Tesis Doctoral

MACHINE LEARNING INTERPRETABLE PARA LA DETECCIÓN DEL FRAUDE CREDITICIO

Autor:

Jacobo Chaquet Ulldemolins

Directores:

Dr. Fco. Javier Gimeno Blanes
Dr. Santiago Moral Rubio

Tutor:

Dr. José Luis Rojo Álvarez

Programa de Doctorado en Tecnologías de la Información y las Comunicaciones

Escuela Internacional de Doctorado

2022

Resumen

Antecedentes

Las empresas cada día dan más facilidades a sus clientes para realizar sus compras, entre estas facilidades está la compra de productos de manera online, esto ha hecho que el comercio electrónico crezca a unas cuotas de uso muy elevadas y por ende el uso de las tarjetas alcance su máxima expresión. Esto es, sin ninguna duda, una oportunidad para que los delincuentes puedan cometer fraudes. En medio de todo esto, están los bancos para asegurar que todas las transacciones son legales y no fraudulentas. Esta es una tarea ardua y complicada, ya que los defraudadores siempre intentan simular legítimas todas las transacciones fraudulentas, lo que convierte la detección del fraude en una tarea muy compleja. El número de transacciones rechazadas erróneamente por sospecha de fraude se estima en torno a 118.000 millones de dólares de pérdidas únicamente en el sector minorista, estas pérdidas suponen una amenaza equivalente al fraude real en el sector de los servicios financieros. En consecuencia, los bancos se ven obligados a dedicar cada vez más recursos a discriminar entre las transacciones legítimas y las fraudulentas para hacer frente al difícil dilema de evitar las acciones de los impostores sin limitar el crecimiento inexorable del comercio online. Por si esto fuera poco, a este reto se le suma la necesidad de transparencia en toda decisión para la determinación de fraude exigida por los organismos reguladores. De hecho en la Unión Europea, en el Reglamento General de Protección de Datos, aprobado en 2018, otorga a sus ciudadanos el derecho a recibir una explicación de las decisiones basadas en el tratamiento automatizado. La justificación de este tipo de regulación radica en el potencial sesgo que se podrían estar aplicando.

Objetivos

Cada una de las problemáticas expuestas anteriormente, se analizan con detenimiento en la presente tesis, donde la meta final consiste en el desarrollo de una metodología fiable, imparcial e interpretable para evaluar automáticamente la detección de fraude de crédito (CFD). Para alcanzar dicha meta, se han definido tres grandes objetivos definidos de la siguiente manera:

- T1: Obtención e interpretación de las características más relevantes para la detección de fraude.
- T2: Comprimir y codificar los datos para aislar las transacciones fraudulentas de las no fraudulentas.

- T3: Proponer y evaluar un modelo interpretable completo para la detección de fraude.

Metodología

La presente tesis se centra en métodos interpretables de Machine Learning (ML) aplicados a la detección de fraude. Basándonos en los resultados y las publicaciones, puede concluirse que la presente tesis contribuye al avance en métodos interpretables. La metodología general puede dividirse en los siguientes puntos:

- (i): revisión extensiva de la literatura.
- (ii): búsqueda, preparación y validación de los datos.
- (iii): definición de las métricas de interés.
- (iv): reducción de dimensionalidad del problema mediante el uso de las características más relevantes.
- (iv): comprimir y codificar eficazmente los datos para aislar las transacciones legítimas de las fraudulentas.
- (v): proponer, y evaluar, nuevas técnicas para ofrecer un modelo interpretable para CFD.

La metodología seguida en el T1 implíca la investigación de métodos capaces de seleccionar características relevantes manteniendo la interpretabilidad. Para ello nos hemos basado en el estado de arte, lo cual nos ha llevado a seleccionar la técnica *Informative Variable Identifier (IVI)*, la cual se ha evolucionado en la presente tesis.

La metodología para T2 se fundamenta en la investigación de métodos de *deep learning*. Entre los métodos existentes para la generación de un espacio latente se encuentra los *autoencoders* los cuales mostraron unas prestaciones superiores a otros métodos clásicos de ML en nuestros test iniciales. Este tipo de algoritmos, como tantos otros de ML, tienen la desventaja de ser complejos, donde es complicado comprender cómo han realizado su procesos de decisión, convirtiéndolos en sistemas opacos o tradicionalmente conocidos como de caja negra.

Consecuentemente, la metodología para T3 se centró en dotar de interpretabilidad a los *autoencoders* y extender esta interpretabilidad o explicabilidad a cualquier algoritmo de caja negra. Para ello, en la presente tesis se ha formalizado la definición del problema de interpretabilidad agnóstica de los métodos empleados, para ello se presenta un novedosa técnica de interpretabilidad *(Single Transaction-level Explanation STE)* centrada en dar

explicabilidad a las transacciones de manera individual. Adicionalmente se completa esta interpretabilidad con la técnica de agrupación (*Individual Transaction Rankings ITR*) que nos permite agrupar transacciones con propiedades similares, lo cual nos permite detectar conjunto de transacciones con patrones de comportamiento comunes.

Resultados

Los resultados relevantes del objetivo T1 consisten en un método fiable, imparcial e interpretable para la selección de características relevantes. El método propuesto se centra en extraer las características relevantes aplicando el algoritmo IVI, el cual ha sido adaptado para amplificar su poder de detección de características usando diversos algoritmos de ML. Adicionalmente para la reducción de posibles sesgos, se han desarrollado procesos de filtrado. Dicha selección de características y filtrado es un paso crucial tanto para mejorar la precisión como para evitar una posible discriminación basada en características no permitidas, como por ejemplo la raza, el sexo o el estado civil. Los resultados obtenidos demuestran que el método propuesto no solo selecciona las características más relevantes sino que mejora la fase de entrenamiento de los modelos ML en términos de eficiencia computacional, al reducir el número potencial de características a sólo las realmente informativas para alcanzar una mayor precisión.

Los resultados relevantes del objetivo T2 consisten en un método que comprima y codifique eficazmente los datos para aislar las transacciones fraudulentas de las no fraudulentas. Esta reducción del espacio real a un espacio latente se realiza a través del uso de algoritmos de *autoencoders* y técnicas de transferencia de aprendizaje (*fine tunning*). Los resultados obtenidos en el espacio latente mejoran la tarea de clasificación al mapear mejor los diferentes tipos de transacciones.

Los resultados relevantes del objetivo T3 consisten en un método que nos permita dotar de interpretabilidad a modelos de caja negra, como son los *autoencoders* empleados en el objetivo T2. En CFD, al ser un sector muy regulado, no son admisibles modelos de caja negra, lo que convierte la interpretabilidad en un factor clave para poder emplear algoritmos de altas capacidades de clasificación en la operativa real. Todo ello lleva a la necesidad de la interpretabilidad como elemento crucial a la hora de romper las barreras de la falta de transparencia en los desarrollos tradicionales de ML. La interpretabilidad generada en la presente tesis se ha focalizado a nivel de transacción individual reflejando la contribución o importancia de cada característica en el proceso de decisión. Este proceso se ha llevado a cabo a partir de modelos sustitutivos para cada transacción. Los resultados obtenidos en la experimentación confirman cómo una pequeña variación en una característica en el espacio de entrada tiene una respuesta diferente en el espacio latente teniendo un alto

impacto en la clasificación de las transacciones permitiendo mejorar la identificación de transacciones legítimas o fraudulentas.

Conclusiones

La presente tesis se centra en métodos interpretables de ML aplicados a la detección de fraude de crédito. Basándonos en los resultados y las publicaciones, puede concluirse que la presente tesis contribuye al avance en métodos interpretables.

Los métodos desarrollados en la tesis se complementan al tratar diferentes fases para la detección de fraude. El primer objetivo consiste en un método capaz de seleccionar las características relevantes, siendo un método computacionalmente eficiente capaz de cuantificar la relevancia de las variables permitiendo reducir la dimensionalidad del problema y minimizar los sesgos en la toma de decisión. Estas propiedades son deseables en el ámbito de CFD, donde el volumen de variables es ingente y la detección temprana de sesgos para evitar posibles casos de discriminación.

El método resultante del segundo objetivo se centra en comprimir y codificar eficazmente los datos para aislar las transacciones fraudulentas de las no fraudulentas. La diferenciación entre transacciones es una tarea compleja dado que los defraudadores intentar replicar el comportamiento de las transacciones legítimas lo que requiere una gran cantidad tiempo y esfuerzo de los analistas teniendo altas tasas de falsos positivos. El método propuesto presenta una correcta separación entre estos tipos de transacciones al comprimir sus características más relevantes en un espacio latente.

Por último, el marco de trabajo resultante de la tercera área de investigación permite interpretar modelos de caja negra como los generados en el objetivo dos. La interpretabilidad a nivel de transacción individual refleja la contribución de cada característica permitiendo justificar la clasificación de las transacciones fraudulentas ante cualquier organismo regulador.

Esperamos que este trabajo contribuya a la adopción de métodos interpretables en CFD. Podemos concluir que nuestra metodología proporciona una evaluación detallada a nivel de transacción individual, dotando de interpretabilidad al proceso de decisión donde visibiliza las características más relevantes. Esta perspectiva individualizada e imparcial proporciona la transparencia necesaria, no sólo para cumplir la normativa, sino también para poder justificar cada operación clasificada ante clientes y autoridades.

Abstract

Background

Companies every day give more facilities to their customers to make their purchases. Online purchase is one of these facilities which has made e-commerce to grow at very high rates of use and therefore the use of cards reaches its maximum expression which is an opportunity for criminals to commit fraud. In the midst of all this are the banks, which must make sure that all the transactions are legal and non-fraudulent. This is an arduous and complicated task, due to fraudsters always trying to make every fraudulent transaction seem legitimate, which makes fraud detection a very challenging and difficult task. To cope with this emerging new reality, financial institutions are forced to devote increasing resources to discriminate between legitimate and fraudulent transactions to address the difficult dilemma of preventing the actions of impostors without limiting the inexorable growth of online commerce. To make matters worse, this challenge is compounded by the need for transparency in all fraud determination decisions required by regulatory agencies. In fact, in the European Union the General Data Protection Regulation aproved in 2018 gives their citizens the right to receive explanations for decisions based solely on automated processing. The justification for this type of regulation lies in the possible bias that could be applied.

Objectives

Each of the above issues are analyzed in detail in this thesis, where the ultimate goal is the development of a reliable, unbiased and interpretable methodology to automatically evaluate the detection of credit fraud (CFD). To achieve this goal, three main objectives have been defined as follows:

- T1: Obtain and interpret the most relevant features for fraud detection.
- T2: Compress and encode the data to isolate fraudulent transactions from non-fraudulent transactions.
- T3: Propose and evaluate a complete interpretable model for fraud detection.

Methodology

This thesis focuses on interpretable machine learning (ML) methods applied to fraud detection. From the results and publications, it can be concluded that the present thesis

contributes to the advancement of interpretable methods. The overall methodology can be divided into the following points:

- (i): extensive literature review of the topic.
- (ii): search, preparation and validation of data.
- (iii): defining metrics of interest.
- (iv): reduction of the dimensionality of the problem by using the most relevant features.
- (iv): effectively compress and encode data to isolate legitimate transactions from fraudulent ones.
- (v): propose, and evaluate, new techniques to provide an interpretable model for CFD.

The methodology followed in T1 involves the investigation of methods capable of selecting relevant features while maintaining interpretability. For this purpose, we have relied on the state of the art, which has led us to select the Informative Variable Identifier (IVI) technique, which has evolved in this thesis.

The methodology for T2 is based on research into deep learning methods. Among the existing methods for generating a latent space are *autoencoders*, which showed superior performance to other classical ML methods in our initial tests. These types of algorithms have the disadvantage of being complex, where it is difficult to understand how they have carried out their decision processes, making them opaque systems or traditionally known as black box systems.

Consequently, the methodology for T3 is focused on providing interpretability to the autoencoders and extending this interpretability or explainability to any black box algorithm. To this end, this thesis has formalized the definition of the agnostic interpretability problem of the methods used, presenting a novel interpretability technique *(Single Transaction-level Explanation STE)* focused on the explanation of transactions individually. Finally, this interpretability is completed with the clustering technique (*Individual Transaction Rankings ITR*) that allows grouping transactions with similar properties, which allows detecting a set of transactions with similar patterns. The relevant results of objective T1 consist of a reliable, unbiased and interpretable method for selecting relevant features. The proposed method focuses on extracting relevant features by applying the IVI algorithm, which has been adapted to amplify its feature detection power using various ML algorithms. In addition, filtering processes have been developed to reduce possible biases. This feature

selection and filtering is a crucial step both to improve accuracy and to avoid possible discrimination based on impermissible features such as race, sex or marital status. The results show that the proposed method not only selects the most relevant features, but also improves the training stage of the ML models in terms of computational efficiency by reducing the potential number of features to only the really informative ones to achieve higher accuracy.

The relevant results of objective T2 consist of a method that efficiently compresses and encodes data to isolate fraudulent transactions from non-fraudulent transactions.

This reduction of the real space to a latent space is performed by using autoencoding algorithms and fine-tuning techniques. This reduction from real space to latent space is accomplished by using autoencoders algorithm and fine-tuning techniques. The results in the latent space improve the classification task by better assigning the different types of transactions.

The results of objective T3 consist of a method to provide interpretability to black box models, such as the autoencoders used in objective T2. In CFD, it is a highly regulated industry, where black box models are not admissible, which makes interpretability a key factor in order to be able to use algorithms with high classification capability in real use. All this leads us to the need for interpretability as a crucial element to break the barriers of lack of transparency in traditional ML developments. The interpretability generated in this thesis has been focused at the individual transaction level reflecting the contribution or importance of each feature in the decision process. The results obtained in the experimentation confirm how a small variation in a feature in the input space has a different response in the latent space having a high impact on the classification of transactions allowing to improve the identification of legitimate or fraudulent transactions.

Conclusions

The present thesis focuses on interpretable ML methods applied to credit fraud detection. From the results and publications, it can be concluded that the present thesis contributes to the advancement of interpretable methods.

The methods developed in the thesis complement each other by addressing different phases of fraud detection. The first objective consists of a method capable of selecting the relevant features. It is a computationally efficient method capable of quantifying the relevance of the features allowing to reduce the dimensionality of the problem and minimize biases in decision making. These properties are desirable in the field of CFD, where the volume of features is huge and early detection of biases to avoid discrimination.

The method resulting from the second objective focuses on effectively compressing and

encrypting data to isolate fraudulent transactions from non-fraudulent ones. To separate fraudulent and non-fraudulent transactions is a complex task, as fraudsters attempt to replicate the behavior of legitimate transactions, which requires a lot of time and effort on the part of analysts and has high false positive rates. The proposed method presents a correct separation between these types of transactions by compressing their most relevant features into a latent space.

Finally, the framework resulting from the third objetive allows for the interpretation of black box models such as those generated in objective two. Interpretability at the individual transaction level reflects the contribution of each features allowing to justify the classification of fraudulent transactions to any regulatory body. We hope that this work will contribute to the adoption of interpretable methods in CFD. We can conclude that our methodology provides a detailed assessment at the individual transaction level, bringing interpretability to the decision process where it makes the most relevant features visible. This individualized and unbiased perspective provides the necessary transparency, not only to comply with regulations, but also to be able to justify each classified transaction to clients and authorities.

"It's better to burn out than fade away."
Kurt D. Cobain

Agradecimientos

Sirvan estas líneas para agradecer a todas las personas que de una forma u otra han hecho posible que esta tesis haya sido posible.

En primer lugar me gustaría agradecer a mi tutor y directores de tesis. Dr. José Luis Rojo Álvarez gracias por la oportunidad que me brindaste para realizar esta tesis, por todo el tiempo dedicado a cualquier hora del día y por los conocimientos que has sabido transmitirme. Dr. Fco. Javier Gimeno Blanes muchísimas gracias por la confianza depositada en mi, la paciencia que has mostrado, tu disposición y tiempo que han sido fundamentales para sacar adelante esta tesis. Dr. Santiago Moral Rubio gracias por la oportunidad de trabajar junto a ti, por la paciencia y las charlas interesantísimas sobre fraude aplicado en el mundo real donde he podido aprender muchísimo. Quiero agradecer también al Dr. Sergio Muñoz Romero que gracias a su tiempo y trabajo, el cual me ayudó muchísimo en sentar las bases de esta tesis.

No puedo olvidarme tampoco de esas personas que han sufrido tanto o más que yo durante estos años. A mis buenos amigos y compañeros de trabajo por siempre motivarme a alcanzar el "*high performance*". A los amigos de toda la vida, por esta ahí aunque no entendieran que estaba haciendo con mi vida. A mi familia, comenzando por mis padres, que debo agradecerles sus constantes muestras de confianza y de apoyo. A mi hermano por ser un referente y motivarme con su ejemplo a seguir este camino. Y por último, pero no por ello menos importante, a mi compañera de vida, gracias por tu compresión, ejemplo y paciencia en los momentos en que he estado ausente, sin ti esto hubiera sido infinitamente más duro.

Lista de Acrónimos y Abreviaturas

2FA Two-factor Authentication.

3DS 3-Domain Secure.

AE Autoencoder.

AI Artificial Intelligence.

ANNs Artificial Neural Networks.

AVS Address Verification Service.

BCE Banco Central Europeo.

CFD Credit Fraud Detection.

CFDIS Credit Fraud Detection Interpretation System.

CME Covariance Multiplication Estimator.

CNP Card Not Present.

COMPAS Correctional Offender Management Profiling for Alternative Sanctions.

DL Deep Learning.

DT Decision Tree.

EMV Europay, Mastercard y Visa.

FS Feature Selection.

GB Gradient Boosting.

GBDT Gradient Boosted Decision Trees.

ITR Individual Transaction Rankings.

IVI Informative Variable Identifier.

KNN K-Nearest Neighbour.

LDA Linear Discriminant Analysis.

LIME Local Interpretable Model-agnostic Explanations.

LR Linear Regression.

MIFF Maximally-informative Features Filter.

ML Machine Learning.

OHE One Hot Encoding.

POS Point Of Sale.

ReLU Rectified Linear Unit.

RF Random Forest.

RFF Recurrent Features Filter.

SHAP SHapley Additive exPlanations.

STE Single Transaction-level Explanation.

SVC Support Vector Classifier.

SVM Support Vector Machines.

SVR Support Vector Regression.

Índice general

Capítulo 1

Introducción

1.1. Motivación

A día de hoy, la mayor parte de las transacciones se realizan de manera online mediante el uso de tarjetas de crédito y sistemas de pago. Estos pagos son bien recibidos tanto por las empresas como por los consumidores, debido a que favorecen el consumo facilitando las compras. En medio de todo esto, están los bancos para verificar que todas las transacciones aceptadas son legítimas. Esta verificación representa un reto en sí misma, no sólo para asegurar las transacciones, sino también para evitar los posibles falsos positivos en los algoritmos de detección de fraude. Se trata de una tarea ardua y complicada, ya que los defraudadores siempre intentan que todas las transacciones fraudulentas parezcan legítimas, lo que convierte la detección del fraude en una tarea difícil y desafiante. Según un informe del Instituto Alan Turing [1], el número de transacciones rechazadas erróneamente por sospecha de fraude puede suponer una amenaza equivalente al fraude real en el sector de los servicios financieros. Otro estudio afirma que las transacciones rechazadas de manera equivocada por sospecha de fraude suponen 118.000 millones de dólares de pérdidas en el comercio minorista [2]. En consecuencia, los bancos se ven obligados a dedicar cada vez más recursos para discriminar entre las transacciones legítimas de las fraudulentas para hacer frente al difícil dilema de evitar las acciones de los impostores sin limitar el crecimiento inexorable del comercio online.

Tradicionalmente, la detección de fraude se ha basado en sistemas expertos [3], los cuales realizaban los controles basados en una serie de reglas y listas de comprobación de factores de riesgo, por ejemplo, repetición de transacciones rechazadas, múltiples intentos fallidos de introducir un número de tarjeta de crédito, compras en ubicaciones diferentes a la ubicación habitual, o gastos repentinos de alto importe no habituales. El gran problema de estos sistemas expertos es que cuanto más especializados son, más caro es su mantenimiento [3]. La Inteligencia Artificial (AI) tiene el potencial disruptivo suficiente

para redefinir la actual industria de Servicios Financieros. En el contexto general de la AI, encontramos tecnologías conocidas globalmente como Machine Learning (ML) que permiten la creación de nuevos algoritmos capaces de proporcionar alternativas y encontrar patrones complejos de las transacciones fraudulentas basadas en registros históricos de transacciones sin prácticamente intervención humana. Sin embargo, uno de los grandes retos y un obstáculo potencialmente bloqueante para su uso en estos modelos es la falta de transparencia en el proceso de decisión. Estos modelos suelen tratarse como cajas negras, en las que sólo conocemos su entrada y salida, pero no el proceso que se ejecuta en su interior. Esto hace que tratar de explicar y validar el resultado sea una tarea difícil, de tal modo que podrían pasar desapercibidas ciertos cálculos o decisiones que fueran erróneas. Este tipo de complejidad puede constituir una barrera importante para el uso de ML en los CFD existentes debido a que son considerados como cajas negras con una nula interpretabilidad [4] [5].

Estas cajas negras tienen implicaciones para los organismos reguladores y supervisores financieros, como el Banco Central Europeo, dado que por un lado son conscientes de que estos métodos pueden mejorar los resultados de manera considerable pero por otro lado, no pueden explicar el por qué, lo cual incumple la regulación. Por ello, han mostrado hasta la fecha importantes reticencias a aceptar un uso generalizado de estas modernas técnicas [5].

Aunque esta realidad se convierte en una clara limitación, el amplio consenso entre los investigadores y las instituciones financieras sugiere que los algoritmos de ML tienen un gran potencial siendo consciente de que existen una serie de desafíos que requieren una atención especial [6]. Por ejemplo, Estados Unidos prohíbe la discriminación basada en varias categorías, como la raza, el sexo o el estado civil. Incluso un algoritmo de préstamos podría ser invalidado incluso si el algoritmo no utiliza directamente ninguna de las categorías prohibidas, sino que utiliza datos que pueden estar altamente correlacionadas con las categorías protegidas. La falta de transparencia se complica aún más en la Unión Europea, donde el Reglamento General de Protección de Datos adoptado en 2016 y que entrará en vigor en 2018 otorga a sus ciudadanos el derecho a recibir una explicación sobre las decisiones basadas únicamente en el tratamiento automatizado. La justificación de este tipo de regulación radica en el potencial sesgo que se podría estar aplicando en las capas ocultas del modelo, dejando así a los analistas, al organismo regulador y a la entidad evaluadora de riesgos, desprovistos de herramientas para identificar cualquier situación indeseable que finalmente pudiera estar produciendo [7, 8]. Más aún, los datos utilizados para entrenar los modelos de ML pueden no ser representativos del problema [6], lo que a veces conduce eventualmente a modelos inexactos, con capacidad de generalización limitada. Dicho esto, y volviendo a las restricciones normativas, las entidades entienden la necesidad

de una regulación que garantice que, el uso de la tecnología no pueda causar un trato discriminatorio a las personas que pase inadvertido, pero también coinciden en la necesidad de una orientación más clara por parte de las autoridades que ofrezca un camino razonable hacia la necesaria y efectiva aplicación de la AI en este ámbito [4]. Teniendo en cuenta que los organismos y autoridades reguladores no permiten que las entidades financieras adopten modelos de AI sin abordar la necesaria descripción del proceso de decisión que se está siguiendo [5], una forma adecuada de superar los problemas regulatorios y la desconfianza respecto a los algoritmos que se están utilizando, es proporcionar entornos y herramientas complementarias que contribuyan de forma efectiva a la interpretabilidad real.

En esta tesis, pretendemos dar una respuesta interpretable a las siguientes preguntas:

- ¿Cuáles son las características más relevantes en el proceso de decisión para discernir las transacciones fraudulentas de las legítimas?
- ¿cómo podemos aislar las transacciones fraudulentas de las no fraudulentas para facilitar su correcta clasificación?
- ¿Cómo podemos explicar que una transacción sea considerada fraudulenta o no?

Estas preguntas representan los problemas habituales a los que se enfrenta los sistemas CFD. Sus respuestas, desde el punto de vista de la interpretabilidad, se ajustan a los tres temas principales de esta tesis, como son: (T1) selección de características relevantes, (T2) aislar las transacciones fraudulentas de las no fraudulentas, y (T3) proporcionar transparencia en el proceso de decisión.

En el objetivo T1, se aborda la selección de características. Este estudio da como resultado una novedosa metodología para abordar el problema de CFD aplicando algoritmos de ML capaces de cuantificar la relevancia de las diferentes variables y sus relaciones. Para ello, en esta tesis se se ha partido de una técnica de selección de características publicada recientemente, conocida como Identificador de Variables Informativas (IVI) [9], la cual es capaz de distinguir entre características informativas, redundantes y ruidosas. A esta técnica se le ha aplicado una serie de mejoras para aumentar su poder de generalización mediante el uso de métodos de ML lineales y filtros. Dicho método se ha aplicado tanto a una base de datos sintéticos, para un mejor modelado descriptivo y un ajuste detallado, como con varios conjuntos de datos reales. Nuestros resultados confirman que nuestra propuesta proporciona una valiosa interpretabilidad al identificar los pesos de las características informativas que vinculan las características originales con la decisión final.

El segundo objetivo de la tesis se centra en comprimir y codificar los datos para aislar las transacciones fraudulentas de las no fraudulentas. El problema de CFD es una tarea compleja debido a que los estafadores hacen todo lo posible para que las transacciones se

diferencien lo menos posible de las reales, tratando de modelar perfiles de comportamiento extremadamente similares. La predicción de estos elementos es especialmente difícil. En el momento de realizar esta investigación, hasta donde sabemos, hay poca literatura sobre el tema del uso de *autoencoder* en CFD. Un *autoencoder* [10, 11] es una red neuronal de múltiples capas que comprime los datos de alta dimensionalidad en una representación del espacio latente de menor dimensión (*encoder*), combinada con una posterior expansión al espacio original (*decoder*). Los resultados de nuestros estudios utilizando el espacio latente mediante el uso de *autoencoder* mejoraron en términos de precisión.

Por último, en el objetivo T3 cubierto en esta tesis, proponemos una metodología interpretable y agnóstica para CFD. En la industria de servicios financieros y, más concretamente, en los CFD, al ser sectores muy regulados, en los que casi no hay lugar para cajas negras, para modelos difíciles de entender o para arquitecturas sin la adecuada transparencia en el uso de los datos. La literatura sobre el tema es escasa, por lo que las instituciones financieras confían en el uso de modelos sencillos e interpretables, como son los árboles de decisión [12] o los modelos de lineales [13]. Este tipo de modelos son fáciles de entender y sus predicciones se explican de forma sencilla en función de las pesos de cada variable. Hasta donde sabemos, hay poca literatura que aborde el uso de modelos avanzados de ML que sean interpretables. Nuestra hipótesis es que la interpretabilidad puede obtenerse en función de la importancia de las características y como solución al problema anterior, en esta tesis proponemos un mecanismo que es capaz de evaluar el peso de cada característica para cada transacción de manera individual en el espacio latente. Los resultados han demostrado ser un método adecuado para interpretar la relación entre la contribución de cada característica y la salida del clasificador en los métodos de caja negra. Los métodos resultantes de esta investigación se basan en datos empíricos recogidos de transacciones reales, que abarcan numerosas categorías y diferentes áreas geográficas.

1.2. Objetivos

La presente tesis Doctoral se centra en el desarrollo de una metodología fiable, imparcial e interpretable para medir automáticamente el riesgo de CFD. Como se indica en la sección 1.1, se centra en tres objetivos principales: (T1) selección de características, (T2) aislar las transacciones fraudulentas de las no fraudulentas, y (T3) metodología interpretable agnóstica para CFD. Para alcanzar dichos objetivos generales se han descompuesto en objetivos específicos los cuales se enumeran a continuación:

- Adquirir, preparar y procesar datos de transacciones reales de distintas geografías.
- Seleccionar las características más relevantes aplicada a los CFD.

- Definir la selección de filtros aplicada a los CFD para minimizar el sesgo de los datos de entrenamiento.

- Construir un *autoencoder* para comprimir el espacio real en un espacio latente de menor dimensionalidad.

- Diseñar un algoritmo que permita obtener la relevancia de una característica concreta en una transacción determinada en el espacio latente.

1.3. Contribuciones y publicaciones

Las contribuciones de esta tesis se organizan según los tres temas principales de investigación en las siguientes secciones, las cuales serán presentadas a lo largo de los diferentes capítulos de esta tesis.

1.3.1. Selección de características y filtros

Nuestra contribución consiste en crear una metodología fiable, imparcial e interpretable, capaz de cuantificar la información de las características y sus relaciones para evaluar automáticamente el riesgo de CFD. Sus principales características son las siguientes:

- Selección de características. Extraer características informativas comunes aplicando el algoritmo IVI.

- Filtración de características recurrentes para la reducción de posibles sesgos y obtener una visión global del problema. Para ello, se desarrolló el Filtro Recurrent Features Filter (RFF) y el Filtro Maximally-informative Features Filter (MIFF).

- Interpretabilidad basada en los pesos de las características. Dichos pesos resumen en realidad las contribuciones de las características en el proceso de decisión y la interacción entre los datos de entrada y salida en el contexto del problema de clasificación concreto de CFD.

Este trabajo se ha publicado como Chaquet-Ulldemolins, Jacobo; Gimeno-Blanes, Francisco-Javier; Moral-Rubio, Santiago; Muñoz-Romero, Sergio; Rojo-Álvarez, José Luis. On the Black-Box Challenge for Fraud Detection Using Machine Learning (I): Linear Models and Informative Feature Selection. Applied Sciences, volume 12, 2022. DOI 10.3390/app12073328. La publicación, junto con un resumen gráfico, se recoge a lo largo de la presente tesis.

1.3.2. Aislar las transacciones fraudulentas de las no fraudulentas dotándolas de interpretabilidad

Nuestra aportación ha consistido en crear una metodología interpretable y agnóstica para todas y cada una de las transacciones que se someten al análisis, mediante un enfoque de explicación a nivel de transacción individual (STE). Esta técnica permite analizar cada transacción individualmente aplicando modelos sustitutivos y pequeñas fluctuaciones del espacio de entrada y evaluando cómo se influye en la salida, arrojando así luz sobre la dinámica subyacente del modelo. A partir de ahí, se puede formular una clasificación individualizada de las transacciones (ITR), aprovechando las contribuciones de cada característica a través de STE. Esta clasificación representa una estimación aproximada de las características más importantes que intervienen en el proceso de decisión. Sus principales características son las siguientes

- Representación en el espacio latente. Comprimir el espacio real en el que hay una alta dimensionalidad a un espacio latente reducido para aislar las transacciones fraudulentas.
- STE. Evaluación del peso de las características para cada transacción a nivel individual.
- Clustering a través de ITR.

Este trabajo se ha publicado como Chaquet-Ulldemolins, Jacobo; Gimeno-Blanes, Francisco-Javier; Moral-Rubio, Santiago; Muñoz-Romero, Sergio; Rojo-Álvarez, José Luis. On the Black-Box Challenge for Fraud Detection using Machine Learning (II): Nonlinear Analysis through Interpretable Autoencoders. Applied Sciences, volume 12, 2022. DOI 10.3390/app12083856. La publicación, junto con un resumen gráfico, se recoge a lo largo de la presente tesis.

1.4. Estructura de la tesis

La organización de la tesis es la siguiente:

- En el capítulo 2, con objeto de contextualizar el ámbito de los servicios financieros de esta tesis, se ofrece una introducción a los CFD. A continuación, repasamos brevemente los métodos de CFD más comunes que se emplean y por último introducimos en que consiste la interpretabilidad.
- En el capítulo 3 se describen los conjuntos de datos empleados, seguidamente se describen los algoritmos de ML lineales y no lineales utilizados en esta tesis.

- En el capítulo 4 se describen los fundamentos de la metodología aplicada y resultados obtenidos para la selección de características para los CFD.
- En el capítulo 5, se presentan los mecanismos aplicados para lograr separar las transacciones fraudulentas de las no fraudulentas mediante el uso de Deep Learning (DL)
- En el capítulo 6, se definen los mecanismos para dotar de interpretabilidad a los procesos de decisión basados en cajas negras como son los algoritmos de DL, empleando para ellos las técnicas presentadas en el capítulo 5.
- Concluimos en el capítulo 7 donde se muestras las conclusiones así como las líneas de trabajo futuro.

Capítulo 2

Contexto

2.1. Detección de fraude de crédito

Las tarjetas de crédito se han vuelto algo imprescindible en la vida cotidiana y son empleadas para adquirir bienes y servicios de manera diaria. Estas tarjetas pueden utilizarse tanto para transacciones físicas como online. En las transacciones físicas, la tarjeta de crédito se introduce en un terminal de pago en los diferentes comercios para adquirir los productos, mientras que las transacciones online se realizan desde cualquier dispositivo conectado a Internet [14]. El fraude podríamos definirlo como cualquier acto intencionado o deliberado para privar a otro de bienes o dinero por medios intencionados, engañosos o injustos. En las transacciones mediante tarjetas de crédito, el fraude es el uso no autorizado y no deseado de una cuenta por alguien que no es el propietario de la misma. En otras palabras, el fraude en tarjetas de crédito puede definirse como un caso en el que una persona utiliza la tarjeta de crédito de otra persona por motivos personales mientras que el propietario y las autoridades emisoras de la tarjeta desconocen el hecho de que se está utilizando indebidamente. Los defraudadores siempre intentan que todas las transacciones fraudulentas parezcan legítimas, para evitar ser identificados y poder reutilizar sus métodos, lo que hace que la detección del fraude sea una tarea muy desafiante y difícil [1].

La introducción del chip EMV (Europay, MasterCard y Visa) en las tarjetas ha proporcionado mayor seguridad a las transacciones físicas en los puntos de venta (POS). EMV es un estándar para hacer pagos seguros interoperables a nivel internacional. El elemento clave de EMV es la inclusión de los datos digitales dinámicos en cada transacción. Esto hace que estos tipos de transacciones resulten totalmente seguras reduciendo el riesgo de fraude. Estos chips, se incrustan en las tarjetas dando protección a los datos del usuario, y desde su introducción en el mercado se convirtió rápidamente en el nuevo estándar mundial para los pagos con tarjetas de crédito y débito [15]. En los más de 80 países en los que se adoptó el estándar EMV, se redujeron significativamente la cantidad de

falsificaciones y fraudes con tarjetas, lo que permitió ahorrar millones de dólares. Tras el lanzamiento de los chips EMV, durante un tiempo la balanza se había inclinado hacia las instituciones financieras reduciendo los intentos de fraude.

Pero esta situación duró poco, las instituciones financieras empezaron a detectar un aumento del fraude en las transacciones sin presencia de tarjeta (CNP). Una transacción de CNP es aquella en la que el comerciante no puede examinar físicamente la tarjeta de crédito, generalmente cuando la compra se realiza a través de canales online o vía telefónica. El aumento de los casos de fraude CNP resultó ser un problema mayor, los bancos empezaron a verse inundados de reclamaciones por disputas. Éstas debían ser evaluadas para detectar la posibilidad de devoluciones de cargos, cuyo objetivo era proteger a los titulares de las tarjetas de cargos no autorizados debidos a transacciones fraudulentas en primer lugar. En 2017 se detectó un archivo en la Dark Web que contenía 1.400 millones de credenciales de inicio de sesión sin cifrar que puso de manifiesto la escala potencial del fraude por suplantación de identidad. Entre los tipos de fraude CNP, el fraude por suplantación de identidad es especialmente peligroso porque el banco no puede distinguir al titular legítimo de la cuenta del defraudador. Este mismo año, sólo en Estados Unidos se produjo un aumento del 81 % del fraude CNP respecto al fraude en los puntos de venta [16], donde se estimó que los bancos incurrieron de nuevo en gastos de procesamiento que oscilaban entre 5$ y 3.011$ por devolución de cargo [17] al que hay que sumar las pérdidas de los comerciantes. Las transacciones CNP están aumentando en todo el mundo, y el fraude CNP está creciendo a un ritmo aún más rápido. Las transacciones CNP representan entre el 60 % y el 70 % de todo el fraude con tarjetas en muchos países desarrollados, en el Reino Unido [18], el gasto en comercio electrónico ha alcanzado los 248 mil millones de libras, y las pérdidas por fraude CNP fueron de 309 millones de libras en 2016. Según Australian Payments Network (AusPayNet), el fraude CNP representó más del 80 por ciento de todo el fraude con tarjetas en Australia entre julio de 2016 y junio de 2017, lo que supuso hasta 443 millones de dólares australianos en pérdidas [19]. En economías emergentes como Sudáfrica, el Centro de Información de Riesgos Bancarios de Sudáfrica informó que el fraude con tarjetas de crédito aumentó un 1 por ciento, donde el fraude CNP sigue lideró las pérdidas por fraude, con un aumento del 7,4 por ciento respecto al año anterior, y representando el 72,9 por ciento de las pérdidas en las tarjetas de crédito emitidas.

Las entidades financieras deben mantenerse alerta e innovar continuamente para detectar y prevenir el fraude online, debido a su complejidad. Por lo general, el fraude CNP es perpetrado por criminales que usan datos de tarjetas de crédito robadas.

2.2. Tipos de fraude

En la industria de servicios financieros, la batalla contra el fraude nunca cesa. Se da de la siguiente manera; si la industria establece nuevas protecciones de seguridad, los defraudadores se las ingenian para evitarlas una y otra vez. Desafortunadamente, el fraude con tarjetas de crédito no terminó con la creación del estándar EMV forzando a estar continuamente innovando. Como se describe en la metodología CyberKillChain [20], los procesos de fraude se estructuran en una serie de pasos que siguen los delincuentes para llevar a cabo sus actividades fraudulentas. Estos pasos, podemos resumirlos en tres pasos genéricos en la anatomía de estos ataques:

1. Recopilación de información. Los delincuentes utilizan los métodos descritos a continuación para poder lanzar una operación de tarjeta suplantando ser el dueño de la misma.

2. Con toda esta información, el atacante lanzara la operación de movimiento de efectivo.

3. La operación es autorizada por los bancos emisor y/o adquirente. La transferencia de fondos se ha realizado por lo que el delito se ha consumado.

El punto en el que intervienen los sistemas de ML en estudio en este trabajo, se ubican en el segundo paso. Una vez lanzada la operación con datos aparentemente reales y realizada por el dueño de la tarjeta, deben identificar el fraude y pararlo antes de que este se produzca. El problema en este punto, es que los defraudadores cada vez son más ingeniosos y han surgido numerosas nuevas técnicas para contrarrestar la seguridad del estándar EMV, estas técnicas se describen a continuación.

Fraude en las solicitudes [21]. El fraude en la solicitud es cuando alguien solicita una tarjeta de crédito con información falsa. Podemos distinguir dos situaciones diferentes, cuando las solicitudes provienen de una persona con los mismos datos, se denominan duplicados, y cuando las solicitudes proceden de individuos diferentes con datos similares, se conocen como suplantadores de identidad. Generalmente los duplicados suelen ser clientes auténticos que bien por coincidencia sus datos coinciden con el de otra persona o bien son datos antiguos sin actualizar. Por el contrario, los suplantadores de identidad, son verdaderos delincuentes que rellenan conscientemente datos erróneos en las solicitudes.

Fraude mediante la técnica de Man-In-The-Middle [22]. Una transacción típica mediante el estándar EMV se desglosa en tres partes: Parte 1: Autenticación de la tarjeta;

Parte 2: Verificación del dueño de la tarjeta; Parte 3: Autorización de la transacción. La vulnerabilidad detectada en 2011 radica en el hecho de que la tarjeta no condiciona la autorización de la transacción (parte 3) en una exitosa verificación del propietario (parte 2); es decir, en este caso, no necesariamente debe validarse la identidad del usuario para que se autorice el pago, lo que abre la posibilidad de que se realice fraude a partir de este fallo. La explotación consiste en dejar que la tarjeta autentica realice las partes 1 y 3, dejando la verificación del usuario a un dispositivo que es sujeto a ataques Man-In-The-Middle. Según los fabricantes esta vulnerabilidad fue resuelta.

Fraude mediante ataque BIN. Los números de tarjetas de crédito se producen en rangos BIN, siendo esto secuencias de números con identificación bancaria que siguen cierta norma en su construcción. Este tipo de fraude consiste en que los defraudadores generan números BIN aleatorios siguiendo la formulación de su construcción. Una vez generados estos números prueban dichos números con transacciones de pequeño importe al que tienen que generar también el código de seguridad CVV, si consiguen lograr una combinación adecuada es cuando los defraudadores pueden realizar el fraude.

Fraude mediante malware o código malicioso. Esta técnica incluye a los coloquialmente conocidos como virus, es un código informático que infecta el equipo y persigue trastornar su funcionamiento, y en este caso robar información. Son habituales spyware los cuales recopilan información sobre la navegación del usuario, contraseñas y demás datos personales y bancarios. Un tipo muy habitual son los keyloggers que registran cada tecla que se pulsa y envía la secuencia de nuestras pulsaciones al defraudador, el cual solo tiene que buscar una secuencia similar a un número de tarjeta de crédito para poder suplantarla.

Fraude mediante skimming. Se trata del robo de información de tarjetas bancarias con intención de clonarlas para utilizarlas de manera fraudulenta. Consiste en el copiado de la banda magnética que se encuentra en el dorso de las tarjetas. Los lugares habituales en los que se lleva está técnica suelen ser cajeros automáticos, bares, restaurantes o gasolineras. En el caso de los cajeros, los delincuentes roban los datos de las bandas magnéticas de las tarjetas de crédito y el PIN de las víctimas. Para lograrlo usan un dispositivo, denominado skimmer, que colocan en la abertura del lector de tarjetas y que parece formar parte del cajero automático. Asimismo, los delincuentes se sirven de cámaras que graban a los usuarios cuando teclean su PIN. Está técnica está en desuso dado que se basaba en la copia de la cinta magnética de las tarjetas, con el uso del estándar EMV en los chips y sus sistemas de seguridad evitan que se puedan clonar de una manera tan sencilla.

Fraude basado en la ingeniería social [23] [24]. Abarca todos los métodos utilizados por los delincuentes para explotar la confianza de una persona con el fin de obtener dinero directamente o información confidencial que les permita cometer un delito posterior. Los medios sociales son el canal preferido para ello, aunque no es inusual que el contacto se realice por teléfono o en persona. Las principales formas para realizar fraude mediante ingeniería social son:

- Phishing, Vishing y SMShing. Este fraude consiste en utilizar correos electrónicos, llamadas de teléfono o mensajes de texto falsos con los que el delincuente se hace pasar por una fuente legítima, como un banco o un sitio de comercio en línea, para inducir a la víctima a revelar información personal o financiera.

- Estafas por medios de telecomunicación. En este fraude el delincuente se pone en contacto con víctimas al azar y afirma ser un amigo, pariente o alguien con un puesto de autoridad, y las embauca para que envíen dinero.

- Estafa via e-mail mediante suplantación de identidad. Con estas estafas los delincuentes penetran en sistemas de correo electrónico para obtener información sobre sistemas de pago corporativos, y posteriormente engañan a empleados para que realicen transferencias a sus cuentas bancarias.

- Estafas sentimentales por Internet. Los delincuentes establecen una relación sentimental falsa con las víctimas a través de los medios sociales con el objetivo final de obtener dinero vía regalos o directamente por robo de información.

El fraude digital es una entidad en evolución y está aquí para quedarse. Las instituciones financieras deben estar atentas para mantenerse a la vanguardia y tratar la la detección y prevención del fraude como iniciativas continuas. Deben desplegar de forma proactiva sistemas inteligentes que puedan detectar el fraude, al tiempo que estos sistemas no afecten a la experiencia del cliente ni obstaculicen el proceso de pago digital. Ante esto las entidades están adoptando las medidas disponibles y tecnologías emergentes para tratar de prevenir el fraude de la mejor manera.

Las capas de seguridad de 3 dominios (3DS) son servicios de autenticación en tiempo real en la comunicación de transacciones que permiten a los bancos emisores y a los comerciantes intercambiar los datos proporcionados por los clientes para su autenticación. Verifiedby-Visa o Mastercard Secure Code son ejemplos de protocolos 3DS en los que las transacciones se inician y se autorizan después del pago mediante una contraseña o una contraseña dinámica de un solo uso (OTP) que se recibe como un mensaje de texto de

texto en el móvil y la cuenta de correo electrónico del usuario. El reto de los protocolos 3DS, sin embargo, es que la información necesaria para el registro está disponible en el mercado negro y puede ser utilizada ilegítimamente por piratas informáticos. Si la tarjeta ya está registrada, un simple registrador de claves puede dar al pirata informático acceso a la contraseña del usuario.

El Servicio de Verificación de Direcciones (AVS) es un mecanismo que puede limitar el fraude y las devoluciones de cargos. El AVS verifica la información proporcionada por titular de la tarjeta con la disponible en el banco emisor. Una vez verificada la información, el banco emisor envía un código AVS a la pasarela de pago del comerciante. El problema, sin embargo, es que incluso las transacciones auténticas pueden ser rechazadas si la información proporcionada por el cliente no coincide con la registrada por el emisor de la tarjeta.

La autenticación de dos factores (2FA) se utiliza ampliamente para asegurar las transacciones en línea. El usuario se conecta a un portal con la ayuda de una contraseña y recibe una OTP dinámica a través de un mensaje de texto en un número de móvil registrado para autenticar la transacción. Esto hace que sea más difícil para un hacker, que necesita tanto la contraseña como el teléfono del titular de la tarjeta para acceder a la cuenta. Cuando el 2FA se utiliza de forma inteligente y con moderación, puede eludir el fraude de suplantación de identidad. Sin embargo, si se utiliza de forma agresiva, puede interferir en la experiencia del usuario

2.3. Sistemas expertos

Los orígenes de los sistemas CFD para CNP se remontan a finales de la década de 1990. Inicialmente, estos sistemas se basaban en reglas definidas por expertos para bloquear los pagos fraudulentos [3]. Las actividades fraudulentas más simplistas pueden detectarse observando las señales superficiales y evidentes. Por ejemplo, las transacciones de gran tamaño o las que se producen en lugares atípicos, suelen llamar la atención, y un equipo de analistas revisan estas transacciones. Los sistemas puramente basados en reglas [25], conocidos también como sistemas expertos, implican el uso de algoritmos que realizan varios escenarios de detección de fraude, definidos por los analistas de fraude. Estos sistemas son una forma de codificar en un sistema automatizado los conocimientos de un experto humano de manera limitada en un sistema automatizado. Estos sistemas cotejan cada transacción con una lista de indicadores determinados por los analistas expertos y cuando cumple una serie de reglas, se marca como posible transacción fraudulenta [26]. La construcción de estos sistemas requiere datos etiquetados evaluados por de expertos y

las propias reglas son de la forma IF (condición) THEN (consecuencia). En la actualidad, los sistemas tradicionales aplican una media de 300 reglas diferentes para aprobar una transacción. Por ello, los sistemas basados en reglas siguen siendo demasiado simples para la detección de fraude y al mismo tiempo demasiado complejos para mantener. Requieren añadir y ajustar escenarios manualmente constantemente y apenas pueden detectar correlaciones implícitas.

Cualquier sistema basado en reglas o experto se compone de algunos elementos básicos y sencillos, como los siguientes: [27]:

1. Un conjunto de hechos. Estos hechos son en realidad las afirmaciones y deben ser cualquier cosa relevante para el estado inicial del sistema.

2. Un conjunto de reglas. Contiene todas las acciones que deben llevarse a cabo dentro del ámbito de un problema especifican cómo actuar. Una regla relaciona los hechos en la parte IF con alguna acción en la parte THEN. El sistema debe contener sólo las reglas relevantes y evitar las irrelevantes porque el número de reglas en el sistema afecta a su rendimiento.

3. Un criterio de terminación. Es una condición que determina que se ha encontrado una solución o que no existe ninguna. Esto es necesario para terminar algunos sistemas basados en reglas que, de lo contrario, se encuentran en bucles infinitos.

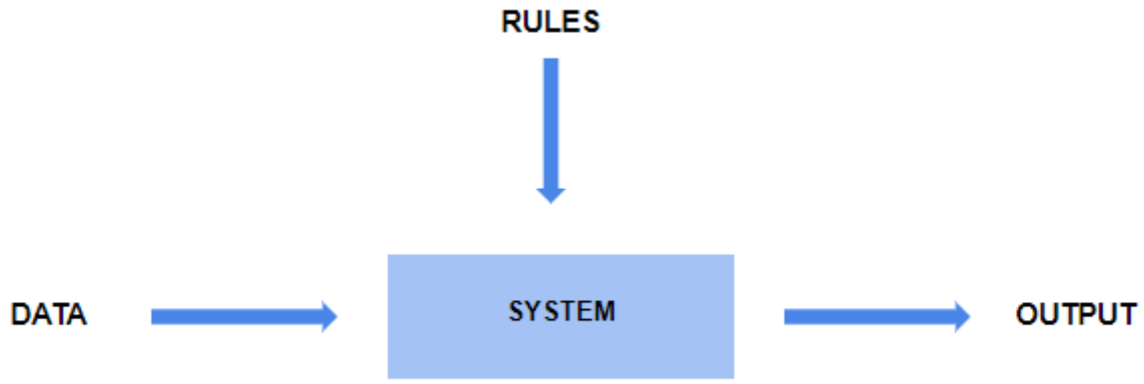


Figura 2.1: Esquema de un sistema experto. El sistema dado unos datos de entrada, en base a las reglas definidas por los analista obtiene el resultado de la predicción.

Los datos asocian el valor de las características a un hecho y a través de las condiciones definidas en las reglas se obtiene otro hecho o la decisión final del sistema. Estos sistemas,

como puede apreciarse, tienen varias limitaciones como: (i) el uso de muchas reglas tiende a dar lugar a un elevado número de falsos positivos; (ii) los umbrales de comportamiento fraudulento pueden cambiar con el tiempo; (iii) cuanto más especializados son, más caros son de mantener. Estas limitaciones, sumado a que cada vez los defraudadores son más imaginativos y veloces en sus técnicas hace que estos sistemas nos sean operativos en la actualidad para su uso en sistemas de CFD.

2.4. ML para la detección de fraude

A día de hoy el enfoque basado en reglas para abordar el problema de CFD requiere una enorme potencia de cálculo y una gran complejidad a la hora de definir y construir la base de reglas, con el fin de identificar con precisión los patrones de fraude. Además, no tiene inteligencia ni capacidad para predecir o analizar los datos de las transacciones en busca de nuevas pautas y estrategias de fraude. Es en este punto donde específicamente la AI tiene un potencial disruptivo capaz de redefinir el actual sector financiero. De hecho, podríamos decir que los sistemas basados en reglas o sistemas expertos son la forma más sencilla de inteligencia artificial. En el contexto general de la AI, destaca las técnicas de ML, que engloba modelos de predicción y reconocimiento de patrones que requieren una intervención humana mínima. Debido a ello, la aplicación de métodos de ML en el sector financiero tiene el potencial de mejorar los resultados tanto para las empresas como para los consumidores, y puede ser una poderosa herramienta contra el fraude crediticio [1, 28–30].

Existen múltiples técnicas de ML que pueden utilizarse para identificar transacciones fraudulentas. A continuación se describen brevemente los diferentes enfoques de ML y sus características.

Random Forest (RF): Los RF son un algoritmo basado en ML que se construyen mediante árboles de decisión (DT), comúnmente utilizado para resolver problemas de clasificación. Ayuda a predecir resultados con gran precisión en grandes conjuntos de datos. La técnica de RF combina varios clasificadores de tipo DT para proporcionar una solución y las predicciones obtenidas se sustentan en la media de los resultados de varios DT. Por lo general, un aumento en el número de árboles tiende a aumentar la precisión del resultado y mediante la combinación de varios DT ayuda a solucionar varias limitaciones de los algoritmo DT de manera individual [31]. Los RF al estar compuestos por varios DT, y siendo estos considerados como clasificadores débiles, el conjunto de todos estos DT hacen que los RF sean clasificador fuerte. La técnica de RF es rápida y eficaz para manejar grandes volúmenes de datos e incluso en situaciones de fuerte desbalanceo entre clases. Sin embargo, los RF tienen limitaciones a la hora de resolver problemas

de regresión. Recientes trabajos como [32] proponen un enfoque híbrido para los CFD utilizando un RF y un DT individuales, para identificar transacciones basadas en anomalías.

Red neuronal artificial (ANN): Los ANN son un algoritmo de ML que pretende imitar el razonamiento del cerebro humano. Normalmente, los ANN pueden ser tanto supervisado como no supervisado. Son algoritmos muy potentes, los cuales se han empleando en numerosos trabajos para su uso en la detección de casos de fraude obteniendo tasas de precisión elevadas [33] [34] [28] [35]. Los métodos ANN son muy tolerantes a los fallos, ya que la generación de la salida se mantiene incluso con fallos en una o varias celdas. Debido a su alta velocidad y a su eficaz capacidad de procesamiento, los ANN puede considerarse cómo una solución eficaz para el CFD. Recientes trabajos han propuesto un modelo basado ANN [36] y la retropropagación para la detección de fraudes con tarjetas de crédito, para ello se usó un conjunto de datos de los clientes, el ID de la transacción y la hora obteniendo una precisión bastante elevada mejorando los estudios previos, si bien es cierto, que los autores no indicaron como de desbalanceados están los conjunto de datos influyendo esto último mucho en la precisión obtenida. Cabe destacar que los autores no mencionaron nada sobre los datos de los clientes empleados pudiendo estar usando características prohibidas debido a la regulación. Los ANN son algoritmos eficaces que pueden utilizarse para los problemas de CFD, su gran limitación es que son modelos totalmente opacos funcionando como cajas negras lo cual impide su interpretabilidad y aceptación por los organismos reguladores.

Support Vector Machine (SVM): SVM pueden utilizarse tanto para problemas de clasificación como de regresión. Para el caso de fraude, es habitual su uso en modo clasificación, donde utilizan el conjunto de características principales para clasificar transacciones fraudulentas de las legítimas obteniendo los principales patrones de cada una de ellas. Para los casos de CFD trabajos como [37] indican que los SVM son eficaces y proporciona resultados precisos cuando se utilizan pocas características. Otros trabajos apuestan por un modelo híbrido mediante SVM junto con RF para fraude con tarjetas de crédito para los datos de mayor dimensionalidad [38]. La idea se inspira en la selección de características de las transacciones fraudulentas mediante RF para reducir su dimensionalidad al que posteriormente aplicar el SVM. Esta aproximación también mejoró el rendimiento de la clasificación sobre sus predecesoras.

K-Nearest Neighbour (KNN): KNN es un método eficaz en el aprendizaje supervisado. Ayuda a mejorar la detección y a reducir la tasa de falsos positivos. Esta técnica supervisada puede detectar la presencia de actividad fraudulenta en las transacciones con tarjetas de crédito [39]. La técnica de detección de fraude KNN se basa en la distancia entre los datos

de las transacciones. Aunque distintos trabajos han realizado aproximaciones progresivas utilizando KNN, detectaron limitaciones del algoritmo ya que KNN es un algoritmo que requiere mucha memoria [40]. Otros trabajos como [41] propusieron el algoritmo híbrido entre k-NN y el modelo oculto de Markov (HMM) para la detección de fraude con tarjetas de crédito tratando de minimizar los falsos positivos y aumentar la tasa de detección. Con esta técnica detectaron que al utilizar los dos modelos juntos, el HMM analizaba el comportamiento del usuario el cuál ayuda a minimizar las tasas de fraude obtenidas por el KNN.

Gradient Boosted (GB): El Gradient Boosted fue sugerido por Freund y Schapire [42]. Un GB implica tres elementos, una función de pérdida a optimizar, un algoritmo de aprendizaje débil para hacer predicciones y un modelo aditivo para añadir los algoritmos de aprendizaje débiles para minimizar la función de pérdida. Generalmente los DT se utilizan como algoritmo de aprendizaje débil. Tras la publicación de Friedman sobre los GB, muchos investigadores han aportado mejoras relacionadas. Un ejemplo es XGBoost [43] [44]. La principal innovación de XGBoost es que los autores agrupan varias optimizaciones en una implementación de GB. Estas optimizaciones incluyen el uso de un algoritmo aproximado para encontrar divisiones DT para conjuntos de datos dispersos, almacenamiento en caché y el uso de la compresión de datos. Trabajos recientes basados en versiones adaptadas de GB como OLightGBM [45] donde estas adaptaciones consisten en la optimización de los hiperparámetros basado en la teoría bayesiana han demostrado su efectividad en el uso de estos tipos de algoritmos para la detección de fraude de tarjetas de crédito.

Todos estos métodos logran en mayor o menor medida distinguir los datos fraudulentos de los auténticos mejorando en gran medida con respecto a los tradicionales sistemas expertos. Por contra, son más complejos de comprender lo que abre el dilema a las entidades financieras entre poder entender sus modelos o tener unos ratios inferiores de fraude.

2.5. Regulación

Los servicios financieros han mostrado un gran interés por el uso de la AI y están en un estado inicial para su adaptación y transformación. De hecho, en la revista especializada The Economist ya lo proclamaba en 2017 en su artículo *Machine learning promises to shake up large swathes of finance* [46]. Las adaptaciones requeridas por la AI, y especialmente en ML, tendrán implicaciones para los supervisores financieros preocupados por la conducta y/o funcionamiento prudente de las entidades financieras. Los supervisores tienen que considerar las oportunidades de mejora que proveen los modelos a través de la AI, así como garantizar el cumplimiento adecuado. Mientras las empresas utilizan la AI para identificar

y reducir el fraude, los reguladores están poniendo sus esfuerzos para detectar y evitar que se produzcan sesgos injustos en los algoritmos y obligando a justificar las decisiones tomadas por los algoritmos, de ahí la importancia de la transparencia.

En 2021, el Parlamento Europeo recibió una solicitud por parte del Consejo de la Unión Europea para su posicionamiento ante la propuesta de ley por el que se establecen normas armonizadas en materia de inteligencia artificial, proponiendo un enfoque basado en los riesgos del uso de sistemas basados en AI en Europa los cuales repercutirá en el uso y desarrollo de los sistemas de AI dentro del sector financiero [47]. Los puntos más destacados de la Ley de IA para las empresas de servicios financieros son:

- La vigilancia del mercado no tiene como objetivo garantizar la seguridad y la solidez de las instituciones financieras, sino que se centra en la protección de los intereses de las personas que podrían verse afectadas por sistemas de AI abusivos, garantizando que dichos sistemas cumplan los requisitos necesarios para asegurar un alto nivel de protección de los intereses públicos, como la salud y la seguridad de las personas.

- Se invita al legislador de la Unión a considerar en qué medida diversos elementos de la evaluación de la conformidad podrían no ser de naturaleza prudencial al relacionarse principalmente con la evaluación técnica de los sistemas de AI para proteger la salud y la seguridad de las personas y velar por el respeto de los derechos fundamentales minimizando el riesgo de procesos asistidos por AI que sean erróneos o sesgados.

- El Banco Central Europeo (BCE) entiende que la evaluación de la conformidad de los sistemas de AI proporcionados por las entidades de crédito para evaluar la solvencia de las personas físicas o establecer su calificación crediticia forma parte de un control interno ex-ante realizado por la entidad de crédito. En este sentido, el reglamento propuesto debería modificarse para reflejar el carácter ex-post de la evaluación específica que debe realizar el supervisor prudencial.

Cómo puede observarse en estos puntos, los reguladores velan por la protección de las personas antes que por las instituciones financieras, en los que priman reducir posibles sesgos que se puedan cometer en cualquier proceso automático. Por ello, es fundamental forzar a un control ex-post de las decisiones tomadas con el objetivo de poder explicar esas decisiones ante cualquier petición por parte de los reguladores. Es por ello, que la propia naturaleza opaca de los métodos de ML impone límites significativos para la redacción de la normativa [6]. Las entidades financieras insisten en la necesidad de una orientación adicional sobre cómo interpretar la normativa vigente. Para romper esta barrera, la interpretabilidad de estos modelos es fundamental.

2.6. Sesgos

La toma de decisiones automáticas basadas en algoritmos tiene claras ventajas a diferencia de las que requieren personas ya que pueden tratar muchas más información por unidad de tiempo sin sufrir de agotamiento ni cometer errores. Sin embargo, al igual que las personas, los algoritmos son vulnerables a los sesgos que hacen que sus decisiones sean injustas [48] [49]. En el contexto de la toma de decisiones, la equidad es la ausencia de cualquier prejuicio o favoritismo hacia un individuo o grupo basado en sus características inherentes o adquiridas [50]. Así pues, un algoritmo injusto es aquel cuyas decisiones están sesgadas hacia un grupo concreto de personas. Para tratar de dar una definición algorítmica, se han realizado varios esfuerzos. Como por ejemplo, basándose en las métricas de equidad basadas en las métricas de clasificación estándar calculadas en diferentes subpoblaciones [51]. Otras aproximaciones se centran en el proceso que lleva a los resultados. Una forma de conseguir esto último es buscar la equidad a través del desconocimiento, es decir, evitar el uso de cualquier característica sensible [52]. Sin embargo, la omisión de características sensibles puede comprometer el rendimiento de la clasificación

Un ejemplo clásico de algoritmo injusto proviene de una herramienta utilizada por los tribunales de Estados Unidos para tomar decisiones sobre la detención y la puesta en libertad. El software, es conocido como COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) empleado en el sistema de justicia penal de Estados Unidos y evalúa el riesgo de que una persona de volver a cometer otro delito. A alto nivel, COMPAS debe ayudar a los jueces a determinar si un acusado debe permanecer en la cárcel o quedar libre mientras espera el juicio. Se entrena con datos históricos de acusados para encontrar correlaciones entre diversas características como la edad, sexo y sus antecedentes. Mediante estas correlaciones calcula una probabilidad de que un acusado sea detenido por otro delito durante el período de espera del juicio. De acuerdo con la ley, COMPAS no incluye la raza en el cálculo de su puntuación de riesgo. Sin embargo, en 2016, una investigación [48] reveló que, aun así, que la herramienta sufría un sesgo contra los afroamericanos. Este sesgo tiene como origen que los detenidos no tienen la misma proporcionalidad entre razas. Las predicciones reflejan los datos utilizados para crearlas y esto es así independientemente del algoritmo que se utilice. Este extraño conflicto de justicia no sólo ocurre en los algoritmos de evaluación de riesgo del sistema de justicia penal. Se han hecho hallazgos similares en otras áreas, como por ejemplo, los modelo de predicción de impagos de tarjetas de crédito que están sesgados para que los clientes que tienen un buen historial de pagos de crédito obtengan una mayor puntuación más alta que los que no lo tienen. Estas predicciones sesgadas se derivan de los sesgos ocultos de los datos

o de los algoritmos. No existe ningún algoritmo capaz de solucionar este problema. De hecho, ni siquiera es un problema algorítmico, las personas toman todo tipo de decisiones en base a sus creencias que se basan en sus vivencias las cuales también pueden estar sesgadas. Muchas tareas de entrenamiento de ML buscan replicar los juicios humanos, y esos juicios pueden estar basados en sesgos conscientes o inconscientes existentes. Es muy probable que cualquier dato de entrenamiento que provenga de un juicio humano contenga prejuicios sociales existentes. Por ejemplo, estudios realizados [53] encontró una fuerte tendencia a describir a los atletas blancos como trabajadores e inteligentes, mientras que los atletas negros eran descritos como físicamente poderosos y atléticos.

Recientemente, se han propuesto diferentes aproximaciones [54] [55] para abordar la injusticia algorítmica de los modelos de ML sin comprometer su rendimiento de clasificación. Se basan en modelos preentrenados que hacen que los algoritmos sean más justos al disminuir su dependencia de las características sensibles. Sin embargo, esto repercute en la precisión de los modelos al no permitir su especialización.

Debido a todo esto los organismos reguladores se cuestionan con la siguiente pregunta, ¿Cómo se puede cuantificar matemáticamente la justicia?. Está es una difícil cuestión, no tiene una respuesta fácil al estar las personas también regidas por sesgos, pudiendo llegar a la paradoja de no poder definir matemáticamente la justicia sin caer sesgos. En este sentido, la interpretabilidad de los algoritmos en su toma de decisión juega un papel fundamental para poder discernir si un algoritmo está demasiado condicionado a alguna característica en concreto y poder razonar a partir de ahí si es un modelo sesgado.

2.7. Interpretabilidad

En la actualidad, diferentes trabajos [1, 28–30] analizan cómo a desarrollar modelos de ML contra el fraude crediticio. De estos trabajos se puede observar que el uso de ML para generar modelos de predicción puede mejorar la eficiencia, reducir los costes, mejorar la calidad y aumentar la satisfacción del cliente [7]. Sin embargo, uno de los grandes retos y un obstáculo potencialmente grande en estos modelos es su falta de transparencia en la toma de decisiones. Estos modelos suelen ser cajas negras, ya que sólo conocemos sus entradas y salidas, pero no los procesos que se ejecutan en su interior. Esto hace que sean difíciles de comprender en su totalidad, y sus propiedades sean complicadas de validar, por lo que ciertas formas de riesgo podrían pasar totalmente desapercibidas.

El posicionamiento de los organismos reguladores se centra en los resultados no deseados, incluidos los sesgos involuntarios que puedan llevar a resultados discriminatorios que pueden surgir de los automatismos y la falta de transparencia de los modelos de AI. En este sentido, podemos considerar la interpretabilidad cómo el proceso o metodología que hace

transparente o explicable cómo un algoritmo de AI obtiene a un determinado resultado. Las expectativas que tienen los reguladores respecto a la transparencia de los modelos empleados y a que dichos modelos no sean cajas negras, no son nuevas. Las autoridades reguladoras están compuestas por personas, y en este sentido las explicaciones deben ser entendidas por y para las personas. Del mismo modo, los modelos de decisión deben ser fácilmente comprensibles y deben permitirnos comprobar qué atributos influyen en las decisiones para producir explicaciones que sean comprensibles [13]. Dichas cajas negras tienen implicaciones para los reguladores financieros, que tendrán que tener en cuenta los potenciales beneficios que aporta los modelos basados en ML sobre sus carencias de opacidad y tratar de dar unas guías claras de cómo deben ser explicados los algoritmos para evitar estos sesgos. Por ejemplo, Estados Unidos prohíbe la discriminación basada en datos tales como la raza, el sexo, el origen, religión, el estado civil o la edad y no sólo eso, ya que incluso si el algoritmo no utiliza directamente ninguno de estos datos pero utiliza información altamente correlacionada con los datos anteriores se considera una infracción. La Unión Europea, no se queda atrás siendo incluso más complicada, donde el Reglamento General de Protección de Datos desde 2016 otorga a sus ciudadanos el derecho a recibir explicaciones sobre las decisiones basadas únicamente en el tratamiento automatizado [6].

La principal diferencia entre la AI y modelos más tradicionales es el nivel de complejidad y la falta de explicabilidad de ciertos tipos de algoritmos de ML. Explicar los complejos algoritmos de ML de una manera que pueda ser entendida por un regulador es todo un reto, dado que los reguladores tienen que capacitar a su personal para que conozca las técnicas de ML. Por su lado, las empresas deben esforzarse más por explicar sus modelos de fraude basados en ML de forma comprensible. La gran mayoría de algoritmos de ML, como pueden ser las redes neuronales o el deep learning (DL), se consideran modelos de caja negra totalmente opacos, ya que producen resultados muy precisos pero que son sumamente difíciles de explicar o probar, a diferencia de los modelos estadísticos tradicionales. Esto se debe a que tales algoritmos de ML trabajan a través de complejas interacciones entre múltiples características o variables en las distintas capas. Esto también dificulta la divulgación a los interesados, en un lenguaje claro y sencillo, de los datos que se utilizan y de cómo afectan al proceso de decisión [56].

La transparencia de un algoritmo de ML es un requisito indispensable para cumplir con la regulación. Si un modelo no es transparente, será difícil evaluar su fiabilidad, rendimiento y equidad, aparte de evaluar el resultado del modelo con respecto a un punto de referencia específico. También sin interpretablididad es difícil establecer la responsabilidad si no está claro qué componentes del algoritmo están causando errores. Las instituciones financieras deben aplicar un buen criterio a la hora de determinar el nivel adecuado de transparencia de sus modelos de ML en función del público al que van dirigidos y de la importancia de

los resultados de los modelos. Los reguladores destacan que cuanto más crítico es el caso de uso, más importante debe ser la transparencia de un algoritmo. De lo contrario, podría refutar los resultados del modelo. Una transparencia inadecuada también podría erosionar la confianza del consumidor o disuadir a los clientes de utilizar soluciones financieras basadas en el ML.

2.8. Interpretabilidad vía modelos sustitutos locales

Entender por qué un modelo realizó una determinada predicción es crucial para lograr la confianza, la equidad, la responsabilidad y la transparencia requerida por los organismos reguladores. Muchos algoritmos de aprendizaje automático, como el DL, ANN, RF y un largo etcétera se tratan como modelos de caja negra debido a su intrincada estructura. Su complejidad puede dar lugar a una alta precisión, pero también a una mala interpretación, lo que suele ser un compromiso fundamental en ML. Para los CFD al tratarse de áreas muy reguladas, no hay lugar para modelos de difícil comprensión o para arquitecturas sin la adecuada transparencia en el uso de los datos. Para hacer frente a esto, y según la literatura existente, las entidades financieras recurren a la utilización de modelos sencillos e interpretables, como los árboles de decisión [12] que es lo más parecido a un sistema experto con reglas o los modelos lineales [13] los cuales pueden explicar una predicción a través de la importancia de las características. Este tipo de modelos son fáciles de entender y sus predicciones pueden explicarse de manera sencilla. En el caso de los árboles de decisión, por ejemplo, la interpretación puede seguirse a través de las ramas, y en el caso de los modelos lineales, las interpretaciones dependen de los pesos de cada característica del modelo. En otra dirección, otros trabajos se centran en nuevas estrategias conocidas como modelos sustitutos locales y, en concreto, en las explicaciones interpretables locales agnósticas del modelos (LIME) [57] [58]. Estos modelos sustitutivos no requieren ningún conocimiento del funcionamiento interno del modelo empleado de tipo caja negra siendo sólo necesario conocer los datos y las predicciones, en contraste con los métodos tradicionales que son específicos para cada modelo, siendo sólo aplicables para un único tipo de algoritmo. En este último método, en lugar de entrenar un modelo sustituto global, utilizan sustitutos locales para aproximar de manera individual cada una de las predicciones del modelo de caja negra subyacente. Para ello, modifican una única instancia ajustando los valores de las características con pequeñas perturbaciones y observando el impacto en el resultado. Al hacerlo, LIME genera una alternativa aproximada interpretable, agnóstica e individual significativa al modelo de datos original de caja negra. Dicho proceso se puede observar en la Figura 2.2. Los creadores de LIME han lanzado un nuevo enfoque de LIME, mediante anclajes [59], que generan conjuntos de reglas de alta precisión en lenguaje sencillo para

describir la predicción de un modelo de aprendizaje automático en términos de entrada del modelo. K-LIME [60] es una modificación de LIME, donde en lugar de realizar pequeñas perturbaciones se generan mediante K clusters.

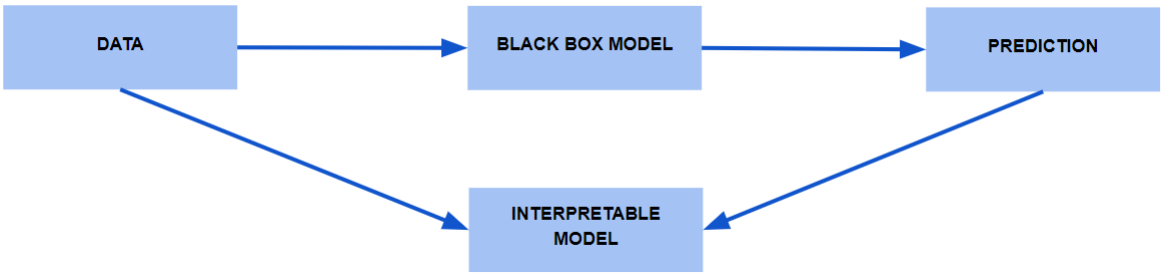


Figura 2.2: Esquema de modelos sustitutivo. Aproximación mediante un modelo sustitutivo para aproximar el modelo de caja negra en una instancia o transacción individual.

Los modelos sustitutos logran detectar las características e interacciones más importantes del modelo complejo, especialmente cuando se combinan con la Dependencia Parcial (DP) [61] y gráficos de Expectativa Condicional Individual (ICE) [62]. Sin embargo, la simplicidad necesaria para que el modelo sea explicable contrasta con la necesidad de fidelidad. La fidelidad y la interpretabilidad comparten una relación muy fina, por lo que este método tiene ciertas limitaciones [63]. Para garantizar una buena fidelidad, suele ser necesario un modelo más complejo y esto nos lleva a la disyuntiva de que un modelo complejo es difícil de explicar.

El proceso de creación de un modelo sustituto también se conoce como extracción del modelo. En trabajos como [64] donde el algoritmo TREPAN y en [65] el algoritmo DeepRed extraen el modelo sustitutivo cómo un árbol de decisión a partir de redes neuronales. Trabajos similares se encuentran en [66], [67] y [68], donde las redes neuronales profundas son objeto de ingeniería inversa. En un trabajo más reciente, el modelos sustitutivo descrito en [69] obtiene un árbol de decisión a partir de un modelo de caja negra mediante remuestreos en el conjunto de entrenamiento para obtener interpretabilidad.

Otro método bien conocido para alcanzar la interpretabilidad es el de las explicaciones de SHapley Additive exPlanations (SHAP) [70], donde las características de un problema de aprendizaje automático se tratan como jugadores en un juego de coalición de la Teoría de Juegos. Lloyd Shapley introdujo en 1953 [71] un concepto de solución para encontrar una distribución de la ganancia total entre N jugadores que cooperan. En el contexto de ML y DL, los valores de Shapley pueden utilizarse para producir un enfoque unificado para interpretar los modelos. En particular, el resultado de un modelo f puede explicarse localmente con un modelo g más sencillo y lineal donde el resultado del modelo puede

verse como la cooperación entre las característica de entrada. A cada característica se le asigna un valor específico, denominado valor de Shapley, que demuestra su contribución al resultado. Aunque los valores de Shapley producen explicaciones de alta calidad, han surgido nuevas implementaciones basadas en modelos en árboles de decisión, utilizando el algoritmo Tree SHAP de [72].

2.9. Sistemas comerciales para CFD

En el mercado hay multitud de estrategias de defensa para evitar el fraude en transacciones y cada uno de ellos se especializa en evitar o dificultar a los delincuentes cometer fraude. Hay productos que se especializan en evitar que los delincuentes puedan recopilar la información que necesitan de los clientes para hacer el ataque. En este grupo destacan los productos conocidos como AntiBots, los cuales intentan evitar que Bots puedan recopilar información de clientes, como el Bot Manager de Akamai [73]. Otros tipos de productos intentan identificar que el usuario que está lanzando la operación es realmente es el usuario auténtico. En este campo destacan productos como BehavioSec [74] o Trusteer de IBM [75]. Otra familia de productos utilizada se centra en validar que la información utilizada sea veraz y/o no empleada anteriormente en otros fraudes. En esta categoría estarían Anti-Fraud Suite de ThreatMark [76] o TruValidate de TransUnion [77]. El cuarto tipo de productos intenta identificar una transacción legítima de una fraudulenta cuando ya se ha lanzado. Este tipo de productos están basados en reglas y/o en ML. De los más actuales destacan Decision Manager de VISA [78], ARIC Risk Hub de FeatureSpace [79]. El clásico en este apartado el Falcon de FICO [80].

En 2021 según el estudio realizado por Gartnet [81] el mercado está integrado por un conjunto de empresas que intenta aglutinar dentro de sus plataformas productos que puedan dar a sus clientes servicios de protección end-to-end. Combinan soluciones y estrategias con las especializaciones descritas anteriormente.

Sólo una parte de estos productos/servicios incorpora AI y de ellas sólo una empresa ofrece ML interpretable. La empresa es seon.io [82] la cual identifica como una de sus ventajas el contar con un modelo de ML interpretable, que podría facilitar los análisis de explicabilidad de sus algoritmos. Si bien es cierto, que la interpretabilidad que proponen está basada únicamente en el algoritmo C5.0 que no deja de ser un DT muy simple siendo un estándar habitual por los analistas de fraude el cuál carece de la potencia de otros algoritmos de ML. Aunque de manera muy incipiente, la explicabilidad de los motores de ML comienzan a usarse como herramienta de marketing siendo una ventaja competitiva dentro los fabricantes del mercado CFD aún no está resuelta limitándose a soluciones sencillas.

Cuadro 2.1: Tabla resumen de las principales empresas que proporcionan plataformas antifraude.

Empresa	**Producto o servicio**
Accertify, an American Express Company	Múltiples productos
ACI Worldwide	Proactive Risk Manager
ai Corporation	Múltiples productos
Akamai	Bot Manager
BAE Systems	NetReveal
Bottomline	Cyber Fraud and Risk Management
Callsign	Múltiples productos
Cequence Security	Bot Defense
Cybersource, a Visa Solution	Decision Manager
DataVisor	dCube
F5	Sin nombre de producto específico
Featurespace	ARIC Risk Hub
FICO	Falcon
IBM	Trusteer
Imperva	Advanced Bot Protection
Kount, an Equifax Company	Múltiples productos
LexisNexis Risk Solutions	BehavioSec
Microsoft	Dynamics 365 Fraud Protection
Netacea	Bot Management
NICE Actimize	IFM-X
NuData Security, a Mastercard solution	NuDetect
PerimeterX	Bot Defender
Ravelin	Fraud solution suite
River Security	River Dynamic Security (Botgate)
seon.io	Seon
SHIELD	Múltiples productos
Sift	Múltiples productos

Capítulo 3

Métodos

Este capítulo está estructurado como sigue. En primer lugar, se presentan y describen todos los conjuntos de datos. En segundo lugar, se presenta una breve referencia de los principales algoritmos de ML tanto lineales como no lineales utilizados en este trabajo. Por último, se presenta la metodología general implementada en la presente tesis.

3.1. Conjuntos de datos

De los principales problemas en la literatura de CFD podemos destacar, por un lado, la falta de información debido a la confidencialidad de los datos por lo que no es fácil encontrar conjuntos de datos representativos, informativos y públicos. Y por otro lado, estos conjuntos de datos suelen estar muy desequilibrados ya que siempre son más numerosas las transacciones legítimas frente a las fraudulentas [1]. Por esta razón, en la presente tesis se ha utilizado primero un conjunto de datos sintético para validar nuestra propuesta, allanando así el camino para el análisis posterior en conjuntos de datos reales.

3.1.1. Conjunto de datos sintético

El primer conjunto de datos consiste en datos sintético preparado para un problema de clasificación lineal con una variable de salida binaria, fue desarrollado en la propuesta original del algoritmo IVI [9], consta con 485 características de entrada, de las cuales hay características informativas, redundantes y no informativas. En la presente tesis por razones de representabilidad y tiempo de ejecución, hemos utilizado un subconjunto de características manteniendo los nombres de las mismas. Este subconjunto se seleccionó con las primeras características de cada grupo. El conjunto de datos utilizado para este trabajo incluye un conjunto de 23 características de entrada distribuidas de la siguiente manera: 11 características con distribución normal, 5 de ellas se utilizan para generar

linealmente la variable de salida binaria, concretamente $f0, f1, f2, f3$, y $f4$. Por lo tanto, estas cinco características serán informativas para el problema de clasificación. Se crea aleatoriamente un conjunto de otras 6 características sin relación con las anteriores, por lo que podrían considerarse como características ruidosas o no informativas. Por último, se generan 6 características redundantes con las características informativas añadiendo cierto grado de ruido.

3.1.2. Conjunto de datos de crédito alemanes

Este conjunto de datos es conocido como German Credit Fraud (Stattog)˜ [83], dicho conjunto de datos contiene datos reales utilizados para evaluar las solicitudes de crédito en un una entidad financiera alemana. Hemos utilizado una versión de este conjunto de datos elaborada por la Universidad de Strathclyde, el cual contiene información sobre 1.000 transacciones de solicitudes de préstamos para créditos alemanes. Cada transacción está definida mediante un conjunto de 20 características diferentes con una variable de salida binaria. De estas 20 características, 17 son categóricas y tres son valores continuos y no contiene valores nulos. Para facilitar el proceso de selección de características principales y con el fin de entrenar los modelos de la mejor manera, los valores de los tres atributos continuos se normalizaron, y para las características discretas se convirtieron mediante la técnica conocida como *one hot encoding* (OHE). Tras estas etapas de preprocesamiento, el conjunto de datos final y empleado para esta tesis consta de 61 dimensiones. Información más detallada de este conjunto de dato puede encontrarse en [83] así como una breve descripción en la Tabla 3.1.

3.1.3. Conjunto de datos PaySim

El conjunto de datos conocido como PaySim simula las transacciones de dinero móvil a partir de una muestra de transacciones reales extraídas de un mes de transacciones de un servicio de transferencia de dinero vía móvil implantado en un país africano [84]. Las transacciones originales fueron proporcionadas por una empresa multinacional, que es la proveedora del servicio financiero móvil que actualmente funciona en más de 14 países de todo el mundo. PaySim cubre cinco de los tipos de transacciones más importantes: entrada de efectivo, salida de efectivo, débito, pago y transferencia. El conjunto de datos de PaySim contiene 6.362.620 transacciones. Cada transacción está definida por un conjunto de 11 características diferentes. Por razones de rendimiento, en esta tesis hemos seleccionado un subconjunto de transacciones con 25.867 transacciones seleccionadas aleatoriamente manteniendo una distribución con un 80 % de transacciones no fraudulentas y un 20 % de transacciones fraudulentas. La información detallada de cada característica se puede

Cuadro 3.1: Tabla resumen de las características del conjunto de datos de crédito alemanes.

Característica	**Descripción**
Status	Estado de la cuenta corriente existente
Duration	Duración del crédito en meses
Credit history	Historial de solicitud de créditos
Purpose	Objeto del crédito
Amount	Importe del crédito
Savings	Cuenta de ahorro
Employment	Años en el último empleo
Personal status	Estado civil
Other parties	Otros deudores
Property magnitude	Bienes inmuebles en propiedad o seguro de vida
Age	Edad
Housing	Vivienda de alquiler o propia
Number of credits	Número de créditos existentes
Job	Trabajo actual
Telephone	Teléfono propio
Foreign worker	Es trabajador extranjero
Other payment plans	Otros préstamos a plazos
Credit balance	Saldo medio de crédito
Location	Ubicación
Overdraft	Histórico de descubierto en cuenta

encontrar en [84] y en la Tabla resumen 3.2.

3.2. Algoritmos

Como se indicaba en capítulos anteriores, las estrategias de CFD existen desde finales de los años 90 y desde entonces han estado basados mayoritariamente en sistemas expertos. En los últimos tiempos, ha comenzado a emerger sistemas basados en ML, debido a que estas nuevas técnicas ofrecen mejores resultados en términos de precisión [1, 3, 7, 28]. Un buen ejemplo de estas técnicas son los clasificadores basados en métodos lineales, los cuales generan una transformación lineal de las entradas para obtener la clase resultado. Esta estrategia relaciona las contribuciones de las características de entrada con el resultado de forma trazable, ofreciendo una interpretabilidad directa del resultado final [9].

En los siguientes párrafos, realizaremos un breve resumen de los principales algoritmos de ML basados en métodos lineales y no lineales los cuales han sido utilizados para el desarrollo de la presente tesis Doctoral. Como punto de partida, introduciremos primero la notación que se utilizará a lo largo de todo el trabajo. Sea $\mathbf{X} \in \mathbf{R}^{N \times L}$ la matriz de datos de entrada la cual está formada por el conjunto de vectores en filas, siendo estas filas N

Cuadro 3.2: Tabla resumen de las características del conjunto de datos de PaySim.

Característica	**Descripción**
Step	Unidad de tiempo
Type	Ingreso, gasto, débito, pago y transferencia (Cash-in, cash-out, debit, payment, and transfer).
Amount	Importe de la transacción en moneda local.
NameOrig	Cliente que inició la transacción.
OldbalanceOrg	Saldo inicial antes de la transacción.
NewbalanceOrig	Saldo final después de la transacción.
NameDest	Cliente destinatario de la operación.
OldbalanceDest	Saldo inicial del destinatario antes de la transacción.
NewbalanceDest	Saldo final del destinatario después de la transacción.
IsFraud	Indicador de si la operación es fraudulenta o no.
IsFlaggedFraud	Marca automática por las reglas de negocio. En esta entidad financiera se marcan las transacciones que tratan de transferir más de 200.000 en una sola transacción para ser analizadas por un analista de fraude.

observaciones de L características, donde $\mathbf{x}_n$ corresponde con la observación n definida como un vector con L características o columnas para $n = 1, \ldots, N$. Consideramos la detección de fraude como un problema de clasificación con una variable de salida binaria agrupada en el vector $\mathbf{y} \in R^N$, tal que $\mathbf{y}_l \epsilon \{-1, +1\}$ para $n = 1, \ldots, N$. La relación entre cada característica de entrada y la clase de salida está representada por los pesos de las características en el vector $\mathbf{w}$ obtenido por un método de clasificación lineal, de modo que $\mathbf{y}_n = \mathbf{w}^T \mathbf{x}_n + b$ representa la función de clasificación, donde el signo de $\mathbf{y}_l$ es utilizado como la salida de la decisión para determinar si una transacción es considerada fraudulenta o no.

3.2.1. Algoritmos lineales

A continuación se presenta el resumen de los algoritmos de ML lineales que han sido empleados en el desarrollo de esta tesis.

Regresión lineal

La regresión lineal (LR) modela la relación entre múltiples variables de entrada ajustándolas mediante una ecuación lineal para obtener la clase resultado. Una de las variables se considera variable dependiente o resultado ($\mathbf{y}_n$), mientras que el resto de variables o características $\mathbf{x}_n$ se consideran variables predictivas [85]. En términos generales,

esta relación puede que no se ajuste para toda la base muestral, por lo que es necesario añadir un término de ruido o términos de error a la fórmula matemática (ε). Podemos definir la ecuación general correspondiente a un modelo de regresión lineal como:

$$\mathbf{y}_l = \mathbf{w}^T \mathbf{x}_n + b + \varepsilon_l \tag{3.1}$$

Análisis lineal discriminante

El análisis lineal discriminante (LDA) es una generalización del modelo de análisis discriminante de Fisher [86]. LDA es capaz de encontrar una combinación lineal de variables de entrada que caractericen dos o más conjuntos con fines de clasificación. La principal diferencia entre LR y LDA es que los modelos de LR tratan con una variable dependiente continua, mientras que los modelos de LDA deben tener una variable dependiente discreta. El algoritmo de LDA tiene como objetivo representar una única variable dependiente como una combinación lineal de las variables predictoras. Dadas dos clases de observaciones multidimensionales, con medias $\mathbf{m}_0, \mathbf{m}_1$ y covarianzas $\mathbf{\Sigma}_0, \mathbf{\Sigma}_1$, la combinación lineal de características $\mathbf{w}^T\mathbf{x}$ con media $\mathbf{w}^T\mathbf{m}_i$ y varianza $\mathbf{w}^T\mathbf{\Sigma}_i\mathbf{w}$, para $i = 0, 1$. La separación entre estas dos distribuciones puede definirse como la relación entre la varianza entre las clases y la varianza dentro de las clases, es decir,

$$S = \frac{\sigma_b^2}{\sigma_w^2} = \frac{\left(\mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_0)\right)^2}{\mathbf{w}^T (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_0) \mathbf{w}} \tag{3.2}$$

donde σ_b^2 representa las varianzas entre las clases y σ_w^2 representa las varianzas dentro de las clases. Se puede demostrar que la máxima separación se produce cuando

$$\mathbf{w} = c (\mathbf{\Sigma}_0 + \mathbf{\Sigma}_1)^{-1} (\mathbf{m}_1 - \mathbf{m}_0) \tag{3.3}$$

donde c es una constante. Para los espacios de características de entrada de alta dimensión con covariables altamente correlacionadas, este algoritmo puede mostrar cierta inestabilidad; por lo tanto, a menudo se aplica la regularización de la inversión de la matriz y utilizamos la siguiente ecuación:

$$\mathbf{w} = c (\mathbf{\Sigma}_0 + \mathbf{\Sigma}_1 + \lambda \mathbf{I})^{-1} (\mathbf{m}_1 - \mathbf{m}_0) \tag{3.4}$$

donde λ es el parámetro de regularización, y $\mathbf{I}$ es la matriz de identidad.

Support Vector Machines

Los clasificadores ML convencionales generalmente se ven muy afectados debido a varios factores como son: la alta dimensionalidad en los datos de entrenamiento, sobreajustarse,

o por un ajuste deficiente debido a realizar la fase de entrenamiento con pocas muestras. En los últimos años, el uso de SVM [87, 88] ha gozado de un gran auge. Los SVM son algoritmos de aprendizaje supervisado que se utiliza en muchos problemas de clasificación (SVC) y regresión (SVR). El objetivo del algoritmo SVM es encontrar un hiperplano que separe con el mayor margen posible clases diferentes de puntos de datos. El margen se define como la anchura máxima de la región paralela al hiperplano que no tiene puntos de datos interiores. Intuitivamente, una buena separación se consigue con el hiperplano que tiene la mayor distancia al punto de datos de entrenamiento más cercano de cualquier clase, ya que en general cuanto mayor sea el margen, mejor será el error de generalización del clasificador [89, 90].

A diferencia de los algoritmos ML anteriores, la SVM mapea el vector de entrada a un espacio de mayor dimensión. Los SVM pueden resolver problemas lineales y no lineales, y funciona bien para muchos problemas prácticos. Para la presente tesis se ha empleado un kernel lineal. Queremos encontrar el hiperplano de margen máximo que divide el grupo de puntos $\mathbf{x}_k$, para el cual $\mathbf{y}_k = 1$ del grupo de puntos $\mathbf{x}_m$,para el cual $\mathbf{y}_m = -1$. Dada un mapeo no lineal $\phi(\cdot)$, el SVM resuelve:

$$\underset{\mathbf{w},\mathbf{b},\beta_l,\rho}{\text{mín}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \beta_l \right\} \tag{3.5}$$

En consideración a

$$\mathbf{y}_l \left(\langle \phi(\mathbf{x}_l), \mathbf{w} \rangle + b \geq 1 - \beta_l \right), \ \forall l = 1 \ldots L \tag{3.6}$$

donde $\mathbf{w}$ y b definen un clasificador lineal en el espacio de características y β son variables de holgura positivas que permiten tratar los errores de clasificación. La elección adecuada del mapeo no lineal ϕ garantiza que las muestras transformadas tienen más probabilidades de ser linealmente separables en el espacio de características. La Ecuación (3.5) se resuelve utilizando counterpart, dejando $\mathbf{w} = \sum_{l=1}^{L} y_l \alpha_l \phi(\mathbf{x}_l)$, y la función de decisión para cualquier vector de test $\mathbf{x}_*$ finalmente se define por

$$f(\mathbf{x}_*) = sgn \left(\sum_{l=1}^{L} y_l \alpha_l K(\mathbf{x}_l, \mathbf{x}_*) + b \right) \tag{3.7}$$

donde α_l son los multiplicadores de Lagrange correspondientes a las restricciones de la Ecuación (3.5), cuyas muestras de entrenamiento $\mathbf{x}_l$ con $\alpha_l \neq 0$ siendo vectores; el término de error b se calcula utilizando los multiplicadores de Langrange no limitados; y K representa los Kernels de Mercer utilizados para manejar las implementaciones de algoritmos no lineales.

Gradient Boosting

El Gradient Boosting (GB) es una técnica de ML para problemas de regresión y clasificación que produce un modelo de predicción mediante la combinación de modelos de predicción débiles. GB construye el modelo por etapas y generaliza los modelos permitiendo la optimización de una función de pérdida [42, 91, 92]. El objetivo es encontrar una aproximación $\hat{F}(\mathbf{x})$ a una función $F(\mathbf{x})$ que minimice el valor esperado de alguna función de pérdida especificada $L(\mathbf{y}, F(\mathbf{x}))$; es decir,

$$\hat{F} = \arg\min_{F} \mathbb{E}_{\mathbf{X},\mathbf{y}}[L(\mathbf{y}, F(\mathbf{X}))] \tag{3.8}$$

El algoritmo GB asume un valor real $\mathbf{y}$ y busca una aproximación de $\hat{F}(\mathbf{X})$ la forma de una suma ponderada de funciones $h_m(\mathbf{x})$ llamadas clasificadores débiles, de tal manera que

$$\hat{F}(\mathbf{x}) = \sum_{m=1}^{M} \gamma_i h_i(\mathbf{x}) + c \tag{3.9}$$

donde M es el número de modelos débiles que se utilizan.

3.2.2. Algoritmos no lineales

A continuación, se presenta el resumen de los algoritmos de ML no lineales que se han empleado en la presente tesis como son los *autoencoder* y sus variantes.

Autoencoder

Un *autoencoder* [93] (AE) es un tipo específico de red neuronal, que está diseñado para codificar una entrada del espacio real en una representación comprimida y significativa, y luego decodificarla de nuevo de manera que la entrada reconstruida sea lo más similar posible a la original. El potencial de los *autoencoder* es comprimir datos de alta dimensión en una de representaciones latentes o de menor dimensionalidad, por eso se definen como dos partes: un *encoder* (codificador) y un *decoder* (decodificador), donde el codificador aprende a mapear el espacio de entrada de alta dimensión a un espacio vectorial de menor dimensión o latente, y el decodificador mapea el espacio vectorial latente al espacio de entrada original sin comprimir. En general, la matriz de datos de salida $\hat{\mathbf{X}}$ es el resultado de reconstruir la matriz de datos de entrada original $\mathbf{X}$. Podemos ver la arquitectura de un *autoencoder* básico en la Figura 3.1.

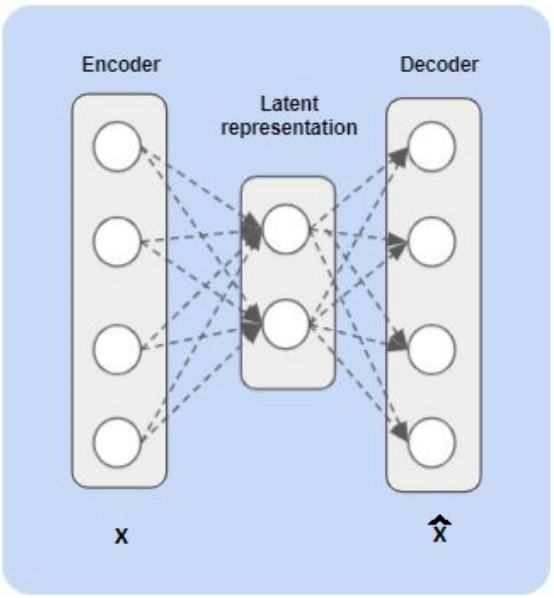


Figura 3.1: Arquitectura de un autoencoder con una sola capa tanto para el en encoder como para decoder.

El problema, definido formalmente en [94], consiste en la transformación desde un dominio de dimensión L o R^L hacia un espacio de menor dimensión, R^P, obtenido mediante el codificador, seguido de una segunda transformación desde el espacio latente R^P hacia el espacio reconstruido R^L, mediante el decodificador.

Autoencoder variacional

En los últimos tiempos, han surgido modelos de *autoencoders* con diferentes enfoques, uno de estas variantes la cual ha tomado gran relevancia es conocida como *autoencoders* variacionales (VAE) [94]. Los VAE son modelos de aprendizaje que mezclan los *autoencoders* con distribuciones de probabilidad. Su principal uso es el de construir modelos generativos que son capaces de producir datos sintéticos que siguen los mismos patrones que los conjuntos de datos con los que se entrenan. Un caso de uso habitual de los VAE, son la generación de imágenes que asemejan, por ejemplo, características conocidas tales como caras, vehículos, habitáculos, etc. En esta tesis hemos usado estos modelos para generar transacciones sintéticas con propiedades realistas y viables que nos permiten dotar de interpretabilidad a transacciones reales como veremos en los próximos capítulos.

En concreto, dado un conjunto de datos observados $\mathbf{X}$, asumimos un modelo generativo para cada dato $\mathbf{x}_i$ condicionado a una variable aleatoria latente no observada $\mathbf{z}_i$, donde $\boldsymbol{\theta}$ son los parámetros que rigen la distribución generativa. Este modelo generativo también es equivalente a un decoder probabilístico. Simétricamente, asumimos una distribución

posterior aproximada sobre la variable latente $\mathbf{z}_i$ dado un dato $\mathbf{x}_i$, lo que equivale a un codificador probabilístico que se rige por los parámetros $\boldsymbol{\phi}$. Finalmente, tenemos una distribución para las variables latentes $\mathbf{z}_i$ denotada por $\mathbf{p}_0(\mathbf{z}_i)$. La verosimilitud marginal se expresa como una suma sobre los puntos de datos individuales como se expresa a continuación,

$$log\, \mathbf{p}_0(\mathbf{x}_i) = D_{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i)\,||(\mathbf{p}_0(\mathbf{z}|\mathbf{x}_i)\right) + \boldsymbol{\phi}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}_i) \tag{3.10}$$

donde el primer término es la divergencia de Kullback–Leibler del modelo aproximado con respecto a la muestra origen y el segundo término se denomina límite inferior variacional de la probabilidad marginal, definido como se expresa a continuación:

$$\boldsymbol{\phi}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}_i) \doteq \mathbf{E}_{\mathbf{q}_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i)}\left[-log\, \mathbf{q}_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) + log\,(\mathbf{p}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{x})\right] \tag{3.11}$$

La inferencia variacional se realiza maximizando $\boldsymbol{\phi}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}_i)$ para todos los puntos de datos con respecto a $\boldsymbol{\theta}$ y $\boldsymbol{\phi}$.

Fine-tuning en autoencoders

Con el uso intensivo de los *autoencoders* y con el objetivo de mejorar su rendimiento se emplea la técnica conocida como ajuste fino o *fine-tuning* la cual es utilizada habitualmente en otros algoritmos y dada su potencia se ha adaptado a los *autoencoders*. El objetivo del *fine-tuning* es realizar un ajuste en los pesos del modelo una vez entrenado para mejorar el resultado de la predicción. Este procedimiento, basado en el concepto de *transfer learning* [95], incluye un paso de pre-entrenamiento inicial seguido de procedimientos de entrenamiento adicionales con un objetivo discriminativo sobre el mismo conjunto de datos [96], pero algunos otros estudios siguen el proceso de reutilización de los pesos de grandes conjuntos de datos como inicializador de los pesos en aplicaciones con acceso limitado a datos etiquetados [97].

Sea $\mathbf{X} \in \mathbf{R}^{N \times L}$ la matriz de datos de entrada, que contiene el conjunto de vectores de entrada en filas, con N observaciones de L características. Consideramos la variable de salida del espacio latente $\mathbf{Y} \in \mathbf{R}^{P \times N}$, siendo P el tamaño del espacio de características de reducción o del espacio latente. Podemos observar el Algoritmo 1 que resume el funcionamiento, donde se ajusta un *autoencoder* para obtener los pesos, tras lo cual se congelan los pesos del *encoder* y se añade la capa softmax para su reajuste.

Algoritmo 1 Fine-tuning.

Require: Conjunto de entrenamiento $\mathbf{X}$ y clase de resultado $\mathbf{Y}$,

1: Inicialización del autoencoder $AE = \{\ \}$.
2: Ajuste del AE.
$AE \leftarrow AE.fit(\mathbf{X})$
3: Congelamos los pesos
4: Separamos AE en encoder (enc) y decoder (dec)
5: Añadimos capa Softmax al enc
$enc' \leftarrow enc + softmaxlayer$
6: Reajustamos enc' los pesos de todas las capas congeladas excepto de la softmax
$enc' \leftarrow enc'.fit(\mathbf{X}, \mathbf{Y})$
7: Ajustamos enc' con todas las capas
$enc' \leftarrow enc'.fit(\mathbf{X}, \mathbf{Y})$

3.3. Métricas

A continuación se presenta un resumen de las métricas empleadas en la presente tesis.

3.3.1. Coeficiente de correlación de Kendall

Para evaluar la similitud entre diferentes transacciones, hemos utilizado el coeficiente de correlación de Kendall. El coeficiente de correlación de Kendall es un estadístico utilizado para medir la concordancia entre conjuntos de datos ordenados como puede ser un ranking de valores.

Sea $(\mathbf{a}_1, \mathbf{b}_1), \ldots, (\mathbf{a}_n, \mathbf{b}_n)$ un conjunto de observaciones de variables aleatorias $\mathbf{A}$ y $\mathbf{B}$, tal que todos los valores de $(\mathbf{a}_i)$ y $(\mathbf{b}_i)$ son únicos. Cualquier par de observaciones $(\mathbf{a}_i, \mathbf{b}_i)$ y $(\mathbf{a}_j, \mathbf{b}_j)$, donde $i < j$, se dice que son pares concordantes si el orden de clasificación de $(\mathbf{a}_i, \mathbf{a}_j)$ y $(\mathbf{b}_i, \mathbf{b}_j)$ coincide, es decir, si tanto $(\mathbf{a}_i > \mathbf{a}_j)$ como $(\mathbf{b}_i > \mathbf{b}_j)$ coinciden o tanto $(\mathbf{a}_i < \mathbf{a}_j)$ como $(\mathbf{b}_i < \mathbf{b}_j)$, en caso contrario se dice que son discordantes. El coeficiente de Kendall τ se define como

$$\tau = \frac{n_c - n_d}{\binom{n}{2}} \tag{3.12}$$

donde n_c es el número de pares concordantes, n_d es el número de pares discordantes, y $\binom{n}{2}$ es el número total de combinaciones de pares. En el coeficiente de correlación de Kendall, el denominador es el número total de combinaciones de pares, por lo que el coeficiente debe estar en el rango $-1 \leq \tau \leq 1$.. Si la concordancia entre las pares es perfecta (es decir, los dos rankings son iguales), el coeficiente tiene valor 1. Si hay discrepancia total entre dos rankings (es decir, un ranking es el inverso del otro ranking), el coeficiente tiene un valor de -1. Si $\mathbf{a}$ e $\mathbf{b}$ son independientes, entonces esperaríamos que el coeficiente

fuera aproximadamente cero.

3.4. Metodología

En esta sección se describen brevemente los cinco pasos de la metodología propuesta para dotar de interpretabilidad a los sistemas de CFD. En la figura 3.2, representamos gráficamente la arquitectura propuesta del proceso descrita paso a paso, como se indica a continuación.

- Paso 1: Selección de características relevantes mediante el algoritmo IVI.
- Paso 2: Aplicación de los filtros MIFF y RFF.
- Paso 3: Aislar las transacciones fraudulentas de las no fraudulentas mediante la compresión del espacio real en espacio latente.
- Paso 4: Interpretabilidad mediante STE. Evaluación del peso de las características para cada transacción individual.
- Paso 5: Generación de cluster mediante ITR.

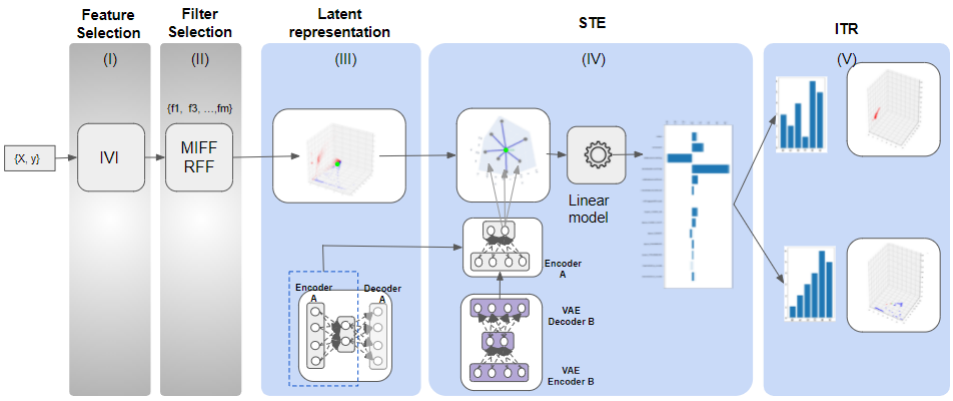


Figura 3.2: Metodología. Los cinco pasos se representan aquí de forma esquemática: (i) selección de características mediante IVI; (ii) filtrado de variables mediante MIFF y RFF; (iii) Aislamiento de las transacciones fraudulentas y evaluación del espacio latente; (iv) modelado de STE; y (v) caracterización de ITR.

Cada uno de los pasos de dicha metodología se verá en detalle tanto formalmente como analíticamente junto con sus correspondientes experimentaciones en los capítulos siguientes.

Capítulo 4

Selección de características

Como se mencionaba en capítulos anteriores, una de las contribuciones de esta tesis consiste en seleccionar las características más relevantes aplicada al problema de CFD. En este capítulo se describen los fundamentos de la metodología aplicada y resultados obtenidos para la selección de características. Para ello, presentamos a continuación en detalle el algoritmo IVI y su utilización, así como las mejoras propuestas. Esta metodología se ha planteado en base a la hipótesis inicial de que la interpretabilidad puede obtenerse en función de la importancia de las características. Con esto en mente, reflejamos la contribución de cada característica en el proceso de decisión calculando sus pesos en diferentes métodos de ML y consolidando estas contribuciones. Una vez seleccionadas las contribuciones de las características, se han desarrollado dos tipos de filtros con el objetivo de evitar sesgos y obtener una visión completa del problema.

4.1. Identificador de variables informativas

En este trabajo se ha utilizado un novedoso método de selección de características (FS) llamado IVI [9], que es capaz de identificar las variables informativas, redundantes y ruidosas. Este método transforma la distribución del espacio de las variables de entrada en un espacio en base a los coeficientes o pesos de las características empleando los clasificadores lineales definidos en el capítulo 3.2.1. Las características informativas y sus relaciones se determinan analizando la distribución conjunta de estos coeficientes con técnicas de remuestreo. Mediante IVI logramos seleccionar las variables informativas para posteriormente poderlas emplear en los clasificadores pertinentes para obtener un incremento en precisión. Los experimentos han demostrado que IVI puede superar a los algoritmos del estado del arte en términos de detección de las características relevantes, e incluso en términos de precisión cuando se emplean dichas características seleccionadas en clasificadores posteriores. El algoritmo IVI se basa en la hipótesis inicial de que los pesos,

$\mathbf{w}$, aprendidos por un método clasificador lineal, $\mathbf{y}_l = \mathbf{w}^T\mathbf{x}_l + b$, son capaces de resumir la relación entre cada característica. El algoritmo IVI, introducido en el trabajo original [9], fue implementado mediante el estimador de covarianza multiplicativa (CME), el cual es un método de generación de pesos diseñado para ser competitivo con respecto a los algoritmos lineales estándar. El CME es un generador de pesos de bajo coste computacional que se basa en las relaciones entre las características de entrada, y en las relaciones entre las variables de entrada y salida. Dado un conjunto de datos de entrada y una variable de clase, $\{\mathbf{X}, \mathbf{y}\}$, donde $\mathbf{X} \in R^{N\times L}$ es la matriz de datos de entrada, que contiene el conjunto de vectores de entrada en filas, con N muestras u observaciones de L características y $\mathbf{y} \in \{-1, +1\}$. $\mathbf{C_{XX}}$ y $\mathbf{C_{Xy}}$, denotan las matrices de covarianza de las muestras estimadas como

$$\mathbf{C_{XX}} = \frac{1}{N}\mathbf{X}^T\mathbf{X} \in R^{L\times L} \tag{4.1}$$

$$\mathbf{C_{Xy}} = \frac{1}{N}\mathbf{X}^T\mathbf{y} \in R^{L\times 1} \tag{4.2}$$

El vector de coeficientes l-dimensional de la CME se define como sigue:

$$\mathbf{w} = \left(sign\left(\mathbf{C_{XX}}\right)^{(g-1)} \odot \mathbf{C_{XX}}^{(g)}\right)\mathbf{C_{Xy}} \tag{4.3}$$

donde $\odot$ denota el producto Hadamard, y g es el exponente de la potencia del producto Hadamard. Mediante el uso del algoritmo IVI, seleccionamos diferentes grupos de características relevantes, para ello, este algoritmo consta de tres etapas diferenciadas que se resumen a continuación:

El primer paso del algoritmo IVI consiste en estimar la distribución estadística de los pesos obtenidos por el criterio que se decida, como se ha comentado en el trabajo original se estableció mediante CME, mientras que en la presente tesis se ha ampliado mediante el uso clasificadores lineales definidos en 3.2.1. Para ello, se transforma el espacio de características de entrada en un espacio de pesos, y mediante remuestreo estadístico del conjunto de datos nos proporciona una estimación de la distribución que nos permite calcular el conjunto de pesos para cada característica de entrada. Esta transformación se resume en el Algoritmo 2.

Se seleccionan N_g filas en $\mathbf{X}$ y $\mathbf{y}$ para obtener el remuestreo $\mathbf{X}^*_{(b)}$, $\mathbf{y}^*_{(b)}$. Para cada remuestreo, hay que estimar la magnitud estadística de interés dado por el vector de pesos $\mathbf{w}^*_{(b)}$. Repitiendo este procedimiento B veces, se obtiene una estimación de la distribución empírica marginal:

$$\widehat{f}^*_{\mathbf{w}}(\mathbf{w}) = \frac{1}{B}\sum_{b=1}^{B}\delta\left(\mathbf{w} - \mathbf{w}^*_{(b)}\right) \tag{4.4}$$

donde $\delta(\mathbf{w})$ muestra la función delta de Dirac. Hecho esto, los resultados del primer paso

del algoritmo IVI son las ponderaciones de las características que posteriormente se utilizan para identificar la relevancia y la redundancia de las características en el algoritmo 2.

Algoritmo 2 IVI-Algorithm Step 1.

Require: Conjunto de entrenamiento $\mathbf{X}$ y $\mathbf{y}$, y número de remuestreos, B.
Ensure: Matriz de pesos remuestreada, $\mathbf{W}^* \in R^{B \times L}$.

1: **for** $b \leftarrow 1$ to B **do**
2: Generar un subconjunto aleatorio del conjunto de entrenamiento $\mathbf{X}_{(b)}$ y $\mathbf{y}_{(b)}$, con tamaño L_b.
3: Calcular el vector de pesos $\mathbf{w}^*_{(b)}$ empleando $\mathbf{X}_{(b)}$ y $\mathbf{y}_{(b)}$ como conjunto de datos de entrenamiento.
4: Guardar el vector con los pesos $\mathbf{w}^*_{(b)}$ en la columna número b columna de la matriz $\mathbf{W}^*$.
5: **end for**

En el segundo paso, se utilizan las propiedades estadísticas de las distribuciones marginales de los pesos para identificar las características informativas y se agrupan las que se consideran mutuamente redundantes. Esto se resume en el Algoritmo 3. Adicionalmente, IVI identifica las características informativas y las redundantes obteniendo las propiedades estadísticas de los pesos. Este proceso se realiza mediante el análisis de los intervalos de confianza de los pesos remuestreados asociados a cada característica de entrada, con ello se utiliza las correlaciones de los pesos para localizar y agrupar las características mutuamente redundantes. En este punto, con estos grupos disjuntos, se pueden distinguir las características informativas de las ruidosas, con lo que se selecciona sólo los grupos disjuntos informativos. Para ello, se considera que los grupos informativos son los grupos disjuntos con al menos una característica identificada. El siguiente paso consiste en descartar los grupos de característica que compartan información pero no contengan ninguna característica relevante, las cuales se consideran características ruidosas. Para ello se define el algoritmo 4 para calcular el umbral y los grupos de características redundantes.

Algoritmo 3 El algoritmo IVI, paso 2. Intervalos de confianza para identificar las características relevantes.

Require: Matriz de pesos $\mathbf{W}^*$ y nivel de confianza para las pruebas de significación, α.
Ensure: Características etiquetadas como relevantes

1: Mediante $\mathbf{W}^*$ se construyen intervalos de confianza de nivel α.
2: Almacenar la lista de características con intervalos de confianza que no se superpongan a cero y se marcan como relevantes.

Algoritmo 4 Paso 2.2 del algoritmo IVI. Identificación de la redundancia de características.

Require: Remuestreo de la matriz de pesos, $\mathbf{W}^* \in R^{B \times L}$; número de hojas, k, para calcular el umbral.

Ensure: Umbral para identificar características redundantes, $p_{\check{z}}$, y grupos disjuntos.

1: Dividir el conjunto$\mathbf{W}^*$ en k subconjuntos, $\mathbf{W}_i^*$, con $i = 1, \ldots, k$.
2: **for** l, m $l \neq m \leftarrow 1$ a N **do**
3: Calcular el valor absoluto del coeficiente de correlación de Pearson de los pesos B remuestreados de las características l y m en $\mathbf{W}^*$ y guardar consecutivamente en el vector $p_{\mathbf{W}}$.
4: Para cada hoja, calcular el valor absoluto del coeficiente de correlación de Pearson de las características l y m en $\mathbf{W}_i^*$, dando $p_{\mathbf{W}}^i$.
5: **end for**
6: Guarda el valor medio de los $p_{\mathbf{W}}^i$ en el vector $\bar{p}_{\mathbf{W}}^k$
7: Ordenar $p_{\mathbf{W}}$ y $\bar{p}_{\mathbf{W}}^k$ en orden descendente por $p_{\mathbf{W}}$.
8: Calcular la diferencia acumulada entre$\bar{p}_{\mathbf{W}}^k$ y $p_{\mathbf{W}}$, y buscar la fila que contenga la mínima diferencia acumulada. Definir el umbral, $p_{\check{z}}$, como el valor de correlación en la fila $r_{\check{z}}$ del vector $p_{\check{z}}$ cumpliendo

$$r_{th} = \arg\,\text{mín}_r \sum_{i=1}^{r} \left(\bar{p}_{\mathbf{W},i}^k - p_{\mathbf{W},i}\right)$$

9: Las características redundantes se definen como pares de características con pesos de correlación superiores al umbral.

El último paso consiste en una clasificación de las características seleccionadas en orden descendente de importancia. Para ello, primero hay que ordenar en orden descendente de importancia las características de cada uno de los grupos disjuntos obtenidos con IVI. Una vez realizado esto, algunos grupos pueden contener más de una característica relevante por lo que hay que separar esos grupos en subgrupos que incluyan sólo una característica relevante cada uno y su relación redundante más directa. Por último, los grupos y subgrupos resultantes se ordenan para producir la clasificación final. Para ello, primero hay que encontrar las características relevantes en los grupos disjuntos identificados por IVI, por lo que utilizaremos una medida de importancia: Imp_l, que se calcula para cada característica l como el valor absoluto del peso medio y dividido por el cuadrado del rango del intervalo de confianza del 95 % de los pesos remuestreados; es decir,

$$Imp_l = \frac{|mean(\mathbf{W}_l^*)|}{\left(\mathbf{w}_l^{h,*} - \mathbf{w}_l^{n,*}\right)^2}, l \,\epsilon\, I \tag{4.5}$$

donde $\mathbf{w}_l^{h,*}$ es el valor del intervalo superior,$\mathbf{w}_l^{n,*}$ que corresponde al intervalo inferior, y I es un conjunto de características informativas. Cuando en un grupo disjunto sólo contiene dos características, la mayor importancia se determina por la característica más relevante del grupo. Las demás características se toman como versiones redundantes de las anteriores.

Para grupos más grandes, puede haber más de una característica relevante. Para resolver esto, exploramos el conjunto completo de subgrupos de características que obtendríamos si aumentáramos el umbral de correlación de pesos utilizado para construir el grupo disjunto inicial. A continuación, definimos la importancia de un (sub)grupo de características G como el valor absoluto de la suma de la importancia de las características del grupo. En consecuencia, es probable que los grupos o subgrupos con una suma de importancia muy superior a 1 incluyan más de un rasgo relevante. Por el contrario, los subgrupos poco informativos sin rasgos relevantes tienen una suma de importancia bastante inferior a 1, mientras que los subgrupos con un solo rasgo relevante tuvieran una suma de importancia cercana a 1. Por lo tanto, para cada grupo disjunto inicial, se seleccionan la configuración de los subgrupos que arroja las sumas más bajas de importancia de los subgrupos. Cada uno de estos subgrupos contiene una característica relevante y su copia redundante. Por último, hay que clasificar los grupos y subgrupos de características resultantes. En este proceso se utilizan una medida de importancia diferente que da más peso a la dispersión de las estimaciones de los coeficientes de las características. Se calcula la importancia de cada característica y se normaliza dividiéndola por la importancia máxima. Los grupos de características se ordenan teniendo en cuenta la suma de sus importancias. La clasificación final de los rasgos se basa en este ordenamiento de los grupos para mostrar primero los rasgos relevantes en cada (sub)grupo informativo, y luego, siguiendo el mismo orden descendente de importancia de los grupos, los rasgos redundantes.

4.2. Adaptación de IVI

En esta tesis, tiene como objetivo crear una metodología fiable, imparcial e interpretable para medir automáticamente el riesgo de CFD, donde uno de los temas principales (T1) consiste en la selección de características relevantes. En este sentido, se ha definido una metodología de tres pasos para obtener las características principales de manera que sea interpretable. El flujo completo del sistema del modelo propuesto se muestra en la Figura 4.1. Esta metodología se describe secuencialmente paso a paso como sigue.

- Paso 1: Extraer las características informativas comunes aplicando el algoritmo IVI.
- Paso 2: Filtrado de las características informativas.
- Paso 3: Interpretabilidad basada en la ponderación de las características.

El primer paso de esta metodología consiste en una etapa de FS, que nos permite encontrar las características relevantes y reducir el ruido en los modelos. Como se ha señalado anteriormente, utilizamos IVI [9], que es capaz de identificar la característica

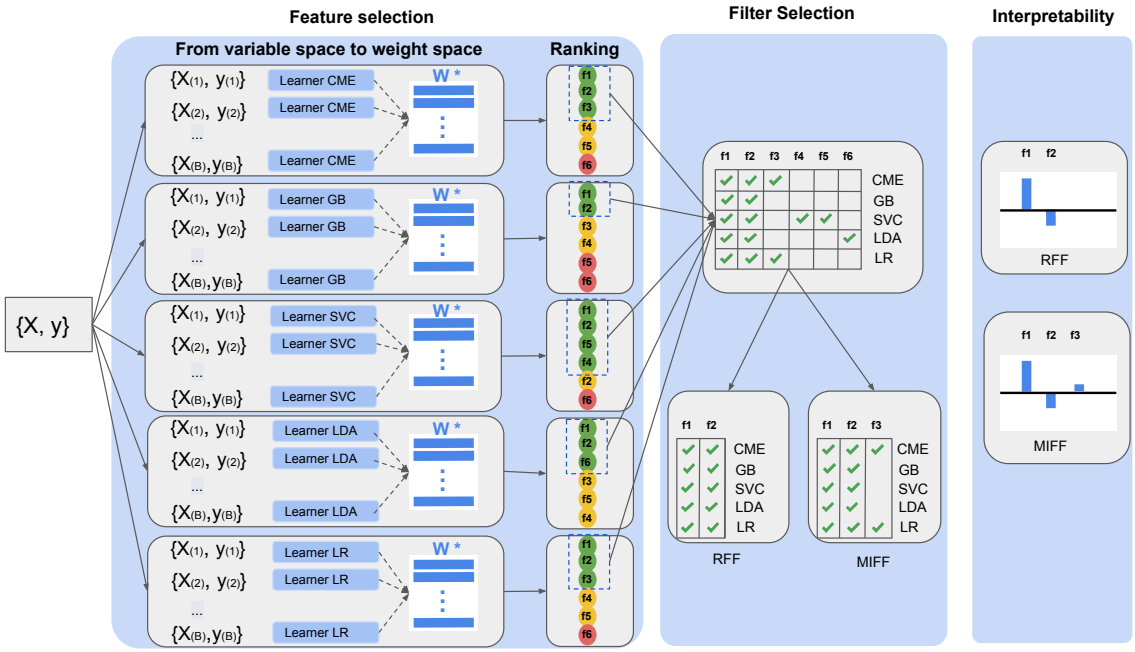


Figura 4.1: Metodología para la selección de características. Las características se transforman en un espacio de coeficientes utilizando IVI con diferentes algoritmos de ML (CME, GB, SVC, LDA y LR). Mediante el uso de intervalos de confianza y relaciones entre características, componemos los rankings de características informativas, a continuación, aplicamos los filtros para detectar las características informativas. Finalmente, en el último paso, se emplean las características obtenidas en el filtrado que entrenamos a través con modelos lineales de los cuales obtenemos los pesos asignados a cada una de estas características. Esos pesos reflejan la relevancia de cada característica en el proceso de decisión.

más relevante, el cual en su trabajo original se implementó con CME como un método de generación de pesos diseñado para ser rápido en la generación de los pesos. Para esta tesis, se ha ampliado el generador de pesos utilizando diferentes algoritmos de clasificación. El motivo de esta ampliación de algoritmos es doble. Por un lado, pretendemos obtener una visión más completa del problema, debido a que cada algoritmo tiene sus particularidades y propiedades específicas. Por otro lado, las características que se han seleccionado en múltiples algoritmos son más consistentes. Para ello, en este trabajo hemos ampliado los métodos a SVM, LDA, LR y GB, además de CME. El algoritmo para obtener las características relevantes por cada uno de los algoritmos de ML se resume como se muestra en el Algoritmo 5.

Algoritmo 5 Metodología paso 1.

Require: Conjunto de entrenamiento **X** y **y** y algoritmos ML
Ensure: Características relevantes para cada Algoritmo ML
1: Inicializar todos los algoritmos de ML $V = \{\text{CME}, \text{GB}, \text{SVC}, \text{LDA}, \text{LR}\}$
2: Inicializar el vector con las características seleccionadas $FS(v) = \{\}\,; \forall v \in V$
3: **for all** $alg \in V$ **do**
4: Ejecutar el algoritmo IVI utilizando alg
5: Obtener el vector con las características relevantes, fr
6: $FS(alg) = fr$
7: **end for**

El segundo paso de nuestra metodología se centra en reducir el sesgo y obtener una visión global del problema. En tal sentido y debido a que los algoritmos de ML pueden caer en principalmente en dos tipos de sesgos hemos desarrollado un proceso de filtrado. Por un lado, tenemos los sesgos basados en la propiedades dependientes de los datos de entrenamiento, por otro, estarían los sesgos basados en los algoritmos de ML utilizados. Para evitar estos sesgos, por un lado, mediante el empleo del algoritmo IVI, estamos haciendo un remuestreo de los datos para entrenar los algoritmos ML, y en este sentido, minimizamos el sesgo de los datos de entrenamiento. En el caso del sesgo por algoritmos ML, y una vez que hemos obtenido el FS con diferentes algoritmos ML (SVM, LDA, LR y GB, además de CME), el siguiente paso es encontrar cuáles de estas características son realmente informativas para todos los casos. Por ello, es necesario someter a las características seleccionadas para cada algoritmo a un proceso de filtrado, al final del cual, sólo las características que son consistentes se incluyen en nuestro modelo. En este sentido, en esta tesis se han establecido dos tipos de filtros sobre la extracción de características relevantes en IVI dejando de lado las características redundantes y ruidosas:

- ***Recurrent Features Filter (RFF)***: Consideramos aquellas características que han sido seleccionadas de forma recurrente en todos los algoritmos de ML empleados en IVI. Este filtro es muy restrictivo, seleccionando sólo las características más representativas, forzando una agresiva reducción de la dimensionalidad, y logrando un aumento precisión de una manera rápida.

- ***Maximally-Informative Features Filter (MIFF)***: Consideramos aquellas características que al menos han sido seleccionadas en dos de los algoritmos de ML utilizados en IVI. Este filtro es menos restrictivo y obtenemos una reducción moderada de la dimensionalidad de las características en comparación con el filtro RFF. En cambio, este filtro es capaz de identificar las relaciones más complejas entre las características logrando una mayor precisión en la predicción

Finalmente, el último paso de nuestra metodología se centra en la interpretabilidad del problema. Los clasificadores lineales basados en ML pueden verse como una transformación del espacio de las variables de entrada o características al espacio de los pesos asignados a cada una de estas características. Estos pesos resumen en realidad la contribución de cada característica en el proceso de decisión y la interacción entre los datos de entrada y salida.

Con esto en mente, se emplean las características seleccionadas tanto en los filtros MIFF y RFF para entrenar de nuevo con modelos lineales y con ello obtener una visión más completa del problema. Los pesos asignados a las características nos muestran la importancia de cada una de ellas.

Esto se muestra en el Algoritmo 6.

Algoritmo 6 Metodología paso 3.

Require: Conjunto de entrenamiento **X** y **y** y algoritmos ML
Ensure: Características seleccionadas en los filtros MIFF y RFF
1: Inicializar todas las características seleccionadas $V_{FS} = \{\mathrm{FS}_{MIFF}, \mathrm{FS}_{RFF}\}$
2: Inicializar todos los algoritmos de ML $V = \{\mathrm{CME}, \mathrm{GB}, \mathrm{SVC}, \mathrm{LDA}, \mathrm{LR}\}$
3: **for all** $alg \in V$ **do**
4: **for all** $fs \in V_{FS}$ **do**
5: Entrenar los modelos de ML alg con fs
6: Obtener los pesos de cada fs
7: **end for**
8: **end for**

4.3. Experimentación

Para la validación del método de selección de características propuesto se han desarrollado un conjunto de experimentos tanto para los conjuntos de datos sintéticos como para los datos reales. Dicha experimentación se ha definido de la siguiente manera, en primer lugar, empleamos el algoritmo IVI incorporando nuevos mecanismos de generación de pesos (SVC, GB, LR y LDA) al generador original CME, que como resultado nos generara diferentes conjuntos y subconjuntos de características relevantes. En segundo lugar, con estos conjuntos de características modelamos diferentes arquitecturas de aprendizaje que nos permiten evaluar y analizar el valor predictivo de cada uno. Este segundo análisis ofrece una visión cuantificada de la capacidad predictiva para calificar adecuadamente las diferentes opciones. En tercer lugar, se presenta una evaluación detallada del valor predictivo incremental para cada uno de los métodos previamente definidos. Para ello, se evalúa la precisión obtenida a medida que se van añadiendo características una a una y se analiza característica por característica, la velocidad de convergencia y la capacidad de predicción del conjunto. Finalmente, se realiza un análisis de los pesos de cada característica aplicado

diferentes métodos lo cual nos permite evaluar la contribución de manera interpretable de las diferentes características sobre la predicción final.

Esta sección se divide en dos subsecciones principales en las que se presentan los experimentos desarrollados sobre los conjuntos de datos sintéticos y reales. La estrategia general que ha guiado la experimentación ha sido la de examinar y ajustar el conjunto de datos sintéticos para validar la metodología, para una posterior evaluación de las capacidades de generalización en casos reales de CFD sobre los conjuntos de datos reales.

4.3.1. Análisis datos sintéticos

En este apartado, tal y como hemos introducido anteriormente, primero aplicamos la novedosa estrategia de FS basada en el algoritmo IVI para identificar las características relevantes. Este experimento se ejecutó de forma intensiva incorporando cinco algoritmos diferentes introducidos anteriormente, como son: CME, SVC, GB, LR y LDA. Este enfoque nos permite lograr una perspectiva imparcial de la efectividad real de las características seleccionadas, así como la coherencia entre los métodos. La figura 4.2 resume el resultado del algoritmo IVI para cada técnica individual de ML.

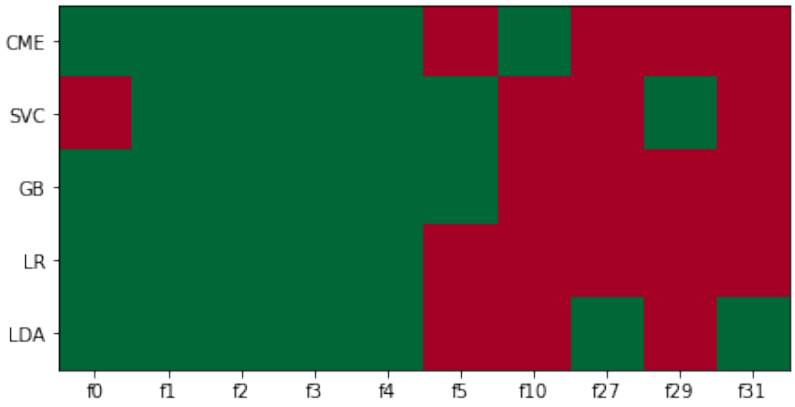


Figura 4.2: Resultados del algoritmo IVI para cada técnica de ML en el conjunto de datos sintéticos. En las filas, las diferentes técnicas de ML: CME, SVC, GB, LR y LDA. En las columnas aparecen las diferentes características definidas anteriormente. El color verde representa los escenarios en los que la característica se identificó como relevante, mientras que en color rojo representa las características que no se identificaron como relevantes durante el análisis.

En este punto, debemos recordar que el filtro RFF selecciona las características comunes identificadas como relevantes para cada uno de los métodos ML, es decir, se mostraron recurrente y consistentemente como relevantes en todos los métodos. En este caso, las características $f1$ a $f4$ se incluyeron en este conjunto, pero $f0$ no se identificó como tal

debido a la clasificación errónea por parte de SVC. En la misma dirección, las características identificadas como relevantes para al menos dos métodos se entendieron como informativos para el análisis posterior y, por tanto, se categorizaron dentro del grupo de variables MIFF. Para este caso concreto, las características $f0$ a $f5$ cumplían los criterios del MIFF y se incluyeron como miembros de este filtro. Estas características coinciden perfectamente con las características relevantes del conjunto de datos sintético ($f0$ a $f4$), añadiendo una de las características redundantes ($f5$). Atendiendo a estos resultados, podemos concluir que el algoritmo IVI fue consistente sobre los diferentes métodos de ML, lo que le confiere una capacidad de selección de características potencialmente válida.

En un intento de verificar y cuantificar los resultados, se calculó la precisión en dieciséis escenarios diferentes, para SVC, GB, LR y LDA, y considerando diferentes conjuntos de características. Estos escenarios son: (i) Clasificación empleando todas las características disponibles; (ii) Clasificación empleando sólo las características relevantes según IVI para cada método ML correspondiente; (iii) Clasificación empleando características clasificadas a través del filtro MIFF; y (iv) Características clasificadas a través del filtro RFF. El CME se descartó para esta tarea dado que no es un clasificador. Con el objetivo de validar estadísticamente los resultados, en la tabla 4.1 resume la media y la desviación estándar para las 100 ejecuciones de remuestreo que se realizaron para cada uno de los 16 escenarios diferentes. Esta tabla muestra cómo la precisión permanece prácticamente invariable a lo largo de todos los métodos de clasificación, ya que como se puede observar las diferentes columnas no reflejan casi ningún cambio en términos de media y desviación estándar. La única excepción se encuentra en el caso para el algoritmo GB, que en todos los casos, muestra una menor capacidad de clasificación en comparación con el resto de los métodos analizados. Igualmente, en el caso de LDA, cuando se aplica a todas las variables disponibles, se aprecia una ligera reducción en su capacidad de predicción.

La comparativa a través de los diferentes modelos, ilustra que los mejores resultados se obtienen cuando se utiliza el filtro MIFF y con un poder predictivo equivalente al que se alcanza cuando se utilizan todas las variables, alcanzando en ambos casos los valores predictivos del 98,8 %. La desviación estándar fue en todos los casos inferior al 4 %. El modelo IVI ofreció en todos los casos, resultados muy similares a los métodos destacados (97 % de precisión). Por el contrario, el filtrado RFF tuvo una menor capacidad predictiva en comparación con el resto de los modelos (90 % de precisión) debido a la no incorporación de todas las variables informativas como consecuencia de la detección incorrecta de las variables relevantes.

Cuadro 4.1: Resultados estadísticos con respecto a la métrica de precisión de los distintos métodos de ML en el conjunto de datos sintéticos, donde se muestra la media y desviación estándar de los resultados de cien ejecuciones de remuestreo. En las filas encontramos los resultados de los distintos métodos de ML, mientras que en las columnas tenemos el análisis para los diferentes conjuntos de características incluidos en el proceso. En las columnas, de izquierda a derecha, se incluyen todas las características disponibles, IVI, MIFF y RFF.

method_Classifier	**Acc_all_features**	**IVI (Relevant)**	**Acc_fs_MIFF**	**Acc_fs_RFF**
GB	**0.9393** $\pm$ 0,005034	0,9319 $\pm$ 0,004903	**0.9393** $\pm$ 0,005209	0,8961 $\pm$ 0,005413
SVC	**0.9870** $\pm$ 0,003195	0,9705 $\pm$ 0,003803	**0.9872** $\pm$ 0,002951	0,9062 $\pm$ 0,005112
LDA	0,9787 $\pm$ 0,003127	0,9709 $\pm$ 0,003638	**0.9877** $\pm$ 0,003356	0,9063 $\pm$ 0,004634
LR	**0.9865** $\pm$ 0,002669	0,9711 $\pm$ 0,003586	**0.9881** $\pm$ 0,003579	0,9062 $\pm$ 0,004958

Como resultado general del análisis, podríamos destacar que aunque nos hemos centrado en un número limitado de familias de algoritmos lineales de ML, cada una de ellas tratada de forma independiente reveló un rendimiento similar, y la precisión estaba estrechamente relacionada con las características incorporadas como variables de entrada. Este resultado suscita dos resultados relevantes distintos: (i) La importancia de las FS como elemento clave para un mejor rendimiento del modelo ML; (ii) La limitada relación entre la precisión y el método ML elegido, despierta la posibilidad de elegir el método centrándose en su eficiencia computacional sin limitar su capacidad de clasificación. Los resultados obtenidos en los experimentos anteriores sugirieron la necesidad de un análisis mayor y más profundo tanto de las técnicas de FS como de las propias variables, para una mejor comprensión. A tal efecto, se realizaron una serie de experimentos considerando todas las variables, para ambos filtros (RFF y MIFF). Los experimentos se diseñaron de forma que se pueda visualizar la contribución de cada variable a medida que se añaden características de forma incremental. Las figuras 4.3 y 4.4, representan los resultados de los experimentos correspondientes. En los experimentos, las características se sumaron en orden secuencial atendiendo a su relevancia, empezando por la más relevante según el método IVI a la de menor relevancia. La figura 4.3a presenta la evolución los resultados secuenciales para todos los experimentos aplicando el filtro RFF, y la figura 4.4 muestra los resultados de aplicar el filtrado MIFF, tanto en la presentación en modo M (perspectiva 3D) como en una representación de perfil. Los gráficos representan una doble imagen, ya que las líneas continuas representan los resultados correspondientes a la representación IVI estándar y, las líneas de puntos superpuestas siguen el proceso utilizando las técnicas de filtrado RFF y MIFF. Las características estándar de IVI se incorporaron a los experimentos de características incrementales en el orden representativo (concretamente, primero las informativas, luego las redundantes y, por último, las ruidosas). En cuanto a los filtros, como sólo se eligieron características filtradas para RFF y MIFF, los componentes restantes se

añadieron según la secuencia estándar de IVI para completar el conjunto de características. En las figuras 4.3 y 4.4, podemos observar la convergencia en términos de precisión utilizando los filtros RFF y MIFF. Podemos ver en estas figuras que una vez que hemos conseguido la máxima precisión con las características relevantes, la precisión no presenta variaciones al añadir nuevas características. Se puede observar que las características redundantes y ruidosas se clasifican correctamente dentro de la estrategia IVI, y no se obtuvo un incremento en precisión cuando estas últimas variables (redundantes y ruidosas) se añadieron al modelo. Como es de esperar y en vista de los resultados, tanto para el filtro MIFF como RFF, las características mal clasificadas como relevantes pueden influir a la secuencia para construir el modelo, retrasando la convergencia, o incluso limitando la potencia de clasificación. En nuestras observaciones, confirmamos que aplicando RFF, no encontramos una fuerte limitación en el poder de clasificación final (ver Figura 4.3 para RFF, y Figura 4.4 para MIFF). Ambos modelos se ajustaron sensiblemente a la precisión alcanzada por IVI, incluso superándola en el caso del filtro MIFF. Como resultado, se presenta una forma más eficiente y rápida de entrenar los modelos finales, ya que hay que incluir un menor número de características para el entrenamiento del modelo. Se realizó un tercer experimento destinado a medir la contribución de cada una de las variables al modelo final. La figura 4.6, recoge los pesos para el análisis MIFF, e indica la contribución para los diferentes experimentos. Y como se pudo apreciar en la figura, aunque todas las características presentaron pesos similares, $f3$ y $f1$ mostraron un ligero incremento entre pares. Una excepción a este razonamiento se encontró para $f5$, mostrando un peso pequeño para todos los algoritmos de ML analizados. Este resultado es coherente con el hecho de que esta característica era efectivamente una característica no relevante y mal clasificada por el algoritmo, permitiendo así un mayor ajuste del modelo.

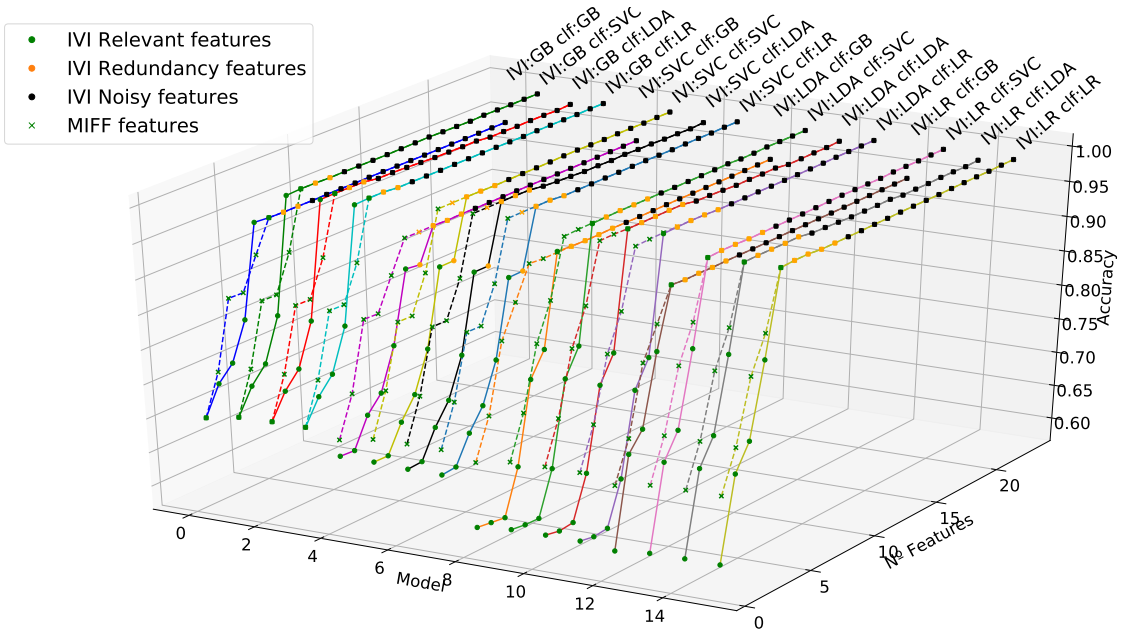


Figura 4.4: Resultados gráficos con respecto a la precisión a través de la incorporación de características en orden secuencial atendiendo a la relevancia para los filtrados MIFF, lo cual nos permiten ver la evolución en la precisión a medida que se añade cada característica. En los gráficos 3D , podemos ver el número de características en el eje x, en el eje y los diferentes algoritmos de ML (GB, SVC, LDA y LR) y la precisión en el eje z. Para cada uno de los algoritmos ML, se presenta la evolución para las características seleccionadas en IVI como línea continua y las características seleccionadas por el filtro como línea discontinua.

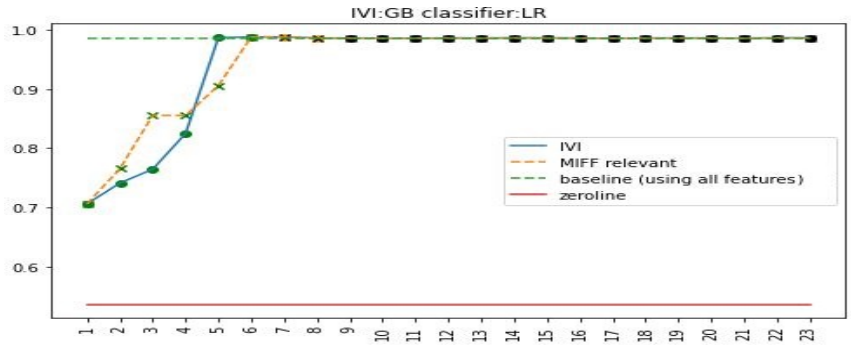


Figura 4.5: Resultados gráficos con respecto a la precisión a través de la incorporación de características en orden secuencial atendiendo a la relevancia para los filtrados MIFF. El gráfico muestra una comparativa entre los filtros y IVI. Para mayor claridad y simplicidad, sólo mostramos una muestra comparativa, donde la línea roja representa un clasificador simple que hace predicciones con la clase más frecuente, y la línea verde discontinua es el resultado del clasificador entrenado con todas las características.

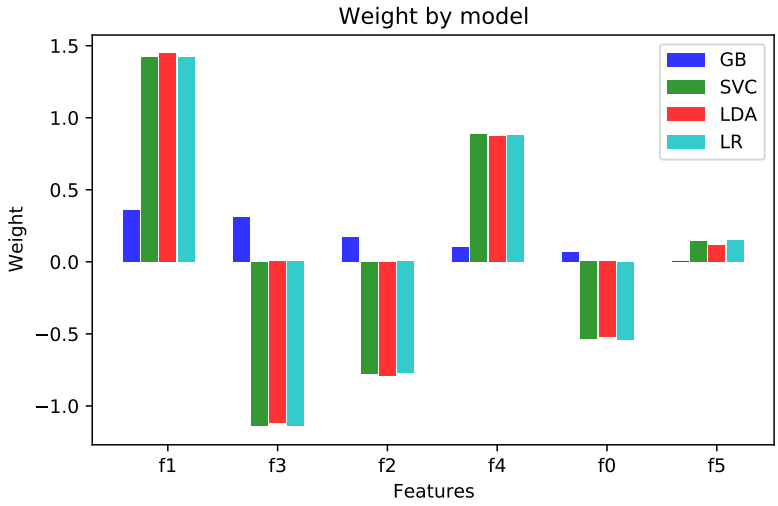


Figura 4.6: Pesos de las características. Pesos en representación gráfica de las características seleccionadas en el filtro MIFF. Cada barra de color representa diferentes técnicas de ML: GB, SVC, LDA, LR.

4.3.2. Análisis datos reales

En este experimento, aplicamos los conocimientos adquiridos en el experimento anterior para el conjunto de datos sintéticos, pero esta vez utilizarlo en datos reales. El objetivo principal de esto es verificar si los resultados obtenidos con datos reales son consistentes, cuando se utiliza la misma metodología definida para los conjuntos de datos sintéticos.

Los resultados del conjunto de datos de créditos alemanes con la técnica IVI y todos los algoritmos de ML se presentan en la Figura 4.7. En la cual, una serie de características fueron identificadas como relevantes para todos los algoritmos de ML, siguiendo el mismo marco observado en el conjunto de datos sintéticos y mostrando consistencia con los resultados descritos previamente en términos de estas características informativas repetidas. De forma equivalente al análisis descriptivo anterior, las características se clasificaron como RFF si la característica había sido seleccionada en todos los algoritmos de ML utilizados, y MIFF si la característica había sido seleccionada al menos en dos de ellos. Adicionalmente, podemos observar en la Figura 4.8 para el conjunto de datos de paysim el mismo efecto. Siguiendo con el esquema anterior, se guiaron cuatro experimentos con cien repeticiones mediante bootstrap para evaluar la significancia estadística. Los resultados de la tabla 4.2 muestran sistemáticamente que la desviación estándar es de muy pequeña magnitud, con independencia del conjunto de datos.

Los modelos emparejados mediante método y característica, mostraron una alta precisión en todos los casos alcanzando valores cercanos al 75 % para todos los casos. Bajo la perspectiva de los métodos, todos y cada uno de ellos mostraron valores similares

para los diferentes conjuntos de característica con la única excepción del método SVC, que presentó un descenso de 4 puntos porcentuales respecto al resto para el conjunto de variables RFF. Cabe destacar que este mismo método (SVC) ofreció el mejor resultado, superando sistemáticamente al resto de métodos y alcanzando un máximo del 76,63 % en el caso del filtro MIFF. Por otro lado, los resultados más bajos se encontraron de forma constante, no por ello dejando de ser buenos resultados, con el método GB, cayendo más de un punto porcentual para los conjuntos de características con respecto a su máximo, con la única excepción del RFF, donde el SVC presentó un descenso incluso mayor. Desde el punto de vista de conjunto de características seleccionados, los resultados fueron muy homogéneos en cada uno de los métodos, y singularmente mejores en el caso del MIFF, con cifras de precisión superiores al 75,10 % en todos los casos. Por el contrario, el filtro RFF no sólo mostró un comportamiento desigual en los distintos métodos, sino también resultados sistemáticamente inferiores en magnitud, además de alcanzar el mínimo. Los resultados mostrados muestran una alta correlación con los encontrados en el caso del análisis realizado con el conjunto de datos sintéticos, validando así las hipótesis formuladas durante el ejercicio.

Cuadro 4.2: Resultados estadísticos de la métrica de precisión de los distintos métodos de ML en el conjunto de datos reales. Se muestra la media y desviación estándar de los resultados para un análisis de cien ejecuciones con remuestreo.

method_Classifier	**Acc_all_features**	**IVI (Relevant)**	**Acc_fs_MIFF**	**Acc_fs_RFF**
GB	$0{,}7433 \pm 0{,}0171$	$0{,}7431 \pm 0{,}01911$	$0{,}7510 \pm 0{,}0191$	$0{,}7446 \pm 0{,}0190$
SVC	$\mathbf{0.7580 \pm 0.0170}$	$\mathbf{0.7598 \pm 0.0205}$	$\mathbf{0.7663 \pm 0.0204}$	$0{,}7256 \pm 0{,}0228$
LDA	$0{,}7533 \pm 0{,}0172$	$0{,}7463 \pm 0{,}0203$	$0{,}7573 \pm 0{,}0193$	$0{,}7443 \pm 0{,}0213$
LR	$0{,}7563 \pm 0{,}0174$	$0{,}7576 \pm 0{,}0190$	$0{,}7563 \pm 0{,}0189$	$\mathbf{0.7476 \pm 0.0209}$

Cabe destacar que todos los algoritmos de ML utilizando el filtro MIFF mostraron una mayor precisión que utilizando únicamente las características IVI, con la única excepción de LR, donde se encontró una pequeña diferencia. Los resultados fueron consistentes con los obtenidos anteriormente, y de nuevo la selección de características mediante MIFF mejoró la fase de entrenamiento en términos de eficiencia computacional, al reducir el número de características para alcanzar una mayor precisión, y así confirmar empíricamente la hipótesis establecida con el conjunto de datos sintéticos.

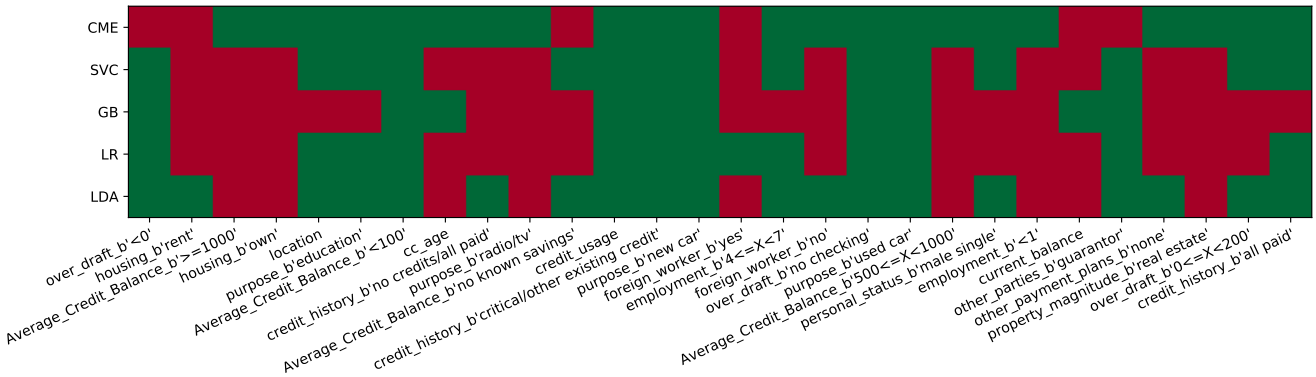


Figura 4.7: Resultados del algoritmo IVI para cada técnica ML en un conjunto con datos reales de créditos alemanes. En filas las diferentes técnicas utilizadas de ML: CME, SVC, GB, LR y LDA. En las columnas, las diferentes características. El color verde representa los escenarios en los que la característica se identificó como relevante. El color rojo representa las características que no se identificaron como relevantes durante el análisis.

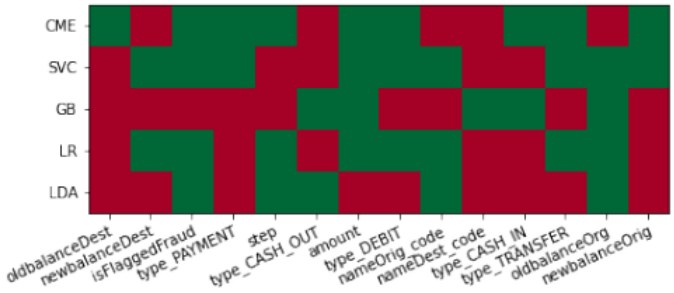


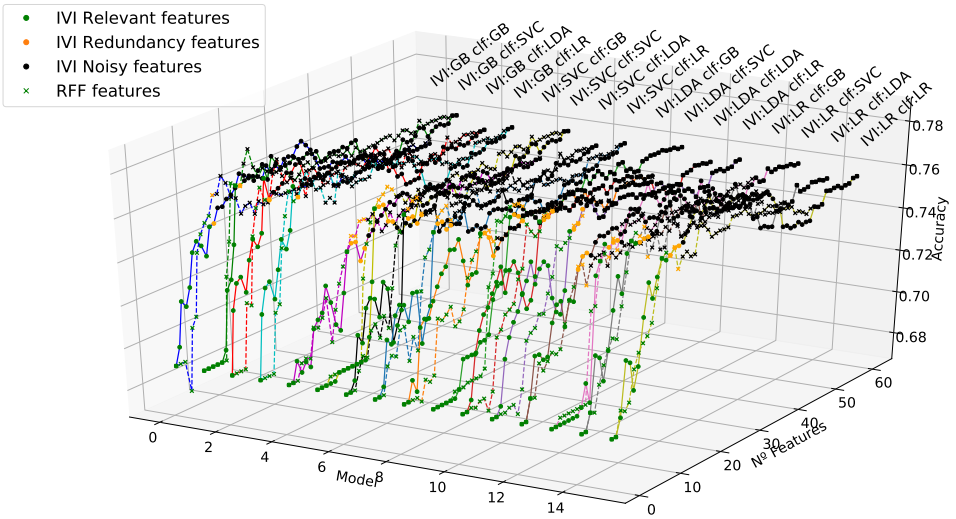
Figura 4.8: Resultados del algoritmo IVI para cada técnica ML en un conjunto de datos real de PaySim. En filas las diferentes técnicas utilizadas de ML: CME, SVC, GB, LR y LDA. En las columnas, las diferentes características. El color verde representa los escenarios en los que la característica se identificó como relevante. El color rojo representa las características que no se identificaron como relevantes durante el análisis.

Se muestran las figuras equivalentes a los analizados anteriormente para los datos sintéticos donde se representaban todos los modelos evaluados y conjunto de características disponibles, pero esta vez para la base de datos real se muestran en la Figura 4.9 y 4.10.

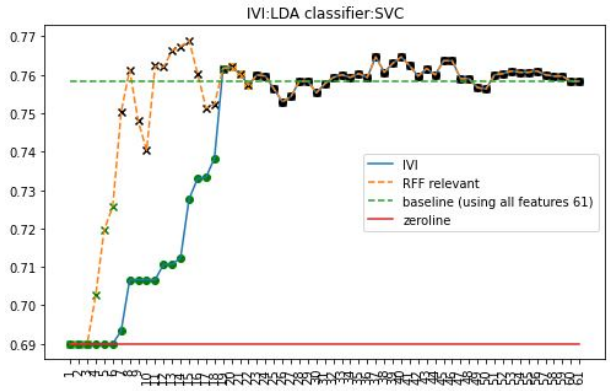
En particular, en la Figura 4.9 vemos el proceso seguido con el algoritmo IVI y las características filtradas RFF, mientras que en la Figura 4.10 vemos el análisis correspondiente para MIFF. Se puede observar en la figura que la precisión aumentó sistemáticamente en todos los casos cuando se añadieron características relevantes hacia el valor máximo, después, la precisión se mantuvo estable a pesar de que se añadieron características redundantes y ruidosas. Además, cuando se utilizaron los filtros MIFF y RFF, se redujo el número de características para alcanzar la precisión más alta frente al enfoque IVI

estándar. Los mejores resultados se obtuvieron una vez más con el filtro MIFF.

La Figura 4.9 muestra una precisión ligeramente inferior en comparación con IVI. Como hemos mencionado anteriormente, esto es debido a que el filtro RFF es muy estricto a la hora de incluir una característica relevante.

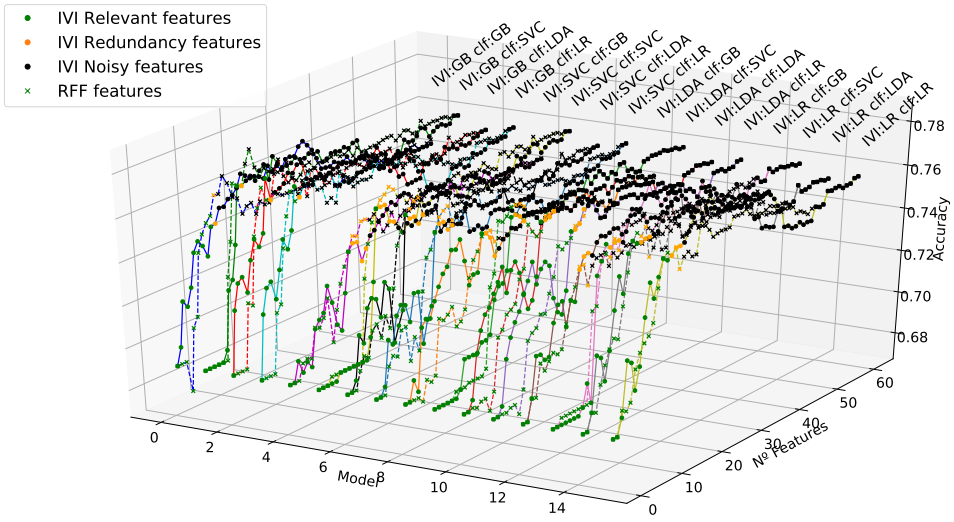


(**a**)

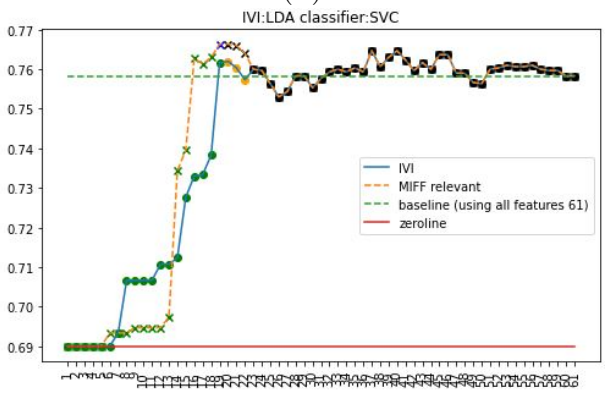


(**b**)

Figura 4.9: Resultados gráficos con respecto a la precisión a través de la incorporación de características en orden secuencial atendiendo a la relevancia para los filtrados RFFEn los gráficos 3D (**a**), podemos ver el número de características en el eje x, en el eje y los diferentes algoritmos de ML y la precisión en el eje z. Para cada uno de los algoritmos ML, se presenta la evolución para las características seleccionadas para IVI con línea continua y las características seleccionadas por el filtro como línea discontinua. En (**b**), muestra una comparativa entre los filtros e IVI. Para mayor claridad y simplicidad, sólo mostramos un caso. En (**b**), la línea roja representa un clasificador sencillo que hace predicciones con la clase más frecuente, y la línea verde discontinua es el resultado del clasificador entrenado con todas las características.



(**a**)



(**b**)

Figura 4.10: Los resultados gráficos de la métrica de precisión a través de la incorporación de características en orden secuencial atendiendo a la relevancia para los modelos MIFF, lo cual nos permiten ver la evolución en la precisión a medida que se añade cada característica. En los gráficos 3D. En los gráficos 3D (**a**), podemos ver el número de características en el eje x, en el eje y los diferentes algoritmos ML (GB, SVC, LDA y LR) y la precisión en el eje z. Para cada uno de los algoritmos ML, se presenta la evolución para las características seleccionadas en IVI como línea continua y las características seleccionadas por el filtro como línea discontinua. En (**b**), muestrar una comparativa entre los filtros y IVI. Para mayor claridad y simplicidad, sólo mostramos una muestra comparativa. En (**b**), la línea roja representa un clasificador simple que hace predicciones con la clase más frecuente, y la línea verde discontinua es el resultado del clasificador entrenado con todas las características.

En la Figura 4.10 en los paneles (a) y (b), podemos ver un efecto interesante. Usando RFF tuvimos un incremento explosivo en términos de precisión, empleando para ello un conjunto relativamente pequeño de características, pero no alcanzamos el máximo. Utilizando RFF se obtiene un incremento explosivo en términos de precisión, utilizando un conjunto relativamente pequeño de características, pero no se llega a alcanzar el máximo.

Sin embargo, al utilizar MIFF, se obtiene un incremento inicial más suave pero por el contrario se alcanza el máximo en precisión, superando la precisión de IVI mientras se utiliza un número menor de características. Esta situación se extiende a todos los modelos representados en la Figura 4.11 en los paneles (a) y (b) en una perspectiva diferente. En esta representación, y para comodidad del lector, la alta dimensionalidad de los datos se restringe sólo mostrando las características relevantes, en un intento de visualizar este efecto a través de los diferentes perspectivas.

De la misma manera que se analizó previamente en el conjunto de datos sintéticos, se evaluó la contribución al proceso de decisión utilizando los pesos de cada característica para cada experimento. Siguiendo con el paralelismo con las representaciones anteriores, se consolida en los diferentes modelos evaluados: (i) la menor contribución o peso de un número de características que fueron clasificadas como ruidosas o redundantes, (ii) la alta contribución de las características clasificadas como relevantes, y (iii) la existencia de un número de ellas con una contribución intermedia fijando las clasificadas redundantes. La figura 4.12 ilustra los coeficientes correspondientes a las características MIFF. La representación ilustra la existencia de características con una alta contribución próxima a $\pm$ *1.0* (*over_draft_<0*, *credit_usage*, *purpose_new_car*, *employment*, *other_parties_guarantor*), así como los de contribución próxima a $\pm$ 0.5 (*credit_history_other_existing_credits, average_credit_balance_<100, location purpose_education*), y una pequeña proporción de características con una contribución significativamente menor de 0.5 (*purpose_rent, current_balance, credit_history_no_credits, foreign_worker*). Este último grupo corresponde muy probablemente a las características redundantes incorrectamente clasificadas y por lo tanto sometidas a ser excluidas en una etapa posterior. Es necesario señalar en este punto la identificación de 3 características destacadas en la clasificación general con contribuciones un 50 % superiores a sus compañeras. En particular, la característica más relevante para todos los algoritmos fue *credit_usage*, y la segunda más relevante fue el *over_draft_<0*. Estos resultados son coherentes con la bibliografía [98, 99]. Este subconjunto, debe ser convenientemente analizado por separado desde una perspectiva de interpretabilidad, dado este interesante comportamiento, que no sólo es efectivamente superior al resto, sino que se reproduce consistentemente en cada método implementado. Cabe mencionar aquí la situación especial en el método GB, que aunque aparentemente ofrece una significación proporcional al resto de los métodos, los valores son inferiores al resto de los métodos siguiendo de nuevo el mismo comportamiento que en el escenario con datos sintético.

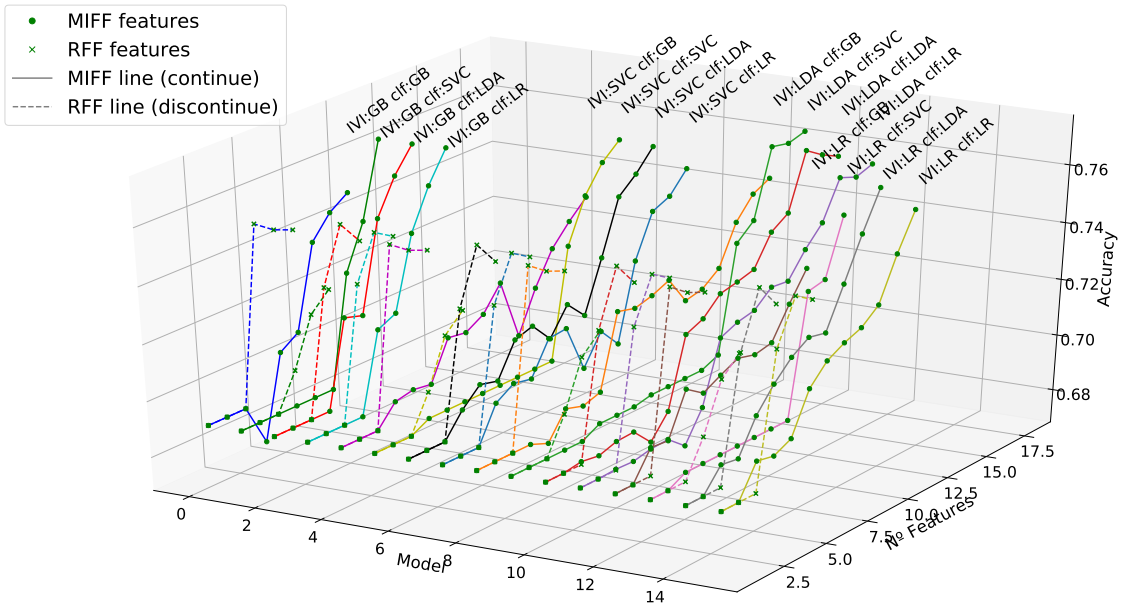


Figura 4.11: Resultados gráficos de la métrica de precisión mediante la incorporación de las características relevantes en orden secuencial para RFF y MIFF. Podemos ver el número de características en el eje x, en el eje y los diferentes algoritmos ML (GB, SVC, LDA y LR) y la precisión en el eje z. Para cada algoritmo ML, se presenta la evolución de las características seleccionadas en el filtro MIFF como línea continua y las características seleccionadas por el filtro RFF como línea discontinua.

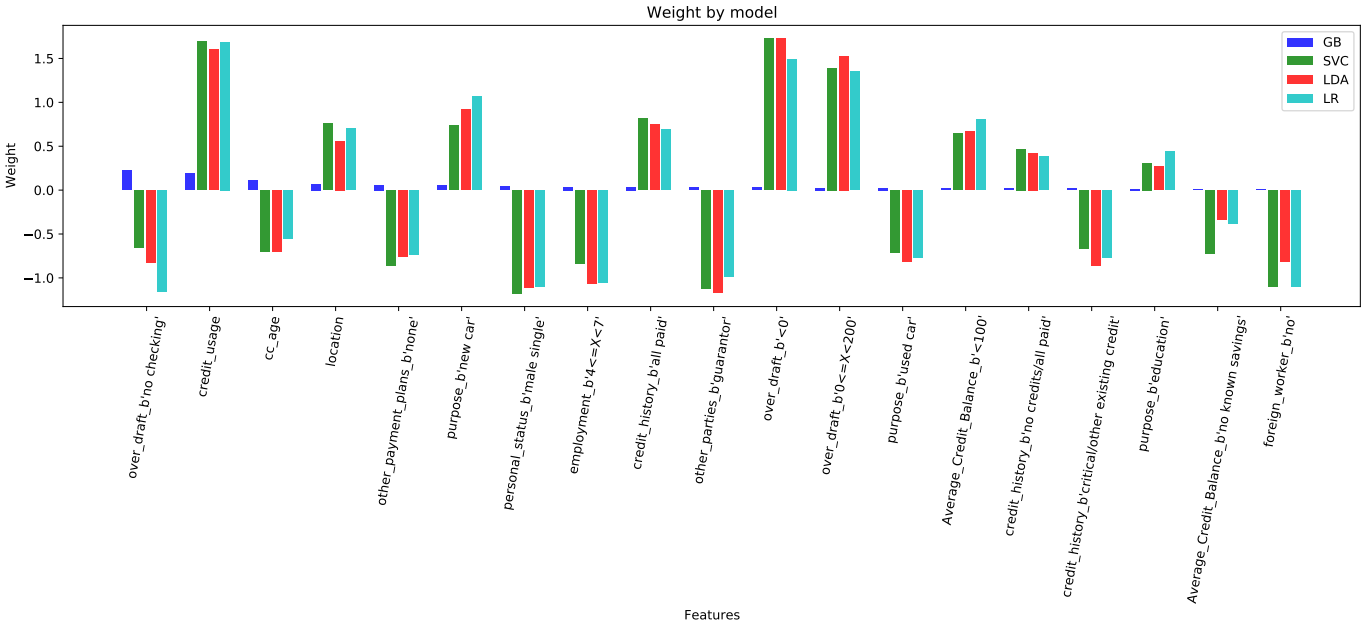


Figura 4.12: Pesos de las características. Pesos en representación gráfica de las características seleccionadas en el filtro MIFF. Cada barra de color representa diferentes técnicas de ML: GB, SVC, LDA, LR.

4.4. Conclusiones del capítulo

En este capítulo, hemos abordado la selección de características relevantes para problemas de CFD aplicando técnicas de ML tratando de dar interpretabilidad al proceso y dejar de lado los problemas de caja negra para este proceso. Para ello hemos definido una metodología novedosa mediante el uso de algoritmos de última generación, los cuales son capaces de cuantificar la información de las características y sus relaciones. Este enfoque ofrece un nuevo concepto de interpretabilidad en el ámbito de la selección de características para poder hacer uso de técnicas de ML en problemas de CFD, dado que la regulación actual determina la obligación expresa de cumplir con las normas de no discriminación y transparencia.

Para ello, se ha presentado un análisis intensivo de una serie de técnicas de ML (GB, SVR, LR y LDA), junto con nuevos procedimientos de selección de características, aplicados tanto en conjuntos de datos sintéticos, para definir y afinar los modelos, como en conjuntos de datos reales, para su validación. A modo de conclusión general de todos los experimentos realizados podemos decir que es posible desarrollar un modelo de ML apoyado en las novedosas técnicas de selección de características presentadas en este trabajo, donde el resultado proporciona, no sólo la detección de CFD, sino que al mismo tiempo permite visualizar la contribución de cada característica en el proceso de decisión al resultado final, ofreciendo así la necesaria interpretabilidad del modelo y los resultados.

Para hacer frente a la compleja dicotomía entre los beneficios de aplicar técnicas de ML y las limitaciones de la regulación, este capítulo ha mostrado como extraer las características informativas y obtener su contribución al proceso de decisión, sin hacer uso de las cajas negras y minimizando los posibles sesgos mediante el uso de técnicas de ML de última generación. Adicionalmente, se ha evidenciado que es posible construir modelos lineales interpretables robustos para cumplir simultáneamente con la restricción reguladora si se utiliza el potencial de las técnicas de ML.

En cuanto a las conclusiones más relevantes del análisis del conjunto de datos sintéticos, podemos decir que haciendo uso del modelo IVI, en primer lugar, utilizando el modelo *IVI*, fuimos capaces de identificar sistemáticamente las características relevantes para el problema. Además, el uso de un subconjunto de características al aplicar los filtros descritos en este trabajo, mejoró el rendimiento en términos de eficiencia computacional al limitar su número de características. Encontramos que todas las características ruidosas y redundantes fueron excluidas sistemáticamente de este método de IVI ampliado. Se propusieron dos procedimientos de filtrado diferentes, *RFF* y MIFF. El primero se verificó que es mucho más restrictivo pero a su favor muestra una forma más rápida de alcanzar un nivel razonable de precisión, pero por contra debido a estas fuertes restricciones falló en la detección de

algunas características relevantes. Por otro lado, el filtro MIFF, aunque incrementó las características incluyendo características redundantes adicionales, no clasificó erróneamente ninguna característica ruidosa como relevante. Como la interpretabilidad de los modelos lineales se propuso en base a los pesos finales obtenidos, y considerando que se encontró una contribución mucho menor para estas características redundantes, la aplicación conjunta de MIFF y la evaluación de pesos podrían considerarse como un modelo eficiente y preciso incluso en los casos en que las características redundantes se clasifican erróneamente. Basándonos en estos resultados, concluimos que la formulación y clasificación mediante IVI, junto con el filtrado *RFF* y MIFF, ofrece una herramienta automática y eficiente que mejora la capacidad de generalización y predicción de CFD en conjuntos de datos sintéticos. Los resultados de la aplicación de estas metodologías sobre el conjunto de datos reales [30, 100] fueron en términos generales consistentes con los hallazgos obtenidos en el conjunto de datos sintéticos. Los resultados obtenidos sugirieron que el uso del método presentado, no sólo mejora los resultados (con un aumento de la precisión del 4 % respecto a los trabajos publicados anteriormente), sino que también mejora la eficiencia computacional al reducir el número de características. Desde el punto de vista metodológico, se confirmó la validez del modelo aplicado, ya que la precisión de todos los modelos emparejados con características del método fue homogénea, con valores cercanos al 75 % en la media y una desviación estándar cercana al 2 %.

Desde el punto de vista computacional, y teniendo en cuenta los cuatro subconjuntos diferentes de características evaluados, como han sido el uso de todas las características, las características seleccionadas por IVI, y las obtenidas mediante los filtros MIFF y RFF, los dos últimos ofrecieron una reducción significativa del número de variables y, por tanto, una mejora significativa de la carga computacional. En ambos casos la precisión fue elevada, aunque el MIFF ofreció una mayor estabilidad en los resultados entre los distintos métodos. El análisis detallado mostró que el RFF consiguió incorporar eficazmente sólo las variables relevantes, pero faltando en algunos casos algunas de ellas, mientras que el MIFF consiguió en todos los casos incluirlas todas, pero por contra incluyo algunas redundantes. La integración de las variables redundantes en el MIFF no generó ninguna falta de precisión o estabilidad en la capacidad de predicción, e incluso pudieron ser eliminadas en una etapa posterior ya que sus contribuciones (pesos) mostraron de forma constante una magnitud mucho menor entre sus pares.

Por el contrario, en el caso de RFF, la rápida convergencia debido a lo restrictivo de su formulación provoca que algunas características significativas no sean tenidas en cuenta, confiriéndole hasta 4 puntos porcentuales de reducción en la precisión respecto a otros modelos analizados. Por lo tanto, podemos argumentar en este punto, que es una estrategia muy estricta en la búsqueda de características verdaderamente informativas,

como es la RFF, aunque acelera intensamente la convergencia y la eficiencia computacional, en ocasiones impide recoger todas las características informativas, provocando falta de convergencia o inestabilidad en los resultados. Al mismo tiempo, la estrategia más laxa de selección de características, como es MIFF, ofrece una mayor flexibilidad, que, aunque a veces permite la selección de las redundantes, maximizan las probabilidades de que se incorporen todas las características relevantes, sin aumentar excesivamente las necesidades computacionales. Por esta razón, el uso de técnicas de clasificación IVI, junto con la aplicación de conjuntos de selección MIFF, puede ofrecer el equilibrio adecuado entre los requisitos computacionales y la precisión. Desde el punto de vista de ML, la metodología utilizada es consistente debido a que todos los métodos mostraron resultados con pocas diferencias para cada uno de los conjuntos de características, si bien es cierto que SVR volvió a consolidarse como el método que proporcionó mejores resultados, mientras que GB tuvo un rendimiento ligeramente inferior entre el resto de métodos.

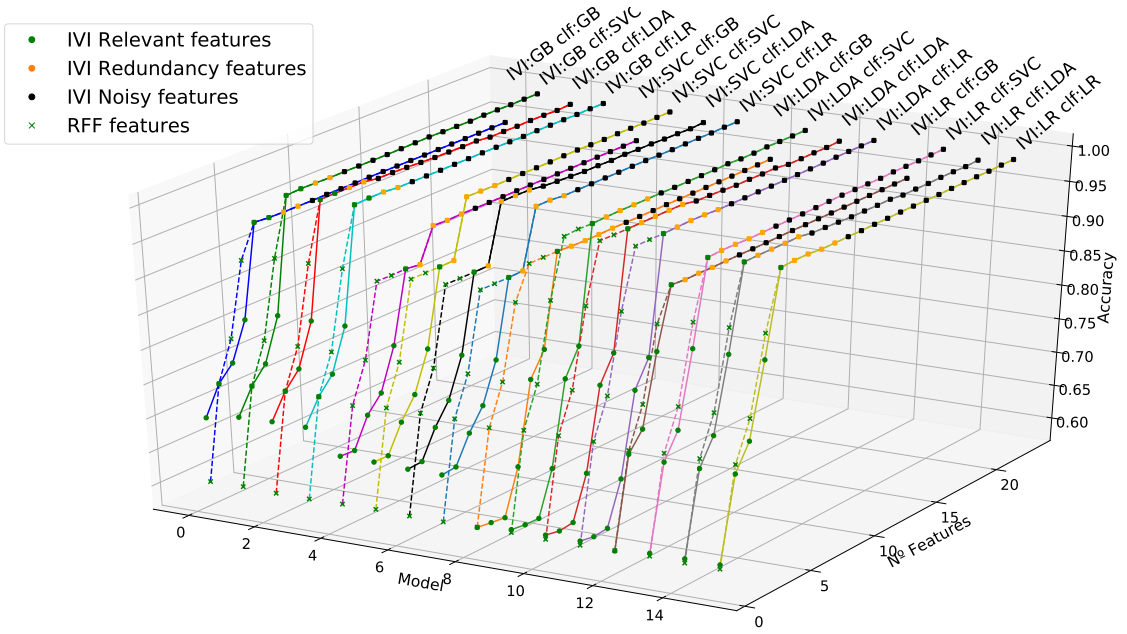
Por último, tal y como se ha introducido anteriormente, los modelos aquí presentados que utilizan estrategias exclusivamente lineales ofrecen una potente técnica interpretable de última generación que supera la predictibilidad de otras aplicaciones de ML más sofisticadas y difíciles de interpretar. Este enfoque allana el camino de la interpretabilidad, ya que la contribución de las diferentes características sensibles finales podría corresponder a los pesos de esas mismas características. Tras el análisis sintético, los coeficientes de las características tienden a ser muy altos en las características relevantes, y bajos o muy bajos en el caso de las redundantes o ruidosas respectivamente. Para el caso concreto del MIFF, se constató la existencia de características de muy alta contribución, así como de las de alta contribución y una pequeña proporción de características con una contribución significativamente menor, que resultaban ser las características redundantes.

Tres características destacaron con contribuciones un 50 % superiores a las de sus compañeros. Este subconjunto especial de características debe evaluarse especialmente como características clave y de apoyo del modelo, ya que no sólo mostraron una gran contribución, sino que reprodujeron sistemáticamente en todos los métodos aplicados. Cabe señalar ahora que estas características (crédito medio con ahorro desconocido, clientes con descubierto y finalidad del crédito para un automóvil nuevo) ya estaban identificadas en la bibliografía [98, 99], con lo que se produce una doble validación: una de los resultados presentados, y dos, la validación del propio modelo presentado.

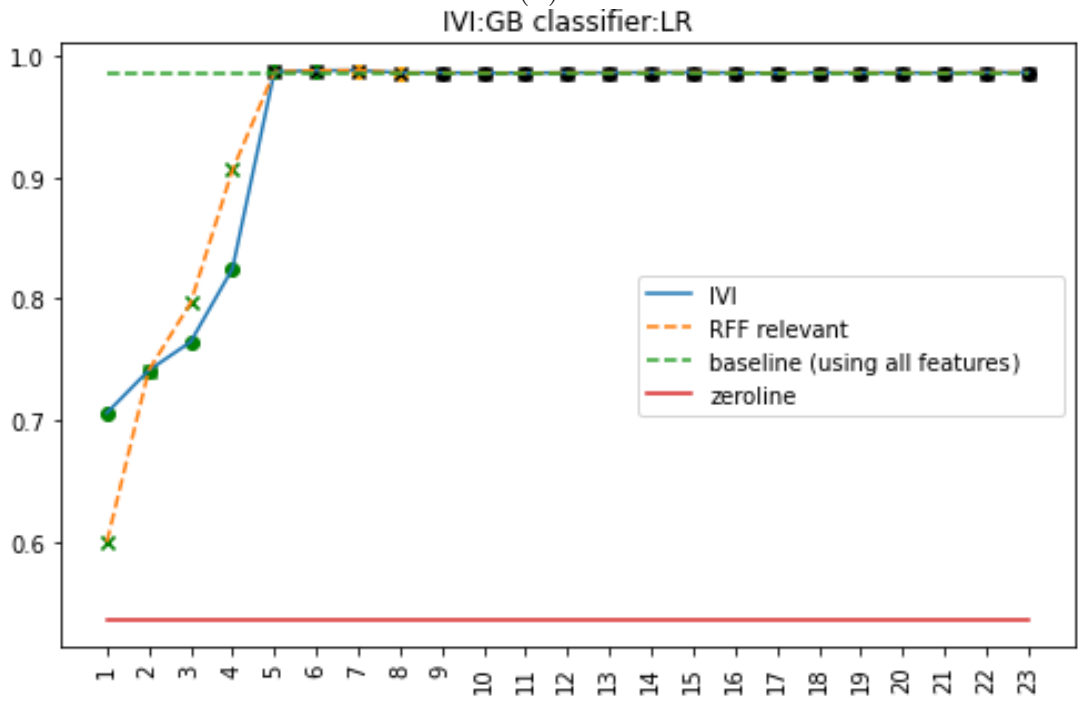
Como conclusión general podemos afirmar que es posible proponer un modelo de selección de características interpretable no sesgado en cinco pasos que incluye (i) un análisis inicial del IVI; (ii) un segundo paso que consiste en la evaluación comparativa de los clasificadores ML para reducir los posibles sesgos; (iii) un tercer de filtrado de las características; (iv) un análisis bootstrap para la estimación de la significación estadística;

(v) y, por último, un cálculo de la significación de las características, basado en los coeficientes, que abre paso o hacia la deseada interpretabilidad del modelo final mediante el uso de características relevantes que han sido seleccionadas de manera que se pueden justificar ante cualquier entidad reguladora.

Por ello, las técnicas innovadoras presentadas en este capítulo, han demostrado ser no sólo más eficaces que los trabajos anteriores, sino también ofrecer métodos computacionalmente más eficientes para el análisis. Con todo ello, reducimos los potenciales sesgos y abrimos el camino para tomar en consideración esta metodología para desarrollar modelos desde el punto de vista legal y en ámbitos éticamente ajustados. En los próximos capítulos abordaremos la obtención del aporte de cada característica en los procesos de decisión mediante técnicas más avanzadas como son el DL para mejorar la precisión en los problemas de CFD sin perder interpretabilidad una vez tenemos las características más relevantes.



(**a**)



(**b**)

Figura 4.3: Resultados gráficos con respecto a la precisión a través de la incorporación de características en orden secuencial atendiendo a la relevancia para los filtrados RFF, que nos permiten ver la evolución en la precisión a medida que se añade cada característica. En los gráficos 3D (**a**), podemos ver el número de características en el eje x, en el eje y los diferentes algoritmos de ML (GB, SVC, LDA y LR) y la precisión en el eje z. Para cada uno de los algoritmos ML, se presenta la evolución para las características seleccionadas en IVI como línea continua y las características seleccionadas por el filtro como línea discontinua. En (**b**), muestra una comparativa entre los filtros e IVI. En (**b**), la línea roja representa un clasificador simple, y la línea verde representa entrenado con todas las características.

Capítulo 5

Aislar las transacciones fraudulentas de las no fraudulentas

Como se comentó previamente, el objetivo T2 consiste en un método que comprima y codifique eficazmente los datos para aislar las transacciones fraudulentas de las no fraudulentas. Esta reducción del espacio real a un espacio latente se realiza a través del uso de algoritmos de última generación como son los autoencoders (AE) y uso de técnicas de ajuste de parámetros, a continuación abordaremos estos puntos.

5.1. Espacio latente

El espacio latente se refiere a un espacio multidimensional o latente que contiene propiedades que no podemos detectar directamente en el espacio original. En este espacio latente de menor dimensión se pueden extraer información muy relevante de las transacciones, por ejemplo, las transacciones que comparten patrones comunes generalmente aparecen representadas próximas unas de otras mientras que en el espacio real pasan inadvertidas.

Para la construcción de este espacio latente, en esta tesis se hace uso de *autoencoders* con una etapa de codificación y otra de decodificación, donde el codificador comprime el espacio real a un espacio latente de 3D, el motivo de escoger 3D es únicamente por razones de representación visual. Para ello, en la experimentación el *autoencoder* se construye con una primera capa cuyo número de celdas es el número de características seleccionadas en el método IVI con el filtro MIFF, dado como se vio en el capítulo anterior fue el método con el que se obtuvo mejor rendimiento de manera consistente. La segunda capa tiene tres celdas (para conseguir representaciones en 3D), y finalmente la tercera capa es la reconstrucción con el espacio latente al espacio de entrada de nuevo.

Como se puede apreciar, el *encoder* se construye con las capas primera y segunda, y el decoder se corresponde con la tercera capa. Las funciones de activación entre capas

utilizadas fueron de tipo ReLU (Rectified Linear Unit). Una vez definida la arquitectura del *autoencoder*, se ajusta con el conjunto de datos de entrenamiento y los resultados de la red neuronal se refinan mediante un proceso de retropropagación en toda la red, este proceso se conoce como "*fine tunning*". Mediante este proceso, se obtiene una mayor dispersión en el espacio latente, y se logra una mayor separación entre las transacciones fraudulentas de la no fraudulentas. Técnicamente se logra este efecto añadiendo una última capa softmax al codificador y ajustando de nuevo habiendo congelado las capas del codificador y permitiendo únicamente que el gradiente se retropropague a través de la capa softmax.

Debemos recordar en este punto que el proceso de compresión va desde el espacio real de cada conjunto de datos definido en la sección 3.1 (de 23, 61 y 14 variables) a un espacio latente de sólo tres dimensiones. Para una mejor comprensión, la figura 5.1 representa una visión general del proceso de fine tunning.

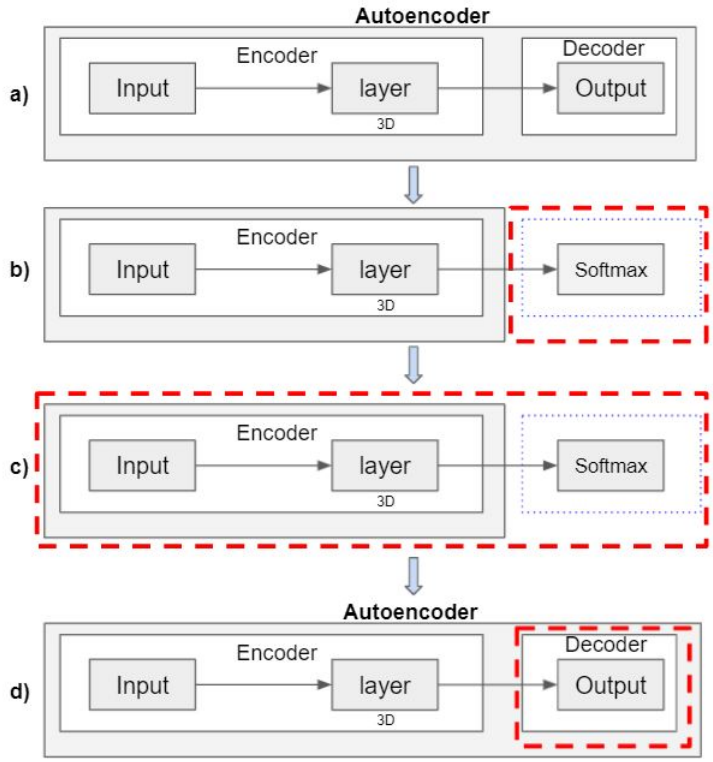


Figura 5.1: Proceso de fine tuning. (**a**) Representa la arquitectura del autoencoder. (**b**) Representa el codificador con una capa softmax adicional. En esta etapa, congelamos los pesos de las capas del encoder y el ajustamos sólo la capa softmax. (**c**) Volvemos a entrenar el codificador y la capa softmax, descongelando las capas del encoder. (**d**) Eliminamos la capa softmax y entrenamos el encoder y decoder congelando las capas codificadoras.

5.2. Experimentación

En la experimentación propuesta para evaluar la capacidad de clasificación con el objetivo de separar las transacciones fraudulentas de las no fraudulentas en un espacio latente reducido. En este caso se ha propuesto una dimensionalidad de 3D para permitir su visualización y facilitar su compresión. Para ello, primero procedemos a realizar la proyección sobre el espacio latente, utilizando un *autoencoder* con las características seleccionadas obtenidas en la experimentación en el capítulo 4. De este modo, y para obtener una mayor dispersión en el espacio latente, aplicamos la técnica de *fine tunning* añadiendo una capa softmax tal y como se comentó en la Sección 5.1.

5.2.1. Representación y clasificación de espacios latentes

A continuación, se emplea un clasificador (SVC) en el espacio latente resultante, lo que nos permite evaluar la capacidad de predicción en este nuevo espacio para los diferentes escenarios en estudio. En otras palabras, el experimento nos permite evaluar cómo la transformación del espacio de entrada al espacio latente contribuye a la posible mejora en términos de precisión. En un intento de verificar y cuantificar los resultados, se calculó la precisión en tres escenarios diferentes para cada conjunto de datos. Los escenarios consideraron diferentes conjuntos de características: (i) conjunto completo de características en el espacio real; (ii) IVI aplicando el filtro MIFF en el espacio real y (iii) IVI con características aplicando el filtro MIFF en el espacio latente. Se implementó el SVC como clasificador para la evaluación comparativa y el análisis. La tabla 5.1 resume la media y la desviación estándar de 100 ejecuciones con remuestreo para los diferentes escenarios. Los resultados mostraron que el espacio latente proporcionó sistemáticamente los mejores resultados para todos los conjuntos de datos. La desviación estándar relativamente pequeña de los resultados obtenidos tras el remuestreo nos valida la estabilidad estadística de los resultados obtenidos.

En la tabla 5.1, las columnas de izquierda a derecha corresponden a la inclusión de todas las características disponibles *Acc_ all_ features*, de IVI con filtro MIFF *Acc_fs_MIFF*, y de IVI con filtro MIFF en el espacio latente (usando *autoencoder*) *Acc_fs_MIFF_LS*. Como podemos ver en la tabla de resultados en esta experimentación, las columnas *Acc_ all_ features* (SVC) y *Acc_fs_MIFF* representan los valores obtenidos en el capítulo 4.2. En este sentido, podemos considerar el *Acc_ all_ features* (SVC) como la línea de base y el *Acc_fs_MIFF* como el gold standard. En la columna *Acc_fs_MIFF_LS*, que representa el espacio latente, podemos observar que obtenemos los mejores resultados. Con estos resultados se confirma que en el uso del espacio latente para los algoritmos ML mejoran la tarea de clasificación al mapear mejor los diferentes tipos de transacciones. Además,

Cuadro 5.1: Resultados estadísticos de la precisión para diferentes conjuntos de datos. Se muestran la media y la desviación estándar de los resultados para el análisis de 100 remuestreos. En las filas están los resultados de los diferentes conjuntos de datos mientras que en las columnas nos encontramos con los análisis para los distintos conjuntos de características incluidos en el proceso. Las columnas de izquierda a derecha corresponden a la inclusión de todas las todas las características disponibles, IVI con filtro MIFF, e IVI con filtro MIFF en el espacio latente (utilizando el autoencoder).

Dataset	**Acc_All_Features (SVC)**	**Acc_fs_MIFF**	**Acc_fs_MIFF_LS**
Synthetic	0,9870 ± 0,003195	0,9872 ± 0,002951	0,9885 ± 0,00039
PaySim	0,9654 ± 0,00011	0,9678 ± 0,0002	0,9758 ± 0,00011
German	0,7580 ± 0,017017	0,7663 ± 0,020409	0,7778 ± 0,00238

en esta columna también observamos una disminución de la desviación estándar de hasta casi 10 veces en los conjuntos de datos sintéticos y PaySim y de 2 veces en el conjunto de datos alemán. Esto indica que el uso del espacio latente no solo mejora la precisión, sino que también aumenta la estabilidad de los resultados.

5.2.2. Análisis de sensibilidad en el espacio latente

A la vista de los resultados presentados en el apartado anterior, se ha considerado de interés estudiar la variabilidad de los resultados de cada característica del espacio de entrada. A tal efecto, este experimento utiliza la puntuación del clasificador SVC definido en el espacio latente para estimar la sensibilidad del resultado a pequeñas variaciones de cada característica en el espacio original. Para ello, realizamos pequeñas variaciones para cada característica en cada transacción de manera individual, aumentando y disminuyendo un pequeño porcentaje al valor de cada característica. La sensibilidad se ha considerado como la relación entre la puntuación obtenida al aplicar un pequeño cambio porcentual en el espacio de entrada y la puntuación sin el cambio porcentual en los valores del espacio de entrada. En la tabla 5.2, podemos ver la sensibilidad media de cada característica en el conjunto de datos PaySim. Este resultado muestra que hay una gran diferencia entre las pequeñas variaciones en cada característica, por ejemplo, la característica *newbalanceOrig* se ve más afectada por las pequeñas variaciones que *step*.

Desde un punto de vista gráfico, estos resultados pueden observarse claramente para mostrar que pequeñas variaciones en algunas características en el espacio de entrada pueden tener un gran impacto en el espacio latente. Podemos ver este efecto en la Figura 5.2, y podemos observar que la misma pequeña variación en una característica en el espacio de entrada puede tener diferente respuesta en el espacio latente; por ejemplo, para la característica *newbalanceOrig*, esta respuesta es más visible que en las características

Cuadro 5.2: Media de la sensibilidad de cada característica en el conjunto de datos PaySim.

Característica	**sensibilidad**	**Característica**	**sensibilidad**
newbalanceOrig	1.118154	bstep	0.000600
amount	0.843496	isFlaggedFraud	−0.000518
newbalanceDest	0.215071	nameOrig_code	−0.001186
type_transfer	0.199242	type_CASH_IN	−0.366544
type_payment	0.104022	type_DEBIT	−0.753180
type_cash_out	0.012399	oldbalanceOrg	−1.086948

type_transfer y *type_payment*, donde esta respuesta no es apreciable.

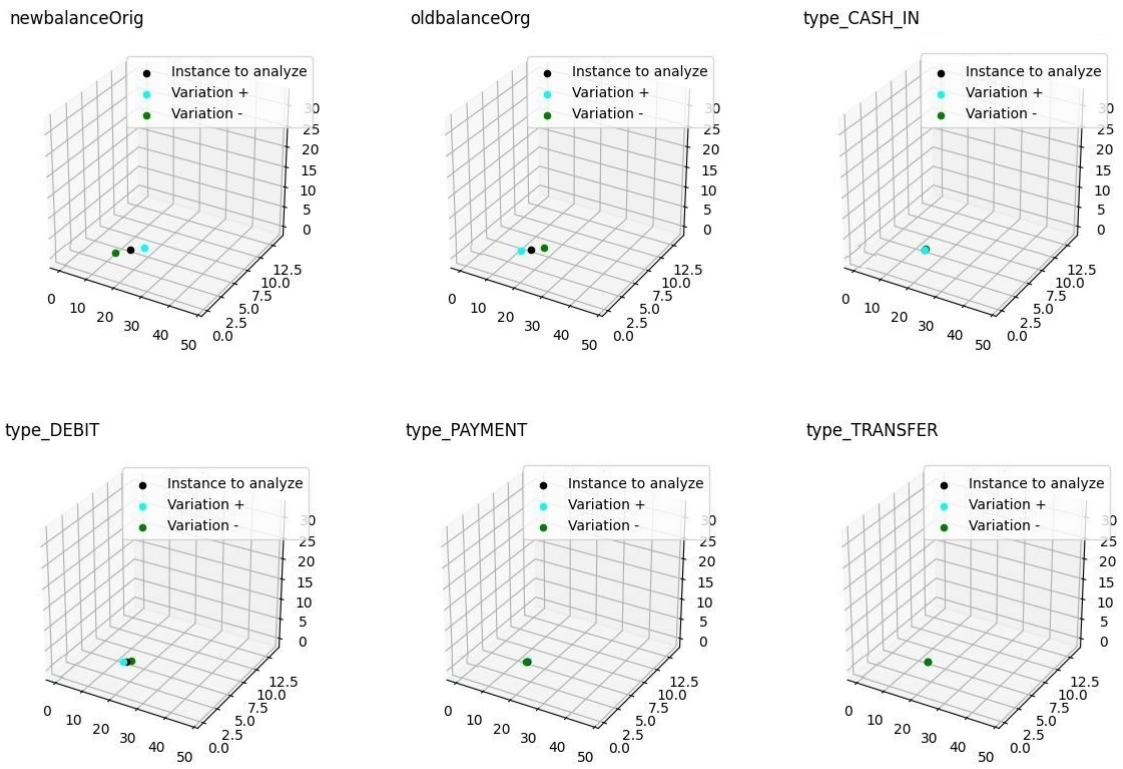


Figura 5.2: Representación para una transacción no fraudulenta en la que mediante pequeñas variaciones de cada característica en el espacio real afectan en el espacio latente con el filtro MIFF en el conjunto de datos PaySim.

En la Figura 5.3, podemos ver las distribuciones de puntuación obtenidas en los espacios latentes cuando aplicamos estas pequeñas variaciones. Por razones de representabilidad sólo hemos representado dos características, *newbalanceOrig* con alta sensibilidad y *step* con baja sensibilidad, según los datos de la Tabla 5.2. En la característica *newbalanceOrig*

podemos ver como las distribuciones se desplazan debido a la sensibilidad, mientras que en el rasgo *step*, al tener baja sensibilidad, se mantiene estático. Además, podemos observar en estas figuras que no tienen distribuciones de tipo normal, sino que son distribuciones multimodales. Este tipo de distribución refuerza nuestra hipótesis, la cual cada transacción puede verse afectada de forma diferente en el espacio latente por las combinaciones de los valores de las características en el espacio real, produciendo así diferentes pesos en cada característica utilizada en el proceso de decisión.

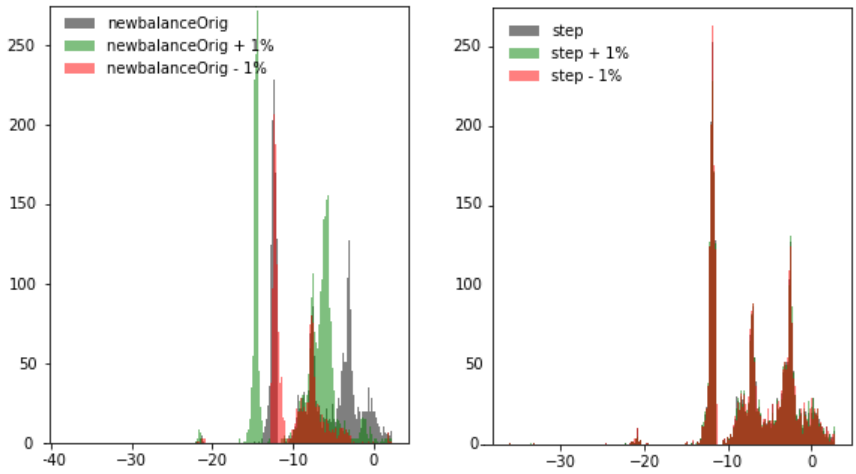


Figura 5.3: Distribuciones del valor del *score* obtenidas en los espacios latentes cuando se aplican pequeñas variaciones para el conjunto de datos PaySim.

5.3. Conclusiones del capítulo

En este capítulo, hemos abordado el problema de cómo aislar las transacciones fraudulentas de las no fraudulentas para ello se ha definido una técnica basada en el uso del espacio latente. De los experimentos se desprende que en el espacio latente los algoritmos de ML mejoran considerablemente la tarea de clasificación al asignar mejor los distintos tipos de transacciones dado que no sólo hemos confirmado una mejora en términos de precisión, sino que hemos verificado la estabilidad de los resultados superando los obtenidos en las experimentaciones previas. Adicionalmente, hemos comprobado empíricamente que pequeñas variaciones sobre las distintas características en el espacio real, tienen distinto efecto en el proceso de decisión en el espacio latente, validando la hipótesis de que en función de la transacción hay características más relevantes que otras. Si bien es cierto, que el uso de espacios latentes es una modalidad de caja negra, en los próximos capítulos abordaremos como dotar de interpretabilidad a este mecanismo.

Capítulo 6

Interpretabilidad

Una vez confirmado la mejora en términos de precisión de los modelos mediante el uso del espacio latente, este capítulo describe cómo lograr la interpretabilidad sobre el proceso de clasificación para diferenciar las transacciones fraudulentas de las que no lo son. Para ello, el capítulo se abordaran diferentes puntos con el objetivo de dotar de interpretabilidad a los modelos de decisión desde varios prismas. El capítulo se ha estructurado en primer lugar con la formulación del problema. En segundo lugar, se introduce la técnica STE la cual permite interpretabilidad basada a nivel de transacción individual. En tercer lugar, presentamos el algoritmo ITR que nos permite ordenar las características por su importancia y por lo tanto generar rankings de relevancia por características para cada transacción. Por último, mediante la correlación de Kendall junto con los rankings generados por cada transacción, construimos perfiles globales que nos permiten agrupar transacciones con propiedades y comportamientos similares.

6.1. Formulación del problema

En las siguientes líneas se describe como alcanzar la interpretabilidad en sistemas de detección del fraude crediticio (CFDIS).

Notación. Dado $CFDIS=\langle \mathbf{T}, \mathbf{F}, \mathbf{C} \rangle$ conforma un sistema de interpretable para los CFD, donde:

- $\mathbf{T}=\{\mathbf{t}_1, \ldots, \mathbf{t}_n\}$ es el conjunto de transacciones participantes en el sistema;
- $\mathbf{F}=\{\mathbf{f}_1, \ldots, \mathbf{f}_m\}$ es el conjunto de características participantes en el sistema;
- $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_g\}$ es el conjunto de clases para la clasificación (fraude y no fraude).

Con esto definimos formalmente una transacción $\mathbf{T}_{ij} \in \mathbf{T}$ como $\mathbf{T}_{ij} = \langle \mathbf{t}_i, \mathbf{c}_j, \mathbf{W}_{ij}, \mathbf{F} \rangle$ donde la transacción $\mathbf{t}_i$ se clasifica como $\mathbf{c}_j$ y su interpretabilidad se basa en los pesos numéricos $\mathbf{W}_{ij}$ para todas las características en $\mathbf{F}$. Normalmente, $\mathbf{c}_j$ se define como las clases de fraude o no fraude adicionalmente se define una función de ranking $\succ$ como la ordenación de un subconjunto de características según su contribución al proceso de decisión. Teniendo en cuenta esta definición, se plantea la siguiente asunción:

Supuesto 1 *Cada transacción $\mathbf{t}_i \in \mathbf{T}$ está compuesta por un conjunto de características, de las cuales se ordenan para su correcta clasificación en el proceso de decisión en función de los pesos, dicha ordenación se denotada por $\boldsymbol{\rho}_i = (\mathbf{F}^i, \succ_{O_i})$, con $\mathbf{F}^i \subseteq \mathbf{F}$ representando sus características. Esto significa que la transacción t_i tiene una ordenación parcial de un subconjunto de características $\mathbf{F}^i \subseteq \mathbf{F}$ según una determinada función de ordenación $\succ_{O_i}$, tal que $\succ_{O_i}: \mathbf{F}^i \rightarrow \mathbf{R}$ permite a la transacción t_i asignar un valor a una determinada característica en $\mathbf{F}^i$, que representa su peso en el proceso de decisión, con respecto a esta transacción en particular.*

Ejemplo 1 Por ejemplo, definimos que $\mathbf{F}^1 = \{$*Vivir por encima de las posibilidades*, *Ausencia de histórico de transacciones*, *Préstamos para automóvil*$\}$ sea el conjunto de características, que están formadas por las características informativas de la transacción $\mathbf{t}_1$ en el proceso de decisión.

- $\succ_{O_i}$ (*Vivir por encima de las posibilidades*) = 4,9.
- $\succ_{O_i}$ (*Ausencia de histórico de transacciones*) = 4,7.
- $\succ_{O_i}$ (*Préstamos para automóvil*) = 5,2.

Entonces, las características más representativas para $\mathbf{t}_1$ en el proceso de decisión son $\boldsymbol{\rho}_1 =$ (*Préstamos para automóvil* $\succ$ *Vivir por encima de las posibilidades* $\succ$ *Ausencia de histórico de transacciones*).

Por lo tanto, si en una transacción el peso de la característica 1 es mayor que el peso de la característica 2, podemos deducir que el proceso de decisión está siendo más influenciado por la característica 1.

Definición 1 *Utilizando la notación anterior, el problema que abordamos en este trabajo se define como sigue:*

1. *Un sistema de interpretabilidad para CFD está representado por $CFDIS = \langle \mathbf{T}, \mathbf{F}, \mathbf{C} \rangle$;*
2. *Un conjunto $\{\boldsymbol{\rho}_i\}$ de pesos internos ordenados W_{ij} para cada característica $\mathbf{F}$;*
3. *Un conjunto $\left\{\mathbf{T}^i \subseteq \mathbf{T}\right\}$ de transacciones donde un subconjunto $\mathbf{F}^i \subseteq \mathbf{F}$ de características, pertenecientes a $\boldsymbol{\rho}_i$, son las características más representativas para el proceso de decisión.*

En este punto, el problema es encontrar cómo construir un conjunto que describa la contribución de los rasgos a partir del conjunto de rasgos **F** dados por tales **T**.

Propuesta 1 Como solución al problema anterior, proponemos incorporar en el CFDIS un mecanismo que sea capaz de evaluar los pesos para cada característica y para cada transacción en el espacio latente. Para ello, construimos un VAE, tal y como se expresa en las ecuaciones (3.10) y (3.11), para obtener la representación del espacio latente de la transacción que queremos interpretar, y generamos un conjunto de datos personalizado con muestras aleatorias utilizando perturbaciones alrededor de la transacción. Para conseguirlo, se propone la técnica STE con este conjunto de datos personalizado con muestras artificiales alrededor de la transacción a evaluar. Estas perturbaciones en el espacio latente se ponderan según su proximidad a la instancia de interés utilizando una función de decisión, con ello obtenemos un ranking (ITR) de las características más influyentes en el proceso de decisión. Una vez que hemos construido el ITR para las características más significativas, repetimos este proceso para todas las transacciones construyendo una ranking global.

Una vez definida la propuesta de la presente tesis, donde más nos focalizaremos es en la construcción de las clasificaciones o *ranking* individuales, lo cual consideramos que tiene un enorme potencial, ya que nos permite descubrir las características más significativas del proceso de decisión, mientras que el análisis más detallado de la clasificación global se planteara y formalizará como línea de trabajo futura.

6.2. STE. Obtención de la relevancia de las características

A modo de síntesis introductoria e ilustrativa, podemos decir que el método STE implementa y valida un modelo lineal sustitutivo. Estos modelos sustitutivos, aproximan el comportamiento de los modelos complejos de tipo caja negra a un modelo lineal equivalente especializado para una transacción concreta. De este modelo lineal se obtendrían los pesos específicos para la transacción analizada e indicarían las contribuciones de cada característica del modelo complejo de tipo caja negra objeto. Teniendo en cuenta este razonamiento, se puede describir dicho proceso comenzando por utilizar las características seleccionadas y filtradas que hemos descrito anteriormente mediante el algoritmo IVI y el filtro MIFF. Seguidamente, implementamos el *autoencoder* del cual empleamos el *encoder* aplicando una técnica de *fine tunning* que encapsula mejor la información relevante del espacio de entrada en un espacio latente 3D. Una vez construido este *encoder*, mediante un *autoencoder* variacional nos permitirá generar transacciones en el espacio real ficticias pero manteniendo su coherencia generando transacciones realistas compatibles con las transacciones existentes. A continuación, utilizando estas nuevas transacciones ficticias

que están lo suficientemente próximas en el espacio latente, se implementa un modelo de regresión lineal donde la variable dependiente es la distancia entre cada transacción ficticia y la transacción bajo estudio. Esta distancia se calculara cómo la diferencia entre la puntuación o score de un modelo de clasificación lineal entrenado específicamente en el espacio latente. Este modelo lineal será considerado como un modelo sustitutivo especializado para cada transacción y que mejor se ajusta al modelo complejo de caja negra, un *autoencoder* en nuestro caso. Estos modelos sustitutivos se generaran a medida para cada transacción que se quiera analizar. Por lo tanto, los pesos de dicho modelo lineal sustitutivo pueden utilizarse cuantitativamente como la contribución local para cada transacción del mencionado modelo de caja negra. Este proceso detallado, se describe en el Algoritmo 7, donde se ajusta un VAE para obtener transacciones aleatorias viables en el espacio latente alrededor de la transacción que queremos interpretar. Con estas muestras, calcularemos la diferencia de puntuación en el espacio latente con respecto a la transacción que queremos interpretar utilizando un clasificador. Estas diferencias de puntuación en el espacio latente se utilizan para obtener los pesos de cada una de las características de entrada en el proceso de decisión mediante un modelo lineal.

Algoritmo 7 Interpretabilidad. Algoritmo STE

Require: El conjunto de entrenamiento en el espacio real es $\mathbf{X}$, $\mathbf{I}_n$ es la transacción a interpretar, el codificador *enc*, el número de remuestreos d, y el número de remuestreos bootstraps s.

1: Dividir el conjunto $\mathbf{X}$ en dos subconjuntos, $\mathbf{X}_{train}$ con $\mathbf{Y}_{train}$ y $\mathbf{X}_{test_i}$ con $\mathbf{Y}_{test_i}$, y número de remuestreos s.
2: Inicializar el $VAE = \{\}$.
3: Ajustar VAE. $VAE \leftarrow VAE.fit(\mathbf{X}_{train})$.
4: Generar datos sintéticos realistas. $X' \leftarrow VAE.predict(\mathbf{X}_{test})$.
5: Ejecutar el codificador para obtener la posición en el espacio latente para la transacción $\mathbf{I}_n$.
$In_{LS} = enc.predict(\mathbf{I}_n)$.
6: Ejecutar el *encoder* para obtener la posición en el espacio latente para $\mathbf{X}_{train}$.
$X_{train_{LS}} = enc.predict(\mathbf{X}_{train})$.
7: Ajustar un modelo de clasificación en el espacio latente CM.
$CM \leftarrow CM.fit(\mathbf{X}_{train_{LS}}, \mathbf{Y}_{train})$.
8: Ejecutar el codificador para obtener la posición en el espacio latente de los datos sintéticos realistas X'.
$Sin_{LS_i} = enc.predict(\mathbf{X}'_i)$ with $i = 1, ..., d$.
9: Calcular la variación de la puntuación en el espacio latente.
$Y_{score_i} = (CM.score(Sin_{LS_i}) - CM.score(In_{LS}))$ with $i = 1, ..., d$.
10: **for** $b \leftarrow 1$ to s **do**
11: Generar un subconjunto aleatorio de datos sintéticos realistas en el espacio real con tamaño N_b, y su distancia en puntuación a la Instancia a interpretar en el espacio latente.
$\mathbf{X}_B = \mathbf{X}'_k$, with $k = 1, ..., N_b$.
$\mathbf{Y}_B = \mathbf{Y}_{score_k}$, with $k = 1, ..., N_b$.
12: Ajustar el modelo lineal.
$Mod \leftarrow LinearModel.fit(\mathbf{X}_B, \mathbf{Y}_B)$.
13: Obtener el vector de pesos $W^*_{(b)}$ empleando $\mathbf{X}_B$ y Y_B.
14: Guardar el vector peso $X^*_{(b)}$ en la columna numero b^{o} de la matriz W^*.
15: **end for**

6.3. Construcción del ITR

Como indicamos anteriormente, la presente tesis proporciona un mecanismo novedoso para entender cómo funciona el proceso de decisión a nivel de transacción individual en un $CFDIS$. En consecuencia, se ha planteado un método agnóstico del mecanismo que se emplee para obtener el espacio latente, de este modo, si aparece un mecanismo más potente en el estado del arte, sería compatible con la presente propuesta. Es decir, si en lugar de utilizar un *autoencoder* utilizamos otro algoritmo, a través de STE y generando transacciones ficticias pero realistas a las que aplicando pequeñas perturbaciones alrededor de la transacción, también podemos determinar la contribución para cada característica a

través de su sensibilidad. Una vez decidido el mecanismo, podemos utilizar esta caja negra para obtener los pesos de cada característica.

Teniendo en cuenta los pesos obtenidos mediante STE para cada transacción, a través de ITR, construimos un ranking individual de características para cada transacción. Este ranking captura el orden de las características, permitiéndonos saber qué características son las más influyentes en el proceso de decisión. Con ello, podremos observar, que para distintas transacciones pueden compartir mismo ranking de características, ya que en este punto, no nos interesa tanto el valor concreto del peso asignado a las característica sino el orden de influencia. Teniendo en cuenta que para cada transacción se ha generado un modelo lineal sustitutivo a medida. Formalmente, el método ITR puede expresarse como sigue.

Definición 2 *Un ITR_i para la transacción t_i que participa en el $CFDIS$ es una estimación Δ_i^t de sus características más representativas ρ_i, tal que:*

$$ITR_i = \Delta_i^t = (F^i, \succ_{O_i'}) \tag{6.1}$$

where:

- *$F^i \subset F$ es un subconjunto de características utilizadas en el proceso de decisión en el t_i;*
- *$\succ_{O_i'}$ es una función de ordenación, tal que $O_i' : F^i \times t_{ij} \times Enc \to \mathbb{R}$ asigna un valor a una determinada característica en F^i teniendo en cuenta el resultado de aplicar un clasificador lineal en el espacio latente utilizando autoencoder a una transacción t_{ij}.*

Ejemplo 2 Ilustremos esta definición con el siguiente ejemplo. Por ejemplo, dados $\mathbf{F}^1 = \{$*Vivir por encima de las posibilidades*, *Ausencia de histórico de transacciones*, *Préstamos para automóvil*$\}$ sea el conjunto de características, que se compone de transacciones $\mathbf{t}_1$, $\mathbf{t}_2$, y $\mathbf{t}_3$ con diferentes pesos obtenidos mediante STE:

- $\mathbf{t}_1$: (*Préstamos para automóvil*) = 5,2 $\succ$ (*Vivir por encima de las posibilidades*) = 4,9 $\succ$ (*Ausencia de histórico de transacciones*) = 4,7
- $\mathbf{t}_2$: (*Préstamos para automóvil*) = 3,9 $\succ$ (*Vivir por encima de las posibilidades*) = 3,7 $\succ$ (*Ausencia de histórico de transacciones*) = 3,2
- $\mathbf{t}_3$: (*Ausencia de histórico de transacciones*) = 4,2 $\succ$ (*Vivir por encima de las posibilidades*) = 3,7 $\succ$ (*Préstamos para automóvil*) = 3,1

Entonces, para las transacciones $\mathbf{t}_1$ y $\mathbf{t}_2$ podemos ver que tienen el mismo ITR (*Préstamos para automóvil*) $\succ$ (*Vivir por encima de las posibilidades*) $\succ$ (*Ausencia de histórico de transacciones*) y $\mathbf{t}_3$ tienen propiedades diferentes con otros ITR (*Ausencia de histórico de transacciones*) $\succ$ (*Vivir por encima de las posibilidades*) $\succ$ (*Préstamos para automóvil*).

Podemos ver el proceso para calcular el ITR resumido como se muestra en el Algoritmo 8.

Algoritmo 8 Interpretabilidad. Obtención de ranking de transacciones a nivel individual

Require: Conjunto de entrenamiento en el espacio real $\mathbf{X}$, número de características L, número de transacciones k

1: Calcular los pesos de todas las instancias
$W_i \leftarrow STE(\mathbf{X}_i)$ with $i = 1, ..., k$.
2: Generar una clasificación individual para cada transacción.
3: **for** $b \leftarrow 1$ to k **do**
4: En función de los pesos de cada característica, obtenemos su posición numérica en el ranking de relevancia, donde el mayor peso es el primero en el ranking y el menor es el último.
$ITR_b = generarRanking(W_b)$.
5: **end for**

6.4. Creación de perfiles globales

Una vez que hemos desarrollado los diferentes rankings con la contribución de las características a nivel de transacción individual, o ITR, podemos hipotetizar que las muestras o transacciones que comparten el mismo ITR también podrían estar compartiendo otras propiedades dado que probablemente estén próximas en el espacio latente. Este razonamiento es coherente con el hecho de que desarrollamos los pesos/contribuciones de las características que guiaron el desarrollo del ITR en base a la proximidad de las muestras en el espacio latente. Esto nos permite considerar que este enfoque no se aleja de la línea argumental, sino que, por el contrario, cierra el bucle, consolidando el modelo propuesto, y puede ser visto como una herramienta para validar las líneas anteriores.

Aunque esto no es necesariamente cierto en ambos sentidos, dado que estar cerca en el espacio latente significaría que muy probablemente podrían estar compartiendo ITR, pero no todas las muestras con el mismo ITR, necesariamente estarán en la misma área en el espacio latente. Recordemos que mediante *fine tunning* logramos separar las transacciones fraudulentas de las legítimas, por lo que un caso posible, sería que una transacción fraudulenta con un ITR equivalente a una transacción legitima con el mismo ITR estarían separadas en el espacio latente, por lo que diferentes áreas podrían compartir mismo ITR. Dicho esto, se propone un análisis basado en la correlación de Kendall para evaluar la similitud entre ITR de diferentes transacciones abriendo la puerta a la agrupación de las transacciones (basadas en muestras con el mismo ITR). El procedimiento para abordar este análisis es el Algoritmo 9, en el que se calcula la correlación de Kendall para todas las transacciones y, para evaluar la similitud, agrupamos con los valores únicos.

Algoritmo 9 Interpretabilidad. Obtención de perfiles

Require: Pesos de las características por cada transacción W, ranking de transacciones individuales ITR, número de transacción k

1: **for** $b \leftarrow 1$ to k **do**
2: $corr_b = kendall(ITR_b, ITR)$.
3: **end for**
4: valores únicos de las correlaciones
5: **for** $c \leftarrow$ valoresUnicos(corr) **do**
6: $X_c = instanciasConMismoRankingDeCaracteristicas(W, ITR_b, c)$.
7: **end for**

6.5. Experimentación

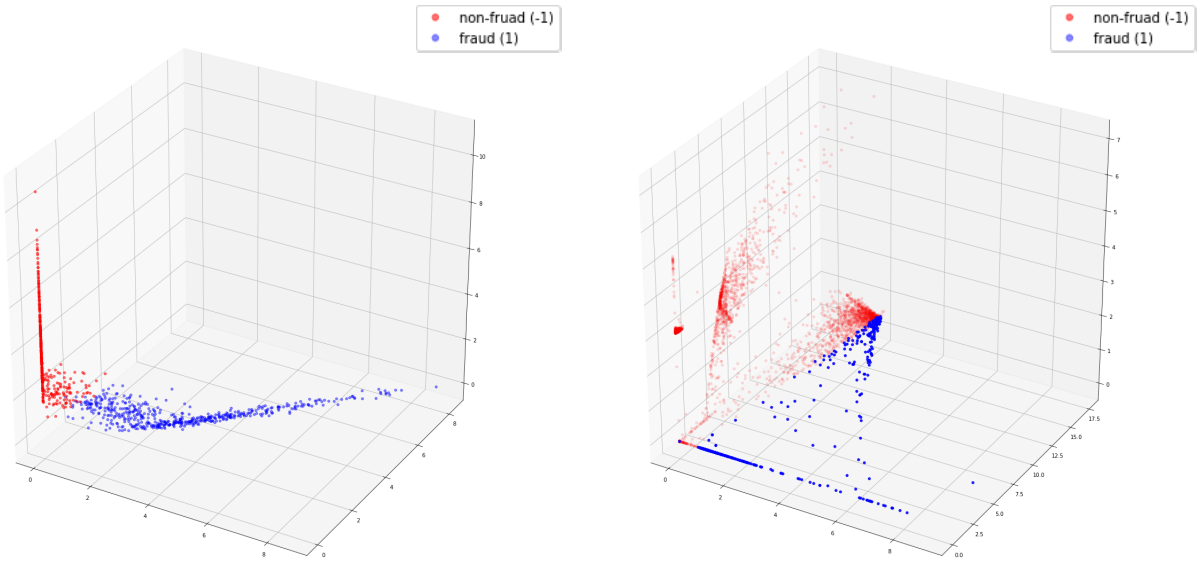
Con objeto de validar las propuestas de la sección anterior, se propone un procedimiento que nos permita afrontar simultáneamente el doble reto de aplicar herramientas potentes y probadas de ML de última generación, así como mantener la interpretabilidad del proceso de decisión subyacente. Este procedimiento podría permitir el cumplimiento de la rigurosa normativa de los organismos reguladores en cuanto a la protección de datos y no discriminación vigente para las entidades financieras. La metodología desarrollada ayuda a los métodos lineales interpretables al capturar las características relevantes, dejando de lado las cajas negras, al tiempo que minimiza los potenciales sesgos. A continuación, se muestran analizarán los diferentes experimentos realizados para alcanzar este objetivo así como los resultados de la metodología descrita.

6.5.1. Caracterización mediante el análisis local STE

Una vez detectado, cómo se vio en el capítulo 5.2, los diferentes niveles de respuesta en cada característica a pequeñas variaciones en el espacio real, pueden surgir las siguientes preguntas. En primer lugar, según el tipo de transacción, ¿son algunos rasgos más relevantes que otros y podemos explicar el proceso de decisión? O, por el contrario, ¿tienen todos los rasgos la misma relevancia y pueden producir alguna, o mala, interpretabilidad?. A partir de los resultados mostrados en la Figura 6.1, podemos observar cómo las clases de fraude y no fraude tienden a ocupar diferentes regiones en el espacio latente una vez que hemos desarrollado el *encoder* con *fine tunning*. La figura 6.1 se generó mediante el codificador definido en la sección 5.1. Como se puede ver en la Figura 6.1, hay diferentes regiones con una concentración de transacciones en el espacio latente para las clases de fraude y de no fraude, que podemos considerar como diferentes perfiles de transacción. Para ello, propusimos realizar el análisis para cada transacción. Siguiendo el modelo descrito anteriormente en la sección 6.2, sólo incorporamos las características que han demostrado

consolidar la información relevante en los experimentos anteriores.

Partimos del modelo de *autoencoder* aplicado en el capítulo anterior. Adicionalmente, con la intención de estudiar el comportamiento individual para cada una de las transacciones mediante STE, utilizamos un modelo VAE para generar un conjunto de muestras ficticias pero viables lo suficientemente cercanas a la transacción en estudio. Finalmente, para cada transacción en estudio y junto con las muestras generadas por el modelo VAE, el resultado del STE propondrá el modelo de regresión lineal que mejor se aproxime a la puntuación del clasificador implementado para este conjunto de datos. Para este conjunto de datos, los coeficientes del regresor se considerarán como los pesos que resumen la contribución de cada característica de esta transacción. Este enfoque nos permite, por tanto, formalizar un modelo lineal, coherente con los experimentos anteriores y específico, válido para la transacción estudiada. La generalización de este experimento sobre todas las transacciones dará lugar a un conjunto de pesos de características para cada transacción, que denominaremos STE.



(**a**) (**b**)

Figura 6.1: Espacio latente para los diferentes conjuntos de datos. (**a**) Conjunto de datos sintéticos. (**b**) Conjunto de datos PaySim. Las muestras rojas corresponden a transacciones no fraudulentas, mientras que las muestras azules corresponden a casos fraudulentos. El fraude y el no fraude se observan fácilmente en zonas separadas, aunque ambos grupos no son siempre claramente separables.

6.5.2. Clustering mediante ITR

Una vez obtenidos los pesos de STE, tenemos la contribución de cada característica para cada transacción. Con esto, es posible desarrollar de forma coherente un ranking de

características según su contribución, basado en la magnitud de los coeficientes siguiendo la estrategia descrita en la Sección 6.3. Podemos establecer, para cada transacción en estudio, la secuencia de características según su relevancia que mejor se aproxima al modelo predicho y su puntuación en la estrategia de clasificación realizada. Para obtener esta secuencia, y su modelización, se describió con detalle en la metodología, Sección 3.4, y se denominó ITR. Este ranking de características o ITR puede considerarse el perfil de la transacción al recoger la secuencia de las contribuciones de cada característica para esa transacción.

La figura 6.2 muestra dos ejemplos (en filas) de un conjunto de muestras que comparten el mismo valor ITR para el conjunto de datos sintéticos. La columna (a) muestra el ITR correspondiente de forma que en la primera fila, podemos ver que para este conjunto de transacciones el ITR muestra que las características informativas en el proceso de decisión se ordenan como $f4, f5, f2, f0, f1, f3$, mientras que para el conjunto de la siguiente fila, se ordenan como $f4, f5, f3, f2, f1, f0$. Para estos ITR podemos observar que el atributo $f3$ para el segundo conjunto tiene una alta relevancia, mientras que para el primer conjunto es el último. En la columna (b), representamos el espacio latente que se ha generado con el codificador definido en la sección 5.1. En dicho espacio latente, las transacciones se marcan en función de si fueron o no correctamente clasificadas por el modelo generado. Así, se puede observar que los elementos en azul corresponden a casos fraudulentos correctamente identificados y los elementos en rojo corresponden a casos no fraudulentos correctamente identificados. Las transacciones de ambas clases que fueron clasificadas incorrectamente se representan en verde y amarillo. Esto es visible en el conjunto de transacciones que comparten el ITR de la primera fila, ya que, en el caso de la segunda fila, el 100 % de las transacciones corresponden a la misma clase y han sido clasificadas correctamente, ya que están suficientemente alejadas del borde visual o frontera que separa las transacciones fraudulentas de las no fraudulentas. Se puede observar cómo las operaciones mal clasificadas, que también se recogen para comodidad del lector en la columna (c), se encuentran en la zona de frontera de las dos clases en el espacio latente, siendo coherentes con la estrategia de clasificación en este espacio. Por último, se incorpora en la columna (d) la matriz de confusión. En la Figura 6.3, se reproducen representaciones similares de las mismas figuras y contenidos que en la Figura 6.2 anterior, pero en este caso, para tres conjuntos de transacciones que comparten el mismo ITR para el conjunto de datos PaySim. Los resultados muestran los mismos patrones de comportamiento que los observados en el conjunto de datos sintéticos, donde se ve una fuerte relación entre las transacciones que comparten el mismo valor de ITR. Y al igual que en el conjunto de datos sintético, no siempre un ITR determinado indica un tipo de transacción, ya que las transacciones situadas en las zonas fronterizas hay transacciones de ambas clases.

(**a**) (**b**) (**c**) (**d**)

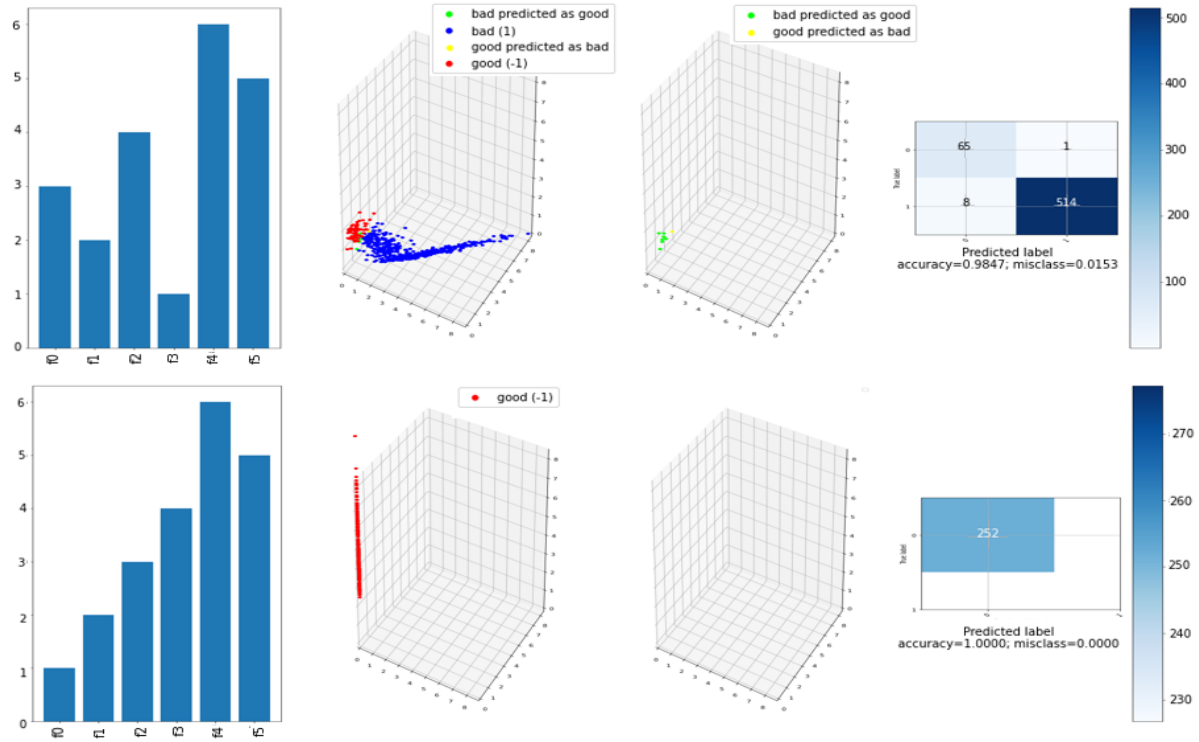


Figura 6.2: Análisis del ITR. Por filas, el análisis de 2 conjuntos de muestras que comparten el mismo ITR para el conjunto de datos sintéticos. Por columnas: (**a**) representa el ITR, (**b**) representación del espacio latente de las muestras que identifican la clase y la predicción AE, (**c**) espacio latente de las muestras mal clasificadas, (**d**) la matriz de confusión para todas las transacciones que comparten el mismo ITR.

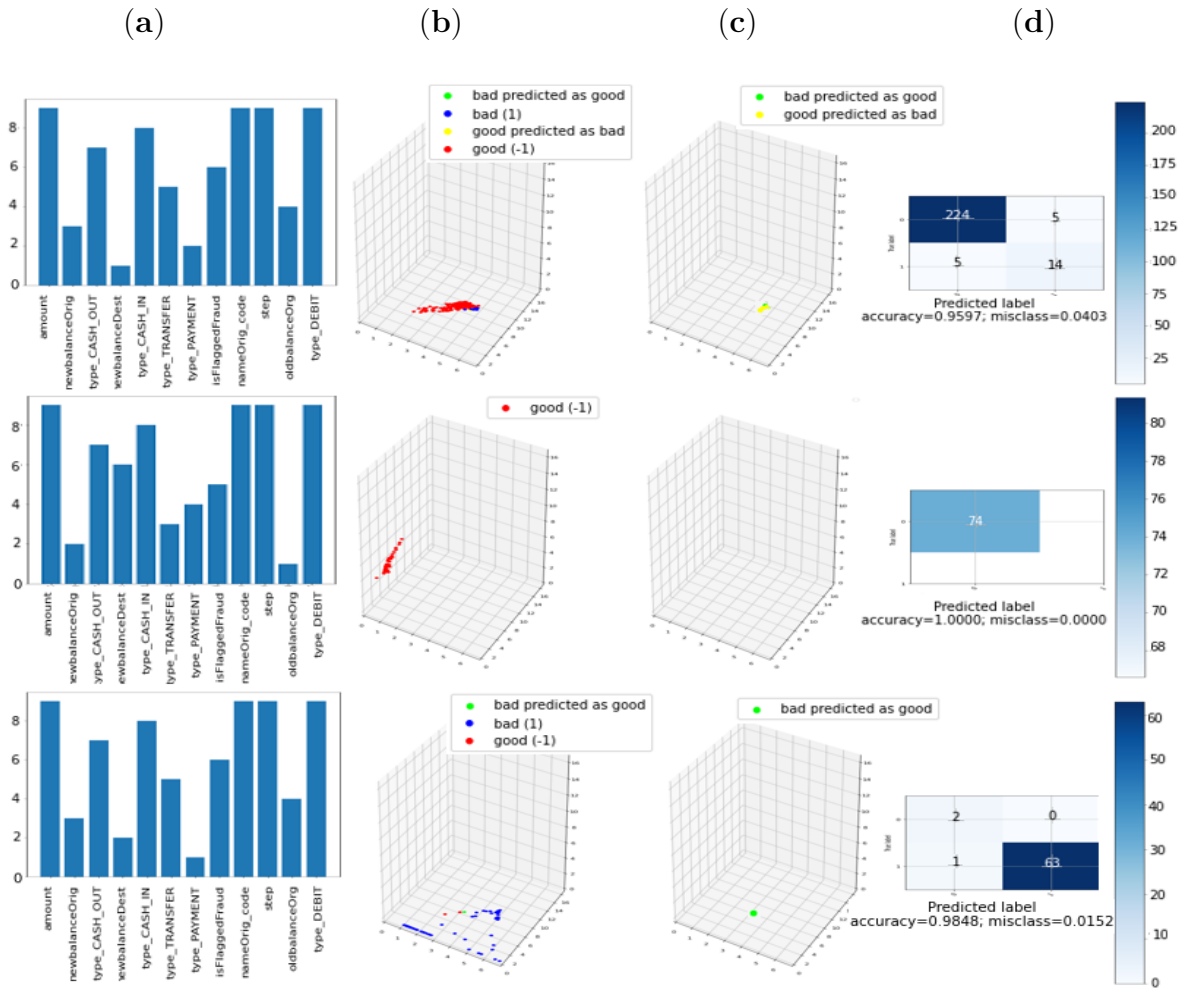
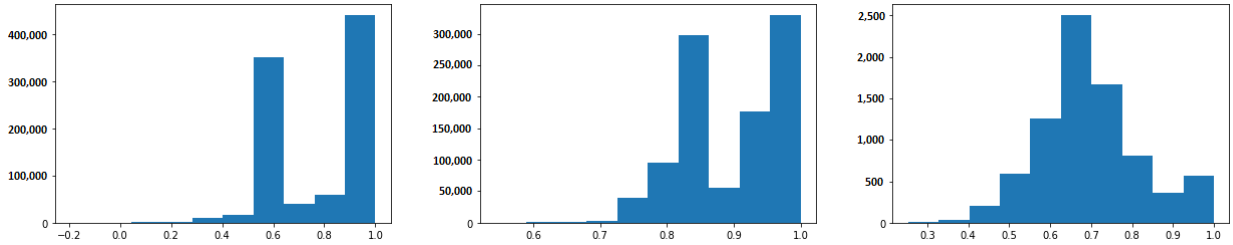


Figura 6.3: Diferentes grupos de ITR en el conjunto de datos de PaySim. La columna (**a**) representa el ITR. En la columna (**b**) podemos ver todas las transacciones con el mismo ITR con la clase predicha con nuestra metodología. La columna (**c**) es similar a (**b**) pero sólo muestra el desajuste. Por último, (**d**) muestra la matriz de confusión para todas las transacciones con el mismo ITR.

6.5.3. Caracterización

Una vez obtenido el ITR de cada transacción de manera individual, podemos realizar un análisis comparativo para evaluar la distribución del ITR. Para ello, procedemos a realizar un análisis de correlación de Kendall de todos los rankings ITR, mediante la comparativa por pares de las secuencias de las posiciones ocupada por cada característica en los rankings. Como resultado, obtenemos, para cada conjunto de datos, una colección de correlaciones de Kendall, cuya distribución se presenta en forma de histograma, tal y como se muestra en la Figura 6.4. Dado que existe un número discreto de combinaciones posibles, la correlación de Kendall alcanza valores discretos correspondientes que pueden corresponder eventualmente a conjuntos de datos que comparten características similares.

En la Figura 6.4, se puede observar para el caso (a) correspondiente al conjunto de datos sintético, cómo se aprecia una clara bimodalidad en los valores 1 y 0,6. Esta bimodalidad se repite en el caso de PaySim en 0,85 y 1, mientras que en el caso del conjunto de datos alemán, el modelo de población se acerca más a una distribución gaussiana. Bajo una perspectiva consolidada, la tabla 6.1 informa del promedio de todas las correlaciones τ para cada uno de los tres conjuntos de datos. Como puede observarse, se aprecia una mayor similitud en cuanto a las características informativas y su contribución al modelo en el caso de PaySim, seguido del conjunto de datos sintético y, con valores más bajos, para el caso del conjunto de datos alemán. Desde este punto de vista, este parámetro proporciona información sobre la dispersión en términos del número de modelos diferentes necesarios para poder caracterizar todo el conjunto de datos en estudio, por lo que puede entenderse en valor absoluto como la inversa del nivel de complejidad necesario para aproximar, por medios lineales, la realidad subyacente.



(**a**) (**b**) (**c**)

Figura 6.4: Distribución. Coeficiente de correlación de Kendall para cada conjunto de datos: (**a**) Conjunto de datos sintéticos; (**b**) Conjunto de datos PaySim; (**c**) Conjunto de datos alemanes.

Cuadro 6.1: Media de todos los coeficientes de correlación de Kendall para cada conjunto de datos. Este valor representa, en valor absoluto, la complejidad inversa necesaria para aproximar la realidad subyacente con un modelo lineal.

Dataset	$\boldsymbol{\tau}$
Synthetic	0,803
PaySim	0,900
German	0,696

6.6. Conclusiones del capítulo

En este capítulo hemos presentado y evaluado una metodología para obtener interpretabilidad en modelos no lineales, particularmente no hemos centrado en trabajar con

autoencoders. Si bien es cierto que el uso de *autoencoders* para construir espacios latentes es un claro ejemplo de un modelo de caja negra, en este capítulo, con el objetivo de mitigar las limitaciones de las cajas negras, hemos introducido dos nuevos mecanismos extensibles a cualquier otro mecanismo de caja negra.

En primer lugar, STE resume la contribución de cada característica para una transacción individual basada en las fluctuaciones a pequeña escala. Esto nos permite identificar de forma efectiva las características más influyentes, así como su relación entre ellas en el proceso de decisión para una transacción determinada, proporcionando así interpretabilidad, y por lo tanto dejando de lado las cajas negras mediante el uso de la tecnología más avanzada en técnicas de ML. En segundo lugar, el método ITR es capaz de construir un ranking de características individuales para cada transacción. Estos rankings representan una estimación más detallada de aquellas características que son más relevantes en comparación al resto en el proceso de decisión de una transacción individual. El fundamento del enfoque basado en el ITR es dar explicación a nivel transacción individual, que nos permite detectar perfiles de transacción similares cuando estas transacciones comparten ITR equivalentes. Con estos perfiles, podemos detectar posibles sesgos en las transacciones causados por dar demasiada importancia a las características no permitidas, y producir entonces una discriminación basada en varias categorías, como por ejemplo, la raza, el sexo o el estado civil.

También podemos confirmar la fuerte relación entre el STE y el ITR, dado que en la experimentación se comprobó cómo una pequeña variación de una característica en el espacio de entrada tiene una respuesta diferente en el espacio latente, por ejemplo, se observó que la característica *nuevobalanceOrig* tiene un alto impacto en el proceso de decisión detectado mediante estas pequeñas variaciones, y esto se confirmó cuando generamos los diferentes perfiles con ITR, que mismos ITR contemplaban tipos de transacciones diferentes.

Además de lo expresado en estas conclusiones respecto a la potencialidad en términos de interpretabilidad mostrada, la evaluación de la correlación de Kendall del ITR a lo largo de los diferentes conjuntos de datos mostró resultados interesantes que animan a profundizar en el análisis propuesto. En este sentido, las diferencias en las medias y distribuciones de la correlación de Kendall, para los distintos conjuntos de datos, pueden interpretarse en varias direcciones. Por un lado está la existencia de modalidades en las distribuciones, que corresponden a la existencia de un número de modelos diferentes necesarios para aproximarse a la realidad subyacente que puede estar relacionado con el número de conjuntos diferentes de transacciones. Este conjunto de transacciones no debe coincidir necesariamente con las clases objeto de estudio, sino con diferentes realidades, o variedades, que deben ser estudiadas individualmente y por separado para una mejor

comprensión de la base muestral y para maximizar la interpretabilidad.

Afirmamos por tanto, que es posible construir modelos interpretables robustos para cumplir simultáneamente con la restricción normativa y utilizar la potencia de las técnicas de ML. Para lograrlo, primero hemos evaluado sobre el conjunto de datos sintéticos con el objetivo de definir y afinar los modelos, y posteriormente se aplicaron los modelos al resto de conjuntos de datos reales para verificar su generalización y consistencia. Podemos concluir que nuestra metodología proporciona una evaluación detallada a nivel de transacción, añadiendo interpretabilidad para cada transacción y haciendo visibles las características más relevantes en el proceso de decisión. Esta perspectiva individualizada, imparcial y rastreable proporciona la transparencia necesaria, no sólo para cumplir con la normativa, sino también para poder justificar cada transacción clasificada ante clientes y autoridades.

Capítulo 7

Conclusiones y Trabajos Futuros

7.1. Resumen

La investigación realizada en esta tesis doctoral se centra en el desarrollo de una metodología fiable, imparcial e interpretable para la detección de fraude en los sistemas de CFD. Como resumen general, podemos afirmar que el objetivo y la contribución de esta tesis se ha dirigido a tres áreas: selección de características principales, aislamiento de las transacciones fraudulentas de las no fraudulentas, y proporcionar transparencia en el proceso de decisión. La investigación en CFD se ha centrado históricamente en sistemas expertos, dejando de lado los algoritmos de ML y todo su potencial debido las dificultades que presentan estos modelos a la hora de interpretar el mecanismo subyacente. La búsqueda de un mecanismo que permita conjugar estas potentes técnicas con la interpretabilidad necesaria, es lo que ha motivado este trabajo. Para abordar estas cuestiones en esta tesis se han empleado conjuntos de datos reales y sintéticos de CFD, permitiéndonos evaluar en primera persona la problemática y retos a los que se enfrentan los analistas de fraude. La metodología usada en este trabajo se ha basado en el desarrollo de modelos sustitutivos independientes y a medida para cada transacción que nos ha permitido dotar de interpretabilidad cada proceso de decisión. Nuestros métodos se han comparado con las técnicas más avanzadas y los resultados indican que los métodos interpretables desarrollados han demostrado no sólo alcanzar una precisión aceptable, sino ofrecer métodos computacionalmente más eficientes para el análisis. Debido a estas características, pueden utilizarse para apoyar la toma de decisiones en la detección de fraude y, por tanto, ayudar a los analistas. Afirmamos que es posible construir modelos interpretables robustos para cumplir simultáneamente con la restricción normativa y utilizar al mismo tiempo la potencia de las técnicas de ML.

7.2. Selección de características

En esta tesis evaluamos y validamos una nueva metodología para abordar la selección de características que nos permite conocer la relevancia y las relaciones de las características seleccionadas con respecto al conjunto total de características.

Esta metodología presenta un análisis intensivo de una serie de técnicas de aprendizaje basado en ML como son GB, SVR, LR y LDA que nos han permitido detectar las características mas relevantes para la detección de fraude. Adicionalmente, se han desarrollado un conjunto de filtros que nos han permitido robustecer el proceso de selección de características. Como conclusión general obtenida a partir de la experimentación llevada a cabo, podemos decir que es posible desarrollar un modelo basado en ML que es interpretable para la de selección de características presentadas en esta tesis. El resultado evidencia que no sólo mejora la detección de CFD, sino que al mismo tiempo permite visualizar la contribución de cada característica en el proceso de decisión, ofreciendo así la necesaria interpretabilidad del modelo y de los resultados. Para hacer frente a la compleja dicotomía de las herramientas de ML y la interpretabilidad, dicha metodología obtiene las características informativas y calcula sus contribuciones en el proceso de decisión para cada una de ellas, dejando de lado las cajas negras debido que evidencia que características influyen en el resultado y en qué cantidad con lo que minimiza los posibles sesgos.

Para concluir podemos afirmar que es posible construir modelos lineales robustos y explicativos que satisfagan simultáneamente las restricciones normativas y utilicen la potencia de las técnicas de ML. Para ello, nuestro trabajo fue doble. En primer lugar, desarrollamos un conjunto de datos sintéticos que nos permitió definir y afinar los modelos en un entorno controlado. En segundo lugar, de los modelos que resultaron satisfactorios se replicaron posteriormente en conjuntos de datos reales para verificar su correcta generalización y coherencia. Desde el punto de vista computacional, y teniendo en cuenta los cuatro subconjuntos diferentes de características evaluados, como fueron, el uso de todas las características, las características obtenidas mediante IVI, características aplicando el filtro MIFF de las obtenidas mediante IVI y las equivalentes aplicando el filtro RFF. Los dos últimos ofrecieron reducciones significativas del número de variables, con lo que se mejoró considerablemente la carga de trabajo de la CPU y por lo tanto la reducción de recursos necesarios. En ambos casos, la precisión fue satisfactoria, destacando que mediante el filtro MIFF ofreció tanto la mejor precisión así como una mayor estabilidad en los resultados entre los métodos.

7.3. Aislar las transacciones fraudulentas de las no fraudulentas

Aislar las transacciones fraudulentas de las no fraudulentas es un problema difícil pero común en los CFD debido a que los estafadores hacen todo lo que está en su mano para emular que las transacciones fraudulentas difieran lo menos posible de las reales, tratando de modelar patrones de comportamiento extremadamente similares.

La contribución de esta tesis a este tema consiste en comprimir y codificar las transacciones utilizando métodos no lineales como los *autoencoders* para aislar eficazmente las transacciones fraudulentas de las no fraudulentas. En nuestra metodología, hemos construido un *autoencoder* con una etapa de codificación y otra de decodificación, donde el codificador comprime el espacio real en un espacio latente en 3D por razones de representación visual. Para ello, el *autoencoder* se construyó con una primera capa cuyo número de celdas es el número de características seleccionadas en el método IVI aplicándole el filtro MIFF. La segunda capa tiene tres celdas para conseguir representaciones en 3D, y finalmente la tercera capa es la reconstrucción con el espacio latente al espacio de entrada de nuevo. El codificador se construye con las capas primera y segunda, y el decodificador se corresponde con la tercera capa. Las funciones de activación entre capas empleadas fueron ReLU. Una vez definida la arquitectura del *autoencoder*, ajustamos el *autoencoder* con el conjunto de datos de entrenamiento. De este modo, y para obtener una mayor dispersión de las transacciones en el espacio latente, aplicamos la técnica conocida como *fine tunning* añadiendo una última capa softmax al codificador y ajustamos de nuevo congelando las capas del codificador y permitiendo únicamente que el gradiente se retropropague a través de la capa softmax. De los experimentos se desprende que el uso del espacio latente nos permite separar de manera eficiente y eficaz las transacciones fraudulentas de las legítimas. Esta separación facilita el proceso de clasificación como demostró las diferentes experimentaciones realizadas. Otro punto a resaltar es que esta reducción de la dimensionalidad, permite observar la sensibilidad de ciertas características ante pequeñas variaciones en el espacio real o de entrada, tienen una respuesta amplificada en el espacio latente, lo cual confirma que ciertas características tienen mayor influencia en los procesos de decisión. Por último, indicar que los *encoder* empleados fueron definidos con una dimensionalidad de 3D para facilitar su interpretación gráfica, quedaría pendiente para trabajos futuros el uso de otras dimensionalidades con el objetivo de mejorar la precisión.

7.4. Transparencia en el proceso de decisión

Nuestra contribución a este tema es un método interpretable basado en transacciones individuales. De los experimentos se desprende que, en el espacio latente, los algoritmos de ML mejoran la tarea de clasificación al diferenciar mejor los distintos tipos de transacciones. El uso de espacios latentes podría seguir considerándose como un modelo de tipo caja negra y con el objetivo de mitigar esta percepción y dotar de explicabilidad las cajas negras, hemos presentado dos nuevos mecanismos.

En primer lugar, STE resume la contribución de cada característica para cada transacción individual basándose en modelos sustitutivos simples construidos a medida para cada transacción que se desee interpretar, y en segundo lugar, el método ITR que es capaz de construir una clasificación de características individuales para cada transacción. Estas clasificaciones representan una estimación más precisa de que características son más importantes que el resto para predecir la clase resultado durante el proceso de decisión para una transacción determinada. El fundamento del enfoque basado en el ITR es una explicación a medida para cada transacción, que nos permite detectar perfiles de transacción similares para aquellas transacciones que compartan ITR equivalentes.

Con estos perfiles, podemos detectar posibles sesgos en las transacciones causados por dar demasiada importancia a las características no permitidas, y producir entonces una discriminación basada en diversas categorías, como por ejemplo, la raza, el sexo o el estado civil. También podemos revelar la fuerte relación entre el STE y el ITR, visible en la experimentación donde comprobamos cómo una pequeña variación de una característica en el espacio de entrada tiene una respuesta diferenciada en el espacio latente. Además de lo expresado en estas conclusiones respecto a la potencialidad en términos de explicabilidad mostrada, la evaluación de la correlación de Kendall del ITR a lo largo de los diferentes conjuntos de datos mostró resultados interesantes que animan a profundizar en el análisis. En este sentido, las diferencias en las medias y distribuciones de la correlación de Kendall, para los distintos conjuntos de datos, pueden interpretarse en varias direcciones, por un lado, la existencia de modalidades en las distribuciones, que corresponden a la existencia de un número de modelos diferentes necesarios para aproximarse a la realidad subyacente que puede estar relacionada con el número de conjuntos diferentes de transacciones que se producen. Este conjunto de transacciones no tiene por qué coincidir con las clases objeto de estudio, sino con diferentes realidades, o variedades, que deben ser estudiadas individualmente y por separado para una mejor comprensión de la base muestral para una mayor interpretabilidad. Por otro lado, la presencia de una única modalidad indicaría que un modelo lineal, único y representativo, sería capaz de evaluar con al menos la misma precisión que el modelo altamente complejo como puede ser un modelo no lineal. En tercer

lugar, la existencia de una distribución no modal, ya sea uniforme, gaussiana o de cualquier otro tipo, podría sugerir diversas interpretaciones que, en todos los casos, podrían sugerir que se afronten nuevos métodos de análisis, ya sea por la existencia de infinitos modelos lineales, equivalentes o un número limitado de modelos no lineales. En esta dirección, es necesario señalar que si bien es posible para todas y cada una de las transacciones obtienen un modelo ITR, que proporcione interpretabilidad a la clasificación, éste será únicamente válida para esa transacción, no siendo posible generalizarlo a otros casos. Esta aproximación local basada en la técnica de STE, podría entenderse como una ventaja a la hora de la interpretabilidad, aunque esta interpretabilidad a nivel individual podría hacer que los reguladores y autoridades sean reticentes a su validación. Es por ello que se propone, como siguiente paso de esta línea de trabajo, avanzar en el conocimiento de estas distribuciones y de los modelos de datos que las originan para poder proponer también modelos no lineales interpretables y generalizables que aseguren la consistencia, si no para el total de muestras del conjunto, al menos para un grupo amplio de ellas que formen parte de subconjuntos que compartan el mismo ITR.

7.5. Líneas futuras

Los resultados y conclusiones presentados en esta tesis para mejorar los sistemas de CFD tanto en precisión como en interpretabilidad abren nuevas líneas de trabajo potenciales para el futuro y en particular:

- Uso de nuevos conjunto de datos reales. Aplicar nuevos conjuntos de datos con transacciones reales y actuales de diferentes geografías. Observar patrones comunes y las características sensibles al fraude.

- Detección de fraude en tiempo real. Desarrollar los métodos STE e ITR accesibles vía servicio web o *API* de tal manera que permita a los analistas de fraude analizar fácilmente las transacciones individuales y poder aplicar sistemas de alertas tempranas mediante el envió de la información de la transacción a esta API.

- Uso de Deep Learning. En esta tesis nos hemos centrado en dar interpretabilidad a los procesos de decisión, sin embargo, por razones de representación hemos construido arquitecturas sencillas para los *autoencoders*. Una futura línea de investigación consistiría en construir modelos de aprendizaje profundo más complejos para mejorar la precisión.

- ITR-clustering. A través de los resultados obtenidos por ITR, explorar los conjuntos de transacciones que comparten las mismas regiones en el espacio latente.

- Durante la experimentación para de selección de características, tres características destacaron con contribuciones un 50 % superiores a las de sus compañeros. Este subconjunto especial de características requiere un análisis detallado, ya que no sólo mostraron sistemáticamente una gran contribución en el proceso de decisión en todos los métodos aplicados. Cabe señalar ahora que estas variables (*credit usage, over draft, purpose of credit for new car* ya estaban identificadas en la bibliografía [98, 99].

- Por otro lado, la interpretabilidad necesaria para la aplicación del CFD donde existe una alta regulación y requiere transparencia, podría obtenerse mediante el análisis de los pesos de cada característica. Cabe el análisis de la posible eliminación de característica, habiéndose observado que las características que participan con menor intensidad podrían ser eliminadas sin efectos relevantes desde un punto de vista práctico.

- Análisis de las distribuciones de puntuación obtenidas en los espacios latentes cuando aplicamos pequeñas variaciones. Dada la presencia de una única modalidad indica la posibilidad de aproximación a un modelo lineal, único y representativo, capaz de evaluar con al menos la misma precisión que el modelo complejo evaluado. Por el contrario, la existencia de una distribución no modal, ya sea uniforme, gaussiana o de cualquier otro tipo, podría sugerir varias interpretaciones que, en todos los casos, podrían sugerir enfrentarse a nuevos métodos de análisis, ya sea por la existencia de infinitos modelos lineales, equivalentes o un número limitado de modelos no lineales. Es por ello que se propone, como siguiente paso de este trabajo, avanzar en el conocimiento de estas distribuciones y de los modelos de datos que las originan para poder proponer también modelos no lineales interpretables y generalizables que aseguren la consistencia, si no para el total de las muestras del conjunto, al menos para un grupo amplio de ellas que formen parte de subconjuntos que compartan el mismo ITR.

Bibliografía

[1] Dornadula, Vaishnavi y S Geetha: *Credit Card Fraud Detection using Machine Learning Algorithms*. Procedia Computer Science, 165:631–641, 2019.

[2] Pascual, Al: *Future Proof Card Authorization*. Informe técnico, Javelin Strategy & Research, 2015.

[3] Buchanan, Bonnie G: *Artificial intelligence in finance*. Informe técnico, The Alan Turing Institute, 2019.

[4] *Machine learning in UK financial services*. Informe técnico, Bank of England, 2019.

[5] Yan, Han y Sheng Lin: *New Trend in Fintech: Research on Artificial Intelligence Model Interpretability in Financial Fields*. Open Journal of Applied Sciences, 09:761–773, 2019.

[6] Wall, Larry: *Some financial regulatory implications of artificial intelligence*. Journal of Economics and Business, 100(C):55–63, 2018.

[7] Ana, Fernandez: *Artificial intelligence in financial services*. Informe técnico, Banco de España, 2019.

[8] Wedge, Roy and Kanter, James Max and Veeramachaneni, Kalyan and Rubio, Santiago Moral and Perez, Sergio Iglesias: *Solving the False Positives Problem in Fraud Prediction Using Automated Feature Engineering*. En Brefeld, Ulf and Curry, Edward and Daly, Elizabeth and MacNamee, Brian and Marascu, Alice and Pinelli, Fabio and Berlingerio, Michele and Hurley, Neil (editor): *Machine Learning and Knowledge Discovery in Databases*, páginas 372–388. Springer International Publishing, 2019.

[9] Muñoz-Romero, Sergio, Arantza Gorostiaga, Cristina Soguero-Ruiz, Inmaculada Mora-Jiménez y José Luis Rojo-Álvarez: *Informative variable identifier: Expanding interpretability in feature selection.* Pattern Recognition, 98, 2020.

[10] Bengio, Y., Pascal Lamblin, Dan Popovici, Hugo Larochelle y U. Montreal: *Greedy layer-wise training of deep networks.* Advances in Neural Information Processing Systems, 19:153—160, 2007.

[11] Bank, Dor, Noam Koenigstein y Raja Giryes: *Autoencoders*, 2020.

[12] Freitas, Alex: *Comprehensible classification models: A position paper.* Association for Computing Machinery. SIGKDD Explorations Newsletter, 15:1–10, 2014.

[13] Carvalho, Diogo, Eduardo Pereira y Jaime Cardoso: *Machine Learning Interpretability: A Survey on Methods and Metrics.* Electronics, 8(8):832–840, 2019.

[14] Goyal, Rahul, Amit Manjhvar y Habeebullah Hussaini Syed: *Review on Credit Card Fraud Detection using Data Mining Classification Techniques and Machine Learning Algorithms.* SSRN Electronic Journal, 7:972–976, Enero 2020.

[15] *The Scope of the Card Not Present (CNP) Fraud Problem.* Informe técnico, Accenture, 2018.

[16] *Secure cards as payment fraud climbs, time for new solutions.* Informe técnico, Accenture, 2019.

[17] *Unmask digital fraud. Today.* Informe técnico, Accenture, 2019.

[18] Research, Juniper: *European card fraud*, 2017. `https://www.juniperresearch.com/researchstore/fintech-payments/online-payment-fraud-research-report`, visitado el 2022-05-22.

[19] *Payment Fraud Statistics.* Informe técnico, Australian Payments Network, Mayo 2016.

[20] Yadav, Tarun y Arvind Mallari Rao: *Technical Aspects of Cyber Kill Chain*, 2015.

[21] Delamaire, Linda, Hussein Abdou y John Pointon: *Credit card fraud and detection techniques: A review.* Banks and Bank Systems, 4, Enero 2009.

[22] Muñoz, Alex: *Countering Man-in-the-Middle Attacks in Point of Sale Credit Card Terminals.* 2015.

[23] Interpol: *Social engineering scams*, 2022. `https://www.interpol.int/en/Crimes/Financial-crime/Social-engineering-scams`, visitado el 2022-05-22.

[24] Bolton, Richard y David Hand: *Statistical Fraud Detection: A Review*. Statistical Science, 17, Agosto 2002.

[25] Leonard, Kevin J.: *Detecting credit card fraud using expert systems*. Computers and Industrial Engineering, 25(1):103–106, 1993.

[26] Baumann, Michaela: *Improving a Rule-based Fraud Detection System with Classification Based on Association Rule Mining*. En *INFORMATIK 2021*, páginas 1121–1134. Gesellschaft für Informatik, Bonn, 2021.

[27] Dheepa, V y R.Dhanapal R: *Analysis of Credit Card Fraud Detection Methods*. SHORT PAPER International Journal of Recent Trends in Engineering, 2, Enero 2009.

[28] Brause, R., T. Langsdorf y M. Hepp: *Neural data mining for credit card fraud detection*. En *Proceedings 11th International Conference on Tools with Artificial Intelligence*, páginas 103–106, 1999.

[29] Chen, Chaofan, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang y Tong Wang: *An Interpretable Model with Globally Consistent Explanations for Credit Risk*, 2018.

[30] Pumsirirat, Apapan y Liu Yan: *Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine*. International Journal of Advanced Computer Science and Applications, 9(1), 2018.

[31] Darwish, Saad: *An intelligent credit card fraud detection approach based on semantic fusion of two classifiers*. Soft Computing, 24, Enero 2020.

[32] Vynokurova, Olena, Dmytro Peleshko, Oleksandr Bondarenko, Vadim Ilyasov, Vladislav Serzhantov y Marta Peleshko: *Hybrid Machine Learning System for Solving Fraud Detection Tasks*. En *2020 IEEE Third International Conference on Data Stream Mining Processing (DSMP)*, páginas 1–5, 2020.

[33] RB, Asha y Suresh Kumar KR: *Credit card fraud detection using artificial neural network*. Global Transitions Proceedings, 2(1):35–41, 2021, ISSN 2666-285X.

[34] Yang, Yang, Rong Chen, Xiao BAI y DeHeng Chen: *Finance Fraud Detection With Neural Network*. E3S Web of Conferences, 214, Enero 2020.

[35] Weston, Jason, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio y Vladimir Vapnik: *Feature Selection for SVMs.* En Leen, T., T. Dietterich y V. Tresp (editores): *Advances in Neural Information Processing Systems*, volumen 13. MIT Press, 2000. `https://proceedings.neurips.cc/paper/2000/file/8c3039bd5842dca3d944faab91447818-Paper.pdf`.

[36] Dubey, Saurabh C., Ketan S. Mundhe y Aditya A. Kadam: *Credit Card Fraud Detection using Artificial Neural Network and BackPropagation.* En *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, páginas 268–273, 2020.

[37] Sriram Sasank, J. V. V., G. Ram Sahith, K. Abhinav y Meena Belwal: *Credit Card Fraud Detection Using Various Classification and Sampling Techniques: A Comparative Study.* En *2019 International Conference on Communication and Electronics Systems (ICCES)*, páginas 1713–1718, 2019.

[38] Rtayli, Naoufal y Nourddine Enneya: *Selection Features and Support Vector Machine for Credit Card Risk Identification.* Procedia Manufacturing, 46:941–948, 2020, ISSN 2351-9789. 13th International Conference Interdisciplinarity in Engineering, INTER-ENG 2019, 3–4 October 2019, Targu Mures, Romania.

[39] Alam, Md, Prajoy Podder, Subrato Bharati y M. Rubaiyat Mondal: *Effective Machine Learning Approaches for Credit Card Fraud Detection.* páginas 154–163, Abril 2021, ISBN 978-3-030-73602-6.

[40] Itoo, Fayaz, Meenakshi Mittal y Satwinder Singh: *Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection.* International Journal of Information Technology, 13, Febrero 2020.

[41] Malini, N. y M. Pushpa: *Article: Investigation of Credit Card Fraud Recognition Techniques based on KNN and HMM.* IJCA Proceedings on International Conference on Communication, Computing and Information Technology, ICCCMIT 2017(1):9–13, June 2018. Full text available.

[42] Friedman, Jerome H.: *Stochastic gradient boosting.* Computational Statistics & Data Analysis, 38(4):367–378, 2002.

[43] Chen, Tianqi y Carlos Guestrin: *XGBoost: A Scalable Tree Boosting System.* En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining, páginas 785—-794, New York, NY, USA, 2016. Association for Computing Machinery, ISBN 9781450342322. https://doi.org/10.1145/2939672.2939785.

[44] Hancock, John y Taghi Khoshgoftaar: *Gradient Boosted Decision Tree Algorithms for Medicare Fraud Detection*. SN Computer Science, 2, Julio 2021.

[45] Taha, Altyeb Altaher y Sharaf Jameel Malebary: *An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine*. IEEE Access, 8:25579–25587, 2020.

[46] *Machine-learning promises to shake up large swathes of finance*. The Economist, 2017. https://www.economist.com/news/finance-and-economics/21722685-fields-trading-credit-assessment-fraud-prevention-machine-learning.

[47] *On a proposal for a regulation laying down harmonised rules on artificial intelligence*. Informe técnico, Official Journal of the European Union, Diciembre 2021.

[48] Angwin, Julia, Jeff Larson, Surya Mattu y Lauren Kirchner: *Machine Bias: there's software used across the country to predict future criminals.* Informe técnico, ProPublica, Mayo 2016.

[49] Bickel, P. J., E. A. Hammel y J. W. O'Connell: *Sex Bias in Graduate Admissions: Data from Berkeley*. Science, 187(4175):398–404, 1975.

[50] Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman y Aram Galstyan: *A Survey on Bias and Fairness in Machine Learning*, 2019.

[51] Makhlouf, Karima, Sami Zhioua y Catuscia Palamidessi: *On the Applicability of Machine Learning Fairness Notions*. SIGKDD Explor. Newsl., 23(1):14—23, 2021.

[52] Kusner, Matt J., Joshua R. Loftus, Chris Russell y Ricardo Silva: *Counterfactual Fairness*, 2017.

[53] *Understanding Bias in Machine Learning. White paper.* Informe técnico, Lexalytics, Inc., Abril 2019.

[54] Bhargava, Vaishnavi, Miguel Couceiro y Amedeo Napoli: *LimeOut: An Ensemble Approach To Improve Process Fairness*, 2020.

[55] Alves, Guilherme, Vaishnavi Bhargava, Fabien Bernier, Miguel Couceiro y Amedeo Napoli: *FixOut: an ensemble approach to fairer models*. HAL, open science, 1, 2020.

[56] Prenio, Jermy y Jeffery Yong: *Humans keeping AI in check – emerging regulatory expectations in the financial sector*. Informe técnico, Financial Stability Institute, 2021.

[57] Ribeiro, Marco, Sameer Singh y Carlos Guestrin: *Why Should I Trust You?: Explaining the Predictions of Any Classifier*, 2016.

[58] Hu, Linwei, Jie Chen, Vijayan N. Nair y Agus Sudjianto: *Surrogate Locally-Interpretable Models with Supervised Machine Learning Algorithms*, 2020.

[59] Ribeiro, Marco Tulio, Sameer Singh y Carlos Guestrin: *Anchors: High-Precision Model-Agnostic Explanations*. AAAI Press, 2018, ISBN 978-1-57735-800-8.

[60] Hall, Patrick, Navdeep Gill, Megan Kurka y Wen Phan: *Machine learning interpretability with h2o driverless ai*. H2O. ai, 2017.

[61] Friedman, Jerome: *Greedy Function Approximation: A Gradient Boosting Machine*. The Annals of Statistics, 29, Noviembre 2000.

[62] Zhao, Qingyuan y Trevor Hastie: *Causal Interpretations of Black-Box Models*. Journal of Business & Economic Statistics, 2019:1–19, Junio 2019.

[63] Bertsimas, D., Arthur Delarue, Patrick Jaillet y S. Martin: *The Price of Interpretability*, 2019.

[64] Craven, Mark W. y Jude W. Shavlik: *Extracting Tree-Structured Representations of Trained Networks*. página 24–30, Cambridge, MA, USA, 1995. MIT Press.

[65] Zilke, Jan, Eneldo Mencía y Frederik Janssen: *DeepRED – Rule Extraction from Deep Neural Networks*. páginas 457–473, Octubre 2016, ISBN 978-3-319-46306-3.

[66] Oh, Seong Joon, Max Augustin, Bernt Schiele y Mario Fritz: *Towards Reverse-Engineering Black-Box Neural Networks*, 2017.

[67] Tramèr, Florian, Fan Zhang, Ari Juels, Michael K. Reiter y Thomas Ristenpart: *Stealing Machine Learning Models via Prediction APIs*, 2016.

[68] Papernot, Nicolas, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik y Ananthram Swami: *Practical Black-Box Attacks against Machine Learning*, 2016.

[69] Bastani, Osbert, Carolyn Kim y Hamsa Bastani: *Interpreting Blackbox Models via Model Extraction*, 2017.

[70] Lundberg, Scott y Su In Lee: *A Unified Approach to Interpreting Model Predictions*, 2017.

[71] Shapley, Lloyd S.: *17. A Value for n-Person Games*. 1953.

[72] Lundberg, Scott M., Gabriel G. Erion y Su In Lee: *Consistent Individualized Feature Attribution for Tree Ensembles*, 2018.

[73] Akamai: *Bot-manager*, 2022. `https://www.akamai.com/es/products/bot-manager`, visitado el 2022-05-22.

[74] LexisNexis: *Behaviosec*, 2022. `https://www.behaviosec.com/`, visitado el 2022-05-22.

[75] IBM: *trusteer*, 2022. `https://www.ibm.com/support/trusteer`, visitado el 2022-05-22.

[76] Threatmark: *AntiFraud-suite*, 2022. `https://www.threatmark.com/products/anti-fraud-suite-afs/`, visitado el 2022-05-22.

[77] Transunion: *Truvalidate*, 2022. `https://www.transunion.com/solution/truvalidate`, visitado el 2022-05-22.

[78] Cybersource: *Decision Manager*, 2022. `https://www.cybersource.com/apac/en/solutions/fraud-and-risk-management/decision-manager.html`, visitado el 2022-05-22.

[79] Featurespace: *Aric risk-hub*, 2022. `https://www.featurespace.com/aric-risk-hub/`, visitado el 2022-05-22.

[80] fico: *Falcon*, 2022. `https://www.fico.com/en/products/fico-falcon-fraud-manager`, visitado el 2022-05-22.

[81] Gartner: *Online fraud detection systems*, 2021. `https://www.gartner.com/reviews/market/online-fraud-detection-systems`, visitado el 2022-05-22.

[82] seon: *Fight-fraud-with-machine-learning/*, 2022. `https://seon.io/`, visitado el 2022-05-22.

[83] Dua, Dheeru y Casey Graff: *UCI Machine Learning Repository*, 2017. data retrieved from UCI, `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`.

[84] Lopez-Rojas, Edgar Alonso, Ahmad Elmir y Stefan Axelsson: *PAYSIM: A FINANCIAL MOBILE MONEY SIMULATOR FOR FRAUD DETECTION*. 2016.

[85] Kaya Uyanık, Gulden y Neşe Güler: *A Study on Multiple Linear Regression Analysis*. Procedia - Social and Behavioral Sciences, 106:234—-240, 2013.

[86] Witten, Daniela M. y Robert Tibshirani: *Penalized classification using Fisher's linear discriminant*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), (5):753–772, 2011.

[87] Vapnik, Vladimir N.: *Statistical Learning Theory*. Wiley-Interscience, 1998.

[88] Schlkopf, Bernhard, Alexander J. Smola y Francis Bach: *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2018, ISBN 0262536579.

[89] Zhang, Yongli: *Support Vector Machine Classification Algorithm and Its Application*. páginas 179–186. International Conference on Information Computing and Applications, 2012.

[90] Rojo-Álvarez, Jose Luis, Manel Martinez-Ramón, Jordi Munoz-Mari y Gustau Camps-Valls: *Digital Signal Processing with Kernel Methods*. Wiley-IEEE Press, 1st edición, 2018, ISBN 1118611799.

[91] Natekin, Alexey y Alois Knoll: *Gradient Boosting Machines, A Tutorial*. Frontiers in neurorobotics, 7:21, 2013.

[92] Bentéjac, Candice, Anna Csörgő y Gonzalo Martínez-Muñoz: *A comparative analysis of gradient boosting algorithms*. Artificial Intelligence Review, 2020.

[93] Schölkopf, Bernhard, John Platt y Thomas Hofmann: *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference (Bradford Books)*. The MIT Press, 2007, ISBN 0262195682.

[94] Baldi, Pierre: *Autoencoders, Unsupervised Learning, and Deep Architectures*. En Guyon, Isabelle, Gideon Dror, Vincent Lemaire, Graham Taylor y Daniel Silver (editores): *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volumen 27 de *Proceedings of Machine Learning Research*, páginas 37–49. PMLR, 2012.

[95] Bengio, Y., G Guyon, V Dror, G Lemaire, D Taylor y Daniel Silver: *Deep learning of representations for unsupervised and transfer learning*. 7, 2011.

[96] Käding, Christoph, Erik Rodner, Alexander Freytag y Joachim Denzler: *Fine-Tuning Deep Neural Networks in Continuous Learning Scenarios*. páginas 588–605, 2017.

[97] Agrawal, Pulkit, Ross Girshick y Jitendra Malik: *Analyzing the Performance of Multilayer Neural Networks for Object Recognition*. Volumen 8695, páginas 329–344, 2014.

[98] Macailao, Manuelito: *Raising the Red Flags: The Concept and Indicators of Occupational Fraud*. Journal of Critical Reviews, 7:26–29, 2020.

[99] DiNapoli, T. P: *Red Flags for Fraud. State of New York Office of the State Comptroller*. State of New York. Office of the State Comptroller, páginas 1–14, 2008.

[100] Gonzalez, Jesus, Lawrence Holder y Diane Cook: *Graph Based Concept Learning*. FLAIRS Conference, 2000.