COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

Optimizando la Calidad de los Datos en Aplicaciones Empresariales: Estrategias de Detección de Outliers

Autor: Carmen Rebollo Monjo
Director: Jenny Alexandra Cifuentes Quintero

MADRID| Junio 2024

Resumen

En la actualidad, la sociedad está rodeada de una cantidad sin precedentes de datos que, si se utilizan adecuadamente, pueden proporcionar información valiosa y servir como una ventaja competitiva en el ámbito empresarial. En esta era del *big data* y la toma de decisiones basadas en datos, la detección de *outliers* se ha convertido en una herramienta de gran utilidad para garantizar la precisión y la integridad de los análisis. Específicamente, en el ámbito del riesgo crediticio, la identificación de valores atípicos es fundamental, ya que puede revelar riesgos no detectados, prevenir fraudes y optimizar las políticas crediticias. Este trabajo se centra en abordar el problema de la detección de *outliers* en el contexto del riesgo crediticio, donde identificar estos valores atípicos permite descubrir riesgos ocultos y prevenir actividades fraudulentas, entre otras aplicaciones.

Para ello, se han seleccionado y aplicado diversas técnicas de detección de outliers, incluyendo *Gaussian Mixture Models* (GMM), Análisis de Componentes Principales Probabilístico (PPCA), distancia de Mahalanobis, *Local Outlier Factor* (LOF) e *Isolation Forest* (iForest). Estas técnicas han sido elegidas por su capacidad para manejar distintos tipos de datos y contextos, permitiendo una evaluación comparativa de su efectividad en la detección de anomalías dentro de conjuntos de datos complejos. La implementación y comparación de estas metodologías proporciona una visión sobre las estrategias con mejor desempeño al identificar y gestionar *outliers* en el ámbito del riesgo crediticio. Para evaluar el rendimiento de las técnicas, se consideran métricas como la sensibilidad, el *accuracy* y la especificidad.

El análisis de los resultados obtenidos muestra que *iForest* es la técnica más adecuada para la detección de *outliers* en este tipo de casos. Esta técnica demostró una alta precisión y sensibilidad, identificando de manera efectiva las anomalías sin incurrir en un gran número de falsos positivos. Por otro lado, tanto LOF como PPCA demostraron ser menos efectivas debido a su baja sensibilidad, lo que resultó en la subdetección de outliers. La baja sensibilidad de LOF y PPCA puede atribuirse a su dependencia en la densidad local y en las estructuras subyacentes de los datos, lo que no siempre captura las anomalías en conjuntos de datos altamente variados o de alta dimensionalidad.

Palabras clave: Detección de *Outliers*, Riesgo Crediticio, Gaussian Mixture Models, Análisis de Componentes Principales Probabilístico, Distancia de Mahalanobis, Local Outlier Factor, Isolation Forest, Datos Multivariados

Abstract

Today, society is surrounded by an unprecedented amount of data that, if used properly, can provide valuable information and serve as a competitive advantage in business. In this era of big data and data-driven decision making, the detection of outliers has become an invaluable tool for ensuring the accuracy and integrity of analysis. Specifically, in the field of credit risk, the identification of outliers is crucial as it can reveal undetected risks, prevent fraud and optimise credit policies. This paper focuses on addressing the problem of outlier detection in the context of credit risk, where identifying these outliers allows uncovering hidden risks and preventing fraudulent activities, among other applications.

To this end, several outlier detection techniques have been selected and applied, including Gaussian Mixture Models (GMM), Probabilistic Principal Component Analysis (PPCA), Mahalanobis distance, Local Outlier Factor (LOF) and Isolation Forest (iForest). These techniques have been chosen for their ability to handle different data types and contexts, allowing a comparative evaluation of their effectiveness in detecting anomalies within complex datasets. The implementation and comparison of these methodologies provides insight into the best performing strategies for identifying and managing outliers in the credit risk domain. To evaluate the performance of the techniques, metrics such as sensitivity, accuracy and specificity are considered.

The analysis of the results obtained shows that iForest is the most suitable technique for the detection of outliers in this type of cases. This technique demonstrated high accuracy and sensitivity, effectively identifying abnormalities without incurring a large number of false positives. On the other hand, both LOF and PPCA proved to be less effective due to their low sensitivity, resulting in the under-detection of outliers. The low sensitivity of LOF and PPCA can be attributed to their dependence on the local density and underlying structures of the data, which does not always capture anomalies in highly varied or high-dimensional data sets.

Keywords: Outlier Detection, Credit Risk, Gaussian Mixture Models, Probabilistic Principal Component Analysis, Mahalanobis Distance, Local Outlier Factor, Isolation Forest, Multivariate Data.

Agradecimientos

Me gustaría agradecer principalmente a mi tutora, Jenny Alexandra Cifuentes Quintero, por toda su ayuda, paciencia y apoyo a lo largo de todo el proceso. Sin ella, no hubiese sido posible.

Agradecer también a mis padres y hermanos por su cariño y apoyo cada día.

Índice general

Índice de figuras

Índice de tablas

Acrónimos

AIC	Criterio de Información de Akaike
BIC	Criterio de Información Bayesiano
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DDoS	Distributed Denial of Service
EM	Expectación-Maximización
FA	Análisis Factorial
GMM	Gaussian Mixture Model
INFLO	Influenced Outliers
IQR	Interquartile Range
KNN	K-Nearest Neighbours
LDOF	Local Distance-based Outlier Factor
LOF	Local Outlier Factor
LRD	Local Reachability Density
LVM	Modelo de Variables Latentes
OPTICS	Ordering Points to Identify the Clustering Structure
PCA	Principal Component Analysis
PPCA	Probabilistic Principal Component Analysis
RD	Reachability Distance
ROC	Receiver Operating Characteristic

Capítulo 1

Introducción

1.1. Justificación

En el panorama empresarial del siglo XXI, las decisiones se encuentran inevitablemente ancladas a la robustez y precisión de la información con la que se cuenta. La llegada de la era digital ha permitido a las organizaciones acceder a una cantidad sin precedentes de datos. Este fenómeno, a menudo descrito como el *Big Data*, no solo ha revolucionado la forma en que las empresas operan, sino que también ha redefinido sus estrategias de análisis y toma de decisiones. Sin embargo, este gran volumen de información viene acompañado de sus propios desafíos. Uno de los más notorios, la presencia de datos anómalos o valores atípicos, conocidos como *outliers*. Un *outlier*, en términos sencillos, se puede describir como una observación que se aleja significativamente del comportamiento típico de un conjunto de datos, exhibiendo valores extremos o inusuales (Ver Figura 1.1). Estos valores atípicos pueden influir en la precisión de los análisis estadísticos y el modelado de datos, ya que tienen el potencial de distorsionar las estimaciones generales y llevar a conclusiones erróneas.

Desde un punto de vista más técnico, un *outlier* se define como una observación cuya distancia con respecto a la distribución de los demás datos en el conjunto es considerablemente mayor que la de la mayoría de las observaciones (Muñoz García y Amón Uribe, 2013). Teniendo en cuenta estas definiciones, no existe una regla única para identificar un *outlier*, ya que la definición puede variar según el enfoque y el dominio de aplicación. No obstante, una práctica común en la detección de estos valores atípicos implica considerar como *outliers* aquellas observaciones que exceden en más de 1.5 veces el Rango Intercuartílico (IQR) por encima o por debajo de los cuartiles del conjunto de datos. El IQR, en este sentido, se calcula como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1), proporcionando una medida de dispersión estadística de los datos.

Además de la regla del IQR para la identificación de outliers, existen múltiples estrategias que se pueden adaptar según el tipo de datos y el contexto específico del análisis. Esta flexibilidad en la elección del método es crucial, considerando que diferentes conjuntos de datos

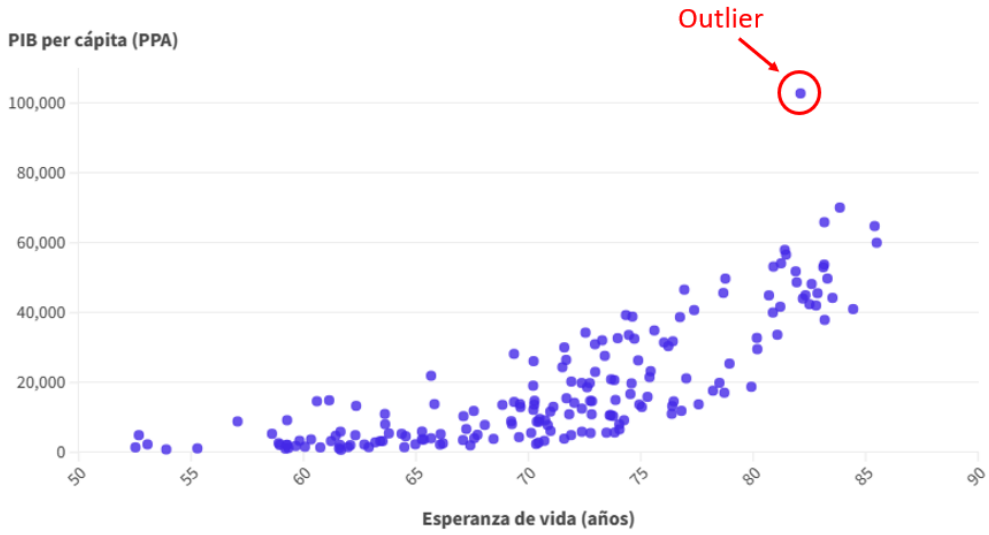


Figura 1.1: Concepto de outlier
Elaboración propia

y escenarios de análisis demandan enfoques personalizados. Por ejemplo, en ciertos escenarios, como en la limpieza y preparación de datos para análisis estadísticos o de aprendizaje automático, identificar y tratar correctamente los *outliers* es un paso crítico para garantizar la calidad y la fiabilidad de los resultados. Por otro lado, en el campo de la detección de fraudes o en el monitoreo de condiciones en maquinaria industrial, la identificación de *outliers* puede ser instrumental para descubrir anomalías significativas o eventos inusuales que requieren atención inmediata. En este sentido, es importante distinguir entre un *outlier* y un dato ruidoso, ya que cada uno tiene un rol y un impacto diferentes en el análisis de datos, y por tanto, requieren estrategias distintas para su manejo. Mientras que un *outlier* es una observación que se desvía significativamente de la norma, aportando a veces información valiosa, un dato ruidoso es simplemente una observación que no añade información relevante al análisis (Smiti, 2020). Estos datos ruidosos pueden distorsionar los resultados, introduciendo variaciones irrelevantes y confusas, y por lo tanto, suelen eliminarse o corregirse para clarificar el análisis.

En el ámbito de una empresa de comercio electrónico que analiza patrones en el comportamiento de compra de sus clientes, la distinción entre un *outlier* y un dato ruidoso adquiere una relevancia particular. Considérese, por ejemplo, un *outlier* manifestado como una compra excepcionalmente grande realizada por un cliente. Esta observación podría señalar un segmento de mercado novedoso o reflejar un cambio emergente en las tendencias de consumo, ofreciendo así pistas valiosas para estrategias de mercado o análisis de comportamiento del consumidor. En contraste, un dato ruidoso en este escenario podría ser un error en la entrada de datos, tal como un precio erróneamente registrado que no corresponde al valor real de una transacción. Mientras que el *outlier* tiene el potencial de revelar oportunidades de mercado

significativas o servir como indicador para investigaciones más detalladas, el dato ruidoso, en cambio, tiende a introducir confusiones y errores en el análisis, como en la interpretación de tendencias de precios o patrones de compra, sin proporcionar insights constructivos.

La generación de datos ruidosos puede obedecer a múltiples factores, tales como la inserción errónea de valores, la inclusión de un tipo de dato incorrecto, o la omisión de ciertos valores. Este fenómeno representa un desafío considerable en el análisis de datos, ya que puede afectar negativamente la capacidad de los algoritmos de aprendizaje automático para interpretar y procesar adecuadamente los datos, comprometiendo así la validez de los resultados obtenidos en el análisis. La identificación y el manejo adecuado de estas dos categorías de datos (*outliers* y datos ruidosos) son esenciales para garantizar la integridad y la precisión de los análisis realizados en diversos campos de estudio y aplicaciones prácticas (Smiti, 2020).

Teniendo en cuenta estas consideraciones previas, el enfoque primordial de este estudio se orientará hacia el análisis de las técnicas de identificación y manejo de *outliers.* Es crucial reconocer que los *outliers* pueden presentarse de múltiples formas dentro de un conjunto de datos, y que su correcta comprensión y tratamiento juegan un rol esencial en la realización de un análisis de datos efectivo. Con el objetivo de proporcionar una mayor claridad en este ámbito, se procederá a detallar las clasificaciones principales de los tipos de *outliers*, cada una de las cuales atiende a distintas características y patrones de aparición en los datos.

- ***Outliers* Univariantes:** Los *outliers* univariantes se destacan cuando se considera una única variable o dimensión del conjunto de datos. Estos valores atípicos se manifiestan en una variable específica y pueden ser observados mediante la comparación de esa variable con el resto de las observaciones. Por ejemplo, si estamos analizando las edades de un grupo de personas, un *outlier* univariado sería una edad significativamente más alta o más baja que el rango típico de edades en el conjunto de datos (Leys, Delacre, Mora, Lakens, y Ley, 2019) (Ver Figura 1.2).

- ***Outliers* Multivariantes:**

  Los *outliers*, en un contexto multivariante, presentan una dinámica particularmente compleja que merece una atención especial. Estas observaciones pueden no ser inmediatamente evidentes cuando se analizan variables de forma aislada, pero emergen claramente al examinar las interacciones entre múltiples variables. Por ejemplo, en un estudio de mercado que incluye variables como el ingreso y el gasto de los consumidores, un *outlier* multivariante podría ser identificado en el caso de un individuo cuyo patrón de gasto no se alinea con su nivel de ingreso. Específicamente, un consumidor con ingresos relativamente bajos pero con gastos desproporcionadamente altos podría constituir un outlier. Este tipo de observación es significativo ya que desafía las expectativas o normas habituales, como la tendencia general de que el gasto de un individuo se correlaciona positivamente con su ingreso (Leys et al., 2019) (Ver Figura 1.3).

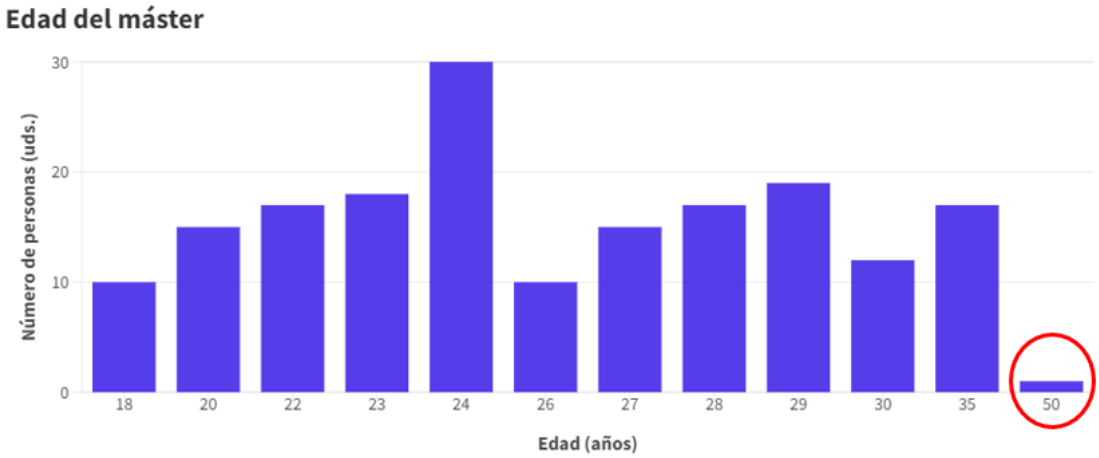


Figura 1.2: Ejemplo Outlier Univariante
Elaboración propia

Figura 1.3: Ejemplo Outlier Multivariante
Elaboración propia

Una clasificación alternativa de outliers ha sido propuesta en la literatura, incorporando las siguientes definiciones (Smiti, 2020):

- ***Outliers* Contextuales**: Los *outliers* contextuales son observaciones que sobresalen de manera significativa dentro de un marco contextual específico, aunque no necesariamente en todos los contextos. La clave de estos valores atípicos reside en su dependencia del entorno o las circunstancias en las que se sitúan. Su naturaleza anómala es, por lo tanto, relativa. Esto implica que mientras en un conjunto de condiciones específicas, una observación puede ser considerada atípica, en otro contexto podría ser completamente normal o incluso esperada. Extendiendo el ejemplo de ingresos mencionado anteriormente, un salario anual de 100,000 dólares puede ser típico en ciudades con

altos costos de vida y economías prósperas, como Nueva York o San Francisco. Sin embargo, en regiones con un costo de vida más bajo, la misma cifra de ingresos podría destacar notablemente, clasificándose como un *outlier* contextual.

- ***Outliers* Globales**: Los *outliers* globales son observaciones que se distinguen notablemente de las normas generales establecidas en un conjunto de datos, impactando de manera significativa en su estructura y análisis global. Estos valores se caracterizan por su desviación extrema de los patrones y comportamientos típicos esperados, influyendo en varias variables de manera simultánea. Su identificación como *outliers* globales proviene de su capacidad para destacar de forma clara en comparación con las tendencias generales observadas en el conjunto de datos. Por ejemplo, en el contexto empresarial, si se considera una empresa que realiza un análisis detallado de sus gastos en publicidad y las ventas asociadas a lo largo de varios años, un *outlier* global podría ser un período específico donde los gastos en publicidad fueron excepcionalmente altos, pero sorprendentemente, este aumento no se correlacionó con un incremento proporcional en las ventas, o viceversa, donde un bajo gasto en publicidad coincidió con un aumento inusual en las ventas (Ver Figura 1.4). Este tipo de *outlier* podría indicar una eficacia anómala de las campañas publicitarias o una disociación entre la inversión publicitaria y su impacto en las ventas, lo cual es crucial para la toma de decisiones estratégicas en marketing y gestión de recursos.

- ***Outliers* Colectivos**: A diferencia de los *outliers* globales, los *outliers* colectivos se refieren a subconjuntos de datos que, en su conjunto, exhiben un comportamiento atípico en comparación con la distribución general del conjunto de datos completo. Lo que los distingue es que, aunque individualmente estas observaciones podrían no ser consideradas como atípicas, es su agrupación y comportamiento colectivo lo que resalta como inusual. Un ejemplo de este concepto podría visualizarse en una empresa que analiza sus ventas a lo largo del tiempo. Si bien las ventas mensuales podrían parecer normales cuando se observan de forma aislada, al agruparlas se podría descubrir que hay períodos específicos donde las ventas son significativamente más altas o más bajas en comparación con la tendencia general. Esto podría ocurrir durante ciertas temporadas o en respuesta a eventos específicos. Como puede evidenciarse en la Figura 1.5, estos períodos representarían *outliers* colectivos, ya que es la agrupación de las ventas en estos meses lo que se aparta del patrón general observado en otros períodos.

Después de haber examinado las diferentes clasificaciones de outliers, resulta esencial analizar las causas subyacentes que conducen a la aparición de estos valores atípicos en los conjuntos de datos. La generación de outliers puede atribuirse a múltiples factores, que varían en función del contexto específico y la naturaleza de los datos. Según (Pelea, 2019), estas causas pueden agruparse en las siguientes categorías principales:

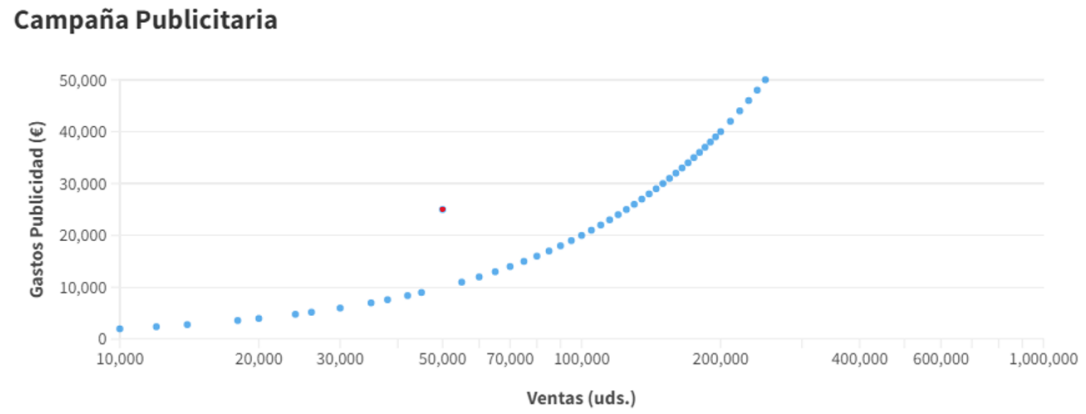


Figura 1.4: Ejemplo Outlier Global
Elaboración propia

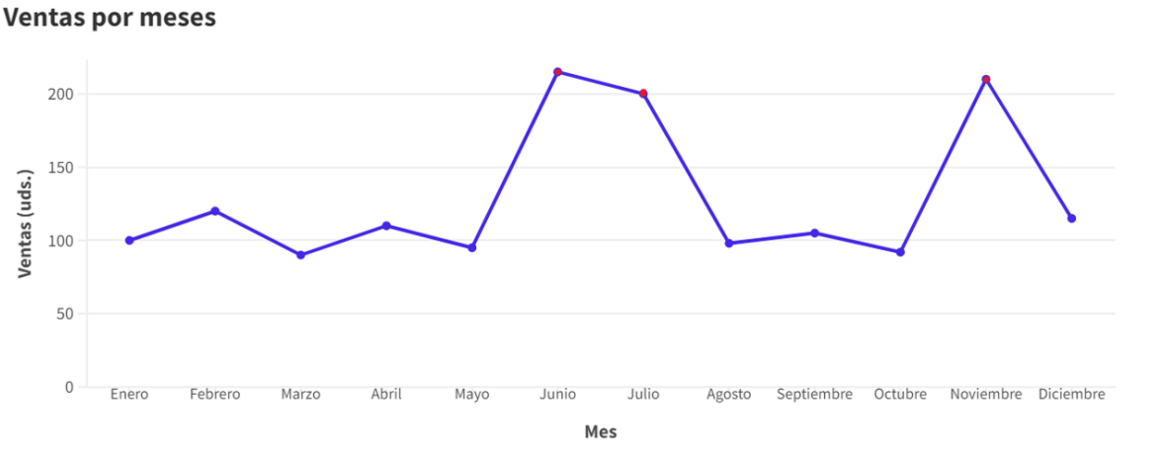


Figura 1.5: Ejemplo Outliers Colectivos
Elaboración propia

- **Errores de Medición:** Estos errores pueden surgir en cualquier etapa del proceso de recopilación o medición de datos. A menudo, son el resultado de fallos en los instrumentos de medición, errores humanos, o deficiencias en el proceso de recolección de datos. La presencia de estos errores introduce valores atípicos que distorsionan la comprensión real del fenómeno en estudio, resultando en observaciones que no reflejan fielmente la realidad.

- **Eventos Inusuales**: En ocasiones, los *outliers* aparecen como consecuencia de circunstancias extraordinarias en el entorno o sistema que está siendo examinado. Estos pueden incluir desastres naturales, fallos técnicos importantes, o incidentes poco comunes. Tales eventos suelen generar observaciones que se desvían significativamente de lo que se considera "normal".

- **Variabilidad Natural**: No todos los *outliers* son indicativos de errores o anomalías. En

determinados contextos, pueden ser simplemente el resultado de la variabilidad inherente dentro de la población o el conjunto de datos estudiados. En estas instancias, los valores atípicos ofrecen una perspectiva importante sobre la diversidad y las características extremas presentes en los datos naturales, aportando información valiosa que puede ser de gran relevancia para comprender la totalidad del fenómeno analizado.

La identificación precisa de la causa subyacente de un *outlier* es fundamental, ya que resalta la importancia de detectar estos valores atípicos de manera eficaz en los conjuntos de datos. De hecho, la detección de *outliers* se considera un paso crítico de la fase de preprocesamiento del análisis de datos, dado que la presencia de estos valores puede influir considerablemente en los resultados del análisis. Además, la identificación y análisis de *outliers* puede llevar a la extracción de nueva información y de conocimientos valiosos. Por estas razones, se han propuesto diversas estrategias para la detección de *outliers*, buscando no solo identificar estos valores, sino también entender su naturaleza y el impacto que pueden tener en la interpretación global de los datos. Entre las categorías de métodos propuestos para la detección de *outliers*, se incluyen (Smiti, 2020):

- **Métodos estadísticos**: Estos métodos consideran como outliers aquellos puntos que se desvían significativamente de la distribución estándar. Se dividen en:
    - Métodos paramétricos: Utilizados cuando se conoce la distribución de los datos. Incluyen técnicas basadas en la distribución gaussiana, como la media-varianza y el diagrama de caja (*box-plot*), así como métodos que se basan en la construcción de modelos de regresión.
    - Métodos no paramétricos: Incluyen técnicas destacadas como aquellas basadas en histogramas, que son ampliamente reconocidas por su facilidad y eficacia en la visualización de datos estadísticos. Asimismo, sobresalen los métodos basados en *Kernels*, los cuales comparan la densidad de cada dato con la de sus vecinos cercanos. Esta comparación de densidades resulta ser una herramienta particularmente efectiva para la identificación de *outliers*, permitiendo una detección precisa basada en las diferencias de densidad local entre los puntos de datos.
- **Métodos basados en la distancia**: Estos métodos representan un enfoque integral en la detección de *outliers*, diferenciándose de los métodos estadísticos tradicionales al enfocarse en la relación espacial entre los puntos de datos. Estos métodos determinan los outliers calculando y analizando las distancias entre los puntos de datos, empleando diversas métricas , como la distancia Euclidiana o de Manhattan. Un ejemplo notable es el método Local Distance-based Outlier Factor (LDOF), el cual identifica el *outlier* basándose en la distancia local entre un punto y sus vecinos más cercanos (Alghushairy, Alsini, Soule, y Ma, 2020). Estos métodos son particularmente útiles en conjuntos de

datos con dimensiones múltiples y complejas, donde la evaluación visual directa no es factible. La efectividad de los métodos basados en la distancia depende en gran medida de la selección adecuada del número de vecinos y de la métrica de distancia.

- **Métodos basados en la densidad**: Son técnicas avanzadas de detección de *outliers* que se centran en la evaluación de la densidad local de los datos. Estas técnicas identifican como *outliers* aquellos puntos cuya densidad local difiere significativamente de la densidad de su entorno inmediato. Entre los enfoques más destacados en esta categoría se encuentran el *Local Outlier Factor* (LOF) y el *Influenced Outlierness* (INFLO). El LOF mide la densidad local de un punto en relación con sus vecinos, permitiendo identificar *outliers* en regiones de densidad variada. INFLO, por su parte, considera no solo la densidad local del punto de interés sino también la densidad de sus vecinos, ofreciendo una perspectiva más amplia del comportamiento atípico.

- **Métodos basados en clústeres**: Son un conjunto de técnicas de detección de *outliers* que se fundamentan en el principio de agrupación de datos. Estos métodos funcionan identificando agrupaciones naturales, o clústeres, dentro de un conjunto de datos, y posteriormente consideran como *outliers* aquellos puntos que no se ajustan claramente a ninguno de estos clústeres. Existe una amplia gama de algoritmos diseñados para implementar esta técnica, entre los que se incluyen K-means, *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) (H. Chen et al., 2020) y *Ordering Points to Identify the Clustering Structure* (OPTICS) (Al Samara, Bennis, Abouaissa, y Lorenz, 2023). Estos métodos son particularmente valiosos ya que no requieren suposiciones previas sobre la distribución de los datos, lo que los hace versátiles y aplicables a una variedad de contextos de datos.

Tras detallar las diversas técnicas de detección de *outliers*, es evidente que cada categoría ofrece enfoques únicos para identificar y manejar valores atípicos, adaptándose a las características específicas de los datos y a los objetivos del análisis (Pelea, 2019). Estas herramientas desempeñan un papel crucial, no solo en la identificación de *outliers*, sino también en su gestión efectiva en una diversidad de contextos y aplicaciones, abarcando desde el ámbito empresarial hasta la investigación científica. No obstante, la detección de *outliers* conlleva desafíos significativos que deben ser abordados cuidadosamente. Entre los retos más significativos en la detección de *outliers*, se encuentran la definición subjetiva de lo que constituye un *outlier*, la dificultad para establecer un comportamiento 'normal' que contemple todas las variaciones posibles, y la frontera a menudo borrosa entre lo que se considera normal y anómalo. Adicionalmente, la presencia de actores malintencionados que intentan camuflar observaciones atípicas como normales introduce una capa adicional de complejidad, especialmente en entornos donde la seguridad y la precisión de los datos son críticas (Chandola y Kumar, 2009).

Por otra parte, en conjuntos de datos con múltiples variables (alta dimensionalidad), la detección de *outliers* se vuelve más compleja. El aumento en la cantidad de variables puede hacer menos clara la distinción entre datos normales y atípicos. Esta dificultad se agrava en campos emergentes o especializados donde la disponibilidad de datos etiquetados para entrenar modelos de aprendizaje automático es limitada, dificultando la identificación automatizada de *outliers*. Otro desafío considerable es la presencia de ruido en los datos, que puede ser confundido con los *outliers*, complicando la distinción entre variaciones genuinas y otro tipo de anomalías (Chandola y Kumar, 2009).

La complejidad y diversidad de los desafíos asociados con la detección de *outliers*, tal como se ha discutido, destacan la relevancia de este trabajo de grado en el ámbito empresarial. En un entorno donde las decisiones estratégicas y operativas dependen cada vez más del análisis preciso de grandes volúmenes de datos, el manejo efectivo de *outliers* se convierte en una tarea fundamental. Este estudio se dedica a abordar estas dificultades, explorando y evaluando críticamente diversas técnicas de detección de *outliers*. De esta manera, al profundizar en la identificación y tratamiento adecuado de los *outliers*, este trabajo no solo busca enriquecer el conocimiento teórico en el análisis de datos, sino también ofrecer soluciones prácticas a los profesionales del sector empresarial.

1.2. Objetivos

1.2.1. Objetivo General

El objetivo general de este estudio consiste en llevar a cabo un análisis detallado y proporcionar una visión integral sobre el estado actual del progreso en las técnicas de detección de *outliers* en el contexto de aplicaciones empresariales. En este proceso, se buscará comprender en profundidad y evaluar de manera crítica los métodos de detección de valores atípicos más ampliamente utilizados y que han demostrado un rendimiento sobresaliente en este campo. La finalidad de esta investigación es proporcionar una visión enriquecida de la forma en la que estas técnicas han evolucionado y se han adaptado para abordar los desafíos específicos del entorno empresarial, destacando tanto sus ventajas como sus limitaciones. Asimismo, se pretende identificar las tendencias emergentes y las áreas que requieren una mayor atención en la detección de *outliers*, contribuyendo de manera significativa al desarrollo continuo de esta área de estudio.

1.2.2. Objetivos específicos

- Proporcionar una descripción detallada de los conceptos fundamentales en la detección de *outliers* y categorizar diversas técnicas según su metodología y enfoque.

- Identificar y evaluar los aspectos positivos y negativos de las técnicas de detección de *outliers* analizadas, ofreciendo un panorama completo de cada técnica y describiendo sus aplicaciones más comunes.
- Identificar los desafíos actuales en el campo de la detección de *outliers* en aplicaciones empresariales y describir posibles direcciones para investigaciones futuras.

1.3. Estructura del Documento

Este Trabajo de Fin de Grado se compone de cinco capítulos que conforman su estructura fundamental. El Capítulo 2 presenta una revisión de la literatura sobre el estado actual en el uso de las técnicas de detección de valores atípicos más destacadas y ampliamente empleadas. Posteriormente, el Capítulo 3 abordará un análisis crítico de estas técnicas, desglosando cada enfoque, identificando sus ventajas y desventajas, y proporcionando una visión detallada de sus posibles aplicaciones en el contexto empresarial. El Capítulo 4 marcará la transición hacia la fase práctica del estudio, donde se analizarán los resultados obtenidos al aplicar estas técnicas en un conjunto de datos empresariales específico. Finalmente, en el último capítulo, se llevará a cabo la conclusión del estudio, destacando los resultados más significativos y señalando las áreas de investigación abiertas en el campo de la detección de datos atípicos.

Capítulo 2

Avances en la Detección de Outliers en el Ámbito Empresarial

La detección eficaz de *outliers* se ha consolidado como un componente fundamental en el análisis de datos dentro del contexto empresarial, adquiriendo una importancia creciente en esta era caracterizada por el auge del big data y las decisiones informadas por datos. En el entorno dinámico de las empresas, donde las decisiones estratégicas a menudo dependen de la interpretación de patrones y tendencias derivados de extensos conjuntos de datos, la identificación de *outliers* se manifiesta como un factor dual, presentando tanto desafíos como oportunidades significativas. El reconocimiento adecuado de estos valores atípicos es necesario no solo para evitar errores en los procesos analíticos y en las decisiones resultantes, sino también para descubrir percepciones valiosas sobre anomalías operativas, tendencias en el comportamiento del consumidor y potenciales nichos de mercado. La habilidad para analizar y responder a la información surgida de la detección de *outliers* se establece como un elemento distintivo en la mejora de la competitividad y la eficiencia empresarial. En este contexto, investigadores de todo el mundo han desarrollado estudios que han contribuido a la implementación de estas estrategias, optimizando así las operaciones, las iniciativas en el *marketing* y la toma de decisiones en un amplio espectro empresarial.

En este contexto, diversas investigaciones han propuesto estrategias para abordar con precisión el proceso de identificación y manejo de *outliers*. Un primer enfoque se centra en los modelos probabilísticos, conocidos por su capacidad para manejar la incertidumbre y variabilidad inherente en los datos. Entre estas estrategias, uno de los métodos más destacados es el *Gaussian Mixture Model* (GMM). Particularmente, este método ha sido aplicado en investigaciones asociadas a la detección de fraudes en empresas de telecomunicaciones. Un ejemplo de ello, es el estudio propuesto por (Yusoff, Mohamed, y Bakar, 2013), en el que se propone un algoritmo basado en GMM que no solo detecta llamadas fraudulentas, sino que también identifica llamadas sospechosas de ser fraudulentas. Este algoritmo es fácil de implementar y tiene un gran potencial para ser extendido a la detección de llamadas

fraudulentas tanto facturadas como salientes, reduciendo así las pérdidas financieras de las compañías de telecomunicaciones. Además, el desarrollo de esta técnica ha ampliado también sus aplicaciones empresariales. En este sentido, esta estrategia ha sido implementada de forma alternativa para simular y detectar ataques a aplicaciones web (Moustafa, Misra, y Slay, 2018), así como en la detección de fraudes con tarjetas de crédito (Zhang, Liu, Li, Yan, y Jiang, 2019), entre otras aplicaciones relevantes.

Otra técnica probabilística muy conocida es el Análisis de Componentes Principales Probabilístico (PPCA) (Tipping y Bishop, 2002), que extiende el PCA tradicional para manejar la incertidumbre en los datos y realizar inferencias probabilísticas sobre los componentes principales. Esta técnica permite no solo la reducción de dimensionalidad, sino también la modelización de la variabilidad de los datos, lo cual es necesario para una identificación más robusta de *outliers* en contextos con alta variabilidad y datos complejos. En un estudio realizado por (Pascoal et al., 2012), se aplicó PPCA en la detección de *outliers* en entornos de red, lo cual es particularmente relevante en sectores donde la seguridad de la información es fundamental, como en la industria financiera o las empresas de tecnología. Basándose en esta técnica, en 2020, (Hussain, Mustafa, Jumani, Baloch, y Saeed, 2020) propusieron un enfoque innovador para detectar fraudes en el consumo de servicios públicos, específicamente en el sector eléctrico. En su estudio, el método propuesto alcanzó una precisión del 94.34 % y una tasa de detección del 92.52 %, superando significativamente a otros métodos evaluados. Este enfoque proporciona una solución efectiva para las empresas de servicios públicos, ayudándolas a prevenir pérdidas financieras y a garantizar la integridad de sus operaciones. Sin embargo, determinar el número óptimo de componentes principales a través de la exploración de datos puede ser un desafío, lo que motiva la búsqueda continua de alternativas y mejoras en este campo.

En cuanto a las técnicas de detección de *outliers* basadas en métodos de distancia, destaca la distancia de Mahalanobis (MAHA), la cual es particularmente adecuada para tareas de detección de *outliers* en conjuntos de datos multivariados compuestos por un solo clúster con forma gaussiana. Este enfoque utiliza parámetros del modelo, incluyendo la media y la matriz de covarianza inversa de los datos, lo que lo hace similar a un modelo de una sola componente GMM con una matriz de covarianza completa. Las observaciones con una gran distancia de Mahalanobis se consideran *outliers*. El estudio realizado por (Hou et al., 2020) analiza las diversas aplicaciones de esta técnica en la detección de intrusiones cibernéticas, fraudes, daños industriales y vigilancia de video. Un ejemplo concreto de su aplicación se encuentra en (Çakmakçı, Kemmerich, Ahmed, y Baykal, 2020), donde se utiliza este método, entre otros, para la detección de ataques DDoS (*Distributed Denial of Service*) en tiempo real. Este enfoque ha demostrado ser altamente efectivo permitiendo una respuesta rápida para proteger los sistemas y servicios afectados. De acuerdo con los resultados presentados, el algoritmo logró detectar el 95 % de los ataques DDoS.

Entre las técnicas de detección de *outliers* más ampliamente implementadas, destacan dos

en particular: el Local Outlier Factor (LOF), basado en el método de vecindad, y el Isolation Forest (iForest), basado en métodos de aislamiento. LOF, introducida por (Breunig, Kriegel, Ng, y Sander, 2000), se centra en las diferencias de densidad entre una observación y sus vecinos más cercanos, permitiendo detectar *outliers* locales comparando la densidad local del dato con la densidad de sus vecinos. En contraste, los *outliers* globales se identifican por su diferencia con respecto a la distribución completa de los datos. *Isolation Forest*, por otro lado, es una técnica que aísla observaciones identificando aquellas que requieren menos particiones para ser aisladas en un espacio de características. Esta técnica es eficiente para grandes conjuntos de datos debido a su capacidad de explorar rápidamente el espacio de datos y detectar *outliers* globales. Sin embargo, cada una de estas técnicas presenta limitaciones. LOF, aunque eficaz para detectar *outliers* locales, tiene una alta complejidad computacional. *Isolation Forest*, aunque eficiente en la detección de outliers globales, puede tener dificultades para identificar outliers locales. Para superar estas limitaciones, se ha propuesto un método de ensamblaje progresivo de dos capas, como se describe en el estudio de (Cheng, Zou, y Dong, 2019). Este método utiliza iForest para explorar rápidamente el conjunto de datos, eliminando datos aparentemente normales y generando un conjunto de candidatos a *outliers*. Posteriormente, se aplica LOF para analizar este conjunto de candidatos con mayor precisión y obtener una detección más precisa de *outliers*. Este enfoque combina la rapidez de iForest con la precisión de LOF, ofreciendo una solución más robusta para la detección de datos atípicos en diversos contextos empresariales.

Específicamente, en el año 2020, (Vijayakumar, Divya, Sarojini, y Sonika, 2020) realizaron un estudio sobre la identificación de fraudes en pagos con tarjeta de crédito, un problema significativo para las instituciones financieras debido al aumento de los pagos con tarjeta. El estudio se centra en identificar las características de comportamiento de transacciones correctas e incorrectas utilizando LOF e iForest. En términos de los resultados, mientras que LOF obtuvo un recall del 2 %, iForest demostró ser significativamente más efectivo bajo las condiciones específicas del estudio con un recall del 29 %, sugiriendo una mayor capacidad para identificar transacciones fraudulentas. Siguiendo esta línea de aplicación empresarial y la combinación de estas técnicas, destaca también el estudio posterior realizado por (Negi, Kumar, Raj, Sahana, y Jain, 2022). Además, recientemente, (Dhulipudi et al., 2024) han llevado a cabo un estudio que busca salvaguardar la eficiencia de los sistemas financieros mediante la detección de transacciones anómalas utilizando el algoritmo de iForest. El estudio describe los procedimientos experimentales para la recopilación de conjuntos de datos, el entrenamiento de modelos y la evaluación del rendimiento utilizando varias métricas, tales como precisión, *recall*, F1-score y AUC-ROC.

En resumen, se han identificado cinco técnicas de detección de *outliers* que representan distintos enfoques y han demostrado ser efectivas en diversos contextos, siendo aplicadas en diferentes ámbitos empresariales y financieros en los estudios mencionados: GMM, PPCA, MAHA, LOF e iForest. Actualmente, la técnica que predomina y ha ganado considerable

popularidad es el iForest, al haber destacado por su capacidad para detectar outliers en conjuntos de datos grandes y de alta dimensionalidad de manera eficiente, superando algunas limitaciones de otras técnicas como el LOF y la distancia Mahalanobis, que pueden volverse computacionalmente costosas en grandes conjuntos de datos. Además, *iForest* es más resistente a valores atípicos globales y puede identificar anomalías en menos pasos, lo que lo hace adecuado para aplicaciones en tiempo real donde la velocidad y la eficiencia son particularmente relevantes. Por otro lado, las técnicas basadas en métodos probabilísticos como GMM y PPCA son de gran utilidad para capturar la estructura subyacente de los datos y detectar *outliers* basados en desviaciones significativas de esta estructura. Sin embargo, estas técnicas pueden ser sensibles a ciertas distribuciones de datos y pueden no ser tan eficientes en conjuntos de datos de alta dimensionalidad. Cada técnica tiene sus propias ventajas y limitaciones, y la elección de la mejor técnica depende del contexto específico de la aplicación y de los requisitos de rendimiento.

En este contexto, la Tabla 2.1 presenta las referencias de las aplicaciones en el ámbito empresarial y financiero mencionadas anteriormente, junto con el objetivo del artículo en el que se presentaron y los resultados concretos de las técnicas de detección implementadas. Al enfocarse en estas cinco técnicas de detección de valores atípicos, el objetivo es ofrecer una descripción detallada y la implementación de un conjunto de herramientas efectivas para la identificación y procesamiento de anomalías en los datos.

Tabla 2.1: Resumen de las aplicaciones de las técnicas de interpretabilidad en el ámbito del riesgo financiero.

Referencia	Técnica de Detección	Objetivo	Resultados
(Moustafa et al., 2018)	GMM	Evaluar el rendimiento del modelo GMM (mediante su variante, Outlier Gaussian Mixture) en la detección de anomalías, con un enfoque específico en su aplicación para detectar ataques web.	Los resultados mostraron que la simulación de datos de ataques web mejoró ligeramente las tasas de detección y falsas alarmas. Esto sugiere que la técnica propuesta es efectiva para detectar tanto ataques web conocidos como desconocidos.
(Zhang et al., 2019)	GMM	Abordar el desequilibrio de clases en conjuntos de datos, mejorando la selección de muestras cerca del borde entre clases mayoritarias y minoritarias, especialmente en la detección de fraudes con tarjetas de crédito	Se demostró una mejora significativa en la clasificación de conjuntos de datos desequilibrados y en la detección de fraudes en transacciones con tarjetas de crédito. Su eficacia se observó en 16 conjuntos de datos públicos, mejorando la tasa de reconocimiento de muestras minoritarias.
(Hussain et al., 2020)	PPCA	Desarrollar un método de detección de fraudes eficiente basado en el PCA para identificar a los consumidores fraudulentos.	El método propuesto logra una precisión del 94.34 %y una tasa de detección del 92.52 %. Además, presenta la menor tasa de falsos positivos y falsos negativos, junto con el mayor puntaje Fscore y tasa de detección de robos.
(Çakmakçı et al., 2020)	MAHA	Desarrollar un esquema de detección de ataques de DDoS que se adapte automáticamente a las tendencias cambiantes del tráfico de red y sea capaz de detectar ataques sofisticados y de día cero, utilizando la distancia de Mahalanobis como métrica principal.	Se demostró su eficacia utilizando el conjunto de datos CICIDS2017, destacando su capacidad para detectar ataques DDoS sin requerir un comportamiento de red normal predefinido o un proceso de reentrenamiento.
(Vijayakumar et al., 2020)	LOF e iForest	Desarrollar un método predictivo para detectar fraudes en pagos con tarjeta de crédito utilizando los algoritmos iForest y LOF.	iForest muestra una precisión del 97 %, demostrando la efectividad del enfoque propuesto para identificar fraudes en pagos con tarjeta de crédito.
(Negi et al., 2022)	LOF e iForest	Desarrollar un modelo de detección de fraudes en tarjetas de crédito utilizando los algoritmos LOF e iForest para obtener la máxima precisión en la detección de transacciones fraudulentas.	El estudio demostró que el algoritmo iForest supera al algoritmo LOF en precisión, con una precisión del 99.74 % frente al 99.65 %. iForest detectó un 30 % más de casos de fraude en comparación con LOF.
(Dhulipudi et al., 2024)	iForest	Evaluar el rendimiento del algoritmo iForest en la detección de fraudes en transacciones con tarjetas de crédito, comparándolo con el algoritmo Random Forest.	El algoritmo iForest logra una alta precisión del 99.79 % en la detección de fraudes en transacciones con tarjetas de crédito. Identifica el 33.67 % de todas las transacciones fraudulentas en el conjunto de datos. Su F1-Score es de 0.3607, demostrando un equilibrio adecuado entre precisión y recall. Además, su coeficiente de correlación de Matthews es de 0.3605, indicando una relación fiable entre las predicciones y las observaciones reales

Capítulo 3

Análisis de las Técnicas de Detección de Outliers en el Ámbito Empresarial

En este capítulo se describirán los fundamentos de las técnicas de detección de *outliers* identificadas en el Capítulo 2, proporcionando también información relevante sobre sus ventajas y limitaciones. Este enfoque permitirá utilizar estas herramientas de manera efectiva para identificar y gestionar anomalías en los datos, mejorando así la precisión de los análisis y la toma de decisiones informadas. A continuación, se explorarán en detalle las siguientes técnicas: Gaussian Mixture Model (GMM), Probabilistic Principal Component Analysis (PPCA), distancia de Mahalanobis (MAHA), Local Outlier Factor (LOF) e Isolation Forest (iForest).

3.1. Gaussian Mixture Model

Un Gaussian Mixture Model (GMM) es un modelo estadístico paramétrico utilizado para representar la densidad de probabilidad de datos continuos. Este modelo se compone de una suma ponderada de distribuciones normales (gaussianas), cada una de las cuales se denomina componente. El GMM es especialmente útil en sistemas biométricos, como el reconocimiento de voz, donde se modelan características espectrales relacionadas con el tracto vocal (Reynolds et al., 2009).

En particular, el GMM se aplica como un método de clustering que adopta un enfoque probabilístico, asumiendo que las observaciones en los datos provienen de una combinación de múltiples distribuciones normales. A diferencia de métodos como *K-means*, que asignan cada observación a un único grupo, el GMM proporciona una distribución de probabilidad para cada observación, indicando su pertenencia a cada uno de los clusters. Esto permite un manejo más flexible y preciso de los datos, especialmente en casos donde los clusters pueden tener formas y tamaños variados.

La función de densidad de probabilidad de un GMM se define como (W. Liu, Cui, Peng,

y Zhong, 2019):

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.1}$$

donde **x** es el vector de datos, K es el número de componentes en la mezcla, π_k es el peso de la k-ésima componente, tal que $\sum_{k=1}^{K} \pi_k = 1$, $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ es la función de densidad de probabilidad de una distribución normal con media $\boldsymbol{\mu}_k$ y matriz de covarianza $\boldsymbol{\Sigma}_k$.

Para ajustar un GMM, es necesario determinar los parámetros asociados a los pesos π_k, la media $\boldsymbol{\mu}_k$ y la matriz de covarianza $\boldsymbol{\Sigma}_k$ para cada componente. Esto generalmente se realiza utilizando el algoritmo de Expectación-Maximización (EM), que optimiza los parámetros del modelo para maximizar la probabilidad conjunta de los datos observados.

El proceso de ajuste del GMM implica los siguientes pasos (CodeEmporium, 2019):

1. **Inicialización**: Establecer valores iniciales para los parámetros π_k, $\boldsymbol{\mu}_k$ y $\boldsymbol{\Sigma}_k$.

2. **Paso E (Expectación)**: Calcular la probabilidad de pertenencia de cada observación a cada componente, dado los parámetros actuales. Esta probabilidad, también conocida como responsabilidad se determina utilizando la fórmula:

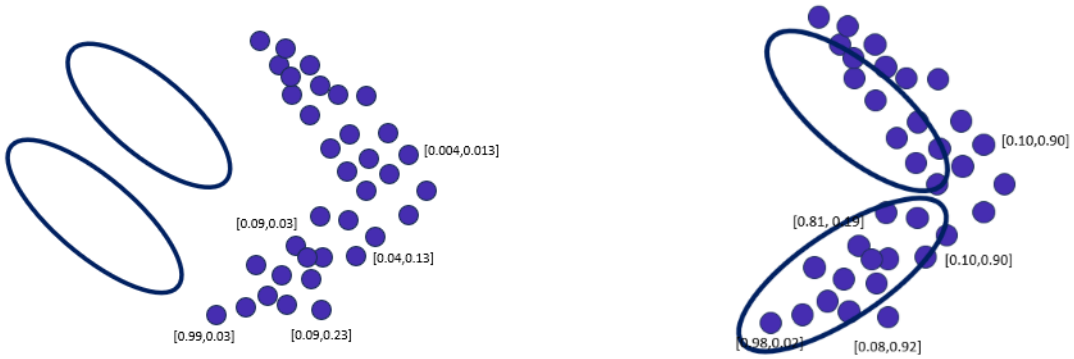
$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{3.2}$$

   En la Figura 3.1a del paso E, se muestra la asignación probabilística inicial de los puntos de datos a los diferentes componentes gaussianos. Los valores en las etiquetas indican las probabilidades de pertenencia de los puntos a cada componente.

3. **Paso M (Maximización)**: Actualizar los parámetros π_k, $\boldsymbol{\mu}_k$ y $\boldsymbol{\Sigma}_k$ para maximizar la probabilidad conjunta. En la Figura 3.1b del paso M, se muestran los nuevos parámetros estimados de las gaussianas, actualizados en función de las probabilidades calculadas en el paso E. Los elipses representan las distribuciones gaussianas ajustadas con los nuevos parámetros.

4. **Iteración**: Repetir los pasos E y M hasta que la convergencia se alcance, es decir, hasta que los cambios en los parámetros sean menores que un umbral predefinido.

El GMM es una de las técnicas de *clustering* más eficaces para la estimación de la densidad de un conjunto de muestras, ampliamente utilizada en el aprendizaje automático. Su flexibilidad permite identificar grupos en los datos y asignar probabilidades de pertenencia a cada observación, lo que es de gran utilidad en diversas aplicaciones prácticas (Wang et al., 2021), (Ghojogh y Toutounchian, 2023). Entre sus ventajas se destaca su capacidad para detectar anomalías al identificar observaciones con densidades de probabilidad muy bajas, clasificándolas como posibles *outliers*. Esta característica es especialmente valiosa en



(a) Paso E (b) Paso M

Figura 3.1: Algoritmo Expectation-Maximization (EM)
Elaboración propia: adaptada de: (Atif Adib, 2023)

áreas como la detección de fraudes y el monitoreo de seguridad, donde es relevante identificar comportamientos inusuales que podrían indicar actividades maliciosas (Chapaneri y Shah, 2021). Además, la capacidad del GMM para proporcionar probabilidades de pertenencia permite una interpretación más detallada de los resultados de *clustering*, en comparación con métodos como *K-means* que asignan observaciones de manera determinista a un único clúster. Según el estudio realizado por (Patel y Kushwaha, 2020), este enfoque resulta en una mayor comprensión de la estructura subyacente de los datos, lo que puede ser aprovechado para mejorar modelos predictivos y estrategias de toma de decisiones.

Sin embargo, el GMM también presenta algunas desventajas (Programmer, 2023). El ajuste del modelo puede ser computacionalmente intensivo, especialmente cuando se trabaja con grandes conjuntos de datos o con un gran número de componentes. El algoritmo EM utilizado para ajustar el GMM puede requerir muchas iteraciones para converger, lo que aumenta el tiempo de cómputo. Además, existe el riesgo de sobreajuste si se seleccionan demasiados componentes, capturando el ruido en lugar de los patrones reales en los datos. Esto puede llevar a modelos que no generalizan bien a nuevos datos, reduciendo la efectividad del GMM en aplicaciones prácticas.

3.2. Análisis de Componentes Principales Probabilístico

El Análisis de Componentes Principales Probabilístico (PPCA) es una extensión del Análisis de Componentes Principales (PCA) tradicional al marco probabilístico, permitiendo manejar la incertidumbre en los datos y realizar inferencias probabilísticas sobre los componentes principales. PPCA es particularmente útil en situaciones donde los datos pueden estar incompletos o contaminados con ruido, proporcionando una metodología más robusta para la reducción de dimensionalidad y la detección de estructuras subyacentes en los datos (Tipping y Bishop, 1999). En PPCA, se asume que los datos observados $\mathbf{X} \in \mathbb{R}^{N \times D}$ (donde N es

el número de observaciones y D es la dimensionalidad de los datos) son generados a partir de un conjunto de variables latentes $\mathbf{Z} \in \mathbb{R}^{N \times d}$ (donde $d < D$) a través de una transformación lineal con la adición de ruido gaussiano. La relación entre las variables observadas y latentes se puede expresar, de acuerdo con (Yu, Yu, Tresp, Kriegel, y Wu, 2006), como :

$$\mathbf{X} = \mathbf{Z}\mathbf{W}^T + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{3.3}$$

donde $\mathbf{W} \in \mathbb{R}^{D \times d}$ es una matriz de pesos, $\boldsymbol{\mu} \in \mathbb{R}^D$ es el vector de medias de las variables observada, y $\boldsymbol{\epsilon}$ es un término de ruido gaussiano con media cero y varianza σ^2. El objetivo de PPCA es encontrar los parámetros $\mathbf{W}$, $\boldsymbol{\mu}$ y σ^2 que mejor expliquen la variabilidad en los datos observados. Esto se realiza maximizando la probabilidad conjunta de los datos observados, lo cual se puede lograr mediante el uso del algoritmo de Expectación-Maximización (EM), similar al GMM (Ver Sección 3.1). Específicamente, la probabilidad de una observación $\mathbf{x}_n$ dada la variable latente $\mathbf{z}_n$ se modela como:

$$p(\mathbf{x}_n|\mathbf{z}_n) = \mathcal{N}(\mathbf{x}_n|\mathbf{W}\mathbf{z}_n + \boldsymbol{\mu}, \sigma^2\mathbf{I}) \tag{3.4}$$

donde $\mathcal{N}(\cdot)$ denota una distribución normal multivariada.

PPCA es particularmente útil para la detección de *outliers* debido a su capacidad para modelar la estructura subyacente de los datos mientras maneja la incertidumbre y el ruido. En un modelo PPCA, los *outliers* pueden identificarse como aquellos puntos de datos que tienen una baja probabilidad de ser generados por las componentes principales estimadas. Estos *outliers* se desvían significativamente de la estructura latente capturada por el modelo, lo que se refleja en sus bajas probabilidades de pertenencia (T. Chen, Martin, y Montague, 2009).

La técnica analizada ofrece dos ventajas significativas, destacadas en el estudio (Akash, 2016). Primero, extiende el alcance del PCA convencional, permitiendo gestionar situaciones con datos faltantes. Segundo, el PPCA puede utilizarse como un modelo de densidad gaussiana general, lo que permite obtener una descripción más precisa de la estructura de datos, ya que las estimaciones de máxima verosimilitud para los parámetros asociados con la matriz de covarianza se calculan de manera eficiente a partir de los componentes principales de los datos. Es especialmente útil en conjuntos de datos con relaciones complejas, ruido y observaciones incompletas.

Sin embargo, PPCA también presenta algunas limitaciones (Zhu, Ge, y Song, 2014). Una de las principales desventajas es su suposición de que los datos siguen una distribución gaussiana, lo cual puede no ser aplicable a todos los conjuntos de datos, especialmente aquellos con distribuciones no normales o con múltiples modos. Otra limitación es que el algoritmo de EM utilizado para la estimación de parámetros puede requerir un número considerable de iteraciones para converger, especialmente en conjuntos de datos de gran tamaño, lo que incrementa el costo computacional. Por último, la selección del número de componentes prin-

cipales no siempre es trivial y puede requerir una exploración exhaustiva o el uso de criterios de selección de modelos adicionales, como el Criterio de Información de Akaike (AIC) o el Criterio de Información Bayesiano (BIC).

3.3. Distancia Mahalanobis

La distancia de Mahalanobis (MAHA) es una técnica efectiva para la detección de *outliers* en datos multivariados introducida por (Mahalanobis, 1930). Esta medida calcula el número de desviaciones estándar en el que se encuentra una observación respecto a la media de una distribución, lo que permite identificar datos que se desvían significativamente del comportamiento esperado. Conceptualmente, la distancia de Mahalanobis ajusta la contribución de cada variable según su variabilidad y las correlaciones existentes entre ellas. Esto la convierte en una métrica más adecuada cuando se trabaja con datos multivariados correlacionados. A diferencia de la distancia euclidiana, que mide la distancia más corta entre dos puntos sin considerar la correlación entre variables, la distancia de Mahalanobis toma en cuenta estas correlaciones, ofreciendo una evaluación más precisa de la "lejanía" de un punto respecto a una distribución.

La fórmula básica para calcular la distancia de Mahalanobis entre un punto $\mathbf{x}$ y la media de la distribución $\boldsymbol{\mu}$ es (De Maesschalck, Jouan-Rimbaud, y Massart, 2000):

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \mu)^T \mathbf{S}^{-1} (\mathbf{x} - \mu)} \tag{3.5}$$

donde $\mathbf{x}$ es el vector de datos, μ es el vector de medias de la distribución, $\mathbf{S}$ es la matriz de covarianza de los datos, y $\mathbf{S}^{-1}$ es la inversa de la matriz de covarianza. Para calcular la distancia de Mahalanobis, es necesario estimar previamente tanto el vector de medias μ como la matriz de covarianza $\mathbf{S}$ a partir de los datos. En la Figura 3.2, se puede observar una comparación entre la distancia euclidiana y la distancia de Mahalanobis. La Figura 3.2a muestra visualmente la forma en la que la distancia euclidiana considera únicamente la distancia lineal entre puntos sin tener en cuenta la forma de la distribución de los datos. En contraste, la Figura 3.2b ilustra la manera en la que la distancia de Mahalanobis ajusta las distancias considerando la variabilidad y las correlaciones de los datos, resultando en una forma elíptica que se adapta mejor a la distribución real de los datos.

En la práctica, esta distancia se utiliza ampliamente en estadísticas multivariadas para identificar *outliers*. Al evaluar la distancia de cada punto respecto a la media multivariada, se pueden detectar observaciones que se alejan significativamente de la estructura principal de los datos. Esta técnica es especialmente útil en conjuntos de datos grandes y complejos, donde la correlación entre variables debe ser considerada para obtener resultados precisos (Ghorbani, 2019). Además, aunque los *boxplots* son ampliamente usados y funcionan bien con datos univariados, para datos multivariados, la distancia de Mahalanobis ofrece una apro-

(a) Euclidiana (b) Mahalanobis

Figura 3.2: Distancia euclidiana vs. Mahalanobis
Elaboración propia: adaptada de: (Leys, Klein, Dominicy, y Ley, 2018)

ximación más precisa. Esta ventaja se hace evidente en aplicaciones donde es fundamental tener en cuenta las interdependencias entre variables para identificar de manera confiable las observaciones atípicas.

Sin embargo, la distancia de Mahalanobis presenta algunas desventajas. Con este enfoque, detectar *outliers* individuales puede ser difícil, y si el número de observaciones en cada clase es menor que el número de dimensiones, se obtiene una matriz de covarianza singular, complicando el análisis de datos ruidosos (Leys et al., 2018). Además, si los clústeres no están bien definidos, la medida de Mahalanobis tampoco lo estará. La presencia de *outliers* extremos también puede influir negativamente en la estimación de la matriz de covarianza, afectando la precisión de la medida. Asimismo, el cálculo de la inversa de la matriz de covarianza puede ser intensiva en recursos para conjuntos de datos de alta dimensionalidad, lo que limita su aplicabilidad en situaciones con grandes volúmenes de datos (Ghorbani, 2019).

3.4. Local Outlier Factor

En el Capítulo 1, se explicó que existen dos tipos de *outliers*: globales y locales. Los *outliers* globales son aquellos puntos que se desvían significativamente cuando se consideran todas las observaciones, mientras que los *outliers* locales se identifican en comparación con los puntos en su vecindad local. En este contexto, los métodos de detección de *outliers* globales asignan un factor de *outlier* a cada registro en función de todo el conjunto de datos, mientras que los métodos locales determinan este factor en relación con los registros vecinos.

El algoritmo Local Outlier Factor (LOF) es un técnica de aprendizaje no supervisado, basada en la vecindad que permite detectar *outliers* locales en conjuntos de datos multidimensionales. A diferencia de los métodos globales, LOF compara la densidad local de un

punto con la densidad local de sus k vecinos más cercanos (Breunig et al., 2000). Esta comparación permite identificar puntos que están en regiones de menor densidad en comparación con sus vecinos, clasificándolos como *outliers* locales.

El algoritmo LOF funciona en varias etapas, que se pueden ilustrar en la Figura 3.3 y se describen a continuación (Paulauskas y Bagdonas, 2015), (Megantara y Ahmad, 2021):

1. **Inicialización**: Se selecciona un valor para k, que representa el número de vecinos considerados para evaluar cada punto. En la Figura 3.3, se ha utilizado $k = 3$.

2. **Distancia de Alcanzabilidad (Reachability Distance)**: Para cada punto A, se calcula la distancia de alcanzabilidad con respecto a sus k vecinos más cercanos, lo que permite determinar la densidad local de puntos de datos cercanos. Si hay muchos otros puntos de datos dentro del área perimetral, se concluye que el punto no es un *outlier*. La distancia de alcanzabilidad de un punto A con respecto a un vecino B (denotada como $RD_k(A, B)$) se define como:

$$RD_k(A, B) = \text{máx}\{d_k(B), d(A, B)\} \tag{3.6}$$

   donde $d_k(B)$ es la distancia al k-ésimo vecino más cercano de B y $d(A, B)$ es la distancia entre A y B.

3. **Densidad de alcanzabilidad local (Local Reachability Density)**: Una vez obtenido el valor de RD, se calcula la Densidad de Alcanzabilidad Local (Local Reachability Density, LRD) con el fin de estimar la relación de distancias para cada vecino más cercano dentro del clúster. La densidad de alcanzabilidad local de un punto A se calcula como la inversa de la media de las distancias de alcanzabilidad de A con respecto a sus k vecinos más cercanos:

$$LRD_k(A) = \left(\frac{\sum_{B \in N_k(A)} RD_k(A, B)}{|N_k(A)|} \right)^{-1} \tag{3.7}$$

   donde $N_k(A)$ es el conjunto de los k vecinos más cercanos de A. Este valor permite comparar la densidad local de un punto con la densidad local de sus vecinos.

4. **Puntuación LOF**: Finalmente, la puntuación LOF de un punto A se calcula comparando la densidad de alcanzabilidad local de A con las densidades de alcanzabilidad local de sus k vecinos más cercanos:

$$LOF_k(A) = \frac{\sum_{B \in N_k(A)} \frac{LRD_k(B)}{LRD_k(A)}}{|N_k(A)|} \tag{3.8}$$

Un valor de $LOF_k(A)$ cercano a 1 indica que el punto tiene una densidad similar a la de sus vecinos, mientras que valores significativamente mayores que 1 indican que el punto es un *outlier* local (Auskalnis, Paulauskas, y Baskys, 2018).

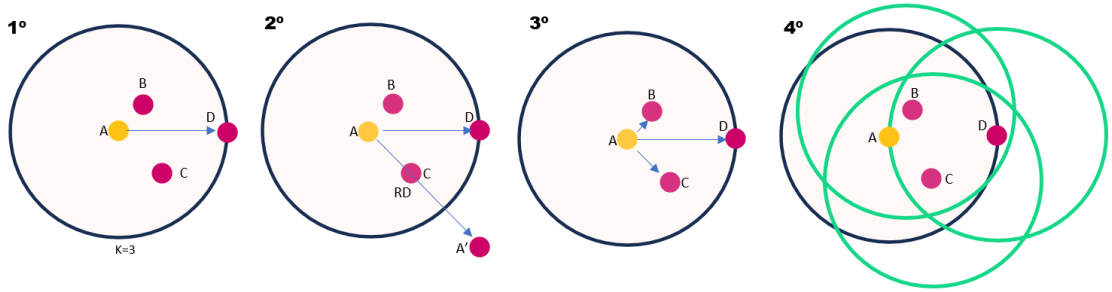


Figura 3.3: Proceso LOF
Elaboración propia: adaptada de: (Megantara y Ahmad, 2021)

El algoritmo LOF tiene varias ventajas (Auskalnis et al., 2018). En primer lugar, se destaca su flexibilidad, ya que no hace suposiciones sobre la distribución de los datos, lo que lo hace adecuado para una amplia variedad de conjuntos de datos. Además, se encuentra su capacidad para identificar outliers locales, lo que permite detectar anomalías en subregiones específicas del conjunto de datos, algo que los métodos globales no pueden hacer. Además, el LOF ajusta las puntuaciones de *outlier* en función de la densidad local, proporcionando una medida más precisa y contextual del comportamiento anómalo. Esta característica es especialmente útil en conjuntos de datos heterogéneos con variaciones locales significativas. También, al comparar la densidad de cada punto con la de sus vecinos, LOF es robusto frente a la variabilidad de densidad, permitiendo identificar outliers en diferentes densidades locales sin ser afectado por las distribuciones globales.

Sin embargo, de acuerdo con el estudio realizado por (Xu y Tian, 2015), es necesario tener en cuenta que el algoritmo también presenta limitaciones. Uno de los principales inconvenientes es su alta complejidad computacional, especialmente cuando se trabaja con grandes conjuntos de datos, debido al cálculo intensivo de distancias entre puntos y la necesidad de determinar múltiples densidades locales. Otra limitación es que este enfoque puede ser sensible al ruido en los datos, lo que puede llevar a falsos positivos si los datos contienen muchas anomalías espurias. Por último, interpretar las puntuaciones LOF puede ser complejo, especialmente en contextos donde la densidad local varía ampliamente.

3.5. Isolation Forest

La técnica *Isolation Forest* fue introducida por (F. T. Liu, Ting, y Zhou, 2008). Este enfoque es una técnica de detección de *outliers* basada en el principio de aislamiento. A

diferencia de otros métodos que se centran en caracterizar el comportamiento normal, *iForest* se enfoca en aislar las observaciones, partiendo de la premisa de que los outliers son "pocos y diferentes", y por lo tanto, más fáciles de aislar. Este enfoque es especialmente eficaz para grandes conjuntos de datos y alta dimensionalidad. Este algoritmo se implementa a través de las siguientes etapas:

- **Inicialización:** : Se selecciona un número t de árboles de aislamiento (*iTrees*) y un número de muestras s a extraer del conjunto de datos para entrenar cada árbol.
- **Construcción de *iTrees*:** Cada árbol se construye de manera recursiva seleccionando al azar una característica y luego seleccionando aleatoriamente un valor de división entre los valores mínimos y máximos de la característica seleccionada. La construcción continúa hasta que cada punto de datos está completamente aislado o se alcanza la profundidad máxima del árbol. La Figura 3.4 ilustra esta estructura, mostrando la forma en la que cada árbol (*iTree*) se construye jerárquicamente. Los nodos representan divisiones aleatorias de las características del conjunto de datos.
- **Cálculo de la Longitud del Camino:** La longitud del camino desde la raíz del árbol hasta una hoja (nodo de terminación), donde una observación es aislada, se utiliza como una medida de anormalidad. Los *outliers* tienden a tener caminos más cortos porque son más fáciles de aislar. En la Figura 3.4, los puntos rojos indican outliers, que se aíslan rápidamente, resultando en caminos cortos dentro de los árboles. Los puntos morados representan puntos de datos normales, que tienden a tener caminos más largos debido a la mayor dificultad para ser aislados.
- **Puntuación de Anomalía**: La puntuación de anomalía para una observación se calcula como la media de las longitudes del camino en todos los árboles, ajustada por un factor de normalización. La puntuación de anomalía $s(x, n)$ para un punto de datos x dado un conjunto de datos de tamaño n se define como:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{3.9}$$

  donde:

  - $E(h(x))$ es la longitud media del camino en los árboles de aislamiento.
  - $c(n)$ es el valor esperado de la longitud del camino para un conjunto de datos de tamaño n.

Cada árbol se construye independientemente utilizando diferentes subconjuntos de datos y divisiones aleatorias, lo que ayuda a capturar diferentes perspectivas de aislamiento. La combinación de múltiples árboles (iForest) mejora la robustez y la precisión de la detección de *outliers*.

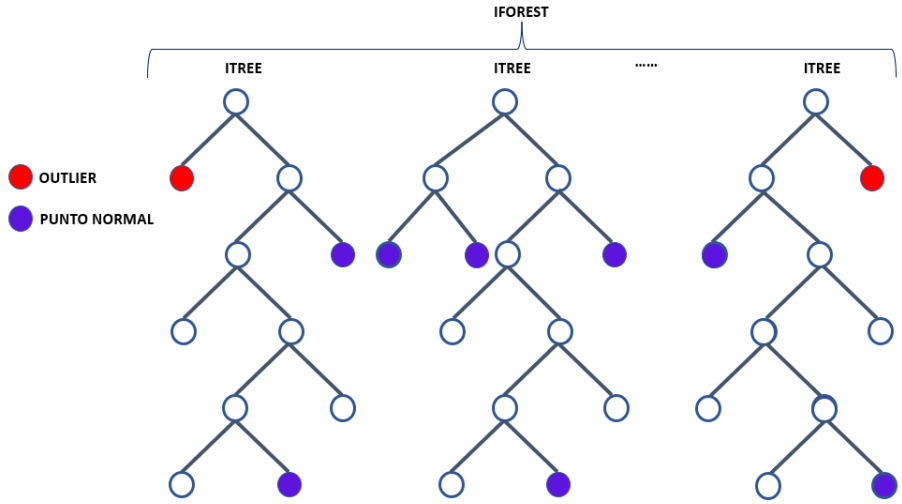


Figura 3.4: Ejemplo Técnica iForest
Elaboración propia: adaptada de: (W.-R. Chen et al., 2016)

Una de las principales ventajas del algoritmo *Isolation Forest* es su eficiencia computacional, ya que no depende de medidas de distancia o densidad para detectar anomalías. En su lugar, el algoritmo calcula la profundidad promedio alcanzada por cada instancia en los árboles del bosque, lo que proporciona una complejidad temporal lineal con una constante baja. Esta característica lo hace altamente eficiente en términos de tiempo de procesamiento y uso de memoria, permitiendo su aplicación en conjuntos de datos grandes y de alta dimensionalidad. Además, el *Isolation Forest* maneja eficazmente atributos irrelevantes, ya que la selección aleatoria de atributos y valores de división ayuda a mitigar su impacto (F. T. Liu et al., 2008).

A pesar de sus numerosas ventajas, el método *Isolation Forest* presenta varias limitaciones. Una de las principales desventajas es su baja precisión en la detección de anomalías condicionales, donde las anomalías solo pueden ser detectadas en relación con ciertas combinaciones de características. Además, aunque el *Isolation Forest* es eficiente para datos de alta dimensionalidad, su precisión puede disminuir en conjuntos de datos específicos con muchas dimensiones, debido a la dificultad para aislar correctamente las observaciones anómalas en estas circunstancias. También se ha observado que el algoritmo puede presentar problemas de sesgo y rendimiento, especialmente cuando el número de árboles no es suficiente para capturar la complejidad de los datos. Estos problemas pueden llevar a falsos positivos en la detección de outliers, comprometiendo la fiabilidad del análisis (Al Farizi, Hidayah, y Rizal, 2021). Asimismo, la dependencia en la selección aleatoria de divisiones puede resultar en variabilidad en los resultados, lo que implica que la consistencia de las detecciones puede variar entre diferentes ejecuciones del algoritmo.

Capítulo 4

Aplicación Práctica de Técnicas de Detección de *Outliers* en un Contexto Empresarial

En este capítulo, se presentará un caso de uso práctico para aplicar las distintas técnicas de detección de outliers analizadas previamente. Se describirá el proceso metodológico seguido, desde la selección y pre-procesamiento del dataset, hasta la implementación y evaluación de cada técnica en un contexto empresarial específico. La finalidad de este capítulo es mostrar la forma en la que estas técnicas pueden ser aplicadas de manera efectiva en un entorno real, proporcionando valor añadido a las empresas mediante la identificación y manejo adecuado de anomalías en los datos.

4.1. Metodología

Para la parte experimental de este trabajo, se aplicarán diversas técnicas de detección de *outliers* a un caso empresarial. El procedimiento seguirá el esquema metodológico descrito a continuación (Ver Figura 4.1):

1. **Selección y Pre-procesamiento del dataset**: Se seleccionará un conjunto de datos relevante para el caso empresarial, asegurando que contenga suficiente cantidad y variabilidad de datos para la aplicación efectiva de las técnicas de detección de *outliers*. Se llevará a cabo la limpieza de datos, eliminando valores nulos, duplicados o erróneos, y se tratarán los valores faltantes. Además, se aplicarán técnicas de normalización de variables y se convertirán las variables categóricas en variables numéricas cuando sea necesario.

2. **Análisis Descriptivo de los datos**: Antes de implementar las técnicas de detección de *outliers*, es necesario comprender en profundidad las características del conjunto

de datos. Este análisis incluye la evaluación de diversos estadísticos descriptivos y la respectiva identificación de patrones relevantes. También se analizan las relaciones entre variables utilizando matrices de correlación para detectar posibles dependencias o asociaciones significativas. Este análisis exploratorio proporciona información relevante, de gran utilidad para la correcta selección e implementación de las técnicas de detección de *outliers*.

3. **Aplicación Técnicas Detección de *Outliers***: Se implementarán las cinco técnicas específicas de detección de *outliers* explicadas en el Capítulo 3: Gaussian Mixture Models (GMM), Probabilistic Principal Component Analysis (PPCA), Mahalanobis Distance (MAHA), Local Outlier Factor (LOF) e Isolation Forest (iForest). Estas técnicas han sido seleccionadas por su capacidad para manejar diferentes tipos de datos y detectar *outliers* en diversos contextos.

4. **Evaluación y Comparación de las Técnicas**: Se evaluarán las técnicas de detección de *outliers* utilizando métricas como especificidad, sensibilidad y *accuracy*. Además de estas métricas, se compararán los resultados obtenidos para determinar la efectividad y eficiencia de cada técnica. Este análisis permitirá identificar el método que proporciona una mejor capacidad de detección de outliers, minimizando tanto los falsos positivos como los falsos negativos.

Durante todo este análisis, se utilizará Python como lenguaje principal debido a su amplia variedad de librerías y paquetes disponibles, lo que facilita significativamente la implementación y el análisis de datos. El código desarrollado está disponible en (Rebollo, 2024).

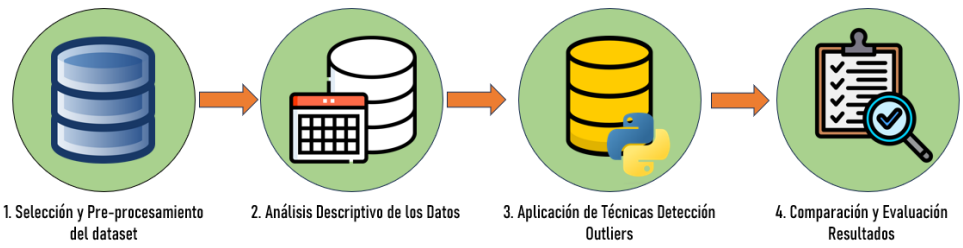


Figura 4.1: Esquema metodológico
Elaboración propia

4.2. Selección, Procesamiento y Descripción del Conjunto de Datos

El riesgo de crédito se define como la probabilidad de que una entidad crediticia experimente pérdidas como resultado del incumplimiento, ya sea total o parcial, de los préstamos

otorgados a sus clientes o deudores durante una transacción financiera o comercial. Es por ello que su detección se considera fundamental para las entidades bancarias y financieras, ya que les permite evaluar el nivel de riesgo asociado a cada solicitud de préstamo. La detección de *outliers* en el campo de la predicción del riesgo de crédito es una herramienta muy interesante para mejorar la calidad de los datos, optimizar los modelos predictivos, y garantizar decisiones de crédito más seguras y justas. Mediante la identificación y tratamiento adecuado de los *outliers*, las instituciones financieras pueden minimizar riesgos, detectar fraudes, y ajustar sus políticas crediticias de manera más efectiva.

En el contexto de este análisis, se ha obtenido un dataset que contiene 1000 observaciones y 20 variables. En el repositorio de (Hofmann, 1994), se proporcionan dos conjuntos de datos: "german.data", que contiene los datos originales tal como fueron adquiridos, incluyendo atributos categóricos y simbólicos; y "german.data-numeric", un archivo editado por la Universidad de Strathclyde, que convierte los atributos en variables numéricas, añadiendo varias variables binarias. Este formato numérico es más adecuado para algoritmos que no pueden manejar variables categóricas, facilitando así su procesamiento y análisis. Inicialmente, se realizará un análisis de variables utilizando el archivo "german.data" para comprender las características de los datos originales e identificar relaciones relevantes dentro de los datos. Posteriormente, para la implementación de diversas técnicas de detección de *outliers*, se ha decidido utilizar el archivo "german.data-numeric". La razón de esta elección es que este archivo ya proporciona las variables tratadas y codificadas en términos numéricos, lo que es necesario para muchos algoritmos de detección de *outliers* que requieren atributos numéricos. Esta decisión asegura que las técnicas aplicadas sean efectivas y que los resultados obtenidos sean relevantes para el análisis.

Cada registro en el conjunto de datos está etiquetado para identificar a los individuos con un riesgo de crédito bueno o malo. Estas observaciones representan a personas que han solicitado un préstamo a una entidad bancaria. El objetivo de este estudio es aplicar diversas técnicas de detección de *outliers* a este conjunto de datos de riesgo de crédito y evaluar cuál estrategia resulta ser la más efectiva. La Tabla 4.1 proporciona una lista de las distintas variables, su tipología y su significado, con el fin de facilitar la comprensión del conjunto de datos analizado y su estructura. Por otra parte, en la Tabla 4.2 se lleva a cabo una descripción de los estadísticos de cada variable, que incluye la media, la desviación estándar, el porcentaje de valores faltantes, el valor máximo y el valor mínimo. Se destacan varios puntos relevantes en el análisis de las variables del conjunto de datos analizado. Primero, todas las variables son numéricas y no presentan valores negativos, lo que es consistente con las expectativas del dominio del problema. Además, no se encontraron valores faltantes, asegurando la integridad de los datos. En particular, las variables 'Loan NurnMonth' y 'CreditHistory' presentan una alta desviación estándar, indicando una gran variabilidad en los datos y la posible presencia de valores atípicos. Sus valores extremos asociados, en particular los valores máximos, se alejan significativamente de la media, sugiriendo que algunos préstamos tienen duraciones

mucho mayores y que algunos historiales crediticios son inusualmente largos.

Finalmente, la Tabla 4.3 muestra las correlaciones entre las distintas variables, con el fin de complementar el análisis descriptivo de las variables. Se observan correlaciones positivas moderadas entre algunas variables, como por ejemplo entre la variable 2 y la variable 4. La correlación moderada de 0.625 indica que existe una relación positiva significativa entre las dos variables: a medida que aumenta el monto del préstamo, tiende a aumentar la duración del préstamo; aunque esta relación no es perfecta y hay otros factores que también pueden influir en la duración de un préstamo. Por último, la mayoría de las variables presentan valores de correlación bajos, cercanos a 0, lo que indica que no existe una relación lineal fuerte entre ellas. Esta baja correlación sugiere que las variables son en gran medida independientes entre sí. En el contexto de detección de *outliers*, esta independencia puede ser ventajosa, ya que permite que las técnicas de detección identifiquen anomalías en las variables individuales sin la interferencia de relaciones complejas entre ellas. Una vez finalizado el análisis descriptivo, se procede a normalizar las diferentes variables con el fin de estandarizar sus escalas. La normalización es necesaria en este estudio porque permite que las técnicas de detección de *outliers* traten cada variable con igual importancia, evitando que las variables con rangos más amplios dominen el análisis.

Teniendo los datos normalizados, la clase anómala se ha seleccionado según las recomendaciones hechas en (Domingues, Filippone, Michiardi, y Zouaoui, 2018). En este contexto, la clase anómala es 'target=2', con una presencia del 30 % de las observaciones. Para modelar los datos anómalos, se han seguido las recomendaciones realizadas por (Emmott, Das, Dietterich, Fern, y Wong, 2016) para el porcentaje de datos anómalos promedio de un dataset de estas características, muestreando 23 observaciones de la clase 2, correspondiente al 3.18 % del total de observaciones.

Nombre de la Variable	**Tipo**	**Descripción**
Status_checking	Cualitativa	Estado de la cuenta corriente: A11 <0 DM, A12: 0 <= ... <200 DM, A13: ... >= 200 DM / asignaciones salariales por al menos 1 año, A14: sin cuenta corriente
Duration_months	Numérica	Duración en meses
Credit_history	Cualitativa	Historial de crédito: A30: sin créditos tomados/-todos los créditos pagados correctamente, A31: todos los créditos en este banco pagados correctamente, A32: créditos existentes pagados correctamente hasta ahora, A33: retraso en el pago en el pasado, A34: cuenta crítica/otros créditos existentes (no en este banco)

Purpose	Cualitativa	Propósito: A40: coche (nuevo), A41: coche (usado), A42: muebles/equipos, A43: radio/televisión, A44: electrodomésticos, A45: reparaciones, A46: educación, A47: (vacaciones - no existe), A48: reciclaje, A49: negocio, A410: otros
Credit_amount	Numérica	Monto del crédito
Savings_account	Cualitativa	Cuenta de ahorros/bonos: A61 <100 DM, A62: 100 <= ... <500 DM, A63: 500 <= ... <1000 DM, A64: ... >= 1000 DM, A65: desconocido/-sin cuenta de ahorros
Present_employment	Cualitativa	Empleo actual desde: A71: desempleado, A72: ... <1 año, A73: 1 <= ... <4 años, A74: 4 <= ... <7 años, A75: ... >= 7 años
Installment_rate	Numérica	Tasa de cuota en porcentaje del ingreso disponible
Personal_status_sex	Cualitativa	Estado personal y sexo: A91: hombre: divorciado/separado, A92: mujer: divorciada/separada/-casada, A93: hombre: soltero, A94: hombre: casado/viudo, A95: mujer: soltera
Other_debtors_guarantors	Cualitativa	Otros deudores/garantes: A101: ninguno, A102: co-solicitante, A103: garante
Present_residence_since	Numérica	Tiempo de residencia actual en años
Property	Cualitativa	Propiedad: A121: bienes raíces, A122: si no A121: acuerdo de ahorro de la sociedad de construcción/seguro de vida, A123: si no A121/A122: coche u otro, no en el atributo 6, A124: desconocido/sin propiedad
Age_years	Numérica	Edad en años
Other_installment_plans	Cualitativa	Otros planes de pago: A141: banco, A142: tiendas, A143: ninguno
Housing	Cualitativa	Vivienda: A151: alquiler, A152: propio, A153: gratis
Number_existing_credits	Numérica	Número de créditos existentes en este banco

Job	Cualitativa	Trabajo: A171: desempleado/no cualificado - no residente, A172: no cualificado - residente, A173: empleado cualificado/oficial, A174: gestión/autoempleado/empleado altamente cualificado/oficial
Number_people_liable	Numérica	Número de personas responsables de proporcionar mantenimiento
Telephone	Cualitativa	Teléfono: A191: ninguno, A192: sí, registrado bajo el nombre del cliente
Foreign_worker	Cualitativa	Trabajador extranjero: A201: sí, A202: no

Tabla 4.1: Descripción de variables

Tabla 4.2: Descriptivos estadísticos de las variables

Variable	Count	Mean	Std	Min	25 %	50 %	75 %	Max
BalanceCheque	1000.0	2.577	1.258	1.0	1.0	2.0	4.0	4.0
Loan NurnMonth	1000.0	20.903	12.059	4.0	12.0	18.0	24.0	72.0
CreditHistory	1000.0	2.545	1.083	0.0	2.0	2.0	4.0	4.0
CreditAmt	1000.0	32.711	28.253	2.0	14.0	23.0	40.0	184.0
SavingsBalance	1000.0	2.105	1.580	1.0	1.0	1.0	3.0	5.0
Mths in PresentEmployment	1000.0	3.384	1.208	1.0	3.0	3.0	5.0	5.0
PersonStatusSex	1000.0	2.682	0.708	1.0	2.0	3.0	3.0	4.0
PresentResidenceSince	1000.0	2.845	1.104	1.0	2.0	3.0	4.0	4.0
Property	1000.0	2.358	1.050	1.0	1.0	2.0	3.0	4.0
AgeInYears	1000.0	35.546	11.375	19.0	27.0	33.0	42.0	75.0
OtherInstallmentPlans	1000.0	2.675	0.706	1.0	3.0	3.0	3.0	3.0
NumExistingCreditsThisBank	1000.0	1.407	0.578	1.0	1.0	1.0	2.0	4.0
NumPplLiablMaint	1000.0	1.155	0.362	1.0	1.0	1.0	1.0	2.0
Telephone	1000.0	1.404	0.491	1.0	1.0	1.0	2.0	2.0
ForeignWorker	1000.0	1.037	0.189	1.0	1.0	1.0	1.0	2.0
Purpose-CarNew	1000.0	0.234	0.424	0.0	0.0	0.0	0.0	1.0
Purpose-CarOld	1000.0	0.103	0.304	0.0	0.0	0.0	0.0	1.0
Otherdebtor-none	1000.0	0.907	0.291	0.0	1.0	1.0	1.0	1.0
Otherdebt-coappl	1000.0	0.041	0.198	0.0	0.0	0.0	0.0	1.0
House-rent	1000.0	0.179	0.384	0.0	0.0	0.0	0.0	1.0
House-owns	1000.0	0.713	0.453	0.0	0.0	1.0	1.0	1.0
Job-unemployed	1000.0	0.022	0.147	0.0	0.0	0.0	0.0	1.0
Jobs-unskilled	1000.0	0.200	0.400	0.0	0.0	0.0	0.0	1.0
Job-skilled	1000.0	0.630	0.483	0.0	0.0	1.0	1.0	1.0
Risk	1000.0	1.300	0.458	1.0	1.0	1.0	2.0	2.0

Tabla 4.3: Correlaciones entre variables

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
1	1.000	-0.072	0.192	-0.043	0.223	0.106	0.043	-0.042	-0.032	0.060	0.047	0.076	-0.014	0.066	-0.027	-0.070	0.064	0.122	-0.051	-0.092	0.129	-0.047	-0.041	0.055	-0.351
2	-0.072	1.000	-0.077	**0.625**	0.048	0.057	0.015	0.034	0.304	-0.036	-0.055	-0.011	-0.024	0.165	-0.138	-0.110	0.145	0.010	0.030	-0.064	-0.075	-0.044	-0.181	0.055	0.215
3	0.192	-0.077	1.000	-0.060	0.039	0.138	0.042	0.063	-0.054	0.147	0.122	0.437	0.012	0.052	0.014	0.042	0.039	0.031	0.008	-0.092	0.101	-0.006	-0.007	0.055	-0.229
4	-0.043	**0.625**	-0.060	1.000	0.065	-0.008	-0.017	0.029	0.312	0.033	-0.045	0.021	0.016	0.277	-0.050	-0.041	0.252	-0.043	0.079	-0.025	-0.117	-0.027	-0.162	-0.093	0.154
5	0.223	0.048	0.039	0.065	1.000	0.121	0.017	0.091	0.019	0.084	0.002	-0.022	0.028	0.087	0.007	-0.002	0.113	0.100	-0.039	-0.008	0.006	0.020	-0.054	0.055	-0.179
6	0.106	0.057	0.138	-0.008	0.121	1.000	0.111	0.245	0.087	0.256	-0.040	0.126	0.097	0.061	-0.027	-0.021	0.039	0.019	-0.037	-0.082	0.008	-0.257	-0.014	0.120	-0.116
7	0.043	0.015	0.042	-0.017	0.017	0.111	1.000	-0.027	-0.007	0.008	-0.037	0.065	0.122	0.027	-0.047	0.015	0.015	-0.047	0.015	-0.100	0.052	-0.029	0.020	0.010	-0.088
8	-0.042	0.034	0.063	0.029	0.091	0.245	-0.027	1.000	0.147	0.266	0.002	0.090	0.043	0.095	-0.054	0.020	0.019	0.021	0.002	0.167	-0.298	-0.035	0.009	-0.001	0.003
9	-0.032	0.304	-0.054	0.312	0.019	0.087	-0.007	0.147	1.000	0.073	-0.090	-0.008	0.012	0.197	-0.132	-0.007	0.173	0.119	0.026	-0.057	-0.308	0.007	0.064	0.036	0.143
10	0.060	-0.036	0.147	0.033	0.084	0.256	0.008	0.266	0.073	1.000	-0.042	0.149	0.118	0.145	-0.006	0.075	0.051	0.031	-0.018	-0.213	0.007	0.060	0.011	-0.148	-0.091
11	0.047	-0.055	0.122	-0.045	0.002	-0.040	-0.037	0.002	-0.090	-0.042	1.000	-0.042	0.023	-0.033	0.003	0.010	0.012	0.048	0.002	0.049	0.002	0.011	-0.053	0.084	-0.110
12	0.076	-0.011	0.437	0.021	-0.022	0.126	0.065	0.090	-0.008	0.149	-0.042	1.000	-0.036	0.001	0.006	0.036	-0.005	0.023	-0.006	0.049	0.041	0.060	-0.010	-0.001	-0.046
13	-0.014	-0.024	0.012	0.016	0.028	0.097	0.122	0.043	0.013	0.118	0.023	-0.036	1.000	0.097	0.065	0.103	0.055	-0.034	-0.033	-0.063	-0.028	-0.007	0.145	-0.107	-0.003
14	0.066	0.165	0.052	0.277	0.087	0.061	0.027	0.095	0.197	0.145	-0.033	0.001	0.097	1.000	-0.049	-0.036	0.137	0.067	-0.016	-0.050	-0.036	-0.040	-0.254	-0.061	-0.036
15	-0.027	-0.138	0.014	-0.050	0.007	-0.027	0.066	-0.054	-0.132	-0.006	0.003	0.006	0.065	-0.049	1.000	0.033	-0.031	-0.017	0.066	0.033	0.019	0.043	0.087	-0.047	-0.082
16	-0.070	-0.110	0.042	-0.041	-0.002	-0.021	0.015	0.020	-0.024	0.075	0.010	0.036	0.103	-0.036	0.033	1.000	-0.187	0.006	0.005	-0.012	-0.010	0.094	0.072	-0.085	0.097
17	0.064	0.145	0.039	0.252	0.113	0.039	0.041	0.019	0.173	0.051	0.012	-0.005	0.055	0.137	-0.031	-0.187	1.000	0.063	-0.053	0.039	-0.141	-0.028	-0.112	-0.033	-0.100
18	0.122	0.010	0.031	-0.004	0.100	0.019	-0.047	0.021	0.119	0.031	0.048	0.023	-0.034	0.067	-0.017	0.006	0.063	1.000	-0.646	-0.048	0.002	0.001	-0.029	-0.010	-0.001
19	-0.051	0.030	0.008	0.079	-0.039	-0.037	0.015	0.002	0.026	-0.018	0.002	-0.006	-0.033	-0.016	0.066	0.005	-0.053	-0.646	1.000	0.048	-0.036	0.038	-0.028	0.065	0.063
20	-0.092	-0.064	-0.103	-0.025	-0.008	-0.082	-0.100	0.167	-0.057	-0.213	0.049	0.033	-0.063	-0.050	0.033	-0.048	0.048	-0.036	0.037	-0.022	0.018	0.092	-0.135	0.006	-0.022
21	0.129	-0.075	0.101	-0.117	0.006	0.008	0.052	-0.298	-0.308	0.007	0.064	-0.040	-0.047	-0.036	0.019	-0.010	-0.141	0.002	-0.036	-0.092	1.000	-0.047	-0.041	0.055	-0.351
22	-0.047	-0.044	-0.006	-0.027	0.020	-0.257	-0.029	-0.035	0.007	0.060	0.011	0.060	-0.007	0.043	0.094	0.072	-0.028	0.001	0.038	0.019	-0.040	1.000	-0.075	-0.196	0.006
23	-0.041	-0.181	-0.007	-0.162	-0.054	-0.014	0.020	0.009	-0.252	0.044	-0.053	-0.010	0.145	-0.254	-0.075	0.036	0.062	-0.075	-0.028	0.014	0.063	-0.074	1.000	-0.652	-0.022
24	0.055	0.055	0.003	-0.093	0.055	0.120	0.010	-0.001	0.036	-0.148	0.084	-0.107	-0.002	-0.061	-0.085	-0.033	-0.010	-0.009	-0.009	0.012	0.013	-0.196	-0.652	1.000	-0.014
25	-0.351	0.215	-0.229	0.154	-0.179	-0.116	-0.088	0.003	0.143	-0.091	-0.110	-0.046	-0.003	-0.036	0.097	0.097	-0.100	-0.001	0.063	0.093	-0.135	0.006	-0.022	-0.014	1.000

4.3. Implementación de la técnica GMM

Teniendo en cuenta las consideraciones descritas previamente, se ha implementado la técnica Gaussian Mixture Model (GMM) utilizando la biblioteca *scikit-learn* en Python. Para ajustar adecuadamente el modelo GMM, es necesario determinar el número de componentes, que representan las diferentes distribuciones gaussianas que el modelo utilizará para capturar la estructura subyacente de los datos. Este parámetro es de gran relevancia en la implementación, ya que un número incorrecto de componentes puede llevar a un sobreajuste o subajuste del modelo. En este estudio, se ha decidido fijar el número de componentes en 1, basándose en investigaciones previas en el campo, que sugieren que este valor es adecuado para detectar anomalías en conjuntos de datos con características similares (Domingues et al., 2018).

La Figura 4.2 muestra la matriz de confusión obtenida al implementar esta técnica. El análisis de la matriz de confusión revela que, aunque el modelo identifica 50 valores como anomalías, lo que representa aproximadamente el 5 % del total de observaciones. Además de las 23 observaciones anómalas, 15 de ellas fueron identificadas como anomalías, mientras que 35 son falsos positivos y 8 son falsos negativos. El modelo muestra una alta *accuracy* del 94 %, lo cual sugiere que la mayoría de las predicciones realizadas por el modelo fueron correctas. Además, la alta especificidad del 95 % indica que el modelo es muy eficiente en identificar correctamente las observaciones que no son anomalías. Sin embargo, la sensibilidad del modelo es relativamente moderada, alcanzando solo el 65.21 %. Esto indica que el modelo no detecta todas las anomalías presentes en los datos, dejando algunas sin identificar (falsos negativos).

4.4. Implementación de la técnica PPCA

En esta sección, se implementará el Análisis de Componentes Principales Probabilístico (PPCA) para la detección de *outliers* en el conjunto de datos de riesgo de crédito. La implementación se realiza utilizando la biblioteca *scikit-learn* en Python. Se aplica el PPCA mediante la función FactorAnalysis, sin especificar el número de componentes, lo que permite al algoritmo determinar automáticamente el número óptimo de factores latentes. Este enfoque facilita la adaptación del modelo a la estructura específica del conjunto de datos, optimizando así su rendimiento. Una vez ajustado el modelo PPCA, se obtienen las matrices de cargas factoriales, que representan las relaciones entre las variables originales y los factores latentes. Estas matrices son necesarias para realizar la reconstrucción de los datos originales a partir de las observaciones transformadas. El siguiente paso consiste en calcular el error de reconstrucción entre los datos originales y los datos reconstruidos. Este error se utiliza para identificar las observaciones que se desvían significativamente de la distribución general, estableciendo un umbral que permite distinguir los *outliers* de las observaciones normales.

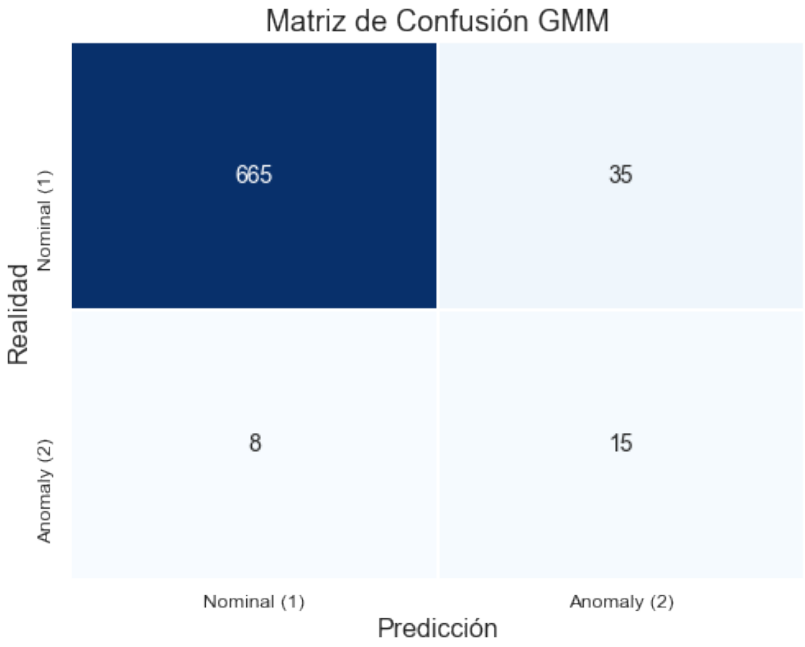


Figura 4.2: GMM Matriz de Confusión
Elaboración propia

En la Figura 4.3 se muestra la visualización de los *outliers* detectados mediante el PPCA. La gráfica muestras varios aspectos importantes: los *outliers* (puntos rojos) están dispersos entre los datos normales (puntos azules), indicando que las anomalías pueden ocurrir en diferentes áreas del espacio de datos. Además, los puntos azules forman ciertos clusters a lo largo de los componentes principales, sugiriendo la presencia de estructuras subyacentes en los datos que PPCA ha intentado capturar. La dispersión de los *outliers* sugiere que las anomalías no están concentradas en una sola región, lo que dificulta su identificación y refuerza la necesidad de un análisis cuidadoso para ajustar el umbral de detección.

Como resultado de la implementación de PPCA, se han detectado 12 *outliers* de manera correcta. Calculada la matriz confusión se observa que hay en realidad 23 *outliers* lo que corresponde al 3.18 % del total de datos de la muestra. Para evaluar el desempeño de la detección, se han calculado varias métricas clave: *accuracy*, sensibilidad y especificidad, obteniendo valores del 93 %, 52 % y 95 % respectivamente. Estos resultados son comparables a los obtenidos con el GMM, aunque presentan pequeñas diferencias en la distribución de las observaciones en la matriz de confusión. En este caso, se han clasificado correctamente 662 observaciones como normales (clase 1) y 12 observaciones como *outliers* (clase 2). Sin embargo, 11 observaciones que eran realmente *outliers* fueron clasificadas incorrectamente como normales, y 38 observaciones que eran normales fueron clasificadas incorrectamente como *outliers*. Esto indica una sensibilidad del 52 %, lo que refleja la proporción de anomalías correctamente identificadas. Esta sugiere una capacidad media del modelo para identificar correctamente las anomalías presentes en el conjunto de datos. Sin embargo, se observan

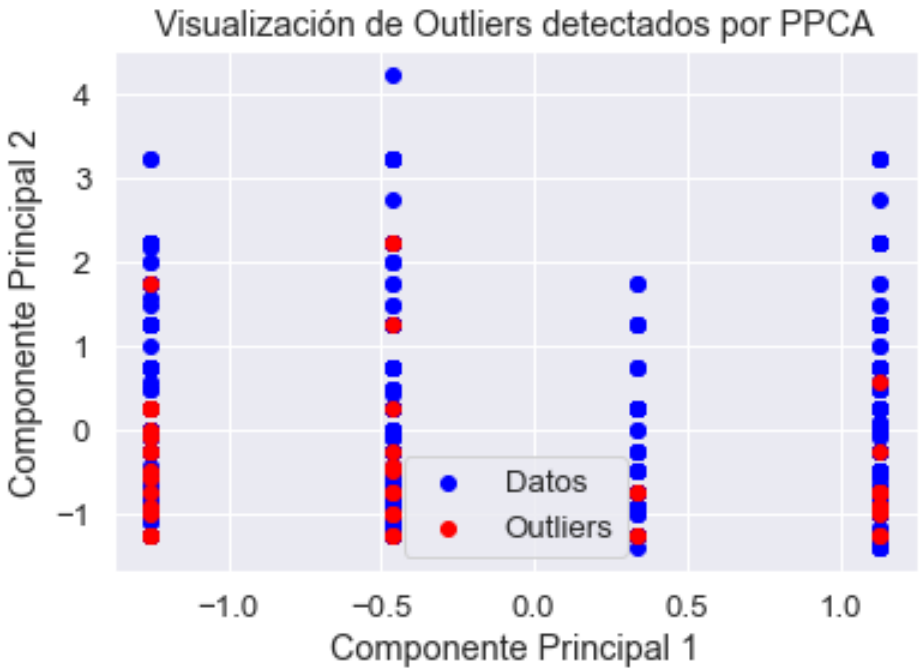


Figura 4.3: PPCA Gráfico de Dispersión
Elaboración propia

limitaciones en la identificación de todas las anomalías, lo que puede atribuirse a la complejidad y variabilidad de los datos, así como a la posible necesidad de ajustar aún más los umbrales de detección.

4.5. Implementación de la técnica basada en la Distancia Mahalanobis

La detección de *outliers* en el conjunto de datos se ha llevado a cabo mediante la aplicación de la técnica de la distancia de Mahalanobis. Este enfoque se fundamenta en la noción de la distancia estadística entre un punto de datos y el centroide del conjunto, teniendo en consideración la estructura de covarianza de los datos. El procedimiento se inicia con el cálculo de la matriz de covarianza del conjunto de datos, la cual proporciona valiosa información sobre la dispersión y las relaciones entre las distintas características presentes. A partir de esta matriz, se deriva la matriz de covarianza inversa, que junto con el centroide del conjunto (calculado como la media de los datos), permite el cálculo de la distancia de Mahalanobis para cada punto en el conjunto. Esta distancia cuantifica la diferencia relativa entre un punto y el centroide, considerando la variabilidad de los datos en todas las direcciones.

Para la identificación de *outliers*, se ha definido un umbral de detección basado en la distribución chi-cuadrado con un nivel de significancia del 5 %. Aquellas distancias de Mahalanobis que superan este umbral son consideradas como *outliers*, indicando puntos que se

desvían significativamente de la distribución general de los datos. El histograma presentado en la Figura 4.4 ilustra la distribución de las distancias de Mahalanobis calculadas para cada punto de datos en el conjunto. Las distancias se muestran en el eje x, mientras que la frecuencia de estas distancias se representa en el eje y. La distribución de las distancias de Mahalanobis se visualiza en azul, mientras que el nivel de significación del 5 % se resalta en rojo. Aquellas distancias que exceden este umbral son identificadas como *outliers*. Se observa que la mayoría de las distancias se concentran en un rango específico, lo que sugiere una agrupación significativa de los datos en torno a su centroide. Sin embargo, algunas distancias superan el umbral de detección, indicando la presencia de *outliers* que se encuentran considerablemente alejados de la distribución general del conjunto de datos.

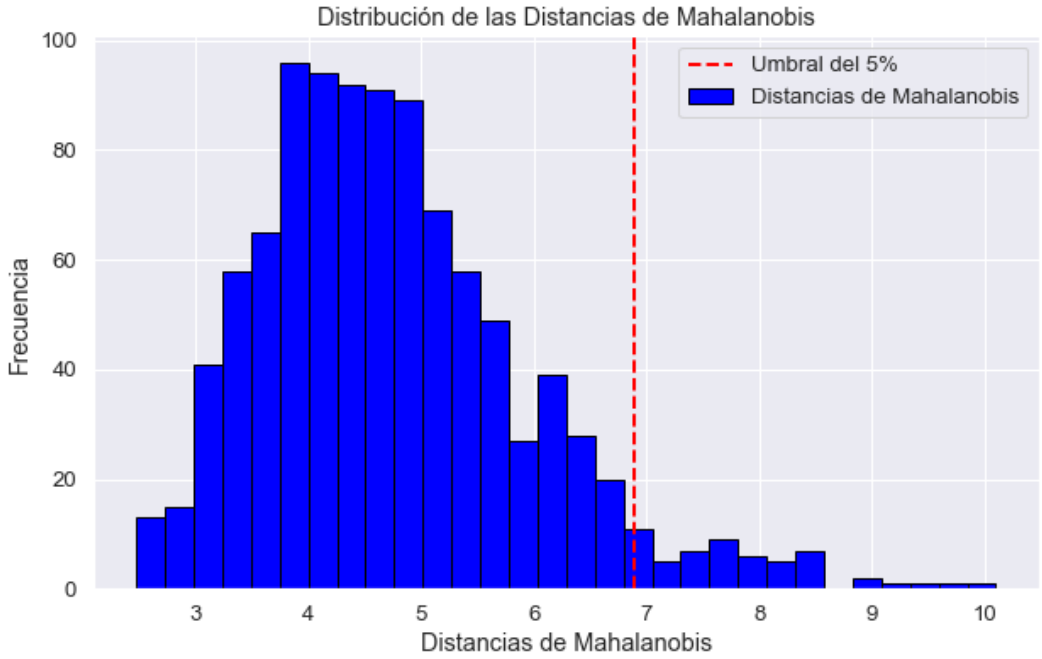


Figura 4.4: Distribución de Distancias Mahalanobis
Elaboración propia

La matriz de confusión muestra la distribución de las predicciones del modelo, donde se observa que de los 50 casos anómalos detectados, 15 fueron clasificados correctamente como *outliers*, mientras que 35 casos fueron erróneamente etiquetados como anómalos. Por otro lado, de los casos que no son *outliers*, 665 fueron correctamente identificados como tales, pero 8 fueron clasificados incorrectamente. En cuanto a las métricas se observa el mismo desempeño que al aplicar la técnica GMM, es decir, una *accuracy* del 94 %, especificidad del 95 % y una sensibilidad del 65.22 %. Esto sugiere que ambas técnicas son igualmente efectivas para la detección de *outliers* en este conjunto de datos, aunque la distancia de Mahalanobis puede tener un menor costo computacional y ser más interpretativa en ciertos contextos.

4.6. Implementación de la técnica LOF

En esta sección se explicará la implementación de la técnica *Local Outlier Factor* (LOF) para detectar *outliers* en el conjunto de datos seleccionado. Este enfoque proporciona una medida de la anormalidad de cada punto de datos en función de la densidad local de su vecindario, comparándola con la densidad local de los vecindarios de los puntos circundantes. El proceso se inicia con la importación de las bibliotecas necesarias, incluyendo 'pandas', 'numpy' y 'sklearn' para el ajuste del modelo LOF, así como para calcular la matriz de confusión y las métricas de evaluación. Adicionalmente, se utilizan las bibliotecas 'seaborn' y 'matplotlib' para la visualización de los resultados. El modelo LOF se ajusta especificando el número de vecinos k que se utilizarán para calcular la densidad local de cada punto (Domingues et al., 2018). La Figura 4.5 muestra la relación entre el número de vecinos k y el número de *outliers* detectados por la técnica LOF. En el eje x, se presentan los diferentes valores de k que se están probando, mientras que en el eje y se indica el número de *outliers* detectados para cada valor de k. Se observa que, a medida que aumenta el número de vecinos, también lo hace el número de *outliers* detectados. Analizando la gráfica, se determina que un valor de k alrededor de 50 es una elección adecuada, ya que en este rango el número de *outliers* detectados muestra un incremento significativo y luego se estabiliza. Esto permite evitar tanto la subdetección como la sobredetección de anomalías. Por ello, utilizaremos este valor de k para la detección de *outliers* en el dataset.

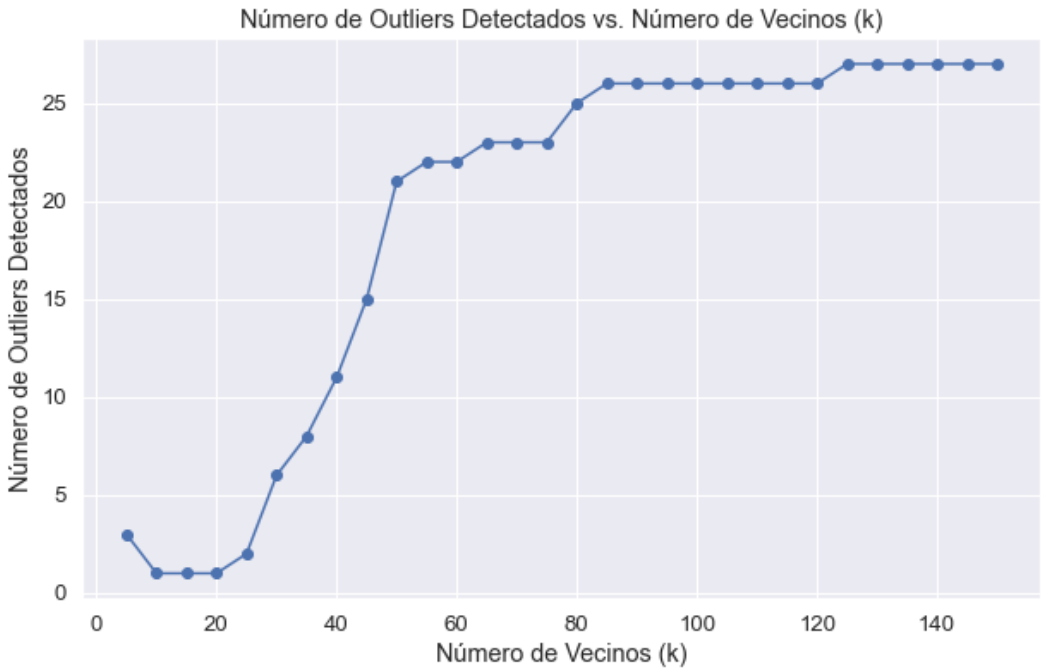


Figura 4.5: Variación del Número de Outliers Detectados en Función de k en el Algoritmo LOF

Elaboración propia

De esta manera, el modelo de detección de *outliers* LOF identificó correctamente solo 8 de los 23 *outliers*. Este resultado sugiere que el modelo tiene limitaciones en su capacidad para identificar eficazmente este tipo de observaciones anómalas. Además es importante destacar que el modelo identificó 29 *outliers* cuando realmente no lo eran y 671 datos como normales que sí lo eran. El desempeño del modelo LOF también se refleja en sus métricas. En términos de *accuracy*, el modelo muestra un rendimiento sólido con un valor del 93.9 %, lo que sugiere que el modelo maneja bien la mayoría de los datos, lo cual es positivo. Sin embargo, al examinar la sensibilidad, observamos que el modelo solo alcanza un 34.8 %, indicando una capacidad limitada para identificar verdaderos *outliers*. En contraste, la especificidad es alta, con un valor del 95.9 %, lo que significa que el modelo es muy efectivo para clasificar correctamente las observaciones normales. La alta especificidad es beneficiosa para minimizar falsos positivos, pero no compensa la deficiencia en la detección de *outliers*, que es el objetivo principal de este tipo de modelos. En conclusión, aunque el modelo LOF demuestra una alta *accuracy* y especificidad, su bajo rendimiento en sensibilidad sugiere que necesita mejoras significativas para ser eficaz en la identificación de *outliers*. Posibles mejoras incluyen la implementación de técnicas complementarias para ajustar los parámetros del modelo, así como realizar un análisis más profundo de las características intrínsecas de los datos.

4.7. Implementación de la técnica iForest

Sobre el conjunto de datos "german.data-numeric", se aplicó la técnica *Isolation Forest* (iForest) para identificar *outliers*. IForest se fundamenta en el principio de aislamiento, identificando observaciones anómalas como aquellas que se encuentran más fácilmente separadas del resto de los datos. Esta metodología destaca por su eficiencia en conjuntos de datos grandes y de alta dimensión gracias a su enfoque basado en árboles de decisión. Para la implementación del modelo, se utilizó la biblioteca *scikit-learn* en Python. El modelo iForest se configuró con 100 estimadores (Ersoy, 2021) y una tasa de contaminación del 5 %, que corresponde a la proporción de datos que se espera sean *outliers*. El proceso comienza con la importación de las bibliotecas necesarias, incluyendo 'pandas' para la manipulación de datos, 'numpy' para operaciones numéricas y 'sklearn' para el ajuste y evaluación del modelo iForest. Una vez configurado el modelo, se procede al entrenamiento utilizando el conjunto de datos completo. IForest crea múltiples árboles de decisión donde cada rama aísla un punto de datos. Los puntos que se aíslan rápidamente, es decir, en menos particiones, son considerados *outliers*.

Tras ajustar el modelo, se identificaron 19 *outliers*, lo que representa aproximadamente el 2.62 % del total de observaciones. La distribución de los puntajes de anomalía, calculados por el modelo iForest para cada punto en el conjunto de datos, se visualiza en la Figura 4.6. En

el eje x, los valores indican el grado de anomalía de cada punto en comparación con el resto del conjunto de datos. Los puntajes más negativos representan mayor anomalía, mientras que aquellos cercanos a cero indican menor anormalidad. Por otro lado, el eje y muestra la frecuencia de los diferentes puntajes de anomalía. La gráfica muestra una distribución asimétrica de los puntajes de anomalía, con una mayor frecuencia de valores alrededor de 0.1 a 0.15, indicando que la mayoría de las observaciones no presentan comportamientos extremadamente anómalos. Sin embargo, la presencia de valores más negativos confirma la detección de unas pocas observaciones que se desvían significativamente de la norma.

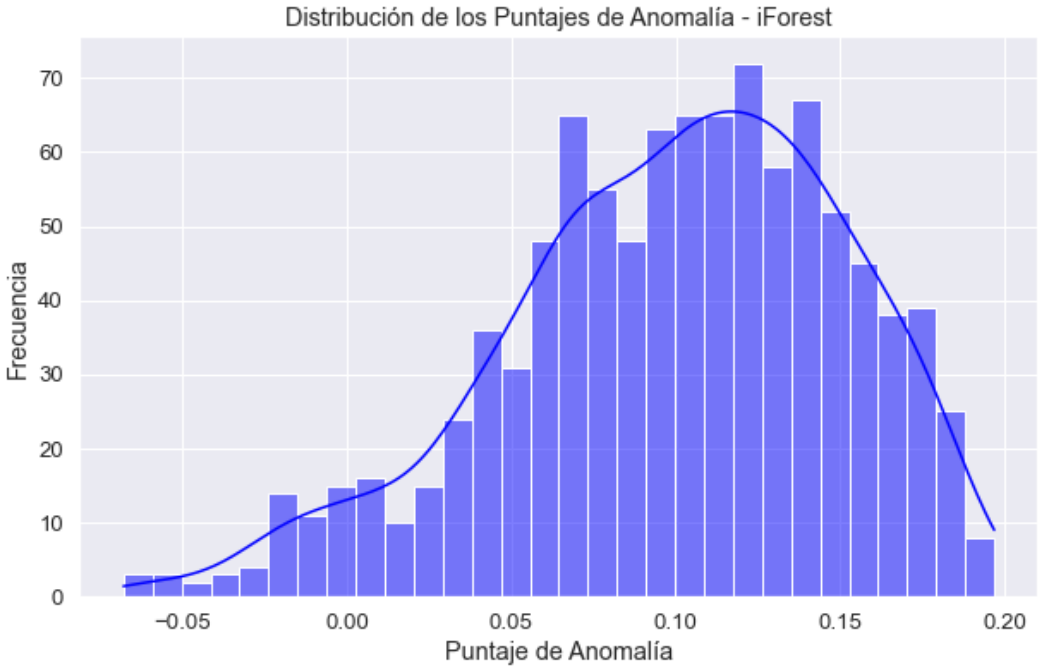


Figura 4.6: Distribución de Puntajes de Anomalía
Elaboración propia

Además, se generó una matriz de confusión donde se muestra que el modelo iForest identificó correctamente 669 observaciones normales (clase 1) y 19 *outliers* (clase 2), pero también clasificó erróneamente 31 observaciones normales como *outliers* y 4 *outliers* como observaciones normales. Posteriormente, se calcularon las métricas para evaluar el desempeño del modelo iForest en la detección de *outliers* y se ha mostrado un rendimiento muy robusto, con altas métricas de *accuracy*, sensibilidad y especificidad. Específicamente, el modelo ha logrado una *accuracy* del 95.16 %, lo que indica que un alto porcentaje de las predicciones totales fueron correctas. La sensibilidad del 82.61 % sugiere que el modelo es bastante eficaz en la identificación de los *outliers*, ya que captura una proporción significativa de las verdaderas observaciones anómalas. Además, la especificidad del 95.57 % muestra que el modelo también es eficaz en la correcta identificación de las observaciones no anómalas, minimizando los falsos positivos. Estos resultados indican que el modelo iForest, en comparación con las técnicas GMM, PPCA, MAHA y LOF, es el más adecuado para la detección

de *outliers*.

4.8. Comparativa de los Resultados en la Detección de *Outliers*

La detección de *outliers* en el conjunto de datos de riesgo crediticio se ha abordado mediante la aplicación de diversas técnicas, cada una con sus propias fortalezas y enfoques. El análisis comparativo muestra tanto similitudes como discrepancias entre los resultados obtenidos por estas técnicas para el conjunto de datos analizado. La figura 4.7, a la que se hará referencia repetidas veces en este apartado, muestra de forma visual una comparativa entre el rendimiento de estas técnicas.

El modelo de Mezcla Gaussiana (GMM), que ofrece flexibilidad en la modelización de la distribución de los datos, obtuvo una sensibilidad del 65.21 %. A pesar de esta mediocre puntuación, la técnica GMM sigue siendo una herramienta valiosa para identificar patrones anómalos en el conjunto de datos. Además, tal y como se puede observar en la figura 4.7, en términos de accuracy y especificidad (representadas por las barras azul y naranja, respectivamente), obtiene unos resultados muy satisfactorios. Por su parte, el Análisis de Componentes Principales Probabilístico (PPCA), al ajustar automáticamente el número óptimo de factores latentes, mostró un desempeño algo más flojo en términos de sensibilidad como bien se puede observar en la figura 4.7, obteniendo una puntuación del 52 %, (métrica representada por la barra de color verde). Esto resalta cierta capacidad del PPCA para capturar patrones anómalos en los datos de riesgo crediticio, aunque de manera menos efectiva que el GMM.

La técnica de distancia de Mahalanobis, basada en la distancia estadística entre los puntos de datos y el centroide del conjunto, obtuvo también una sensibilidad del 65.21 %, igualando el rendimiento del GMM. De hecho, se puede observar que en términos de especificidad y *accuracy*, la figura 4.7 muestra que ambas técnicas son también exactamente igual de efectivas al aplicarse en este campo. Al implementar el algoritmo de Factor de Vecindad Local (LOF), se observa que este considera la densidad local de los vecindarios de los puntos de datos y muestra una sensibilidad variable al ajuste del parámetro k. Con una puntuación de sensibilidad del 34.8 %, el LOF pasa a ser, con diferencia, el que peor rendimiento obtuvo, y es por ello que en la figura 4.7 se muestra el último por ser el método que menos capacidad de detección de valores atípicos correctamente tiene. Finalmente, el *Isolation Forest* obtuvo una puntuación de sensibilidad del 82.61 %, situándose en primer lugar en la figura 4.7 por su capacidad para seleccionar cuidadosamente las observaciones anómalas. Esta técnica es la que mejor rendimiento presentó en la detección de *outliers*.

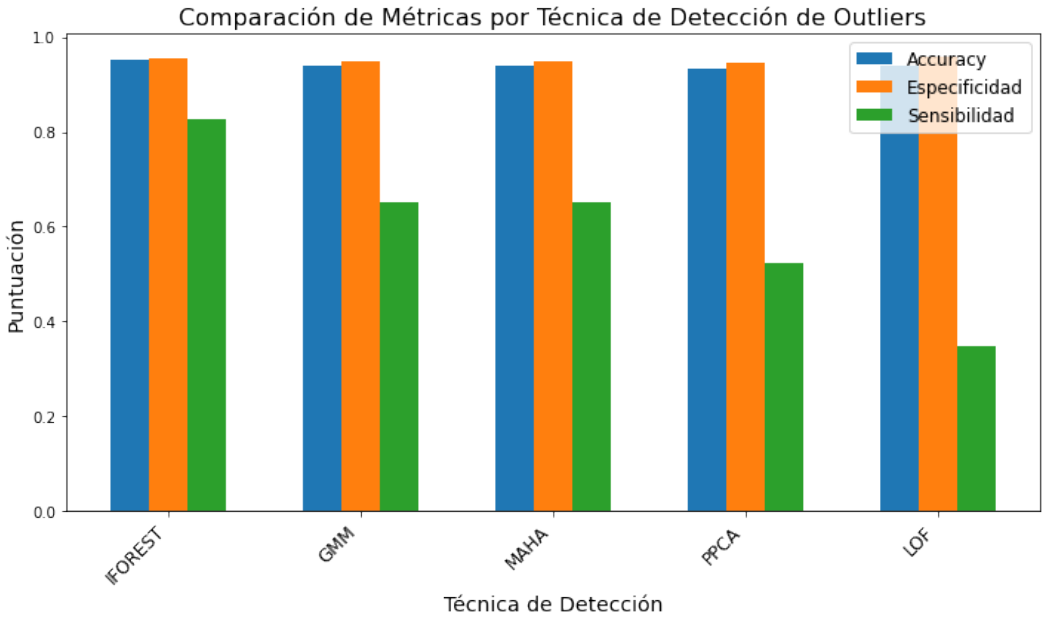


Figura 4.7: Comparativa Rendimiento Técnicas de Detección *Outliers*
Elaboración propia

Capítulo 5

Conclusiones

La detección de *outliers* es una tarea necesaria para asegurar la calidad y fiabilidad de los análisis de datos en entornos empresariales. Con el aumento continuo del volumen de datos, esta práctica ha ganado relevancia con el tiempo, convirtiéndose en una herramienta valiosa para diversas aplicaciones empresariales. Gracias a la identificación de estos valores anómalos, se pueden detectar fraudes, optimizar la gestión de riesgos, mejorar la segmentación de clientes, refinar estrategias de marketing y garantizar la integridad de los datos, entre otros beneficios. La motivación principal de este trabajo ha sido abordar diversas técnicas de detección de *outliers* con el fin de comprender cada una de ellas e identificar las que ofrecen un mejor rendimiento. Para ello, este estudio ha incluido las siguientes técnicas de detección: Gaussian Mixture Model (GMM), Análisis de Componentes Principales Probabilístico (PPCA), Distancia Mahalanobis (MAHA), *Local Outlier Factor* (LOF) y *Isolation Forest* (iForest). Cada técnica presenta ventajas y limitaciones específicas en términos de eficiencia, sensibilidad y capacidad para detectar *outliers* en conjuntos de datos de alta dimensionalidad.

Una vez realizada una descripción de las distintas técnicas, así como un análisis comparativo de sus principales características, se procedió a aplicarlas en un caso de uso. Para ello, se adquirió un dataset de riesgo crediticio específico para clasificar solicitudes de préstamos como “buenas” o “malas”. El análisis y la implementación de las técnicas de detección de outliers se llevaron a cabo utilizando Python, aprovechando sus bibliotecas especializadas en análisis de datos y aprendizaje automático, como ‘numpy’ y ‘scikit-learn’. Durante el estudio, se observó que el análisis exploratorio de datos, incluyendo la evaluación de estadísticos descriptivos y la identificación de patrones relevantes, es una etapa de gran relevancia para entender la naturaleza y complejidad de los datos. Posteriormente, se evaluaron las técnicas de detección de *outliers* utilizando métricas como especificidad, sensibilidad y *accuracy*. La comparativa de resultados permitió determinar la efectividad y eficiencia de cada técnica, minimizando falsos positivos y falsos negativos.

Los resultados obtenidos indican que aunque algunas técnicas muestran un rendimien-

to satisfactorio en la detección de *outliers*, otras no tienen tan buen desempeño. iForest se destaca como la técnica más efectiva, alcanzando una sensibilidad del 82.61 %. Además, es altamente eficiente en términos de tiempo de ejecución y consumo de memoria, lo que la hace ideal para grandes conjuntos de datos y especialmente adecuada para la detección de *outliers* en riesgo crediticio. En contraste, GMM y Mahalanobis Distance muestran exactamente el mismo desempeño con una sensibilidad del 65.21 %. Se podría considerar preferible implementar la Distancia Mahalanobis debido a su capacidad para modelar relaciones entre variables y su robustez frente a correlaciones en los datos. Por otro lado, tanto LOF como PPCA obtuvieron puntuaciones muy bajas de sensibilidad en la identificación de valores anómalos, 34 % y 52 % respectivamente, por lo que no serían recomendables para casos similares al expuesto. Durante el estudio, también se identificaron limitaciones, como la necesidad de ajustar cuidadosamente los hiperparámetros y la dependencia del tamaño y la naturaleza del dataset para obtener resultados óptimos. Por lo tanto, para el caso práctico se seleccionó una muestra de datos que incluyó todos los datos de la clase 1 (nominal) y una muestra de 23 observaciones de la clase 2 (anomalía), asegurando así resultados representativos.

En cuanto a posibles direcciones para investigaciones futuras, se podría explorar la combinación de técnicas supervisadas y no supervisadas para mejorar la precisión en la identificación de *outliers*. Este enfoque híbrido podría aprovechar las fortalezas de ambos tipos de métodos, aumentando la capacidad de detección en diferentes contextos. Además, la aplicación de técnicas de aprendizaje profundo, como las redes neuronales convolucionales y las redes neuronales recurrentes , podría ofrecer nuevas perspectivas y capacidades para manejar datos complejos y de alta dimensionalidad. Otra área prometedora es el desarrollo de algoritmos adaptativos que ajusten automáticamente sus parámetros en función de las características del conjunto de datos en tiempo real. Esto permitiría una detección de *outliers* más dinámica en entornos empresariales cambiantes.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que *ChatGPT* u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Carmen Rebollo Monjo, estudiante de Derecho y Business Analytics (E-3 Analytics) de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Optimizando la Calidad de los Datos en Aplicaciones Empresariales: Estrategias de Detección de Outliers', declaro que he utilizado la herramienta de Inteligencia Artificial Generativa *ChatGPT* u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.

2. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.

3. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

4. Traductor: Para traducir textos de un lenguaje a otro.

5. Solución de errores de código: Para identificar y corregir problemas en código informático programado en Python, proporcionando sugerencias y soluciones.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado *ChatGPT* u otras herramientas similares). Soy consciente de

las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: Junio 2024

Firma: Carmen Rebollo Monjo

Referencias

Akash, A. K. (2016). *Application of probabilistic pca* (Tesis Doctoral no publicada). Indian Institute of Technology Bombay Mumbai 400076, India.

Al Farizi, W. S., Hidayah, I., y Rizal, M. N. (2021). Isolation forest based anomaly detection: A systematic literature review. En *2021 8th international conference on information technology, computer and electrical engineering (icitacee)* (pp. 118–122).

Alghushairy, O., Alsini, R., Soule, T., y Ma, X. (2020). A review of local outlier factor algorithms for outlier detection in big data streams. *Big Data and Cognitive Computing*, *5*(1), 1.

Al Samara, M., Bennis, I., Abouaissa, A., y Lorenz, P. (2023). Complete outlier detection and classification framework for wsns based on optics. *Journal of Network and Computer Applications*, *211*, 103563.

Atif Adib. (2023). *Gaussian Mixture Models Explained | Basics of ML*. Descargado de `https://www.youtube.com/watch?v=hJLaHWaLsyg` (Consultado el 24 de mayo de 2024)

Auskalnis, J., Paulauskas, N., y Baskys, A. (2018). Application of local outlier factor algorithm to detect anomalies in computer network. *Elektronika ir Elektrotechnika*, *24*(3), 96–99.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., y Sander, J. (2000). Lof: identifying density-based local outliers. En *Proceedings of the 2000 acm sigmod international conference on management of data* (pp. 93–104).

Çakmakçı, S. D., Kemmerich, T., Ahmed, T., y Baykal, N. (2020). Online ddos attack detection using mahalanobis distance and kernel-based learning algorithm. *Journal of Network and Computer Applications*, *168*, 102756.

Chandola, V., y Kumar, V. (2009, 01). Outlier detection : A survey. *ACM Computing Surveys*, *41*.

Chapaneri, R., y Shah, S. (2021). Multi-level gaussian mixture modeling for detection of malicious network traffic. *The Journal of Supercomputing*, *77*(5), 4618–4638.

Chen, H., Yu, G., Liu, F., Cai, Z., Liu, A., Chen, S., … Cheang, C. F. (2020). Unsupervised anomaly detection via dbscan for kpis jitters in network managements. *Computers, Materials & Continua*, *62*(2).

Chen, T., Martin, E., y Montague, G. (2009). Robust probabilistic pca with missing data and contribution analysis for outlier detection. *Computational Statistics & Data Analysis*, *53*(10), 3706–3716.

Chen, W.-R., Yun, Y.-H., Wen, M., Lu, H.-M., Zhang, Z.-M., y Liang, Y.-Z. (2016). Representative subset selection and outlier detection via isolation forest. *Analytical methods*, *8*(39), 7225–7231.

Cheng, Z., Zou, C., y Dong, J. (2019). Outlier detection using isolation forest and local outlier factor. En *Proceedings of the conference on research in adaptive and convergent systems* (pp. 161–168).

CodeEmporium. (2019). *Understanding gaussian mixture models.* Descargado de `https://www.youtube.com/watch?v=wT2yLNUfyoM&t=2s` (Fecha de acceso: 24 de mayo de 2024)

De Maesschalck, R., Jouan-Rimbaud, D., y Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, *50*(1), 1–18.

Dhulipudi, S., Siram, S., Pasha, M. J., Yadlapalli, Y., Kadiravan, G., y Subramanyam, M. M. (2024). Optimizing fraud detection in financial transactions: A comprehensive exploration of the effectiveness of random forest and isolation forest algorithms in detecting anomalies within credit card transactions. *Educational Administration: Theory and Practice*, *30*(4), 9146–9157.

Domingues, R., Filippone, M., Michiardi, P., y Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern recognition*, *74*, 406–421.

Emmott, A., Das, S., Dietterich, T., Fern, A., y Wong, W.-K. (2016). Anomaly detection meta-analysis benchmarks.

Ersoy, P. (2021). Evolution of outlier algorithms for anomaly detection. *Manchester Journal of Artificial Intelligence and Applied Sciences*, *2*(1).

Ghojogh, B., y Toutounchian, M. A. (2023). Probabilistic classification by density estimation using gaussian mixture model and masked autoregressive flow. *arXiv preprint arXiv:2310.10843*.

Ghorbani, H. (2019). Mahalanobis distance and its application for detecting multivariate outliers. *Facta Universitatis, Series: Mathematics and Informatics*, 583–595.

Hofmann, H. (1994). *Statlog (German Credit Data).* UCI Machine Learning Repository. (DOI: https://doi.org/10.24432/C5NC77)

Hou, Y., Chen, Z., Wu, M., Foo, C.-S., Li, X., y Shubair, R. M. (2020). Mahalanobis distance based adversarial network for anomaly detection. En *Icassp 2020 - 2020 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 3192-3196). doi: 10.1109/ICASSP40776.2020.9053206

Hussain, S., Mustafa, M. W., Jumani, T. A., Baloch, S. K., y Saeed, M. S. (2020). A novel unsupervised feature-based approach for electricity theft detection using robust pca and

outlier removal clustering algorithm. *International Transactions on Electrical Energy Systems*, *30*(11), e12572.

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., y Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, *32*(1).

Leys, C., Klein, O., Dominicy, Y., y Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the mahalanobis distance. *Journal of experimental social psychology*, *74*, 150–156.

Liu, F. T., Ting, K. M., y Zhou, Z.-H. (2008). Isolation forest. En *2008 eighth ieee international conference on data mining* (pp. 413–422).

Liu, W., Cui, D., Peng, Z., y Zhong, J. (2019). Outlier detection algorithm based on gaussian mixture model. En *2019 ieee international conference on power, intelligent computing and systems (icpics)* (pp. 488–492).

Mahalanobis, P. C. (1930). On test and measures of group divergence: theoretical formulae.

Megantara, A. A., y Ahmad, T. (2021). A hybrid machine learning method for increasing the performance of network intrusion detection systems. *Journal of Big Data*, *8*(1), 142.

Moustafa, N., Misra, G., y Slay, J. (2018). Generalized outlier gaussian mixture technique based on automated association features for simulating and detecting web application attacks. *IEEE Transactions on Sustainable Computing*, *6*(2), 245–256.

Muñoz García, J. A., y Amón Uribe, I. (2013). Técnicas para detección de outliers multivariantes. *Revista en telecomunicaciones e informática*.

Negi, K., Kumar, G. P., Raj, G., Sahana, S., y Jain, V. (2022). Degree of accuracy in credit card fraud detection using local outlier factor and isolation forest algorithm. En *2022 12th international conference on cloud computing, data science engineering (confluence)* (p. 240-245). doi: 10.1109/Confluence52989.2022.9734123

Pascoal, C., de Oliveira, M. R., Valadas, R., Filzmoser, P., Salvador, P., y Pacheco, A. (2012). Robust feature selection and robust pca for internet traffic anomaly detection. En *2012 proceedings ieee infocom* (p. 1755-1763). doi: 10.1109/INFCOM.2012.6195548

Patel, E., y Kushwaha, D. S. (2020). Clustering cloud workloads: K-means vs gaussian mixture model. *Procedia computer science*, *171*, 158–167.

Paulauskas, N., y Bagdonas, A. F. (2015). Local outlier factor use for the network flow anomaly detection. *Security and Communication Networks*, *8*(18), 4203–4212.

Pelea, L. P. (2019). Valores atípicos en los datos,¿ cómo identificarlos y manejarlos? *Revista del Jardín Botánico Nacional*, *40*, 99–107.

Programmer, T. L. (2023). *Gaussian mixture model (gmm)*. Descargado de `https://lazyprogrammer.me/mlcompendium/clustering/gmm.html` (Accessed: 2024-05-29)

Rebollo, C. (2024). *Detección de outliers*. `https://github.com/carmenrebollo/deteccion_outliers`.

Reynolds, D. A., y cols. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*(659-663).

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, *38*, 100306.

Tipping, M. E., y Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *61*(3), 611–622.

Tipping, M. E., y Bishop, C. M. (2002, 01). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *61*(3), 611-622. Descargado de `https://doi.org/10.1111/1467-9868.00196` doi: 10.1111/1467-9868.00196

Vijayakumar, V., Divya, N. S., Sarojini, P., y Sonika, K. (2020). Isolation forest and local outlier factor for credit card fraud detection system. *International Journal of Engineering and Advanced Technology (IJEAT)*, *9*, 261–265.

Wang, R., Zhou, J., Jiang, H., Han, S., Wang, L., Wang, D., y Chen, Y. (2021). A general transfer learning-based gaussian mixture model for clustering. *International Journal of Fuzzy Systems*, *23*(3), 776–793.

Xu, D., y Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of data science*, *2*, 165–193.

Yu, S., Yu, K., Tresp, V., Kriegel, H.-P., y Wu, M. (2006). Supervised probabilistic principal component analysis. En *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (pp. 464–473).

Yusoff, M. I. M., Mohamed, I., y Bakar, M. R. A. (2013). Fraud detection in telecommunication industry using gaussian mixed model. En *2013 international conference on research and innovation in information systems (icriis)* (p. 27-32). doi: 10.1109/ICRIIS.2013.6716681

Zhang, F., Liu, G., Li, Z., Yan, C., y Jiang, C. (2019). Gmm-based undersampling and its application for credit card fraud detection. En *2019 international joint conference on neural networks (ijcnn)* (p. 1-8). doi: 10.1109/IJCNN.2019.8852415

Zhu, J., Ge, Z., y Song, Z. (2014). Robust modeling of mixture probabilistic principal component analysis and process monitoring application. *AIChE journal*, *60*(6), 2143–2157.