrega una penalización L1.
- RF: en regresión (Segal, 2004) combina múltiples DTs para lograr predicciones numéricas más precisas y robustas. Su capacidad para controlar la variabilidad y su flexibilidad lo convierten en una elección popular en una amplia variedad de aplicaciones, ya que es eficaz para manejar relaciones no lineales en los datos.
- Regresión de Vectores de Soporte (en inglés Support Vector Regression - SVR): (Parbat & Chakraborty, 2020) es similar a las SVM para clasificación, y busca encontrar una "zona de tolerancia" alrededor de la línea de regresión que minimice el error.

Las SVR se emplean para predecir valores numéricos en lugar de realizar clasificaciones de etiquetas. De acuerdo con este enfoque, las SVR buscan encontrar un hiperplano que se ajuste de manera óptima a los datos, permitiendo cierto margen de error. Este margen se controla mediante un parámetro específico. El objetivo principal es minimizar tanto el error en el margen como el error de predicción, lo que conduce a un modelo de regresión robusto, considerado caja negra.

- Regresión de K-NN: (Magnussen & Tomppo, 2014). Cuando se trata de predecir un valor numérico, se ubican los K vecinos más próximos y se promedia sus valores de salida. La predicción se calcula como el promedio de estos valores. No es paramétrico, de ahí que sea tan efectivo en modelos de regresión. Se considera caja negra.

- Modelo de Regresión Polinómica (en inglés Polynomial Regression – PR): en la regresión polinómica (Cheng et al., 2018), se utiliza un modelo matemático que se expresa como una función polinómica. Es considerado caja blanca. En lugar de ajustarse a una línea recta, la regresión polinómica se adapta a un polinomio, lo que permite modelar relaciones no lineales entre las variables de entrada y salida.

La regresión polinómica es un enfoque en el ML que se utiliza para modelar relaciones no lineales entre las variables de entrada y la variable de salida. A diferencia de la LR, que asume una relación lineal entre las variables, la PR permite ajustar curvas polinómicas a los datos.

La elección del grado del polinomio es importante que sea el adecuado para evitar el sobreajuste y garantizar predicciones precisas. La forma general de este modelo es:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \ldots\ldots + + \beta_n X^n + \varepsilon \quad (5)$$

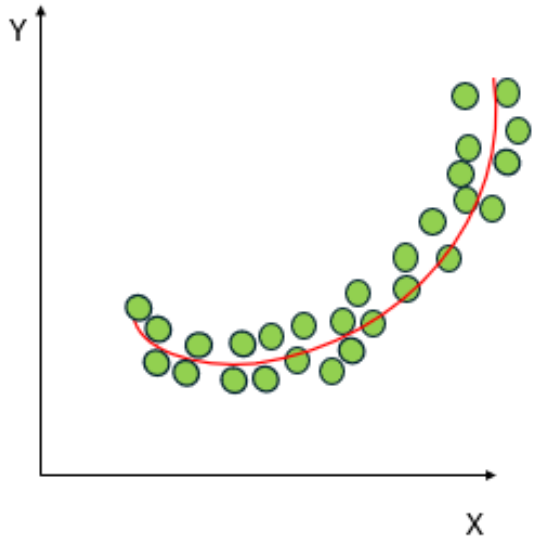


Figura 11. Regresión Polinómica.

Donde:

- Y es la variable de salida que se está tratando de predecir.

- X es la variable de entrada o predictor.

- β_0, β_1, β_2, ….., β_n son los coeficientes del modelo que se estiman a partir de los datos.

- X^2, X^3,…, X^n representan términos polinómicos de orden superior.

- ε es el término de error que captura la variabilidad no explicada por el modelo.

Ventajas de la PR: puede modelar relaciones no lineales en los datos, lo que la hace útil en casos en los que la relación entre las variables no es lineal y puede ajustarse de manera precisa a los datos si se selecciona el grado del polinomio adecuado. A pesar de su capacidad para modelar relaciones no lineales, los coeficientes del modelo siguen siendo interpretables, lo que permite analizar la influencia de cada término polinómico en las predicciones.

Desventajas de la PR: si se utiliza un polinomio de alto grado, el modelo puede sobreajustar los datos de entrenamiento y no generalizar bien a nuevos datos por lo que se debe seleccionar el grado del polinomio de manera cuidadosa, ya que un grado inapropiado puede conducir a un modelo poco preciso.

- La Regresión de DTs, (Pekel, 2020) es un modelo de ML donde el objetivo es predecir un valor numérico en lugar de asignar una observación a categorías predefinidas. Este enfoque se basa en la estructura de un DT, que se compone de nodos y ramas, y se utiliza para modelar relaciones entre las variables de entrada y la variable de salida de manera no lineal. Se puede considerar caja blanca o negra en función de la profundidad de los árboles utilizados.

Para realizar una predicción con DTs de regresión, se comienza en el nodo raíz y se sigue un camino descendente a través de las divisiones y reglas de decisión hasta

llegar a un nodo hoja. El valor numérico en ese nodo hoja se toma como la predicción para la observación.

Las ventajas de este modelo es que es altamente interpretable, ya que cada nodo representa una decisión basada en una característica específica. Esto facilita la comprensión de cómo se llega a una predicción. Además, puede capturar relaciones no lineales en los datos, lo que lo hace adecuado para problemas en los que las relaciones son complejas. Es resistente a valores atípicos y datos ruidosos.

Las desventajas de este modelo es que tiende a sobreajustarse a los datos de entrenamiento si no se controla su crecimiento. Para evitar el sobreajuste, se pueden aplicar técnicas como la poda del árbol o el uso de árboles ensemble, como RF o GB. Debido a esto, pequeñas variaciones en los datos de entrenamiento pueden dar lugar a árboles diferentes, lo que puede hacer que el modelo sea inestable. No es eficaz para problemas con relaciones lineales simples.

- GB Regresión: Similar al GB para clasificación, este modelo construye varios árboles secuencialmente para mejorar la precisión de la regresión, su naturaleza de ensamblaje y su capacidad para capturar relaciones complejas. Estas características permiten que se considere como un modelo de caja negra.
- Las ANN: son modelos altamente flexibles que pueden utilizarse tanto para regresión como para clasificación y son considerados caja negra. Pueden aprender patrones complejos en los datos. La elección de un tipo específico de red neuronal depende de la tarea y los datos a los que se aplique, lo que requiere una cuidadosa selección del modelo más adecuado en cada caso. A continuación, se describen algunos de los tipos más destacados en regresión:

Las Redes Neuronales de Alimentación hacia Adelante (en inglés Feed-Forward Neural Networks - FNN): (Sağ & Abdullah, 2021) la información fluye en una sola dirección, desde la entrada hasta la salida, sin bucles ni conexiones hacia atrás. Utiliza una función de activación lineal para predecir valores continuos.

Las CNN: (Yamashita et al., 2018) operan a través de sus capas en una estructura que emula la organización del córtex visual humano, lo que le permite identificar diversas características en las entradas y, en última instancia, desempeñar una función de percepción visual. Esta CNN consta de múltiples capas ocultas que siguen una jerarquía de procesamiento. En términos más concretos, las primeras capas se especializan en la detección de elementos visuales elementales, como líneas y curvas, mientras que a medida que avanzamos en las capas, se produce una especialización creciente que permite el reconocimiento de formas más complejas, como rostros humanos o la silueta de animales. Este proceso de jerarquización y especialización en las capas de la CNN se asemeja al método de procesamiento visual en el cerebro humano. Pueden analizar datos con estructura espacial, como series temporales y señales.

Las RNN (Lipton et al., 2015) fueron ya concebidas en la década de 1980 pero por sus requerimientos computacionales eran muy difíciles de entrenar. Incluye conexiones que apuntan "hacia atrás", una especie de retroalimentaciones entre las neuronas dentro de las capas como muestra la Figura 12. Son adecuadas para datos secuenciales y pueden utilizarse en problemas que involucran series temporales o datos de secuencia.

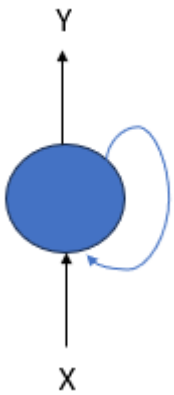


Figura 12. Redes Neuronales Recurrentes.

Las Redes Neuronales AutoEncoders (en inglés AutoEncoders Neural Networks -AE): (Pathirage et al., 2018) son conocidas por la reducción de

dimensionalidad y la representación de datos. Generan una reconstrucción de la entrada mediante una estructura compuesta por dos redes más pequeñas: un codificador y un decodificador. Durante el proceso de entrenamiento, el codificador se encarga de aprender un conjunto de características, conocido como representación latente, a partir de los datos de entrada. Esta representación latente captura aspectos fundamentales de los datos, permitiendo una reconstrucción eficiente de las entradas originales por medio del decodificador. Los autocodificadores son una herramienta esencial en el ML y la extracción de características, ya que contribuyen a representar de manera efectiva la información subyacente en los datos de entrada.

Redes Neuronales Siamesas para Regresión (en inglés Siamese Neural Networks – SNN): Una red siamesa, (Chicco, 2021) consiste en replicar parte de la arquitectura de una red de neuronas, para posteriormente fusionarse en una o varias capas comunes que permiten obtener resultados a través de todo el proceso previo de ambas réplicas. Con ello se permite comparar dos inputs, extraer las características de cada uno de los inputs y realizar cualquier tipo de método (clasificador) que defina como resultado un output fácilmente interpretado, afirmativo o negativo.

Redes Neuronales Generativas Antagónica (en inglés Generative Antagonistic Neural Networks - GAN): (Wang et al., 2020) es una arquitectura de aprendizaje profundo que se compone de dos redes neuronales entrenadas de manera que compitan entre sí, un generador y un discriminador. Su principal objetivo es generar nuevos datos que sean más auténticos y se asemejen a los datos de entrenamiento originales.

Redes Neuronales Recurrentes Bidireccionales (en inglés Bidirectional Recurrent Neural Networks - Bi-RNN): (Cheng et al., 2020) procesan secuencias en ambas direcciones por lo que tienen la capacidad de examinar tanto datos pasados como futuros, lo que conlleva a una mejora sustancial en su capacidad predictiva.

LSTM: es el modelo utilizado en nuestro caso de uso de Ciudades Inteligentes. Las LSTM (Gers et al., 2000; Hochreiter & Schmidhuber, 1997), una variante de las

RNN, son especialmente útiles cuando se necesita capturar dependencias a largo plazo en datos secuenciales. Fueron desarrolladas para abordar el problema del desvanecimiento del gradiente.

Son un tipo de RNN en las que cada neurona recibe retroalimentación de las demás y aparte tiene un estado interno cuyo valor se modifica en función de la experiencia: cada neurona aprende a detectar qué patrones concretos activan su memoria particular, cuáles la resetean y qué información memorizar. De este modo, una neurona concreta puede decidir conservar la información sobre un patrón que detectó hace tiempo, hasta que otro patrón concreto actualice esta información. La riqueza de dinámicas que pueden ser detectadas por este tipo de redes es por tanto enorme, y además elimina el efecto de "gradiente evanescente" que tienen las RRNs tradicionales (imposibilidad de mantener en memoria información más allá de una determinada ventana temporal).

Mediante una red LSTM podemos clasificar y predecir series temporales que presentan dinámicas rápidas mezcladas con dinámicas lentas y cuyas escalas temporales desconocemos en un primer momento.

Las principales características y componentes de las redes LSTM son:

- La celda de memoria es la característica central de una LSTM, puede retener información a largo plazo. Estas celdas permiten a la red aprender y recordar patrones a lo largo de secuencias largas.
- Las LSTM utilizan puertas para controlar el flujo de información dentro y fuera de la celda de memoria. Hay tres tipos principales de puertas:
    - La Puerta de Olvido (en inglés Forget Gate) determina qué información debe ser olvidada o mantenida en la celda de memoria.
    - Puerta de Entrada (en inglés Input Gate) controla qué nueva información se debe agregar a la celda de memoria.
    - Puerta de Salida (en inglés Output Gate) regula la salida de información de la celda de memoria.

- Cada LSTM tiene un estado oculto que se utiliza para realizar predicciones o para comunicar información a otras capas de la red. El estado oculto es una versión filtrada de la celda de memoria y captura la información relevante para la tarea actual.

Al igual que las RNN tradicionales, las LSTM se pueden apilar en múltiples capas para construir redes profundas. Esto les permite aprender representaciones jerárquicas más complejas.

Las LSTM ajustan sus pesos durante el entrenamiento para minimizar la diferencia entre las predicciones y los valores reales en las secuencias de entrenamiento y utilizan el algoritmo de retropropagación a través del tiempo (en inglés BackPropagation Through Time - BPTT) para calcular los gradientes y actualizar los pesos. A continuación, en la Figura 13 se muestra el funcionamiento de este método:

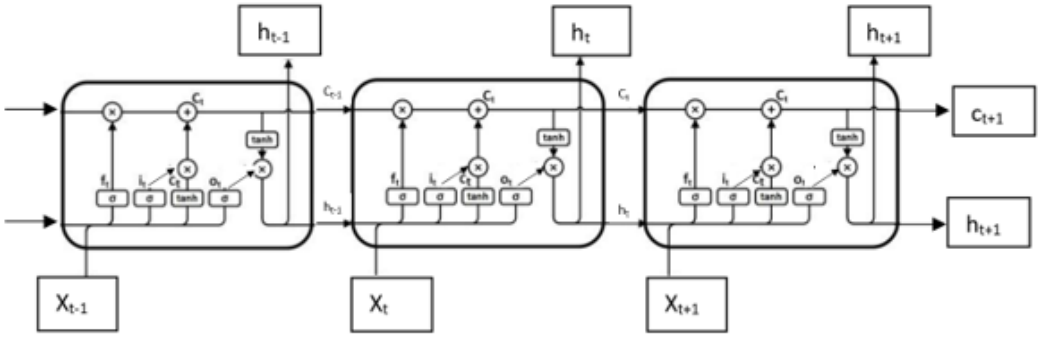


Figura 13. LSTM.

En resumen, las redes LSTM son una poderosa herramienta en el campo del aprendizaje profundo para modelar datos secuenciales. Su capacidad para capturar dependencias a largo plazo las hace especialmente adecuadas para aplicaciones que involucran series temporales, como es nuestro caso de uso.

4.2. Modelo lingüístico difuso 2-tuplas

El paradigma de la lógica difusa emerge como una alternativa a la lógica convencional con el propósito de introducir una dosis de incertidumbre en las decisiones tomadas por el agente decisor. En 1965, el distinguido matemático Lotfi A. Zadeh formuló el concepto de conjuntos difusos (Zadeh, 1975), un aporte de gran relevancia dado que la información que tratamos en el mundo real exhibe grados variados de apreciación y sus atributos pueden ser de naturaleza diversa. De este modo, en numerosas ocasiones, la cuantificación de juicios no se erige como el método más preciso para evaluar una alternativa o, sencillamente, resulta inaccesible cuantificar. Por el contrario, la valoración desde una perspectiva cualitativa se convierte en una alternativa más idónea. En tales casos, un enfoque basado en lenguaje natural suele proporcionar mejores resultados que un enfoque puramente numérico.

Ciertas circunstancias imponen la necesidad de evaluar información a través de términos lingüísticos, debido a la incapacidad intrínseca de cuantificarla. Por ejemplo, al evaluar la probabilidad de ocurrencia de un evento, la asignación de una puntuación numérica puede empobrecer la apreciación subjetiva de dichas opciones. La información se enriquece de manera considerable cuando estas ocurrencias se valoran a través de términos lingüísticos como "muy bajo," "medio" o "muy alto."
En otras ocasiones, trabajar con información precisa de manera cuantitativa no es factible. Esto puede deberse a la falta de elementos necesarios para medir la información de manera exacta o, a que el proceso computacional para obtener valores precisos resulta demasiado costoso, y un valor aproximado es suficiente. En este contexto, el uso de conjuntos difusos, tal como los definidos por Zadeh, ha demostrado ser eficaz para el tratamiento cualitativo de información, (Zadeh, 1975)
Por todo ello, el modelado lingüístico difuso constituye una herramienta que posibilita la representación de atributos cualitativos. Estos modelos se fundamentan en el concepto de variables lingüísticas, cuyos atributos carecen de naturaleza

numérica y se expresan en forma de palabras, enunciados, juicios, etc., empleando tanto el lenguaje natural como el artificial. Cada atributo lingüístico se encuentra definido por una etiqueta o aspecto sintáctico y un significado o valor semántico. La etiqueta representa la palabra o expresión que se incorpora en un conjunto de términos lingüísticos, mientras que el significado se configura como un subconjunto difuso dentro del contexto del discurso.

No obstante, surge una limitación inherente en el modelo, derivada de su metodología para la representación de información y la ejecución de cálculos. Esta restricción tiene su origen en la necesidad de expresar los resultados en un conjunto inicial de términos lingüísticos discretos, lo que implica una pérdida sustancial de información. Con el fin de abordar este desafío, (Herrera & Martínez, 2000, 2001a) se ha concebido un modelo de representación lingüística difusa en el que se emplean pares (s, α), conocidos como 2-tuplas, para representar información de naturaleza lingüística. Cada 2-tupla consta de s, que representa la etiqueta lingüística, y α, que denota el valor de traducción simbólica. En su propuesta, consideran que dado un conjunto de términos lingüísticos, $S = \{s_0, \ldots, s_k\}$, y un valor que refleja el resultado de una operación de agregación simbólica, $\beta \in [0, g]$, la representación de la información equivalente a β se obtiene mediante la siguiente expresión:

$$\Delta_S[0, g] \rightarrow S\ [-0.5, 0.5) \tag{6}$$
$$\Delta_S(\beta) = (s_i, \alpha_i), con \{ i = round(\beta)\ \alpha = \beta - i, \alpha \epsilon [-0.5, 0.5),$$

Donde [·] es la operación de redondeo, si tiene la etiqueta de índice más próxima a β, y α es el valor de la traducción simbólica. Además, a partir de una representación lingüística 2-tuplas siempre es posible volver a su representación lingüística equivalente $\beta \in [0, g]$ mediante la función Δ^{-1}.

Observación. La conversión de un término lingüístico a lingüístico de 2 tuplas consiste en sumar un valor 0 como traducción simbólica, de tal manera que si $\in S \Rightarrow (s_i, 0) \in S$.

Otra presentación de este concepto se exhibe en la Figura 14. Supongamos que disponemos de un valor $\beta = 1.275$, que representa el resultado de una operación de agregación simbólica realizada sobre un conjunto de etiquetas lingüísticas, $S = \{s_0 =$ Muy baja, $s_1 =$ Baja, $s_2 =$ Media, $s_3 =$ Alta, $s_4 =$ Muy alta$\}$. En este contexto, la representación en forma de 2-tupla que transmite la información equivalente a β se expresa como (Bajo, +0,025).

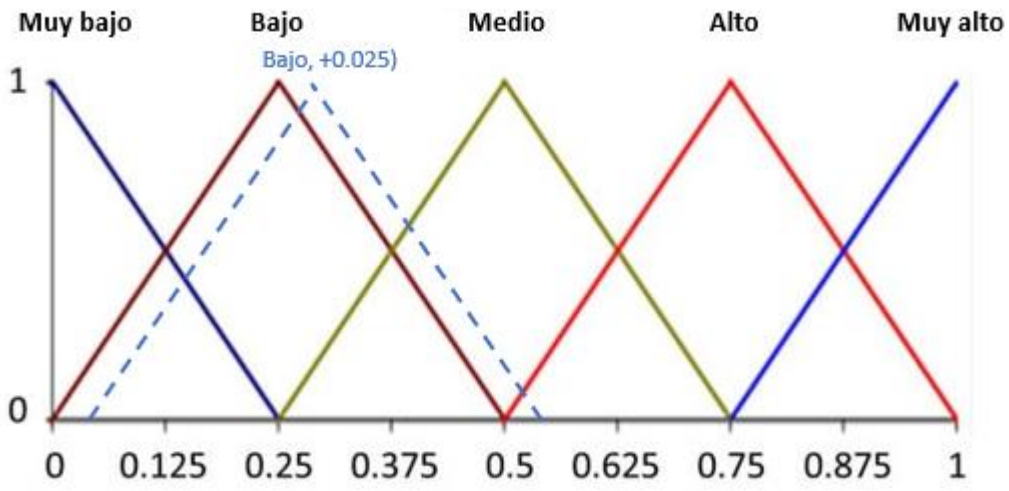


Figura 14. Definición del conjunto S para la variable probabilidad de impago temprano

Es posible emplear cualquier función de pertenencia que sea coherente con la semántica de los términos lingüísticos en el contexto del modelado lingüístico 2-tuplas, lo que conlleva a mejoras significativas en la precisión en comparación con los enfoques simbólicos tradicionales. Sin embargo, (Herrera & Martínez, 2000, 2001a) han indicado que las funciones de pertenencia de forma triangular producen los resultados más precisos.

Por último, el modelado lingüístico 2-tuplas incorpora varios operadores que resultan útiles dependiendo del propósito para el cual se aplique el modelo (Marín Díaz et al., 2021). En este contexto, se presentan dos de estos operadores con el objetivo de facilitar la comprensión del trabajo: el operador de negación (como se muestra en la Ecuación 2) y el operador de agregación de la media ponderada, que definimos a continuación.

$$neg((s_i, \alpha)) = \Delta(g - \Delta^{-1}(s_i, \alpha)) \quad (7)$$

Definición. Sea $S^- = \{(s_1, \alpha_1), ..., (s_n, \alpha_n)\}$ un conjunto lingüístico 2-tuplas y $\Omega = \{\omega_1, ..., \omega_n\}$ sus pesos asociados. Entonces, la media ponderada 2-tuplas ($S^{-\omega}$) se obtiene tal que:

$$S^{-\omega}[(s_1, \alpha_1), ..., (s_n, \alpha_n)] = \Delta\left(\frac{(\Sigma_{i=1}^{n} \beta i \cdot \omega i)}{\Sigma_{i=1}^{n} \omega i}\right) \quad (8)$$

4.3. Técnicas de XAI

La XAI se refiere a un conjunto de enfoques y técnicas diseñados para proporcionar comprensión y claridad en las operaciones y decisiones de los modelos IA. Esto surge como respuesta a la complejidad inherente de muchos modelos de IA que operan como "cajas negras" (Corchado et al., 2021), dificultando la explicación de su lógica subyacente.

Existen dos enfoques básicos en este ámbito. Por un lado, están los enfoques globales que se centran en explicar las decisiones de los modelos de caja negra de forma genérica. Los métodos globales se expresan a menudo como valores esperados basados en la distribución de los datos. Dado que los métodos de interpretación global describen el comportamiento medio, resultan especialmente útiles cuando el modelizador desea comprender los mecanismos generales de los datos o depurar un modelo. Las principales técnicas son:

1. Importancia de Características (en inglés Feature Importance): (König et al., 2021). Esta técnica evalúa la influencia de cada característica o variable en la predicción general del modelo. Puede ayudar a identificar las características más relevantes y, por lo tanto, comprender qué información es crucial para las decisiones.

Si una característica es crucial para las decisiones de un modelo de IA, la precisión predictiva debería disminuir si esa característica se vuelve irreconocible. En el

conjunto de datos, los valores de una característica se mezclan aleatoriamente, eliminando así cualquier relación significativa entre el nuevo valor y el valor objetivo. Si de este proceso resulta un aumento significativo en el error de las predicciones del modelo de IA, se concluye que la característica es importante. En caso contrario, o si el impacto es mínimo, se considera que la característica tiene menor relevancia.

2. Importancia de características mediante permutación (en inglés Permutation feature Importance): (Altmann et al., 2010) mide el aumento en el error de predicción del modelo después de permutar los valores de la característica, lo que rompe la relación entre la característica y el resultado real.

Una característica es "importante" si al barajar sus valores se incrementa el error del modelo, porque en este caso el modelo depende de la característica para la predicción. Una característica es poco importante si al barajar sus valores el error del modelo permanece sin cambios, porque en este caso el modelo ignora la característica para la predicción. La medida de importancia de características mediante permutación fue introducida por Breiman en 2001 para bosques aleatorios. Basándose en esta idea, (Fisher et al., 2019) propusieron una versión independiente del modelo de la importancia de características que llamaron dependencia del modelo.

3. Gráfico de Dependencia Parcial (en inglés Partial Dependence Plot - PDP): (Moosbauer et al., 2021), muestra el efecto marginal que tienen una o dos características sobre el resultado previsto de un modelo de ML. Un PDP puede mostrar si la relación entre el objetivo y una característica es lineal, monótona o más compleja, muestra cómo cambia la predicción del modelo cuando se varía una característica específica mientras se mantienen las demás constantes. Ayuda a comprender la relación entre una característica y la variable objetivo, revelando tendencias y patrones.

4. Gráfico de Efectos Locales Acumulados (en inglés Accumulated Local Effects – ALE). Los gráficos ALE (Gkolemis et al., 2023) se consideran una alternativa más rápida e imparcial a los PDP.

Los gráficos ALE tienen como objetivo proporcionar información sobre cómo cambian, en promedio, las predicciones de un modelo de ML a medida que varía el valor de una característica particular. Los gráficos ALE se calculan tomando el efecto promedio de variar una sola característica mientras se mantienen constantes todas las demás. Esto se hace en diferentes instancias del conjunto de datos.

Para crear gráficos ALE, los pasos son los siguientes:

- Para una característica específica, el conjunto de datos se ordena según los valores de esa característica.
- Luego, se calculan las predicciones del modelo para cada instancia en el conjunto de datos ordenado.
- Los valores ALE se calculan como la diferencia entre la predicción promedio para las instancias con un valor específico de la característica y la predicción promedio para las instancias con un valor ligeramente inferior de la característica.

Los valores ALE positivos indican que aumentar el valor de la característica tiene, en promedio, un efecto positivo en la predicción del modelo y, los valores negativos sugieren que aumentar el valor de la característica tiene, en promedio, un efecto negativo.

Los gráficos ALE ayudan a identificar relaciones no lineales entre características y predicciones.

La ventaja que tienen los gráficos ALE respecto de los PDP es que se consideran más rápidos y son imparciales, ya que no dependen de técnicas de agrupamiento o suavizado, lo que los hace menos sensibles a la elección de hiperparámetros.

5. Interacción de Características (en inglés Feature Interaction - FI): (Oh, 2019). Esta técnica busca identificar las interacciones entre diferentes características. Puede revelar si el modelo considera las relaciones no lineales o las combinaciones de características para la toma de decisiones.

Si un modelo de aprendizaje automático realiza una predicción basada en dos características, podemos descomponer la predicción en cuatro términos: un término constante, un término para la primera característica, un término para la segunda característica y un término para la interacción entre las dos características. La interacción entre dos características es el cambio en la predicción que ocurre al variar las características después de considerar los efectos individuales de cada una.

La ventaja, dado que la estadística es adimensional, es comparable entre características e incluso entre modelos. La desventaja es que computacionalmente es muy caro.

6. Los Modelos Surrogados Globales (en inglés Global Surrogate Models) son una parte fundamental en nuestros casos de uso. Estos modelos son modelos simplificados y altamente interpretables, y tienen la capacidad de representar el comportamiento del modelo complejo subyacente en términos más simples y comprensibles, siendo posible comprender cómo se toman las decisiones. Son considerados modelos de caja blanca. Suelen ser empleados en situaciones donde se deben interpretar modelos considerados de caja negra, es decir, modelos cuyo funcionamiento interno no es fácilmente comprensible.

La generación de un modelo surrogado global a menudo se traduce en la creación de un conjunto de reglas que describen el proceso de toma de decisiones. Varios estudios se han centrado en la extracción de estas reglas (Bologna, 2019; Keneni et al., 2019; Singh et al., 2019), lo que contribuye a mejorar la comprensión de los modelos de ML y su aplicabilidad en diversas disciplinas.

El propósito fundamental de los modelos surrogados interpretables es, lograr la máxima aproximación entre la función de predicción del modelo de ML y la función de predicción del modelo surrogado, sin sacrificar la interpretabilidad de este último.

Para entrenar un modelo surrogado, se adopta un enfoque agnóstico respecto al modelo subyacente, lo que significa que no se requiere información sobre su funcionamiento interno. Solo se necesita acceso a los datos y a la función de predicción del modelo complejo. Esto permite que el método surrogado sea versátil y aplicable a diferentes modelos considerados caja negra sin la necesidad de ajustes significativos. La elección tanto del tipo de modelo de caja negra como del modelo surrogado no está condicionada, por lo que aumenta la flexibilidad de la metodología.

Una forma común de evaluar la calidad de ajuste entre el modelo surrogado y el modelo de caja negra es mediante la métrica de R-cuadrado (R^2). Esta métrica proporciona información sobre lo bien que el modelo surrogado replica el comportamiento del modelo original, lo que es esencial para medir la eficacia y precisión del modelo surrogado.

$$R^2 = 1 - \frac{SSE}{SST} \tag{9}$$

Donde, R^2 es el coeficiente de determinación., SSE (suma de cuadrados de error) representa la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo y, SST (suma de cuadrados totales) denota la suma de los cuadrados de las diferencias entre los valores observados y la media de sus valores.

Si R^2 es cercano a 1 (SSE bajo), entonces el modelo interpretable se aproxima muy bien al comportamiento del modelo de caja negra. Si el modelo interpretable se aproxima mucho, es posible que se desee sustituir el modelo complejo por el modelo interpretable. Si el R^2 es cercano a 0 (alto SSE), entonces el modelo interpretable falla en la explicación del modelo de caja negra.

El rendimiento del modelo de caja negra no influye en la formación del modelo alternativo. La interpretación del modelo alternativo sigue siendo válida porque hace afirmaciones sobre el modelo y no sobre el mundo real. Pero, la interpretación del modelo alternativo es irrelevante, si el modelo de caja negra es poco predictivo.

Por otro lado, tenemos enfoques locales que tratan de explicar decisiones particulares del modelo (para una entidad concreta). Estos métodos tienen como objetivo arrojar luz sobre cómo un modelo de ML en particular llega a una decisión específica o a una salida para una entrada dada.

Algunas técnicas que se incluyen en esta categoría son: LIME (Explicaciones Locales Model-Agnostic Interpretativas) (Nóbrega et al., 2019; Ribero et al., 2016) y SHAP (Lundberg, 2017; Sundararajan & Najmi, 2020). Estos métodos generan explicaciones locales mediante la aproximación del comportamiento del modelo en las proximidades de la instancia en cuestión.

1. Modelo surrogado local (en inglés Local Surrogate Model - LIME): es una técnica que se utiliza para explicar el funcionamiento de un modelo de ML mediante la creación de un modelo interpretable y local que aproxima el comportamiento del modelo de caja negra en torno a una instancia específica. LIME se centra en entrenar modelos surrogados locales para explicar predicciones individuales.

Opera generando un conjunto de datos perturbado alrededor de la instancia de interés y observando cómo el modelo de caja negra realiza predicciones en estos datos perturbados. Luego, ajusta un modelo interpretable, como una LR o un DT, a estos datos perturbados para explicar las predicciones del modelo de caja negra en ese entorno local. Matemáticamente, los modelos surrogados locales con restricciones de interpretabilidad se pueden expresar de la siguiente:

$$explanation(x) = \arg \min_{g \varepsilon G} L(f, g, \pi_x) + \Omega(g) \quad (10)$$

2. SHAP: Los valores Shapley son un marco matemático que se utiliza para asignar una contribución específica a cada característica en una predicción de

modelo. A través de los valores Shapley, se puede entender cómo cada característica influye en la predicción, y cómo se distribuye la contribución entre las características. Para calcular los valores Shapley, se evalúa cómo cambia la predicción al introducir diferentes combinaciones de características. Esto implica evaluar todas las posibles permutaciones de características y determinar su impacto en la predicción. Proporcionan explicaciones precisas y cuantificables sobre el papel de cada característica en la toma de decisiones del modelo. La fórmula se simplifica a:

$$g(x') = \phi_0 + \sum_{j=1}^{M} \phi j \quad (11)$$

Donde la representación como un modelo lineal de coaliciones es para el cálculo de los ϕ. Para x, la instancia de interés, el vector de coalición x es un vector de todos 1, es decir, todos los valores de las características están presentes.

3. Expectativas Condicionales Individuales (en inglés Individual Condicional Expectation - ICE). (Goldstein et al., 2015). Muestran una línea por instancia que indica cómo cambia la predicción de la instancia cuando se altera una característica. Un gráfico ICE visualiza la dependencia de la predicción de una característica para cada instancia por separado, dando como resultado una línea por instancia, en comparación con una línea en total en los PDP. Un PDP es el promedio de las líneas de un gráfico ICE. Los valores para una línea (una instancia) se pueden calcular manteniendo todas las demás características iguales, creando variantes de esta instancia al reemplazar el valor de la característica con valores de una cuadrícula y haciendo predicciones con el modelo de caja negra para estas instancias recién creadas. El resultado es un conjunto de puntos para una instancia con el valor de la característica de la cuadrícula y las predicciones respectivas.

Los gráficos PDP pueden mostrar cómo se ve la relación promedio entre una característica y la predicción. Esto funciona bien solo si las interacciones entre las características para las cuales se calcula el PDP y las demás características son débiles. Si hay interacciones entre las características, los gráficos ICE proporcionan una

perspectiva más detallada que los PDP. A continuación, en la Tabla 4 se muestra un resumen con las diferentes técnicas de XAI (Keneni et al., 2019; Singh et al., 2019):

Tabla 4. Principales características de las distintas técnicas de XAI

Técnicas	**Global o Local**	**Ventajas**	**Desventajas**
Importancia de características	Global	Interpretación global muy comprimida. Considera las interacciones entre las características	No está claro si se puede utilizar en el conjunto de datos de entrenamiento o en el conjunto de datos de prueba.
PDP	Global	Interpretación intuitiva y clara	Hipótesis de independencia entre características
Interacción de características	Global	Detecta todas las interacciones entre características	Costoso desde el punto de vista informático
Importancia de características mediante permutación	Global	Facilidad de Implementación	Costo computacional
ALE	Global	Independencia de la Distribución de las Características	Complejidad Computacional
Modelos surrogados globales	Global	Es fácil medir la bondad del modelo surrogado utilizando la medida R-cuadrado	No está claro cuál es el mejor límite de R-cuadrado para confiar en el modelo surrogado resultante
LIME	Local	Explicación breve y comprensible Explica diferentes tipos de datos	Puntos muy cercanos pueden tener explicaciones totalmente diferentes
SHAP	Local	La explicación se basa en el teorema de la teoría de juegos fuertes	Costoso desde el punto de vista informático
ICE	Local	Intuitivo y fácil de entender	La gráfica puede volverse demasiado saturada para comprenderla.

El propósito de este estudio es mejorar la interpretabilidad y para ello utilizaremos el siguiente algoritmo (basado en Molnar, 2020):

1. Seleccionar el conjunto de datos X utilizado para entrenar el modelo de caja negra.

2. Para el conjunto de datos X seleccionado, obtener las predicciones del modelo de caja negra.

3. Seleccionar un modelo de árbol de regresión.

4. Entrenar el árbol de regresión en el conjunto de datos X y sus predicciones.

5. Obtener las reglas del árbol de regresión.

6. Medir lo bien que el modelo surrogado replica las predicciones del modelo de caja negra.

El conjunto de datos de entrada para el modelo surrogado es el utilizado para entrenar el modelo de ML inicial. Este conjunto de datos debe ser representativo de las condiciones en las que se espera que el modelo haga predicciones. Por lo tanto, la selección del conjunto de datos para entrenar el modelo de caja negra es un paso crucial en el proceso de interpretación de modelos. La Figura 15 muestra cada uno de los pasos seguidos:

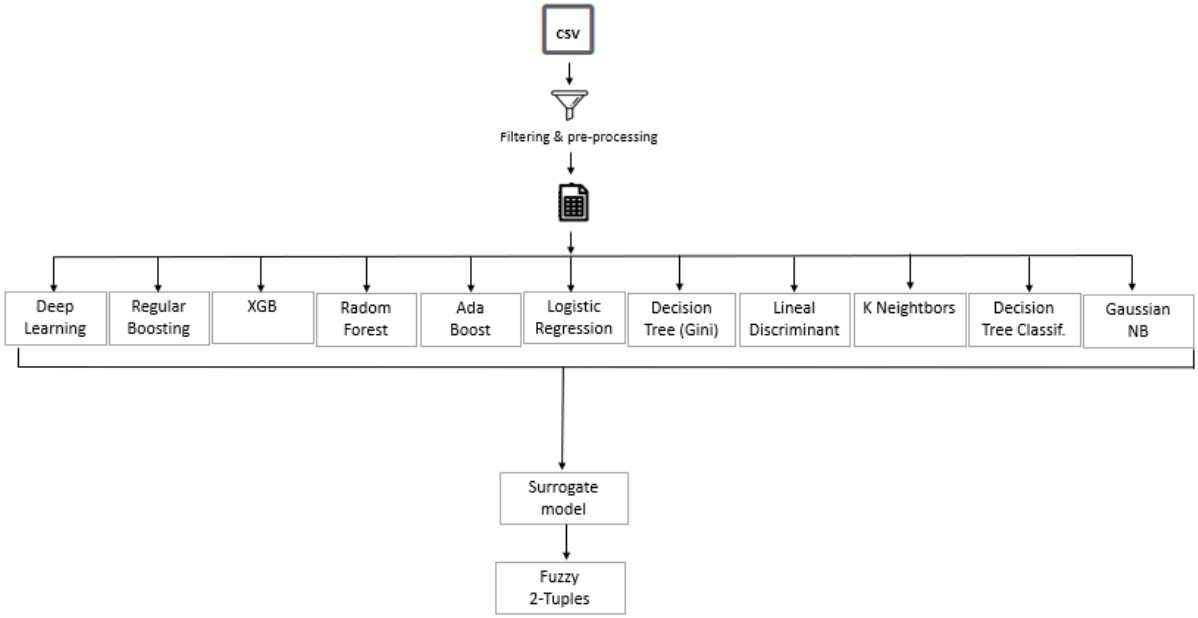


Figura 15. Proceso XAI Machine Learning.

Cuando se aborda la interpretación de un modelo de caja negra, es esencial elegir cuidadosamente el conjunto de datos utilizado para entrenar ese modelo, ya que la selección cuidadosa del conjunto de datos garantiza que el modelo de caja negra se entrene y evalúe en condiciones que reflejen con precisión su aplicación en el mundo real, y esto es fundamental para una interpretación significativa y una mejor comprensión de cómo el modelo toma decisiones.

El conjunto de datos seleccionado para ambos casos de uso fueron lo suficientemente amplios en términos de tamaño y, abarcaron una gama significativa de escenarios y contextos relevantes para las aplicaciones previstas. Esta amplitud aseguró que los modelos seleccionados no solo se ajustasen a nuestros casos específicos, sino que también son capaces de adaptarse eficazmente a situaciones diversas y emergentes.

Al considerar la representatividad del conjunto de datos, es esencial tener en cuenta las condiciones del entorno, y esto implica tener en cuenta factores como la variabilidad estacional, los cambios en las tendencias a lo largo del tiempo y las fluctuaciones en los patrones de comportamiento. Al capturar estas variaciones en el conjunto de datos, se garantiza que los modelos estén mejor preparados para

enfrentar los desafíos y complejidades del entorno real. En última instancia, la representatividad del conjunto de datos se traduce en una mayor robustez y aplicabilidad del modelo predictivo.

La correcta división del conjunto de datos en conjuntos de entrenamiento y pruebas es un componente esencial para evaluar la capacidad de generalización de un modelo de caja negra a datos no observados. En el análisis del comportamiento de los pasajeros, se implementó una partición del 80% para el conjunto de entrenamiento (enero 2015 - septiembre 2016) y el 20% para el conjunto de pruebas (septiembre 2016 – febrero 2017). En contraste, en el ámbito de la investigación sobre préstamos entre particulares, se optó por una proporción del 70% para entrenamiento y el 30% para pruebas.

Esta selección de proporciones específicas buscó lograr un equilibrio entre la necesidad de que el modelo aprendiese patrones representativos durante el entrenamiento y la capacidad de evaluar su desempeño de manera efectiva en datos no vistos. La elección de estas proporciones se basó en consideraciones prácticas y buscó evitar el sobreajuste o subajuste del modelo a través de una distribución adecuada de datos de entrenamiento y prueba.

Una vez que ya tenemos los modelos más predictivos para los distintos casos de uso, XGBoost para Fintech y LSTM para Ciudades Inteligentes, el siguiente paso es la elección del modelo interpretativo.

El modelo elegido como modelo interpretativo fue un DT regresión (modelo surrogado) que nos explicó la probabilidad de incumplimiento temprano en préstamos P2P y la probabilidad de demanda de pasajeros en la línea 1 de autobuses. La razón detrás de esta elección de modelo radica en la naturaleza inherentemente interpretable de este modelo, que contrasta con la opacidad de los modelos de caja negra. La elección de un modelo surrogado interpretable busca proporcionar una comprensión más clara a través de reglas y transparencia de las relaciones existentes entre las variables de entrada y las predicciones del modelo. Los DTs son

especialmente idóneos para este propósito, ya que representan de manera visual y estructurada las decisiones lógicas tomadas por el modelo desde el punto de vista de negocio, en función de las características del conjunto de datos. Además, la elección de un modelo surrogado interpretable permite crear puentes entre la complejidad de los modelos de caja negra utilizados y la necesidad de comprensión y transparencia en la toma de decisiones. La interpretación clara de las relaciones en el modelo surrogado contribuye a la confianza en las predicciones de los modelos predictivos utilizados inicialmente y proporciona información significativa sobre cómo las variables de entrada influyen en las salidas del modelo. En consecuencia, este enfoque busca equilibrar la precisión predictiva con la interpretabilidad, esencial en situaciones donde la comprensión de las decisiones del modelo es crítica. Para ello, la variable objetivo de nuestro modelo surrogado en cada caso de uso ha sido las predicciones obtenidas de los modelos iniciales, en el LSTM (previsión de demanda de pasajeros) y el XGBoost (predicción del incumplimiento temprano).

Las entradas al árbol fueron el subconjunto de variables de nuestros modelos que vamos a utilizar para explicar. Estas variables proporcionan información clave sobre la situación financiera, historial crediticio y detalles del préstamo. En el caso de uso de previsión de demanda de pasajeros las variables de entrada proporcionan la información sobre día de la semana, mes y tipo de día del año.

A continuación, se proporciona una breve explicación de las variables presentes en el conjunto de datos de Lending Club:

- Term_60_months: indica el plazo del préstamo en meses. En este caso, específicamente, se refiere a préstamos con un plazo de 60 meses (5 años). Puede ser utilizada para evaluar la duración del compromiso financiero asociado al préstamo.
- Application_type: describe el tipo de aplicación presentada para el préstamo. Puede tener dos valores principales: "Individual", si la solicitud

es presentada por una persona sola, y "Joint", si la solicitud es conjunta, es decir, presentada por más de una persona.

- Delinq_2yrs: representa el número de pagos atrasados durante los últimos dos años. Esta variable es un indicador del historial de pagos y puede influir en la evaluación del riesgo crediticio del solicitante.
- Inq_last_6mths: indica la cantidad de consultas al historial crediticio realizadas en los últimos 6 meses. Un mayor número de consultas puede sugerir mayor actividad financiera.
- Bc_util: refiere a la tasa de utilización de la línea de crédito en relación con el límite disponible. Es un indicador de cuánto del crédito disponible está siendo utilizado por el prestatario.
- Acc_open_past_24mths: representa el número de cuentas abiertas en los últimos 24 meses. Esta variable ofrece una perspectiva sobre la actividad crediticia reciente del prestatario.
- Grade: esta variable clasifica el riesgo crediticio asignado al préstamo por Lending Club. Las clasificaciones de grado son letras, como A, B, C, etc., donde A representa menor riesgo y G representa mayor riesgo. Esta variable es crucial en la evaluación del riesgo crediticio asociado al préstamo.

Para la línea de investigación de predicción de demanda de pasajeros hemos utilizado las variables:

- Días de la semana: lunes, martes, miércoles, jueves, viernes, sábado y domingo.
- Meses: enero, febrero, marzo, abril, mayo, junio, julio, agosto, septiembre, octubre, noviembre y diciembre.
- y vacaciones.

La interpretación precisa de estas variables es esencial para una evaluación efectiva del riesgo crediticio y la toma de decisiones financieras, de la misma manera que para la toma de decisiones en función de la demanda de autobuses.

En ambos casos de estudio hemos empleado un árbol de regresión. Utilizaremos el coeficiente de determinación R^2, es una medida estándar que evalúa la calidad del modelo, qué tan bien el árbol de regresión explica las predicciones del modelo de caja negra que estamos intentando explicar. En la línea de investigación de Fintech obtuvimos un R^2 en entrenamiento = 0,806, R^2 en test = 0,808 y, en el caso de uso de demanda de pasajeros, R^2 entrenamiento = 0,95 y R^2 test = 0,95. Por lo tanto, conocemos cualitativamente la razón de la predicción y hemos obtenido una explicación fácil de las predicciones de los modelos subyacentes. El modelo pudo interpretarse desde el punto de vista empresarial para ambos casos de uso mediante un conjunto de 9 reglas.

Posteriormente, procedimos a la extracción de reglas del modelo surrogado. En este contexto, cada nodo y cada rama del árbol se transforman en reglas condicionales que se fundamentan en las características inherentes al conjunto de datos. La combinación de nodos y ramas permitió la construcción de reglas más complejas que abarcan múltiples condiciones. Las reglas condicionales extraídas se interpretaron para comprender cómo el modelo surrogado utiliza las características específicas del conjunto de datos para realizar predicciones. Esto proporcionó claridad sobre los patrones y relaciones fundamentales capturadas por el modelo.

4.4. Metodología KDD

Uno de los marcos de trabajo más reconocidos para el proceso de KDD fue propuesto por Fayyad (Fayyad, et al., 1996). Este marco divide el proceso general en una serie de fases ordenadas, las cuales involucran la comprensión del dominio y la definición del alcance de la aplicación, el preprocesamiento de datos (que incluye selección, limpieza y transformación), la minería de datos y la interpretación/evaluación de los resultados obtenidos.

Fayyad et al. resaltan que este proceso es complejo y altamente iterativo, con retroalimentación continua sobre las distintas etapas, y subraya la necesidad de supervisión y evaluación de un agente humano, quien toma decisiones sobre el siguiente paso a seguir en el proceso.

El motivo de haber seleccionado esta metodología es porque tiene una fase explícita que nos habla de interpretación de patrones de extracción, en línea con XAI. Para lograr nuestros objetivos, nos basaremos en las fases del proceso KDD que se muestran en la Figura 16.

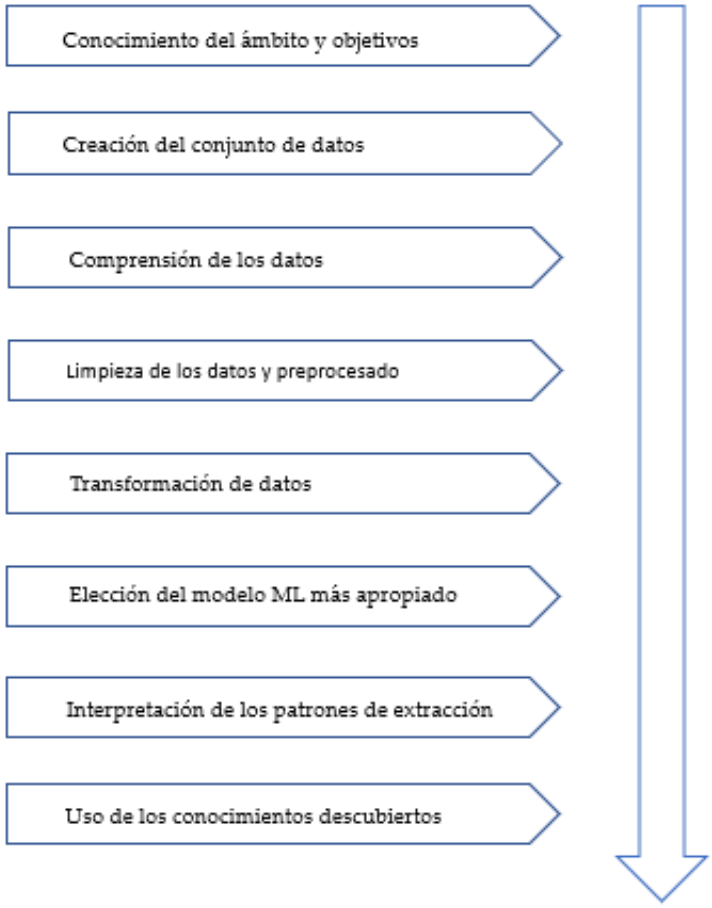


Figura 16. Fases del proceso KDD

Fase 1. Conocimiento del ámbito y objetivos: es la primera etapa del proceso KDD. Esta fase es esencial para establecer una base sólida para el resto del proceso, asegurando que los objetivos estén alineados con las necesidades del negocio, y que se tengan en cuenta todos los aspectos relevantes del dominio de la aplicación.

Fase 2. Creación de un conjunto de datos: Esta segunda fase es esencial para establecer la base sobre la cual se aplicarán técnicas de minería de datos y aprendizaje automático en las etapas posteriores del proceso KDD. La calidad y la preparación adecuada del conjunto de datos son críticas para obtener resultados precisos y significativos en el análisis.

Fase 3. Comprensión de los datos: esta tercera fase es crucial para obtener información significativa y preparar los datos para análisis más avanzados. La comprensión adecuada de los datos es esencial para garantizar que el análisis y la interpretación siguiente se realice de manera significativa. Esta etapa sienta las bases para el diseño y la implementación de modelos y técnicas de minería de datos en las etapas posteriores del proceso KDD.

Fase 4. Limpieza de los datos y preprocesado: Se trata de la cuarta fase del proceso KDD, enfocada en la limpieza y preprocesamiento de datos. Esta etapa permite asegurar la calidad y la preparación adecuada de los datos antes de aplicar técnicas de análisis más avanzadas.

Fase 5. Transformación de datos: Esta etapa se centra en la transformación de los datos para mejorar su utilidad en el análisis y así los algoritmos puedan implementarse de manera efectiva.

Fase 6. Elección del modelo de ML más apropiado: En esta sexta fase, procederemos a construir el sistema predictivo. Antes de seleccionar un modelo, es necesario tener claridad sobre los objetivos específicos del proyecto de KDD. Dependiendo de estos objetivos, se podrán elegir modelos que se adapten mejor a la naturaleza del problema. Con este propósito, hemos desarrollado diversos tipos que permitan seleccionar el más predictivo.

Fase 7. Interpretación de los patrones de extracción. Esta etapa es esencial para cerrar el ciclo del proceso KDD y convertir los resultados en información valiosa y accionable. Hemos empleado la técnica del modelo surrogado, considerado modelo de caja blanca, para obtener un conjunto de reglas que expliquen las predicciones del modelo de ML construido y seleccionado en cada caso.

Fase 8. Uso de los conocimientos descubiertos. Esta fase constituye el último y definitivo paso del proceso KDD, donde el conocimiento descubierto posee utilidad para diversos casos de uso.

5. Aplicación del modelo lingüístico difuso 2-tuplas en las técnicas XAI de árboles surrogados: 2-Tuplas, árboles de surrogación lingüística difusa y reglas.

El objetivo de esta investigación es mejorar la interpretabilidad de los modelos de XAI, centrándonos en el uso de técnicas surrogadas como hemos visto en la sección 4.3

Se ha observado que, hasta el momento, no se ha empleado la interpretabilidad lingüística difusa para mejorar la interpretabilidad de los propios modelos de XAI, los cuales consisten a menudo en reglas difíciles de entender. Estas reglas pueden proporcionar soluciones numéricas cuya interpretación, en términos de magnitud, resultaría ambigua. Además, las reglas pueden referirse a múltiples variables con diferentes escalas y significados comerciales, lo que dificulta su comprensión. En este estudio, la interpretabilidad se aborda mediante el uso de etiquetas lingüísticas, lo que las hace completamente interpretables para los humanos sin sacrificar la precisión del modelo.

5.1. Propuesta de mejora de la interpretabilidad lingüísticas de un modelo XAI surrogado

Dado que queremos comprender la predicción en los distintos casos de uso de este estudio, aplicado el modelo surrogado (sección 4.3), utilizaremos el modelo de 2-tuplas para conseguir interpretar lingüísticamente nuestro modelo. Con este propósito, redefinimos el algoritmo (basado en Molnar, 2020) propuesto en la sección 4.3, incorporando los pasos 7, 8 y 9, integramos el modelo de 2-tuplas en XAI y el algoritmo se incluye en la fase 7 de KDD (sección 4.4): interpretación de los patrones de extracción. Se presenta el algoritmo redefinido:

1. Seleccionar el conjunto de datos X utilizado para entrenar el modelo de caja negra.

2. Para el conjunto de datos X seleccionado, obtener las predicciones del modelo de caja negra.

3. Seleccionar un modelo de árbol de regresión.

4. Entrenar el árbol de regresión en el conjunto de datos X y sus predicciones.

5. Obtener las reglas del árbol de regresión.

6. Evaluar la capacidad del modelo surrogado para replicar las predicciones del modelo de caja negra.

7. Fuzzificar la variable a predecir usando el modelo lingüístico difuso de 2-tuplas.

8. Definición de las etiquetas.

9. Interpretar el modelo lingüístico difuso surrogado.

A continuación, seleccionamos una serie de variables a interpretar. Como ya se ha comentado en la sección 4.2, si queremos mejorar la interpretabilidad, podemos representar las variables numéricas continuas, independientes y dependientes, con su correspondiente valor de 2-tupla. En primer lugar, realizaremos una transformación para representar las variables numéricas en el intervalo [0, 1], como paso previo y necesario para la transformación a un valor 2-tuplas (véase la Figura 14). Seguidamente, aplicamos un proceso de fuzzificación para reemplazar sus valores numéricos por términos lingüísticos en las reglas correspondientes. Esto permite representar las variables originales de manera más comprensible, manteniendo la precisión total y sin pérdida de información.

Teniendo en cuenta las reglas extraídas y la fuzzificación de las variables de interés, se interpreta el modelo surrogado fuzzificado.

Imaginemos un árbol surrogado utilizado para interpretar la demanda de pasajeros. Este árbol genera reglas que contienen variables numéricas cuya escala es desconocida. La regla obtenida es,

IF (sat = 1) and (y (-1) ≤ 3110) Predicción = 1627

si el día de la semana es sábado, y los pasajeros del día anterior fueron igual o inferior a 3110, entonces la predicción es 1627 pasajeros. Desconocemos la escala de la variable número de pasajeros, y no sabemos si 3110 o 1627 es mucho o poco. Como la gente entiende mejor etiquetas que números, vamos a realizar la transformación a un valor de 2-tuplas, pero previamente realizaremos la conversión lineal min-max para representar la variable de número de pasajeros en el intervalo [0, 1]

IF (sat = 1) and (y (-1) ≤ 0,70194465) Predicción = 0,1473448,

Aplicamos el modelo 2–tuplas y obtenemos la conversión a un término lingüístico fácilmente interpretable por los humanos:

IF (sat = 1) and (y (-1) ≤ (Alta, -0,048)) Predicción = (Baja, −0,103)

Por lo tanto, proporciona una explicación más comprensible y lingüísticamente interpretable de cómo las características del conjunto de datos influyen en las predicciones del modelo de caja negra.

El enfoque de Molnar tiene como objetivo hacer que los modelos complejos sean más comprensibles y accesibles, facilitando así la interpretación de sus resultados y aumentando la confianza en el modelo original. En aras de incluir la mejora de la interpretabilidad lingüística, el algoritmo presentado se integrará en la fase de interpretación de patrones del proceso KDD (sección 4.4). A continuación, se detallan los diferentes casos de uso en los que se aplicará este enfoque.

5.2. Caso de uso en Fintech. Predicción del incumplimiento temprano (clasificación)

En el ámbito de este estudio, nos centramos en el análisis de los préstamos P2P en el período que abarca desde 2007 hasta el tercer trimestre de 2020. Para ello desarrollamos distintas fases KDD (sección 4.4).

El propósito central de nuestro estudio en el ámbito Fintech es lograr la interpretabilidad de la probabilidad de impago temprano. El primer paso consiste en desarrollar un modelo capaz de predecir el impago temprano, lo que es de vital importancia para cualquier entidad financiera. Igualmente, relevante es comprender los determinantes subyacentes de estos impagos, de manera que, la empresa pueda tomar decisiones informadas sobre la concesión de préstamos, basándose en el conocimiento anticipado de los riesgos, y de los patrones y reglas que guían el comportamiento de los clientes.

Los datos se proporcionaron en formato CSV, en forma de archivo plano, con un total de 2.925.493 préstamos. Este archivo contenía información detallada sobre los solicitantes de préstamos, incluyendo variables relevantes para la solicitud y eventos posteriores. El procesamiento de los datos se llevó a cabo utilizando el lenguaje de programación Python. En última instancia, se seleccionaron aquellas variables que contenían información relevante para predecir el impago temprano y que no incluían datos posteriores a la aprobación del préstamo.

Se consideraron exclusivamente los préstamos de los que se disponía de información detallada sobre su estado de liquidación, incluyendo tanto los que fueron completamente pagados como aquellos que presentaron retrasos en el pago de entre 31 y 120 días.

Una de las variables de interés, denominada "Delinq_2yrs," refleja la cantidad de impagos que superan los 30 días en el historial crediticio del prestatario durante los dos años anteriores.

La variable estado del préstamo, en el dataset “loan status”, es la variable dependiente y adopta una naturaleza binaria, donde el valor 1 corresponde a los

préstamos que en algún momento presentaron un impago de más de 30 días. Esto incluye tanto los préstamos en estado de impago, los morosos que acumularon un retraso de 31 a 120 días, así como aquellos cuyo período de pago excedió el plazo acordado inicialmente, aunque posteriormente liquidaran por completo. Por otro lado, el valor 0 se asigna a los préstamos que fueron completamente pagados y cumplieron con el plazo acordado.

De un conjunto de 806.161 préstamos analizados, un 20,20% (171.066) se catalogaron como operaciones con impago temprano, como se ilustra en la Figura 17

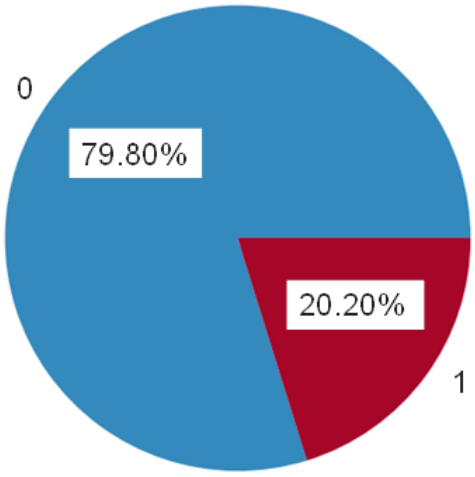


Figura 17. Porcentaje de impago temprano.

Para llevar a cabo el proceso de limpieza de los datos y preprocesado, se implementaron varias estrategias iniciales. En primer lugar, se eliminaron las variables que contenían información posterior a la formalización del préstamo, así como aquellas que no resultaban relevantes para nuestro estudio. Posteriormente, se procedió a eliminar las columnas que presentaban más del 40% de valores vacíos.

Para lograr el objetivo de transformación de datos, se llevaron a cabo las siguientes transformaciones en las variables categóricas:

- Codificación One Hot: Para cada variable categórica L (por ejemplo, verification_status, purpose, etc.), se creó una columna por cada valor distinto presente en la característica. La columna LZ, correspondiente a la variable categórica L, toma el valor 1 o 0 en función de si el valor de L es igual a Z o no.

- Codificación de etiquetas: En este paso, las etiquetas se convirtieron a un formato numérico. Por ejemplo, para la variable "grade," se asignaron los siguientes valores: A:1, B:2, C:3, D:4, E:5, F:6, G:7. Para subgrados de "A," se asignaron valores secuenciales: A1:1, A2:2, A3:3, A4:4, etc. De manera similar, se realizaron conversiones numéricas para otras variables, como "emp_length."

- Eliminación del símbolo "%": Se eliminó el símbolo "%" de las variables "int_rate" y "revol_util" para convertirlas a formato numérico.

A continuación, la distribución de algunas de las variables transformadas:

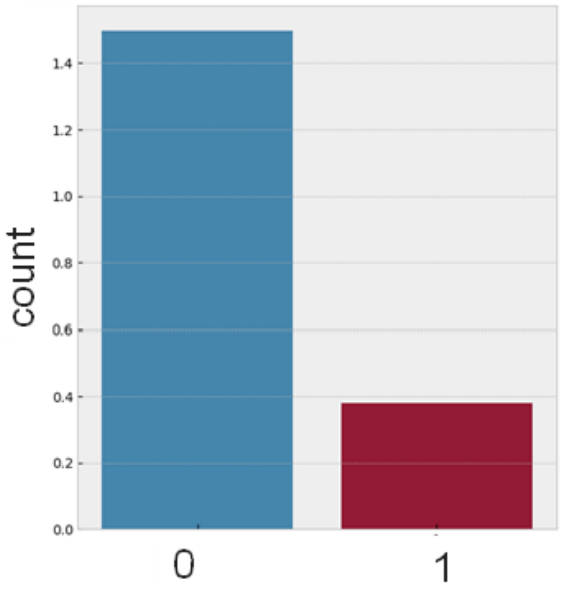


Figura 18. Frecuencia del impago temprano.

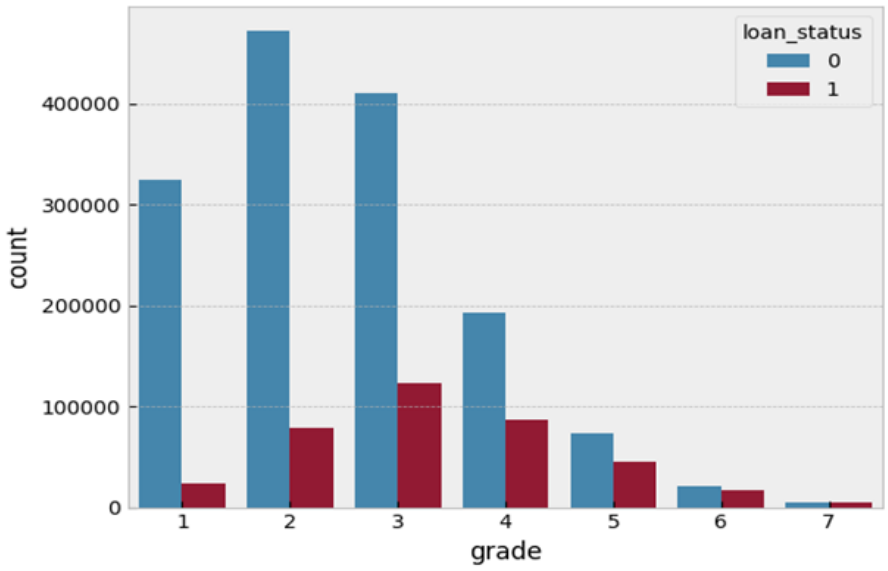


Figura 19. Distribución del estado del préstamo en función del grado.

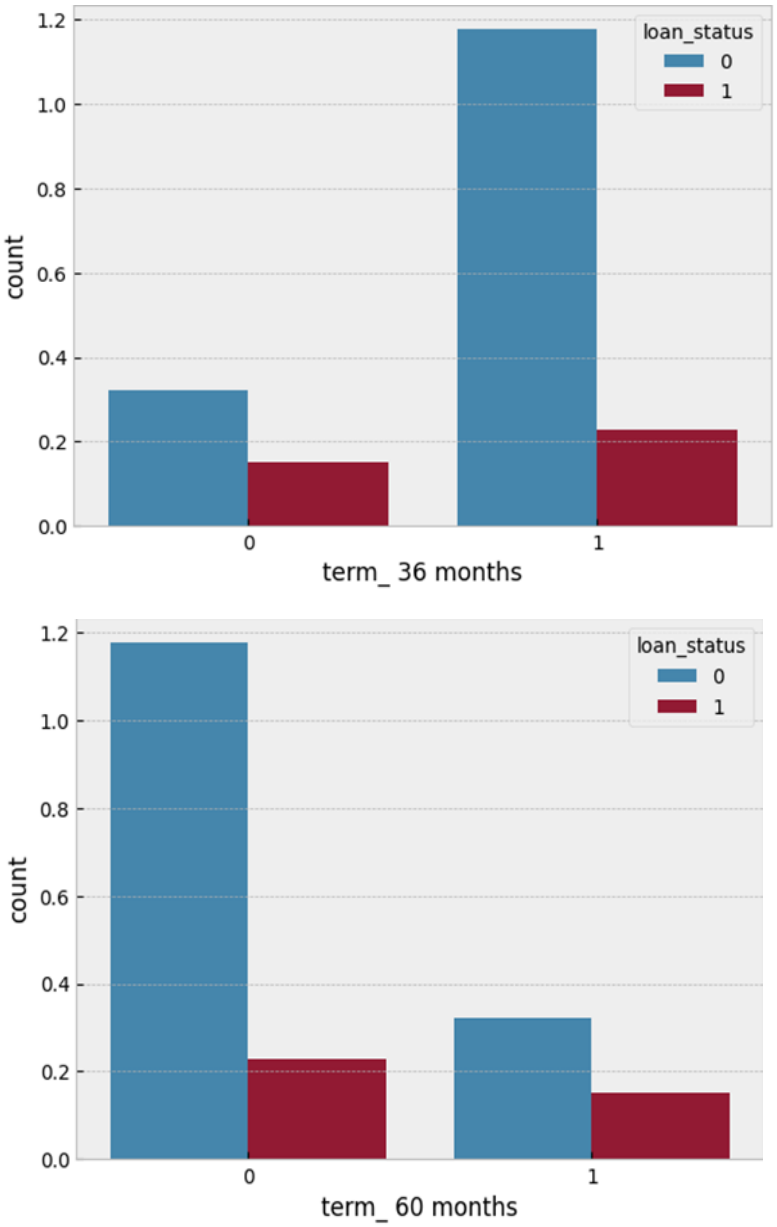


Figura 20. Distribución del estado del préstamo en función del plazo del préstamo.

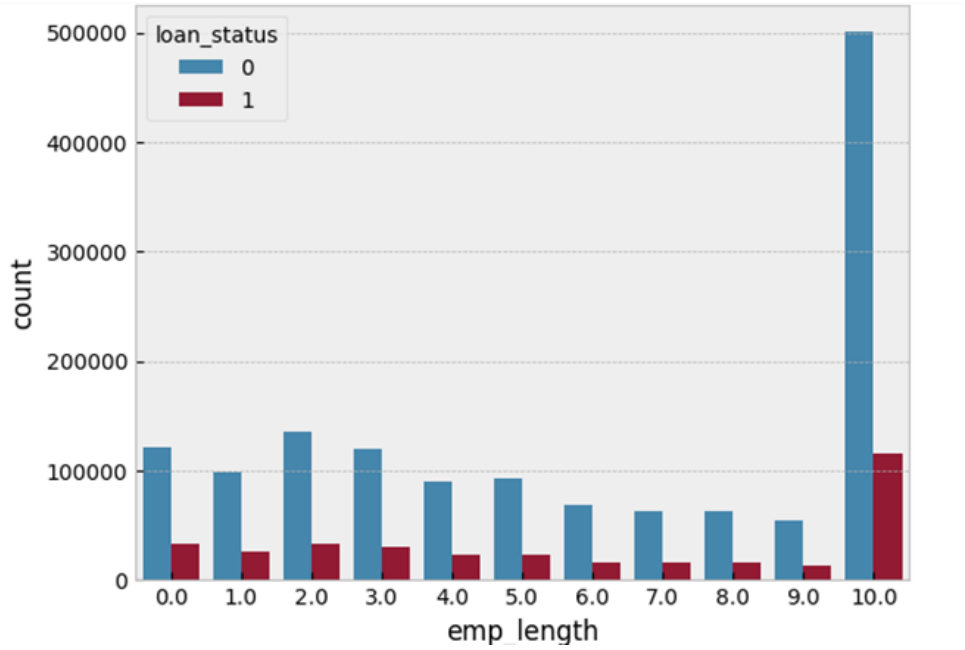


Figura 21. Distribución del estado del préstamo en función de la antigüedad en el trabajo.

Posteriormente, se realizó la selección del modelo de ML, que en este caso de uso fue XGBoost, y se optimizó mediante un grid search para encontrar los mejores hiperparámetros. La comparación de modelos se muestra en la Tabla 5

Tabla 5. AUC de los distintos modelos considerados

	without grid search		with grid search		Smote		Oversampling		Undersampling	
MODELS	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
XGBoost	0.73	0.794	**0.731**	**0.795**	0.729	0.793	0.662	0.72	0.663	0.744
Ada Boost Classifier	0.715	0.793	0.721	0.791	0.699	0.78	0.642	0.715	0.714	0.645
Logistic Regression	0.714	0.792	0.715	0.798	0.679	0.752	0.642	0.598	0.71	0.658
?andom Forest Classifie	0.711	0.793	0.714	0.791	0.71	0.788	0.713	0.785	0.714	0.651
DecisionTreeClassifier	0.694	0.788	0.703	0.787	0.692	0.779	0.551	0.697	0.572	0.574
GaussianNB	0.687	0.308	0.687	0.272	0.643	0.401	0.657	0.409	0.658	0.536
ılanced Bagging Classifi	0.657	0.714	0.713	0.651	0.668	0.711	0.6	0.766	0.665	0.658
KNeighbors Classifier	0.642	0.768	0.672	0.785	0.569	0.588	0.551	0.569	0.568	0.548
Deep Learning	0.542	0.79	0.725	0.793	0.704	0.778	0.563	0.783	0.62	0.215

La configuración de los parámetros del modelo seleccionado fue la siguiente:

- Suma mínima de peso de instancia (hessian): 5.
- Profundidad máxima de un árbol: 6.
- Probabilidad de salida: 'binary: logistic'.

- Gamma: 10. Este valor representa la reducción mínima de pérdida requerida para realizar una partición adicional en un nodo hoja del árbol; el algoritmo se mostrará más cauteloso a medida que se incremente el valor de gamma.

El conjunto de datos se dividió en un 70% de entrenamiento y un 30% de prueba. Asimismo, desarrollamos un modelo XGBoost, que exhibe un AUC de 0,731 en la predicción de incumplimiento temprano en préstamos P2P. La Figura 22 presenta la curva ROC en la prueba del modelo construido. (Krzanowski & Hand, 2009). ROC curves for continuous data. CRC Press).

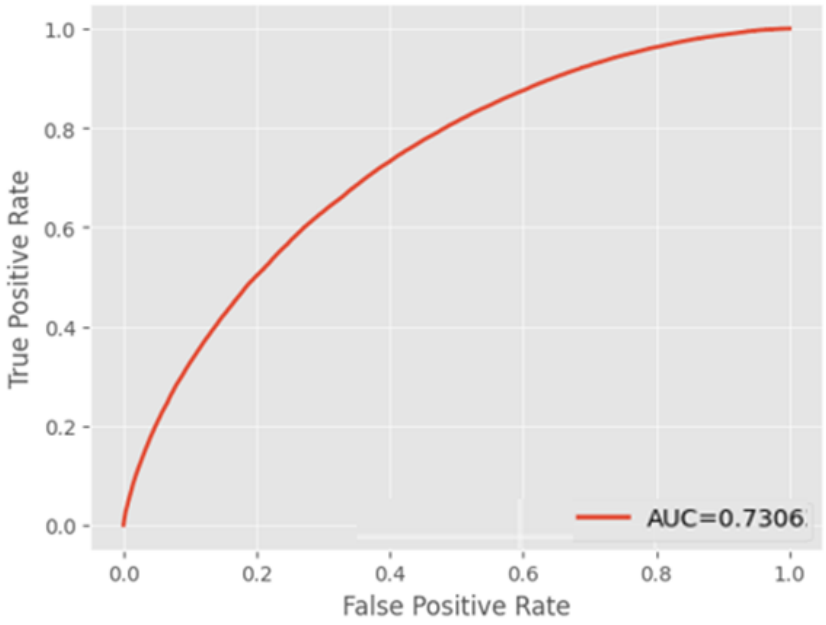


Figura 22. Curva ROC.

Para mejorar su interpretabilidad, empleamos un modelo surrogado, (R^2 en entrenamiento = 0,806, R^2 en prueba = 0,808), que nos proporciona un conjunto de reglas que explican las predicciones del modelo XGBoost. Esto nos permite comprenderlas desde una perspectiva empresarial, por lo que, conocemos cualitativamente la razón que hay detrás de ellas y obtenemos una explicación comprensible para las predicciones del modelo de caja negra. Posteriormente, aplicamos los pasos 7 y 8 del algoritmo explicado anteriormente en esta sección. Primero, transformamos las variables independientes numéricas y variable dependiente con la función Rank para representarla en el intervalo [0, 1], y obtenemos

las reglas de la Tabla 6, que es un paso previo necesario para la transformación a un valor de 2-tuplas (ver Figura 314).

Tabla 6. Reglas surrogadas obtenidas para el modelo XGBoost.

Regla	Antecedente	Consecuencia (entonces)
1	IF (term_ 60 months = 0) and (application_type ≤ 1) and (delinq_2yrs ≤1) and (inq_last_6mths ≤ 0) and (bc_util ≤ 0,5) and (acc_open_past_24mths ≤ 6)	Predicción = 0,00
2	IF (term_ 60 months = 0) and (application_type ≤ 1) and (delinq_2yrs ≤ 1) and (inq_last_6mths ≤ 0) and (bc_util > 0,5)	Predicción = 0,125
3	IF (term_ 60 months = 0) and (application_type ≤ 1) and (delinq_2yrs ≤ 1) and (inq_last_6mths ≤ 0) and (bc_util ≤ 0,5) and (acc_open_past_24mths > 6)	Predicción = 0,250
4	IF (term_ 60 months = 0) and (application_type ≤ 1) and (delinq_2yrs ≤ 1) and (inq_last_6mths > 0) and (acc_open_past_24mths ≤ 6)	Predicción = 0,375
5	IF (term_ 60 months = 0) and (application_type ≤ 1) and (delinq_2yrs > 1)	Predicción = 0,5
6	IF (term_ 60 months = 0) and (application_type ≤ 1) and (delinq_2yrs ≤ 1) and (inq_last_6mths > 0) and (acc_open_past_24mths > 6)	Predicción = 0,625
7	IF (term_ 60 months = 1) and (grade ≤ 4 and (application_type ≤ 1) and (inq_last_6mths ≤ 0)	Predicción = 0,750
8	IF (term_ 60 months = 1) and (grade ≤ 4) and (application_type ≤ 1) and (inq_last_6mths > 0)	Predicción = 0,875
9	IF (term_ 60 months = 1) and (grade > 4) and (application_type > 1)	Predicción = 1,0

Definimos para variable x los correspondientes conjuntos de etiquetas S_j (sección 4.2) donde:

S se define como Figura 14.

Δs se define como ecuación (1).

Utilizando un modelo de 2-tuplas, obtenemos una interpretación lingüística de la variable original que es completamente precisa, sin pérdida de información. Siguiendo este proceso, generamos las reglas que se muestran en la Tabla 7.

Tabla 7. Reglas surrogadas obtenidas para el modelo XGBoost con representación lingüística de 2 tuplas.

Regla	Antecedente	Consecuencia (entonces)
1	IF Δ_s (term_ 60 months = 0) and Δ_s (application_type ≤ 1) and Δ_s (delinq_2yrs ≤ 1) and Δ_s (inq_last_6mths ≤ 0) and Δ_s (bc_util ≤ Medium) and Δ_s (acc_open_past_24mths ≤ 6)	Predicción = (Muy bajo)
2	IF Δ_s (term_ 60 months = 0) and Δ_s (application_type ≤ 1) and Δ_s (delinq_2yrs ≤ 1) and Δ_s (inq_last_6mths ≤ 0) and Δ_s (bc_util > Medium)	Predicción = (Muy bajo, 0,125)
3	IF Δ_s (term_ 60 months = 0) and Δ_s (application_type ≤ 1) and Δ_s (delinq_2yrs ≤ 1) and Δ_s (inq_last_6mths ≤ 0) and Δ_s (bc_util ≤ Medium) and Δ_s (acc_open_past_24mths > 6)	Predicción = (Bajo)
4	IF Δ_s (term_ 60 months = 0) and Δ_s (application_type ≤ 1) and Δ_s (delinq_2yrs ≤ 1) and Δ_s (inq_last_6mths > 0) and Δ_s (acc_open_past_24mths ≤ 6)	Predicción = (Medio, -0,125)
5	IF Δ_s (term_ 60 months = 0) and Δ_s (application_type ≤ 1) and Δ_s (delinq_2yrs > 1)	Predicción = (Medio)
6	IF Δ_s (term_ 60 months = 0) and Δ_s (application_type ≤ 1) and Δ_s (delinq_2yrs ≤ 1) and Δ_s (inq_last_6mths > 0) and Δ_s (acc_open_past_24mths > 6)	Predicción = (Medio, +0,125)
7	IF Δ_s (term_ 60 months = 1) and Δ_s (grade ≤ 4 and Δ_s (application_type ≤ 1) and Δ_s (inq_last_6mths ≤ 0)	Predicción = (Alta)
8	IF Δ_s (term_ 60 months = 1) and Δ_s (grade ≤ 4) and Δ_s (application_type ≤ 1) and Δ_s (inq_last_6mths > 0)	Predicción = (Muy alta, -0,125)
9	IF Δ_s (term_ 60 months = 1) and Δ_s (grade > 4) and Δ_s (application_type > 1)	Predicción = (Muy alta)

En la Regla 1, la solicitud de préstamo (application_type) es realizada por una sola persona, y el plazo de concesión es de tres años (term_ 60 months =0), lo cual lo clasifica como préstamo a corto plazo, reduciendo así el riesgo de exposición de la entidad al impago. El cliente no ha experimentado retrasos superiores a 30 días (delinq_2yrs), con un máximo de uno en los últimos 2 años, lo que indica que no tienen un historial de incumplimientos recurrente. La relación de utilización de crédito se mantiene igual o por debajo del promedio y, el historial crediticio del cliente no presenta consultas en los últimos seis meses (inq_last_6mths), indicando que no ha solicitado ningún producto en ese periodo. Asimismo, el número de cuentas abiertas en los últimos 24 meses (acc_open_past_24mths) es igual o menor a

6, lo que sugiere una búsqueda controlada de crédito en forma de préstamos, tarjetas de crédito u otras formas de financiación. En conjunto, estos factores confirman un uso prudente del crédito y, por lo tanto, una probabilidad muy baja de incumplimiento temprano en relación con el préstamo.

En la Regla 2, se mantienen las condiciones anteriores, pero la relación de utilización de crédito (bc_util) es más alta que la "media", indicando una gestión deficiente del crédito y, por lo tanto, aumentando la probabilidad de incumplimiento temprano, aunque de manera muy baja (+0,125).

En la Regla 3, se conservan las condiciones de la Regla 1, excepto que el cliente tiene más de seis cuentas abiertas en los últimos 24 meses. Un exceso de cuentas abiertas podría señalar un mayor riesgo crediticio, sugiriendo una búsqueda activa de crédito en un corto período, lo que puede implicar dificultades financieras y, en consecuencia, una probabilidad "baja" de incumplimiento.

La Regla 4 establece que, si la solicitud de préstamo es realizada por una sola persona para un plazo de tres años, con un máximo de 30 días de retraso en los últimos dos años, lo que indica un historial de pagos aceptable, y menos de seis cuentas abiertas, pero con alguna consulta en su historial crediticio en un corto período, esto podría indicar una búsqueda de crédito en diversas entidades, sugiriendo dificultades financieras y, por ende, una probabilidad "media" (-0,125) de incumplimiento temprano.

La Regla 5 establece que, si el número de incumplimientos de más de 30 días en los últimos dos años es mayor que 1, la probabilidad de incumplimiento temprano se eleva a "media".

En la Regla 6, si el cliente tiene más de seis cuentas abiertas en los últimos 24 meses y consultas en su historial crediticio en un corto período, esto indica búsqueda activa de crédito debido a dificultades financieras y deterioro del cliente, lo que eleva la probabilidad de incumplimiento temprano a "media" (+0,125).

La Regla 7 destaca que, aunque el cliente que solicita el préstamo no tiene consultas de riesgo en los últimos seis meses y tiene una buena calificación (grade)

(A-B-C-D), el plazo del préstamo solicitado es de cinco años. El incremento de la vida del préstamo hace que aumente la probabilidad de incumplimiento a "alta", ya que la entidad se expone a un posible impago durante un período más prolongado.

En la Regla 8, se mantienen las condiciones de la Regla 7, pero el cliente tiene consultas en su historial en los últimos seis meses, lo que podría indicar búsqueda activa de crédito. Se interpreta como un riesgo potencial de sobreendeudamiento, por lo que la probabilidad de incumplimiento es "muy alta" (-0,125).

Finalmente, la Regla 9 establece que, si el préstamo solicitado tiene un plazo largo de 5 años, la calificación crediticia es considerada de alto riesgo (E-F-G) y la solicitud es conjunta (dos solicitantes), esto implica mayor riesgo, especialmente si uno de los solicitantes tiene un historial crediticio débil, por lo que la probabilidad de incumplimiento temprano es "muy alta".

Es crucial resaltar que el modelo de 2-tuplas posibilitó, por un lado, una buena interpretación lingüística del modelo XGBoost y, por otro, permitió precisión en la interpretación de las reglas, facilitando la subdivisión de la predicción de incumplimiento temprano en categorías como Muy Baja, Baja, Media, Alta y Muy Alta, con distintas traducciones simbólicas que permiten distinguir la gravedad de cada caso.

Hemos construido un modelo XGBoost que predice el impago temprano con gran precisión, y hemos aplicado la metodología explicada en la sección 5.1 construyendo un árbol surrogado y transformándolo en un 2-tuplas. Las reglas finalmente obtenidas tienen interpretabilidad lingüística sin pérdida de precisión.

Por todo ello, el proceso de KDD es fundamental para convertir datos en información útil y conocimiento que pueda beneficiar a las organizaciones y a la toma de decisiones. Sus etapas representan un ciclo continuo, ya que la retroalimentación de las etapas posteriores puede influir en las anteriores.

5.3. Caso de uso en ciudad inteligente. Predicción de demanda de pasajeros (regresión)

Para el caso de uso de Ciudades Inteligentes, nuestro objetivo fue desarrollar un modelo que permitiera predecir y entender la demanda real de pasajeros según los datos proporcionados por la Empresa Municipal de Transportes (EMT). Esta información es crucial para la empresa, ya que le permite conocer la demanda de pasajeros en cualquier línea de autobús de la ciudad en un día determinado, pero igual de importante es comprender por qué se genera esta demanda. De esta manera, la empresa podría tomar decisiones estratégicas relacionadas con su servicio de transporte, basándose en el conocimiento de lo que se espera y de los factores y reglas que influyen en esta demanda.

Los datos utilizados en este estudio se corresponden con la línea 1 de autobuses, incluyendo toda la flota de autobuses, durante el período comprendido entre enero de 2015 y abril de 2017.

Estos se proporcionaron en forma de archivos planos en formato CSV y contenían eventos registrados cada vez que un pasajero validaba su billete o tarjeta de viaje en la máquina. Cada evento incluía información sobre la fecha, la hora, el ID del autobús, el ID de la parada, la dirección del autobús y el tipo de billete utilizado. Dada la gran cantidad de datos transaccionales proporcionados, se optó por utilizar plataformas de Big Data. Por lo tanto, estos datos se cargaron en una plataforma Hadoop dentro de un clúster, con el objetivo de disponer de una gestión coherente e integrada de los datos para su posterior análisis. En este proceso, se utilizaron herramientas como Hive para la inserción de datos procedentes de archivos CSV en la base de datos Hadoop, y Spark para el procesamiento en paralelo de los datos ya insertados en la base de datos Hadoop.

Las variables de salida incluidas en el análisis son las siguientes:

Fecha: Este campo representa la fecha en formato año, mes y día (por ejemplo, 20160101).

Pasajeros: Número de pasajeros en la franja horaria seleccionada.

Para la comprensión de los datos, se llevó a cabo un análisis del comportamiento de los mismos de manera mensual, semanal y horaria. Se observó que los meses de febrero y diciembre presentan la mayor demanda, mientras que agosto registra la menor. Los lunes experimentan la mayor afluencia de pasajeros, mientras que los jueves presentan el menor número de pasajeros. Además, se identificaron patrones de demanda por franjas horarias (0-6, 6-12, 12-18 y 18-24), donde los picos se localizan al mediodía y al atardecer, mientras que la hora del almuerzo muestra una disminución en el número de pasajeros. En promedio, se registraron 3.065,6 pasajeros.

En la Figura 23 se presenta una serie temporal que ilustra el número diario de pasajeros durante la tarde, mientras que la Figura 24 muestra el histograma correspondiente a esta variable.

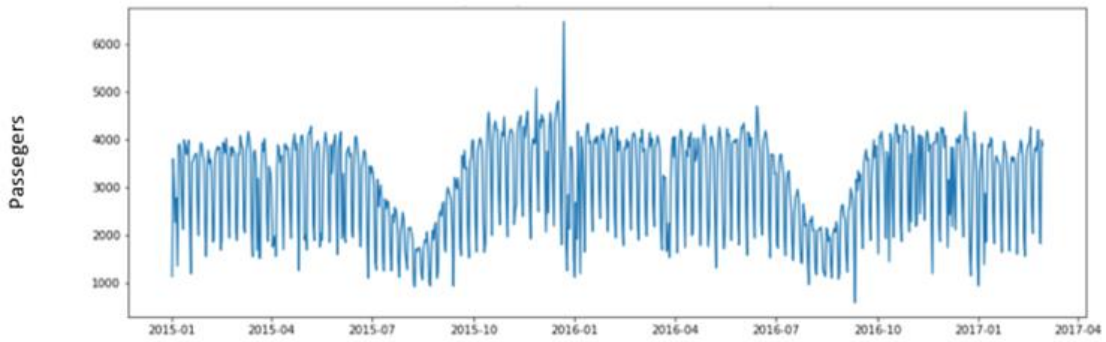


Figura 23. Número diario de pasajeros por la tarde.

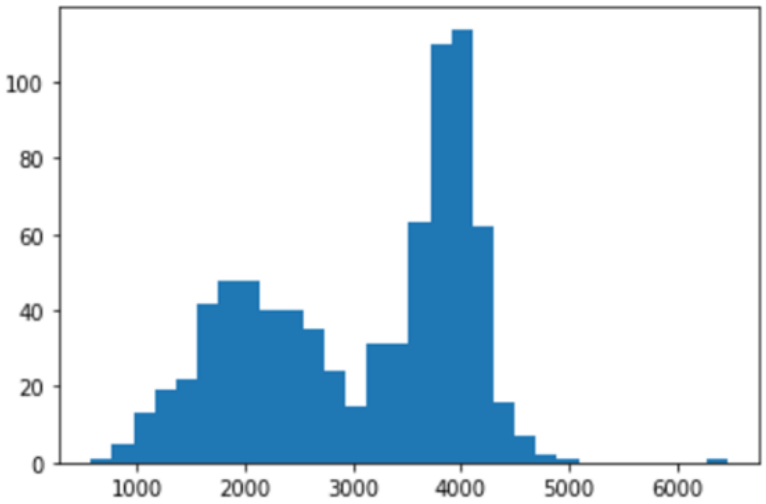


Figura 24. Histograma de pasajeros.

Tras la integración de diversas fuentes de datos, se han identificado las siguientes variables:

- Fecha: Este campo refleja la fecha en un formato que comprende año, mes y día.

- Mes: Representa el mes del año, siendo enero igual a 1, febrero a 2, y así sucesivamente.

- Día festivo: Se trata de una variable binaria que toma el valor 1 si corresponde a un día festivo y 0 en caso contrario.

- Día de la semana: Indica el día de la semana, donde lunes se codifica como 0, martes como 1 y así sucesivamente.

- Pasajeros: Esta variable numérica refleja el número de pasajeros en el contexto de estudio.

Para la limpieza de los datos y preprocesado, representamos la línea 1 en función de franjas horarias y si el día corresponde a un día laborable o de fin de semana. Durante este proceso, identificamos la presencia de valores atípicos o errores, como los casos de pasajeros en autobuses diurnos en horarios nocturnos, y

viceversa. Estos casos se eliminaron de la base de datos. Para las líneas nocturnas, conservamos únicamente los valores que correspondían al horario entre las 23:00h y las 07:00h.

Nuestra base de datos se compone de dos fuentes principales: los datos generados por los autobuses, a los que llamamos "fuentes internas", y todos los datos que no provienen de los autobuses, que denominamos "fuentes externas".

En primer lugar, cargamos los datos en formato CSV u otros formatos en el sistema de archivos distribuidos Hadoop, para su posterior inserción en nuestra base de datos. Con este propósito, se desarrollaron dos secuencias de comandos en lenguaje Bash encargadas de cargar los datos desde el propio clúster al sistema HDFS (en inglés Hadoop Distributed File System). Estas secuencias de comandos se dividieron en función del origen de los datos, diferenciando entre fuentes internas y fuentes externas.

Las fuentes internas provienen de las máquinas de los autobuses, que registran la entrada de pasajeros. Cada vez que un pasajero introduce su billete o tarjeta de viaje en la máquina, se genera una nueva entrada en la base de datos, registrándose como un nuevo viaje. Finalmente, obtuvimos el calendario de días festivos de la Comunidad de Madrid a través de la fuente http://www.calendarioslaborales.com (consultada el 28 de febrero de 2022).

Dado que algunos modelos son sensibles a la escala de los datos de entrada, se procedió a realizar una transformación de escala dividiendo la variable a predecir entre 5000.

Para todos los modelos, excepto LSTM, se realizó una codificación tipo "one-hot" de las variables categóricas "day of week" y "month," lo que generó las siguientes variables ficticias:

- "day of week": lunes, martes, miércoles, jueves, viernes, sábado y domingo.

- "month": enero, febrero, marzo, abril, mayo, junio, julio, agosto, septiembre, octubre, noviembre y diciembre.

Para el modelo LSTM, las variables categóricas se procesaron utilizando capas de incrustación (embedding layers). Finalmente, se calcularon las versiones rezagadas de las variables hasta 5 días.

Para nuestro caso de uso, se desarrollaron diversos modelos con el fin de identificar aquel que exhibía el mayor nivel de precisión, y se optó por el modelo LSTM, utilizando el marco de trabajo Keras, (Chollet, 2018). En la Tabla 8 se presenta un análisis comparativo de los distintos modelos desarrollados.

Tabla 8. Comparación de modelos.

	Models	R^2 (test)
1	LSTM	0,89
2	SVM (linear kernel)	0,88
3	XGBoost	0,87
4	Random Forest	0,87
5	SVM (RBF kernel)	0,80
6	Shallow Dense Neural Network	0,80
7	SVM (polynomial kernel)	0,75

Los componentes de una red LSTM consisten en módulos con compuertas internas diversas que regulan la transferencia de información, como se ilustra en la Figura 13. Estas compuertas determinan si la información fluye a través de ellas o no. Cuando el valor de una compuerta se acerca a 0, la información se bloquea, mientras que un valor cercano a 1 permite el flujo de información a través de la compuerta.

Una de las compuertas (Compuerta de Olvido) se dedica a olvidar y controla el grado en que un valor permanece en la memoria de la neurona, determinando cuánta información sobre el estado pasado debe descartarse.

$$f_t = \sigma\, W_f\, x\, [h_{t-1}, x_t] + b_f \qquad (12)$$

donde f_t es la compuerta de olvido; σ es la función sigmoide; W_f es la matriz de pesos de la compuerta de olvido; h_{t-1} es la salida de la celda LSTM anterior; x_t es la entrada; b_f es el sesgo de la compuerta.

La compuerta de entrada, i_t, determina cuánta información sobre el estado actual de la red debe almacenarse en el estado interno.

$$i_t = \sigma\,(W_i\, x\, [h_{t-1}, x_t] + b_i) \tag{13}$$

La compuerta de salida, o_t, determina cuánta información del estado interno actual debe enviarse al estado externo.

$$o_t = \sigma\,(W_o\, x\, [h_{t-1}, x_t] + b_o) \tag{14}$$

El nodo "tanh" puede considerarse como la función de activación de la capa oculta, y su salida es:

$$C_t = tanh\,(W_C\, x\, [h_{t-1}, x_t] + b_C) \tag{15}$$

En el modelo LSTM final, utilizamos una configuración con 5 neuronas LSTM, un tamaño de ventana de 5, un tamaño de lote de 32 y 2000 épocas de entrenamiento. Cada una de las variables de entrada categóricas (día de la semana y mes) se procesa mediante una capa de incrustación.

Finalmente, el conjunto de datos se divide en un conjunto de entrenamiento (80%, enero de 2015 - septiembre de 2016) y un conjunto de prueba (20%, septiembre de 2016 - febrero de 2017). En la Figura 25, podemos ver la predicción día a día de un rango de días en el conjunto de prueba, y observar que la predicción se ajusta bastante bien (R^2 entrenamiento = 0,92, R^2 prueba = 0,89).

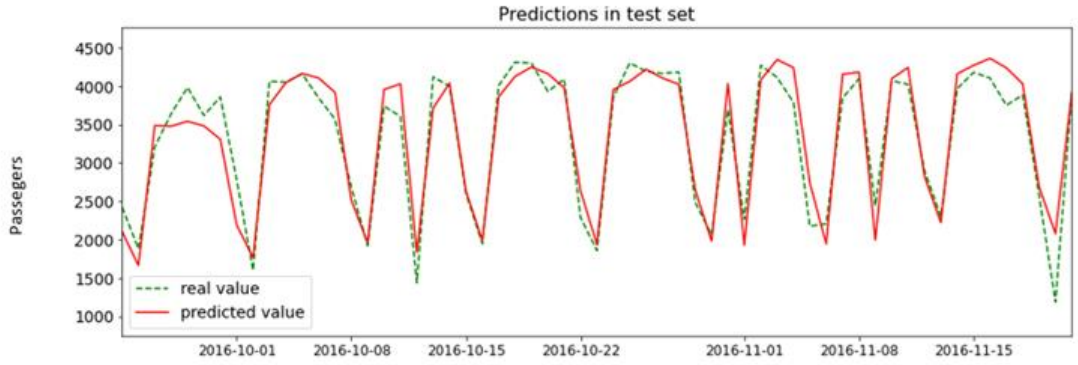


Figura 25. Ampliación de la predicción en el conjunto de pruebas. Verde discontinuo: valor real. Rojo sólido: predicción.

Hemos desarrollado un modelo LSTM para la predicción de la demanda de pasajeros en autobuses, alcanzando una notable precisión. Posteriormente, aplicamos la metodología detallada en la sección 5.1, que implica la creación de un árbol surrogado y su transformación en un modelo de 2-tuplas.

De este modo, obtenemos las reglas del árbol surrogado que se presentan en la Tabla 9.

Tabla 9. Reglas surrogadas obtenidas para el modelo LSTM.

Regla	**Antecedente**	**Consecuencia (entonces)**
1	IF (sat = 0) and (sun = 0) and (holiday = 0) and (jul = 0) and (aug = 0) and (sep = 0)	Predicción = 3907
2	IF (sat = 0) and (sun = 0) and (holiday = 0) and (sep = 1)	Predicción = 3150
3	IF (sat = 0) and (sun = 0) and (holiday = 0) and (jul = 1)	Predicción = 2694
4	IF (sat = 0) and (sun = 0) and (holiday = 1) and (aug = 0)	Predicción = 1819
5	IF (sat = 0) and (sun = 0) and (aug = 1)	Predicción = 1968
6	IF (sat = 1) and (y (-1) ≤ 3110)	Predicción = 1627
7	IF (sat = 1) and (y (-1) > 3110)	Predicción = 2438
8	IF (sun = 1) and (y (-1) ≤ 1832)	Predicción = 1233
9	IF (sun = 1) and (y (-1) > 1832)	Predicción = 1874

Como se discutió en la sección 4.2, si deseamos mejorar la interpretabilidad de la variable número de pasajeros (variable de predicción), podemos representarla con su correspondiente valor de 2-tuplas. Con este propósito, aplicamos los pasos 7 y 8 del algoritmo explicado anteriormente en la sección 5.1. Primero, transformamos esta variable con la típica transformación lineal min-max para representarla en el intervalo [0, 1], que es un paso previo necesario para la transformación a un valor de 2-tuplas (ver Figura 3). Los resultados de estas transformaciones se muestran en la Tabla 10:

Tabla 10. Representación de 2 tuplas de la variable objetivo predicha.

Predicción	**Predicción Min-Max escala**	**Predicción 2-Tuple Representación**
1233	0	Muy bajo
1627	0,1473	(Bajo, −0,103)
1819	0,2191	(Bajo, -0,031)
1832	0,2240	(Bajo, -0,026)
1874	0,2397	(Bajo, -0,01)
1968	0,2749	(Bajo, 0,025)
2438	0,4506	(Medio, -0,049)
2694	0,5464	(Medio, 0,046)
3110	0,7019	(Alta, -0,048)
3150	0,7169	(Alta, -0,033)
3907	1	Muy alta

Una vez obtenida la representación lingüística de las variables numéricas continuas mediante un proceso de fuzzificación, se sustituyeron los valores de dichas variables en las reglas correspondientes. Las variables independientes también se

reemplazaron por sus respectivos valores de 2-tuplas con la finalidad de mejorar la interpretabilidad lingüística de las reglas. Después de este proceso, obtenemos las reglas que se presentan en la Tabla 11. Cabe destacar que el uso del modelo de 2-tuplas permite la representación de la variable original de manera lingüística pero completamente precisa, sin pérdida de información.

Tabla 11. Reglas surrogadas obtenidas para el modelo LSTM con representación lingüística de 2 tuplas.

Regla	Antecedente	Consecuencia (entonces)
1	IF (sat = 0) and (sun = 0) and (holiday = 0) and (jul = 0) and (aug = 0) and (sep = 0)	Predicción = (Muy alta, 0)
2	IF (sat = 0) and (sun = 0) and (holiday = 0) and (sep = 1)	Predicción = (Alta, -0,033)
3	IF (sat = 0) and (sun = 0) and (holiday = 0) and (jul = 1)	Predicción = (Medio, 0,046)
4	IF (sat = 1) and (y (-1) > (High, -0,048))	Predicción = (Medio, -0,049)
5	IF (sat = 0) and (sun = 0) and (holiday = 1) and (aug = 0)	Predicción = (Bajo, -0,031)
6	IF (sat = 0) and (sun = 0) and (aug = 1)	Predicción = (Bajo, 0,025)
7	IF (sat = 1) and (y (−1) ≤ (High, -0,048))	Predicción = (Bajo, -0,103)
8	IF (sun = 1) and (y (-1) > (Low, -0,026))	Predicción = (Bajo, -0,01)
9	IF (sun = 1) and (y (-1) ≤ (Low, -0,026))	Predicción = (Muy bajo, 0)

Por todo ello, teniendo en cuenta las reglas extraídas y la fuzzificación de las variables de interés, se interpreta el modelo surrogado fuzzificado. Esto proporciona una explicación más comprensible y lingüísticamente interpretable de cómo las características del conjunto de datos afectan en las predicciones del modelo de caja negra.

Este enfoque de Molnar busca ofrecer una interpretación más accesible y comprensible de modelos complejos y opacos, mejorando así la confianza y comprensión de los resultados del modelo original.

Las reglas surrogadas resultantes para el modelo LSTM, expresadas en forma lingüística de 2-tuplas que se presentan en la Tabla 11 están sujetas a las siguientes interpretaciones:

Regla 1: Indica una demanda muy alta en días no veraniegos que no son ni fines de semana ni días festivos, denotando días laborables típicos.

Regla 2: Indica una demanda alta en días de septiembre que no son fines de semana ni días festivos, un fenómeno observado en Madrid debido al reinicio de la actividad laboral y al comienzo del año académico.

La demanda de pasajeros de nivel medio se clasifica bajo las siguientes condiciones:

Regla 3: Designa días de julio, que no son fines de semana ni días festivos, correlacionado con una disminución en la demanda de pasajeros a medida que los estudiantes y los trabajadores inician sus vacaciones desde la segunda quincena del mes.

Regla 4: sábados, cuando la demanda de pasajeros en la víspera fue muy alta. Una gran parte de la población no trabaja los sábados por la tarde. En esos días, el transporte público suele utilizarse para actividades de ocio, por lo que la demanda de pasajeros es menor que en la víspera.

La demanda de pasajeros de nivel bajo se clasifica bajo las siguientes condiciones:

Regla 5: Días festivos que no son fines de semana ni días de agosto, sugiriendo una preferencia por quedarse en casa o viajar a otros lugares durante estos periodos.

Regla 6: Designa días de agosto que no son fines de semana, un período marcado por el cierre de empresas, vacaciones obligatorias para los trabajadores y una disminución de la actividad estudiantil, contribuyendo a una demanda de pasajeros reducida.

Regla 7: sábados en los que la demanda de pasajeros en la víspera fue alta/media. Misma interpretación que en la Regla 4.

Regla 8: Identifica los domingos donde la demanda de pasajeros en el día anterior no se clasifica como baja o muy baja, reflejando el cierre del comercio de calle y una demanda disminuida en consecuencia.

La demanda de pasajeros de nivel muy bajo se clasifica bajo las siguientes condiciones:

Regla 9: Domingos con una demanda de pasajeros en el día anterior que fue muy baja, análoga a las condiciones descritas en la Regla 8.

El modelo de 2-tuplas proporciona una efectiva interpretabilidad lingüística del modelo LSTM, junto con precisión en la interpretación de las reglas. Notablemente, las Reglas 5, 6, 7 y 8 indican una previsión de baja demanda, cada una con traducciones simbólicas diferenciadoras que facilitan distinciones matizadas en los niveles de demanda previstos.

Por lo tanto, hemos conseguimos un modelo LSTM que predice la demanda de pasajeros, como se muestra en la Figura 25, con una precisión notable (R^2 entrenamiento = 0,92, R^2 prueba = 0,89)

El modelo LSTM puede interpretarse mediante nueve reglas desde el punto de vista empresarial con una buena precisión (R^2 entrenamiento = 0,95, R^2 prueba = 0,95), por lo que obtenemos una explicación comprensible para las predicciones del modelo de caja negra.

El proceso KDD juega un rol esencial en la transformación de datos en información valiosa y conocimiento, que puede beneficiar a las organizaciones y a la toma de decisiones. Las etapas del KDD forman un ciclo continuo.

6. Conclusiones y trabajo futuro

El planteamiento abordado en la línea de investigación de predicción del incumplimiento temprano para préstamos P2P representa un desafío crucial para instituciones financieras y plataformas Fintech. Se centra en la obtención de interpretabilidad lingüística de un modelo de caja negra que estima la probabilidad de incumplimiento temprano en préstamos. Esta aproximación busca facilitar la adopción de medidas preventivas y la búsqueda de soluciones viables para ambas partes antes de que se materialice un incumplimiento significativo. Con tal propósito, hemos empleado un modelo como XGBoost, reconocido por ser caja negra. La interpretabilidad resulta crucial para garantizar la transparencia, trazabilidad y comprensión de cada decisión en la concesión de créditos, aspecto relevante para los reguladores sobre la calidad de los modelos.

En la literatura, encontramos trabajos enfocados en la predicción de incumplimientos y en la búsqueda de métodos para hacer que estos modelos sean interpretables. Hasta donde alcanza nuestro conocimiento, no existe un trabajo específicamente orientado al problema de la predictibilidad e interpretabilidad del incumplimiento temprano en préstamos P2P, como se presenta en este trabajo.

En este estudio, hemos desarrollado un algoritmo XGBoost capaz de predecir el incumplimiento temprano de préstamos y lo hemos validado empíricamente con un conjunto de préstamos otorgados entre 2007 y 20203T, utilizando una amplia cantidad de datos. Finalmente, hemos logrado un modelo de caja negra interpretable gracias al modelo surrogado considerado caja blanca y al modelo lingüístico 2-tuplas difuso. La aplicación del modelo de 2-tuplas (técnica difusa) ha mejorado la interpretabilidad lingüística del modelo XAI generado sin sacrificar precisión, y lo hemos integrado en la metodología KDD.

No hemos encontrado en la literatura el uso conjunto de ambos enfoques.

Desde una perspectiva económica, la interpretación del incumplimiento temprano en modelos no interpretables por humanos permite a las instituciones financieras detectar este incumplimiento antes de que ocurra. Esto posibilita la toma de medidas preventivas, como la ampliación del plazo del préstamo o la concesión de un período de gracia, buscando soluciones viables para ambas partes antes de que se produzca el incumplimiento. Por otro lado, se puede explicar al regulador la denegación de un préstamo solicitado por un cliente, permitiendo el uso de modelos considerados cajas negras que son más precisos, lo que implica un ahorro en provisiones y una mejora del capital.

Finalmente, esta metodología podría aplicarse en futuros trabajos para abordar la explicabilidad del incumplimiento en otros tipos de préstamos que no sean P2P.

En nuestro segundo caso de uso sobre Ciudades Inteligentes, el problema abordado es de gran importancia, ya que se centra en hacer más eficiente y controlar la red de transporte a través de la predicción de la demanda de pasajeros en autobús. La mayoría de los algoritmos de ML utilizados para series temporales en este contexto son considerados cajas negras, es decir, proporcionan buenas predicciones, pero no son interpretables por humanos. Sin embargo, la interpretabilidad del modelo es crucial para la transparencia, la trazabilidad y la auditabilidad, especialmente en la asignación de recursos públicos.

En la literatura, existen trabajos orientados a la predicción y comprensión del flujo de tráfico. Hasta donde sabemos, no hay ningún trabajo específicamente orientado al problema de la predictibilidad e interpretabilidad de la demanda de usuarios de autobuses públicos, como el que presentamos aquí.

En este estudio, hemos desarrollado un algoritmo LSTM, considerado caja negra, capaz de predecir la demanda de pasajeros en la red de transporte público, validándolo empíricamente con un caso de uso y utilizando una gran cantidad de datos de fuentes internas y externas. Finalmente, hemos logrado que el modelo sea interpretable mediante el uso de un modelo surrogado considerado caja blanca y un

modelo lingüístico difuso de 2-tuplas. La aplicación de la técnica difusa llamada modelo 2-tuplas ha mejorado la interpretabilidad lingüística del modelo de XAI sin perder precisión y lo hemos integrado en la metodología KDD. No hemos encontrado en la literatura el uso conjunto de ambos enfoques.

Desde un punto de vista económico, la predicción de la demanda de pasajeros puede ser utilizada para la planificación óptima de recursos, generando ahorros en energía y costos para las empresas de transporte público. Además, esta metodología podría aplicarse en trabajos futuros para predecir la demanda de pasajeros en otros modos de transporte, como el aéreo, ferroviario o marítimo.

Sin embargo, a pesar de que el modelo de 2-tuplas ha mejorado la interpretabilidad en variables numéricas, facilitando la generación de etiquetas lingüísticas precisas, su aplicabilidad presenta limitaciones en las variables categóricas, carecen de un orden natural que permita asignar etiquetas lingüísticas con una relación gradual. En estos casos, la ausencia de una estructura ordinal impide la aplicación efectiva del modelo de 2-tuplas.

7. Bibliografía

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intelligent systems in accounting, finance and management, 18(2-3), 59-88.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.

Aghaeipoor, F., Javidi, M. M., & Fernández, A. (2021). IFC-BD: an interpretable fuzzy classifier for boosting explainable artificial intelligence in big data. IEEE Transactions on Fuzzy Systems, 30(3), 830-840. DOI: 10.1109/TFUZZ.2021.3049911.

Aghaeipoor, F., Sabokrou, M., & Fernández, A. (2023). Fuzzy rule-based explainer systems for deep neural networks: From local explainability to global understanding. IEEE Transactions on Fuzzy Systems, 31(9), 3069-3080.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10),1340-1347. DOI: 10.1093/bioinformatics/btq134.

Akpan, A. G., Nkubli, F. B., Ezeano, V. N., Okwor, A. C., Ugwuja, M. C., & Offiong, U. (2022). XAI for medical image segmentation in medical decision support systems. Explainable Artificial Intelligence in Medical Decision Support Systems, 137.

Apostolopoulos, I. D., & Groumpos, P. P. (2023). Fuzzy Cognitive Maps: Their Role in Explainable Artificial Intelligence. Applied Sciences, 13(6), 3412. DOI: 10.3390/app13063412.

Arnaut, D., & Bećirović, D. (2023). FinTech innovations as disruptor of the traditional financial industry. In Digital Transformation of the Financial Industry: Approaches and Applications (pp. 233-254). Cham: Springer International Publishing. DOI: 10.1007/978-3-031-23269-5_14.

Arora, N., & Kaur, P. D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. Applied Soft Computing, 86, 105936. DOI: 10.1016/j.asoc.2019.105936.

Artelt, A., & Hammer, B. (2022). Efficient computation of counterfactual explanations and counterfactual metrics of prototype-based classifiers. Neurocomputing, 470, 304-317. DOI: 10.1016/j.neucom.2021.04.129.

Awotunde, J. B., Adeniyi, E. A., Ogundokun, R. O., & Ayo, F. E. (2021). Application of big data with fintech in financial services. In Fintech with artificial intelligence, big data, and blockchain (pp. 107-132). Singapore: Springer Singapore. DOI: 10.1007/978-981-33-6137-9_3.

Bai, Y., Sun, Z., Zeng, B., Deng, J., & Li, C. (2017). A multi-pattern deep fusion model for short-term bus passenger flow forecasting. Applied Soft Computing, 58, 669-680. DOI: 10.1016/j.asoc.2017.05.011.

Bárcena, J. L. C., Ducange, P., Marcelloni, F., Nardini, G., Noferi, A., Renda, A., ... & Virdis, A. (2023). Enabling federated learning of explainable AI models within beyond-5G/6G networks. Computer Communications, 210, 356-375. DOI: 10.1016/j.comcom.2023.07.039.

Barredo-Arrieta, A., Laña, I., & Del Ser, J. (2019, October). What lies beneath: A note on the explainability of black-box machine learning models for road traffic forecasting. In 2019 IEEE Intelligent Transportation Systems Conference (ITSC) (pp. 2232-2237). IEEE. DOI: 10.1109/ITSC.2019.8916985.

Basha, S. A., Elgammal, M. M., & Abuzayed, B. M. (2021). Online peer-to-peer lending: A review of the literature. Electronic Commerce Research and Applications, 48, 101069. DOI: 10.1016/j.elerap.2021.101069.

Bedoui, A., & Lazar, N. A. (2020). Bayesian empirical likelihood for ridge and lasso regressions. Computational Statistics & Data Analysis, 145, 106917. DOI: 10.1016/j.csda.2020.106917.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967. DOI: 10.1007/s10462-020-09896-5.

Berrar, D. (2019). Bayes' Theorem and Naive Bayes Classifier. DOI: 10.1016/B978-0-12-809633-8.20473-1.

Biedma-Rdguez, C., Gacto, M. J., Anguita-Ruiz, A., Alcalá-Fdez, J., & Alcalá, R. (2022). Transparent but accurate evolutionary regression combining new linguistic fuzzy grammar and a novel interpretable linear extension. International Journal of Fuzzy Systems, 24(7), 3082-3103. DOI: 10.1007/s40815-022-01324-w.

Bologna, G. (2019). A simple convolutional neural network with rule extraction. Applied Sciences, 9(12), 2411.

Bueno, I., Carrasco, R. A., Porcel, C., & Herrera-Viedma, E. (2022). Profiling clients in the tourism sector using fuzzy linguistic models based on 2-tuples. Procedia Computer Science, 199, 718-724. DOI: 10.1016/j.procs.2022.01.089.

Chen, S., Wang, Q., & Liu, S. (2019, June). Credit risk prediction in peer-to-peer lending with ensemble learning framework. In 2019 Chinese Control And Decision Conference (CCDC) (pp. 4373-4377). IEEE, DOI:10.1109/CCDC.2019.8832412.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794, DOI:10.1145/2939672.2939785.

Chen, T., Fang, J., Xu, M., Tong, Y., & Chen, W. (2022). Prediction of Public Bus Passenger Flow Using Spatial–Temporal Hybrid Model of Deep Learning. Journal of Transportation Engineering, Part A: Systems, 148(4), 04022007. DOI: 10.1061/JTEPBS.00006.

Chen, T. (2021, December). Credit default risk prediction of lenders with resampling methods. In 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) (pp. 123-127). IEEE.

Cheng, X., Khomtchouk, B., Matloff, N., & Mohanty, P. (2018). Polynomial regression as an alternative to neural nets. arXiv preprint arXiv:1806.06850. DOI: 10.48550/arXiv.1806.06850.

Cheng, Z., Lu, R., Wang, Z., Zhang, H., Chen, B., Meng, Z., & Yuan, X. (2020, August). BIRNAT: Bidirectional recurrent neural networks with adversarial training for video snapshot compressive imaging. In European Conference on Computer Vision (pp. 258-275). Cham: Springer International Publishing. DOI: 10.1007/978-3-030-58586-0_16.

Chicco, D. (2021). Siamese neural networks: An overview. Artificial neural networks, 73-94. DOI: 10.1007/978-1-0716-0826-5_3.

Chishti, S. (2016). How peer to peer lending and crowdfunding drive the fintech revolution in the UK. Banking beyond banks and money: A Guide to Banking Services in the Twenty-First Century, 55-68, DOI:10.1007/978-3-319-42448-4_4.

Cho, P., Chang, W., & Song, J. W. (2019). Application of instance-based entropy fuzzy support vector machine in peer-to-peer lending investment decision. IEEE Access, 7, 16925-16939, DOI:10.1109/ACCESS.2019.2896474.

Chollet, F. Deep Learning with Python; Manning: Shelter Island, NY, USA, 2018.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. Journal of the American Society for information Science and Technology, 63(8), 1609-1630. DOI: 10.1002/asi.22688.

Corchado, J. M., Chamoso, P., Hernández, G., Gutierrez, A. S. R., Camacho, A. R., González-Briones, A., ... & Omatu, S. (2021). Deepint. net: A rapid deployment platform for smart territories. Sensors, 21(1), 236. DOI: 10.3390/s21010236.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. DOI: 10.1007/BF00994018.

Cristóbal, T., Padrón, G., Quesada-Arencibia, A., Alayón, F., de Blasio, G., & García, C. R. (2019). Bus travel time prediction model based on profile similarity. Sensors, 19(13), 2869. DOI: 10.3390/s19132869.

Csiszár, O., Pusztaházi, L. S., Dénes-Fazakas, L., Gashler, M. S., Kreinovich, V., & Csiszár, G. (2023). Uninorm-like parametric activation functions for human-understandable neural models. Knowledge-Based Systems, 260, 110095. DOI: 10.1016/j.knosys.2022.110095.

Elshawi, R., Al-Mallah, M. H., & Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. BMC medical informatics and decision making, 19(1), 1-32. DOI: 10.1186/s12911-019-0874-0.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

Fernández, A., Herrera, F., Cordon, O., del Jesus, M. J., & Marcelloni, F. (2019). Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?. IEEE Computational intelligence magazine, 14(1), 69-81. DOI: 10.1109/MCI.2018.2881645.

Fernández, G., Aledo, J. A., Gamez, J. A., & Puerta, J. M. (2022). Factual and counterfactual explanations in fuzzy classification trees. IEEE Transactions on Fuzzy Systems, 30(12), 5484-5495. DOI: 10.1109/TFUZZ.2022.3179582.

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J. Mach. Learn. Res., 20(177), 1-81.

Fu, X., Ouyang, T., Chen, J., & Luo, X. (2020). Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks. Information Processing & Management, 57(4), 102236, DOI:10.1016/j.ipm.2020.102236.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. Neural computation, 12(10), 2451-2471. DOI: 10.1162/089976600300015015.

Ghenimi, A., Chaibi, H., & Omri, M. A. B. (2017). The effects of liquidity risk and credit risk on bank stability: Evidence from the MENA region. Borsa Istanbul Review, 17(4), 238-248. DOI: 10.1016/j.bir.2017.05.002.

Ghosh, J., & Nag, A. (2001). An overview of radial basis function networks. Radial basis function networks 2: new advances in design, 1-36. DOI: 10.1007/978-3-7908-1826-0_1.

Gkolemis, V., Dalamagas, T., & Diou, C. (2023, April). DALE: Differential Accumulated Local Effects for efficient and accurate global explanations. In Asian Conference on Machine Learning (pp. 375-390). PMLR.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. journal of Computational and Graphical Statistics, 24(1), 44-65. DOI: 10.1080/10618600.2014.907095.

Gummadi, R., & Edara, S. R. (2019). Prediction of passenger flow of transit buses over a period of time using artificial neural network. In Third International Congress on Information and Communication Technology: ICICT 2018, London (pp. 963-971). Springer Singapore. DOI: 10.1007/978-981-13-1165-9_88.

Han, Y., Wang, C., Ren, Y., Wang, S., Zheng, H., & Chen, G. (2019). Short-term prediction of bus passenger flow based on a hybrid optimized LSTM network. ISPRS international journal of geo-information, 8(9), 366. DOI: 10.3390/ijgi8090366.

Hasan, M. M., Hossain, M. M., Rahman, M. M., Azad, A. K. M., Alyami, S. A., & Moni, M. A. (2023). FP-CNN: Fuzzy pooling-based convolutional neural network for lung ultrasound image classification with explainable AI. Computers in Biology and Medicine, 165, 107407. DOI: 10.1016/j.compbiomed.2023.107407.

Havrylchyk, O., & Verdier, M. (2018). The financial intermediation role of the P2P lending platforms. Comparative Economic Studies, 60, 115-130. DOI: 10.1057/s41294-017-0045-1.

Heng, Y. S., & Subramanian, P. (2022, October). A Systematic Review of Machine Learning and Explainable Artificial Intelligence (XAI) in Credit Risk Modelling. In Proceedings of the Future Technologies Conference (FTC) 2022, Volume 1 (pp. 596-614). Cham: Springer International Publishing], DOI:10.1007/978-3-031-18461-1_39.

Herrera, F.; Martínez, L. A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Trans. Fuzzy Syst. 2000, 8, 746–752. Available online: https://ieeexplore.ieee.org/document/890332 (accessed on 28 February 2022).

Hidayat, T., Masyita, D., Nidar, S. R., Ahmad, F., & Syarif, M. A. N. (2021). Early warning early action for the banking solvency risk in the COVID-19 pandemic era: A case study of Indonesia. Economies, 10(1), 6. DOI: 10.3390/economies10010006.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780. DOI: 10.1162/neco.1997.9.8.1735.

Hu, Q., Gois, F. N. B., Costa, R., Zhang, L., Yin, L., Magaia, N., & de Albuquerque, V. H. C. (2022). Explainable artificial intelligence-based edge fuzzy images for COVID-19 detection and identification. Applied Soft Computing, 123, 108966. DOI: 10.1016/j.asoc.2022.108966.

Islam, M. A., Anderson, D. T., Pinar, A. J., Havens, T. C., Scott, G., & Keller, J. M. (2019). Enabling explainable fusion in deep learning with fuzzy integral neural networks. IEEE Transactions on Fuzzy Systems, 28(7), 1291-1300. DOI: 10.1109/TFUZZ.2019.2917124.

Jiao, F., Huang, L., Song, R., & Huang, H. (2021). An improved STL-LSTM model for daily bus passenger flow prediction during the COVID-19 pandemic. Sensors, 21(17), 5950. DOI: 10.3390/s21175950.

Jin, W., Li, P., Wu, W., & Wei, L. (2019). Short-term public transportation passenger flow forecasting method based on multi-source data and shepard interpolating prediction method. In Man-Machine-Environment System Engineering: Proceedings of the 18th International Conference on MMESE 18 (pp. 281-294). Springer Singapore. DOI: 10.1007/978-981-13-2481-9_33.

Kamal, M. S., Dey, N., Chowdhury, L., Hasan, S. I., & Santosh, K. (2022). Explainable AI for Glaucoma Prediction Analysis to Understand Risk Factors in Treatment Planning. *IEEE Transactions on Instrumentation and Measurement, 71*, 1-9. DOI: 10.1109/TIM.2022.3171613.

Keneni, B. M., Kaur, D., Al Bataineh, A., Devabhaktuni, V. K., Javaid, A. Y., Zaientz, J. D., & Marinier, R. P. (2019). Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. IEEE Access, 7, 17001-17016. DOI: 10.1109/ACCESS.2019.2893141.

Kim, J. Y., & Cho, S. B. (2019). Towards repayment prediction in peer-to-peer social lending using deep learning. Mathematics, 7(11), 1041, DOI:10.3390/math7111041.

Kim, J. Y., & Cho, S. B. (2022). Ensemble of diverse deep neural networks with pseudo-labels for repayment prediction in social lending. Science Progress, 105(3), 00368504221124004, DOI:10.1177/00368504221124004.

Ko, P. C., Lin, P. C., Do, H. T., & Huang, Y. F. (2022). P2P Lending Default Prediction Based on AI and Statistical Models. Entropy, 24(6), 801, DOI:10.3390/e24060801.

Kohonen, T. (1995). Learning vector quantization. Self-organizing maps, 175-189. DOI: 10.1007/978-3-642-97610-0_6.

Kohonen, T. (1997, June). Exploration of very large databases by self-organizing maps. In Proceedings of international conference on neural networks (icnn'97) (Vol. 1, pp. PL1-PL6). IEEE. DOI: 10.1109/ICNN.1997.611622.

König, G., Molnar, C., Bischl, B., & Grosse-Wentrup, M. (2021, January). Relative feature importance. In 2020 25th International Conference on Pattern

Recognition (ICPR) (pp. 9318-9325). IEEE. DOI: 10.1109/ICPR48806.2021.9413090.

Kostopoulos, G., Panagiotakopoulos, T., Kotsiantis, S., Pierrakeas, C., & Kameas, A. (2021). Interpretable models for early prediction of certification in MOOCs: A case study on a MOOC for smart city professionals. IEEE Access, 9, 165881-165891. DOI: 10.1109/ACCESS.2021.3134787.

Koyuncugil, A. S., & Ozgulbas, N. (2012). Financial early warning system model and data mining application for risk detection. Expert systems with Applications, 39(6), 6238-6253. DOI: 10.1016/j.eswa.2011.12.021.

Krämer, W., & Sonnberger, H. (2012). The linear regression model under test. Springer Science & Business Media. DOI: 10.1007/978-3-642-95876-2.

Kriebel, J., & Stitz, L. (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. European Journal of Operational Research, 302(1), 309-323. DOI: 10.1016/j.ejor.2021.12.024.

Krzanowski, W. J., & Hand, D. J. (2009). ROC curves for continuous data. Crc Press

Lee, J. W., Lee, W. K., & Sohn, S. Y. (2021). Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers. Expert Systems with Applications, 168, 114411, DOI:10.1016/j.eswa.2020.114411.

Lenz, R. (2016). Peer-to-Peer Lending: Opportunities and Risks. European Journal of Risk Regulation, 7(4), 688-700. DOI:10.1017/S1867299X00010126.

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. Risks, 7(1), 29, DOI:10.3390/risks7010029.

Li, B. (2022). Online Loan Default Prediction Model Based on Deep Learning Neural Network. Computational Intelligence and Neuroscience, 2022. DOI:10.1155/2022/4276253.

Li, H., Cao, Y., Li, S., Zhao, J., & Sun, Y. (2020). XGBoost model and its application to personal credit evaluation. IEEE Intelligent Systems, 35(3), 52-61, DOI:10.1109/MIS.2020.2972533, DOI:10.1109/MIS.2020.2972533.

Li, W., Ding, S., Chen, Y., & Yang, S. (2018). Heterogeneous ensemble for default prediction of peer-to-peer lending in China. Ieee Access, 6, 54396-54406. DOI:10.1109/ACCESS.2018.2810864.

Li, W.; Sui, L.; Zhou, M.; Dong, H. Short-term passenger flow forecast for urban rail transit based on multi-source data. EURASIP J. Wirel. Commun. Netw. 2021, 2021, 9. DOI: 10.1186/s13638-020-01881-4.

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019. DOI: 10.48550/arXiv.1506.00019.

Liu, L., & Chen, R. C. (2017). A novel passenger flow prediction model using deep learning methods. Transportation Research Part C: Emerging Technologies, 84, 74-91. DOI: 10.1016/j.trc.2017.08.001.

Liu, H., Ning, H., Mu, Q., Zheng, Y., Zeng, J., Yang, L. T., ... & Ma, J. (2019). A review of the smart world. Future generation computer systems, 96, 678-691. DOI: 10.1016/j.future.2017.09.010.

Liu, Y., Lyu, C., Liu, X., & Liu, Z. (2020). Automatic feature engineering for bus passenger flow prediction based on modular convolutional neural network. IEEE Transactions on Intelligent Transportation Systems, 22(4), 2349-2358. DOI: 10.1016/j.trc.2017.08.001.

Liu, K., Hu, X., Zhou, H., Tong, L., Widanage, W. D., & Marco, J. (2021). Feature analyses and modeling of lithium-ion battery manufacturing based on random forest classification. IEEE/ASME Transactions on Mechatronics, 26(6), 2944-2955. DOI: 10.1109/TMECH.2020.3049046.

Liu, W., Tan, Q., & Wu, W. (2020). Forecast and early warning of regional bus passenger flow based on machine learning. Mathematical Problems in Engineering, 2020(1), 6625435. DOI: 10.1109/TITS.2020.3004254.

Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic

review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine, 107161. DOI: 10.1016/j.cmpb.2022.107161.

Lundberg, S. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874. DOI: 10.48550/arXiv.1705.07874.

Luo, D., Zhao, D., Ke, Q., You, X., Liu, L., Zhang, D., ... & Zuo, X. (2020). Fine-grained service-level passenger flow prediction for bus transit systems based on multitask deep learning. IEEE Transactions on Intelligent Transportation Systems, 22(11), 7184-7199. DOI: 10.1109/TITS.2020.3002772.

Lv, W., Lv, Y., Ouyang, Q., & Ren, Y. (2022). A bus passenger flow prediction model fused with point-of-interest data based on extreme gradient boosting. Applied Sciences, 12(3), 940. DOI: 10.3390/app12030940.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electronic Commerce Research and Applications, 31, 24-39, DOI:10.1016/j.elerap.2018.08.002.

Maaroof, N., Moreno, A., Valls, A., Jabreel, M., & Romero-Aroca, P. (2023). Multi-Class Fuzzy-LORE: A Method for Extracting Local and Counterfactual Explanations Using Fuzzy Decision Trees. Electronics, 12(10), 2215. DOI: 10.3390/electronics12102215.

Magnussen, S., & Tomppo, E. (2014). The k-nearest neighbor technique with local linear regression. Scandinavian Journal of Forest Research, 29(2), 120-131. DOI: 10.1080/02827581.2013.878744.

Manibardo, E. L., Laña, I., & Del Ser, J. (2021). Deep learning for road traffic forecasting: Does it make a difference?. IEEE Transactions on Intelligent Transportation Systems, 23(7), 6164-6188. DOI: 10.1109/TITS.2021.3083957.

Marques-Silva, J., & Ignatiev, A. (2023). No silver bullet: interpretable ML models must be explained. Frontiers in Artificial Intelligence, 6, 1128212. DOI: 10.3389/frai.2023.1128212.

Marín Díaz, G., Carrasco, R. A., & Gómez, D. (2021). RFID: A fuzzy linguistic model to manage customers from the perspective of their interactions with the contact center. Mathematics, 9(19), 2362. DOI: 10.3390/math9192362.

Monje, L., Carrasco, R. A., Rosado, C., & Sánchez-Montañés, M. (2022). Deep learning XAI for bus passenger forecasting: A use case in Spain. Mathematics, 10(9), 1428. DOI: 10.3390/math10091428.

Moosbauer, J., Herbinger, J., Casalicchio, G., Lindauer, M., & Bischl, B. (2021). Explaining hyperparameter optimization via partial dependence plots. Advances in Neural Information Processing Systems, 34, 2280-2291.

Nagaraj, N., Gururaj, H. L., Swathi, B. H., & Hu, Y. C. (2022). Passenger flow prediction in bus transportation system using deep learning. Multimedia tools and applications, 81(9), 12519-12542. DOI: 10.1007/s11042-022-12306-3.

Neuberger, D., & Rissi, R. (2024). Macroprudential banking regulation: Does one size fit all?. Journal of Banking and Financial Economics, 2014(1), 4-27. DOI: 10.7172/2353-6845.jbfe.2014.1.1.

Nóbrega, Caio, and Leandro Marinho. "Towards explaining recommendations through local surrogate models." Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. 2019. DOI: 10.1145/3297280.3297443.

Flórez, J. J. M., España, G. M., & Cárdenas, R. B. (2008). Evaluating a k-nearest neighbours-based classifier for locating faulty areas in power systems. Ingeniería e Investigación, 28(3), 81-86. DOI: 10.15446/ing.investig.v28n3.15124.

Martin, L., Wittmann, M., & Li, X. (2021). The influence of public transport delays on mobility on demand services. Electronics, 10(4), 379. DOI: 10.3390/electronics10040379.

Mora-Florez, J., Morales-España, G., & Barrera-Cárdenas, R. (2008). Evaluación del clasificador basado en los k vecinos más cercanos para la localización de la zona en falla en los sistemas de potencia. Ingeniería e Investigación, 28(3), 81-86.

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. Expert Systems with Applications, 165, 113986, DOI:10.1016/j.eswa.2020.113986.

Moayedi, H., Mosallanezhad, M., Rashid, A. S. A., Jusoh, W. A. W., & Muazu, M. A. (2020). A systematic review and meta-analysis of artificial neural network application in geotechnical engineering: theory and applications. Neural Computing and Applications, 32, 495-518.

Murray, B. J., Islam, M. A., Pinar, A. J., Anderson, D. T., Scott, G. J., Havens, T. C., & Keller, J. M. (2020). Explainable ai for the choquet integral. IEEE Transactions on Emerging Topics in Computational Intelligence, 5(4), 520-529. DOI: 10.1109/TETCI.2020.3005682.

Namvar, E. (2014). An introduction to peer-to-peer loans as investments. Journal of Investment Management First Quarter, DOI:10.2139/ssrn.2227181.

Niu, K., Zhang, Z., Liu, Y., & Li, R. (2020). Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. Information Sciences, 536, 120-134. DOI:10.1016/j.ins.2020.05.040.

Schober, P., & Vetter, T. R. (2021). Logistic regression in medical research. Anesthesia and analgesia, 132(2), 365. DOI: 10.1213/ANE.0000000000005247.

Sundararajan, M., & Najmi, A. (2020, November). The many Shapley values for model explanation. In International conference on machine learning (pp. 9269-9278). PMLR.

Oh, S. (2019). Feature interaction in terms of prediction performance. Applied Sciences, 9(23), 5191. DOI: doi.org/10.3390/app9235191.

Ouyang, Q., Lv, Y., Ma, J., & Li, J. (2020). An LSTM-based method considering history and real-time data for passenger flow prediction. Applied Sciences, 10(11), 3788. DOI: 10.3390/app10113788.

Parbat, D., & Chakraborty, M. (2020). A python based support vector regression model for prediction of COVID19 cases in India. Chaos, Solitons & Fractals, 138, 109942. DOI: 10.1016/j.chaos.2020.109942.

Pekel, E. (2020). Estimation of soil moisture using decision tree regression. Theoretical and Applied Climatology, 139(3-4), 1111-1119. DOI: 10.1007/s00704-019-03048-8.

Pérez-Rave, J., & González Echavarría, F. (2018). Árboles de clasificación vs regresión logística en el desarrollo de competencias genéricas en ingeniería. Computación y Sistemas, 22(4), 1519-1541. DOI: 10.13053/cys-22-4-2804.

Pathirage, C. S. N., Li, J., Li, L., Hao, H., Liu, W., & Ni, P. (2018). Structural damage identification based on autoencoder neural networks and deep learning. Engineering structures, 172, 13-28. DOI: 10.1016/j.engstruct.2018.05.109.

Pradhan, B., Jena, R., Talukdar, D., Mohanty, M., Sahu, B. K., Raul, A. K., & Abdul Maulud, K. N. (2022). A New Method to Evaluate Gold Mineralisation-Potential Mapping Using Deep Learning and an Explainable Artificial Intelligence (XAI) Model. Remote Sensing, 14(18), 4486. DOI: 10.3390/rs14184486.

Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., ... & Wrede, B. (2020). Explanation as a social practice: Toward a conceptual framework for the social design of AI systems. IEEE Transactions on Cognitive and Developmental Systems, 13(3), 717-728. DOI: 10.1109/TCDS.2020.3044366.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386-408. DOI: 10.1037/h0042519.

Sağ, T., & Abdullah Jalil Jalil, Z. (2021). Vortex search optimization algorithm for training of feed-forward neural network. International Journal of Machine Learning and Cybernetics, 12, 1517-1544. DOI: 10.1007/s13042-020-01252-x.

Salimi-Badr, A., Hashemi, M., & Saffari, H. (2023). A type-2 neuro-fuzzy system with a novel learning method for Parkinson's disease diagnosis. Applied Intelligence, 53(12), 15656-15682. DOI: 10.1007/s10489-022-04276-8.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression.

Singh, N., Singh, P., & Bhagat, D. (2019). A rule extraction approach from support vector machines for diagnosing hypertension among diabetics. Expert Systems with Applications, 130, 188-205. DOI: 10.1016/j.eswa.2019.04.029.

Song, Y., Wang, Y., Ye, X., Wang, D., Yin, Y., & Wang, Y. (2020). Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending. Information Sciences, 525, 182-204, DOI:10.1016/j.ins.2020.03.027.

SS Júnior, J., Mendes, J., Souza, F., & Premebida, C. (2023). Survey on Deep Fuzzy Systems in regression applications: a view on interpretability. International Journal of Fuzzy Systems, 1-22. DOI: 10.1007/s40815-023-01544-8.

Stevens, A., Deruyck, P., Van Veldhoven, Z., & Vanthienen, J. (2020, December). Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1241-1248). IEEE, DOI:10.1109/SSCI47803.2020.9308371.

Tan, Q., Ling, X., Chen, M., Lu, H., Wang, P., & Liu, W. (2019). Statistical analysis and prediction of regional bus passenger flows. International Journal of Modern Physics B, 33(11), 1950094. DOI: 10.1142/S0217979219501017.

Taud, H., & Mas, J. (2018). Multilayer Perceptron (MLP). En M. Camacho Olmedo, M. Paegelow, J. F. Mas y F. Escobar (Eds.), Geomatic Approaches for Modeling Land Change Scenarios (pp. 493-502). Springer. DOI: 10.1007/978-3-319-60801-3_27.

Thakor, A. V. (2015). The financial crisis of 2007–2009: Why did it happen and what did we learn?. The Review of Corporate Finance Studies, 4(2), 155-205. DOI: 10.1093/rcfs/cfv001.

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems, 32(11), 4793-4813. DOI: 10.1109/TNNLS.2020.3027314.

Toqué, F., Khouadjia, M., Come, E., Trepanier, M., & Oukhellou, L. (2017, October). Short & long term forecasting of multimodal transport passenger flows with machine learning methods. In 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC) (pp. 560-566). IEEE. DOI: 10.1109/ITSC.2017.8317939.

Wang, H., Chen, K., Zhu, W., & Song, Z. (2015). A process model on P2P lending. Financial Innovation, 1(1), 1-8.], DOI:10.1186/s40854-015-0002-9.

Wang, T., Zhao, S., Zhu, G., & Zheng, H. (2021). A machine learning-based early warning system for systemic banking crises. Applied Economics, 53(26), 2974–2992. DOI: 10.1080/00036846.2020.1870657.

Wang, Y., Currim, F., & Ram, S. (2022). Deep learning of spatiotemporal patterns for urban mobility prediction using big data. Information Systems Research, 33(2), 579-598. DOI: 10.1287/isre.2021.1072.

Wang, Z., Chen, L., Wang, L., & Diao, G. (2020). Recognition of audio depression based on convolutional neural network and generative antagonism network model. IEEE Access, 8, 101181-101191. DOI: 10.1109/ACCESS.2020.2998532.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10), 1550-1560. DOI: 10.1109/5.58337.

Wickramasinghe, I., & Kalutarage, H. (2021). Naïve Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Journal of Big Data, 8(1), 1-29. 10.1007/s00500-020-05297-6.

Wu, W., Xia, Y., & Jin, W. (2020). Predicting bus passenger flow and prioritizing influential factors using multi-source data: Scaled stacking gradient boosting decision trees. IEEE Transactions on Intelligent Transportation Systems, 22(4), 2510-2523. DOI: 10.1109/TITS.2020.3035647.

Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: a machine learning methodology. Scientific Reports, 11(1), 18759, DOI:10.1016/j.eswa.2020.114411.

Yang, R., Wang, P., & Qi, J. (2023). A novel SSA-CatBoost machine learning model for credit rating. Journal of Intelligent & Fuzzy Systems, 44(2), 2269-2284, DOI:10.3233/JIFS-221652.

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. Insights into imaging, 9, 611-629. DOI: 10.1007/s13244-018-0639-9.

Yeganejou, M., Dick, S., & Miller, J. (2019). Interpretable deep convolutional fuzzy classifier. IEEE transactions on fuzzy systems, 28(7), 1407-1419. DOI: 10.1109/TFUZZ.2019.2946520.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning—I. Information sciences, 8(3), 199-249. DOI: 10.1016/0020-0255(75)90036-5.

Zaman, M., & Hassan, A. (2021). Fuzzy heuristics and decision tree for classification of statistical feature-based control chart patterns. Symmetry, 13(1), 110. DOI: 10.3390/sym13010110.

Zeltner, D., Schmid, B., Csiszár, G., & Csiszár, O. (2021). Squashing activation functions in benchmark tests: Towards a more eXplainable Artificial Intelligence using continuous-valued logic. Knowledge-Based Systems, 218, 106779. DOI: 10.1016/j.knosys.2021.106779.

Zhai, H., Cui, L., Nie, Y., Xu, X., & Zhang, W. (2018). A comprehensive comparative analysis of the basic theory of the short term bus passenger flow prediction. Symmetry, 10(9), 369. DOI: 10.3390/sym10090369.

Zhai, H., Tian, R., Cui, L., Xu, X., & Zhang, W. (2020). A Novel Hierarchical Hybrid Model for Short-Term Bus Passenger Flow Forecasting. Journal of Advanced Transportation, 2020(1), 7917353. DOI: 10.1155/2020/7917353.

Zhang, N., Chen, H., Chen, X., & Chen, J. (2016). Forecasting public transit use by crowdsensing and semantic trajectory mining: Case studies. ISPRS International Journal of Geo-Information, 5(10), 180. DOI: 10.3390/ijgi5100180.

Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. Physica A: Statistical Mechanics and its Applications, 534, 122370, DOI:10.1016/j.physa.2019.122370.

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. Procedia Computer Science, 162, 503-513, DOI:10.1016/j.procs.2019.12.017.

Zou, L., Shu, S., Lin, X., Lin, K., Zhu, J., & Li, L. (2022). Passenger flow prediction using smart card data from connected bus system based on interpretable XGBoost. Wireless Communications and Mobile Computing, 2022, 1-13. DOI: 10.1155/2022/5872225.

Anexo

a) Artículos publicados

1. Deep Learning XAI for Bus Passenger Forecasting: A Use Case in Spain

Título: Deep Learning XAI for Bus Passenger Forecasting: A Use Case in Spain

Autores: Leticia Monje, Ramón A. Carrasco, Carlos Rosado, Manuel Sánchez-Montañés

Año publicación: 2022

DOI: 10.3390/math10091428

Link: https:// www.mdpi.com/2227-7390/10/9/1428

Publicación: MDPI - Mathematics

Fuente	Nº citas	Actualización
Scopus	19	23-4-2025
Web of Science	13	8-1-2025
Dimensions (totales)	18	29-12-2024
Dimensions (recientes < 2 años)	15	29-12-2024
OpenCitations	7	

Impacto del medio de difusión

FWCI: 3.07

Journal Impact Factor – JCF (JCR)

Año 2022

Factor de impacto de la revista: 2,4

Factor de impacto sin autocitas: 2,0

Cuartil mayor: Q1

Área: MATHEMATICS Cuartil: **Q1** Posición en el área **23/330** (Edicion: SCIE)

Journal Citation Indicator – JCI (JCR)

Año 2022

JCI de la revista: 2,08

Cuartil mayor: Q1

Área: MATHEMATICS Cuartil: **Q1** Posición en el área **13/485** (Edicion: SCIE)

Article Influence score (JCR)

Año 2022

Article Influence score: 0,369

SCImago Journal Rank

Año 2022

Impacto SJR de la revista: 0,446

Cuartil mayor: Q2

Área: Computer Science (miscellaneous) Cuartil **Q2** Posición en el área **125/351**

Área: Engineering (miscellaneous) Cuartil **Q2** Posición en el área **154/482**

Área: Mathematics (miscellaneous) Cuartil **Q2** Posición en el área **183/463**

Scopus CiteScore

Año 2022

CiteScore de la revista: 3.5

Área: Mathematics (all) Percentil: **88**

Área: Engineering (miscellaneous) Percentil: **63**

Área: Computer Science (miscellaneous) Percentil: **62**

mathematics

MDPI

Article

Deep Learning XAI for Bus Passenger Forecasting: A Use Case in Spain

Leticia Monje [1], Ramón A. Carrasco [2,*], Carlos Rosado [3] and Manuel Sánchez-Montañés [3]

[1] Faculty of Statistics, Complutense University Puerta de Hierro, 28040 Madrid, Spain; lmonje01@ucm.es
[2] Department of Marketing, Faculty of Commerce and Tourism Complutense, University of Madrid, 28003 Madrid, Spain
[3] Computer Science Department, Universidad Autónoma de Madrid, 28049 Madrid, Spain; carlos.rosado.moral@gmail.com (C.R.); manuel.smontanes@uam.es (M.S.-M.)
* Correspondence: ramoncar@ucm.es

Abstract: Time series forecasting of passenger demand is crucial for optimal planning of limited resources. For smart cities, passenger transport in urban areas is an increasingly important problem, because the construction of infrastructure is not the solution and the use of public transport should be encouraged. One of the most sophisticated techniques for time series forecasting is Long Short Term Memory (LSTM) neural networks. These deep learning models are very powerful for time series forecasting but are not interpretable by humans (black-box models). Our goal was to develop a predictive and linguistically interpretable model, useful for decision making using large volumes of data from different sources. Our case study was one of the most demanded bus lines of Madrid. We obtained an interpretable model from the LSTM neural network using a surrogate model and the 2-tuple fuzzy linguistic model, which improves the linguistic interpretability of the generated Explainable Artificial Intelligent (XAI) model without losing precision.

check for updates

Citation: Monje, L.; Carrasco, R.A.;

Keywords: deep learning; LSTM; XAI; time series; passenger forecasting; smart city; surrogate model; 2-tuple fuzzy model

2. **Machine Learning XAI for early loan default prediction**

Título: Machine Learning XAI for early loan default prediction

Autores: Leticia Monje, Ramón A. Carrasco, Manuel Sánchez-Montañés

Año: 2025 (en proceso de publicación)

Estado: Aceptado

Publicación: Computational Economics- Springer

Impacto del medio de difusión

Journal Impact Factor – JCF (JCR)

Año 2023

Factor de impacto de la revista: 1.9

Factor de impacto sin autocitas: 1.6

Área: Economics, Econometrics and Finance (miscellaneous) Cuartil: **Q2** Posición en el área **224/600** (Edicion: SCIE)

Journal Citation Indicator – JCI (JCR)

Año 2023

JCI de la revista: 0.57

Área: ECONOMICS Cuartil: **Q2** Posición en el área **269/603** (Edicion: SCIE)

Article Influence score (JCR)

Año 2023

Article Influence score: 0,353

SCImago Journal Rank

Año 2024

Impacto SJR de la revista: 0,535

Área: Economics, Econometrics and Finance (miscellaneous) Cuartil **Q2** Posición en el área **131/514**

Scopus CiteScore

Año 2023

CiteScore de la revista: 4.0

Área: Economics, Econometrics and Finance (miscellaneous) Posición en el área **53/242** Percentil: **78.**

SPRINGER NATURE

Leticia Monje García
Complutense University of Madrid
Faculty of Statistical Studies
Avda. Puerta de Hierro, s/n
Madrid 28040
Spain

Editorial Office
Computational Economics

Amsterdam School of Economics
Roetersstraat 11
1018 WB Amsterdam
The Netherlands

https://www.springer.com/journal/10614

Date	Your reference	Our reference
April 24, 2025	Request	CSEM-D-23-00993R1
Contact person	**Telephone**	**Email**
Hans Amman	+31651532162	h.m.amman2@contact.uva.nl

Subject
Submission acceptance

To whom it may concern: Herewith I can confirm that the submission

CSEM-D-23-00993R1
with title
Machine Learning XAI for early loan default prediction

authored by

Leticia Monje
Ramón Alberto Carrasco
Manuel Sánchez-Montañés

has been accepted for publication in the scientific journal *Computational Economics*.

Computational Economics is the official journal of the Society for Computational Economics. The journal presents new research in a rapidly growing multidisciplinary field that uses advanced computing capabilities to understand and solve complex problems from all branches in economics. *Computational Economics* is published by Springer Nature, Heidelberg, Germany. ISSN: 0927-7099 (Print) 1572-9974 (Online). ISI impact factor 1.90 (2024).

Sincerely,

Hans Amman
Editor-in-Chief