UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE
ESCUELA POLITÉCNICA SUPERIOR DE ELCHE
GRADO EN INGENIERÍA DE TECNOLOGÍAS DE TELECOMUNICACIÓN

"Análisis de la relación dinámica entre la comunicación en X y los CDs"

TRABAJO FIN DE GRADO

Septiembre –2024

AUTOR: Miguel Ángel García Martínez
DIRECTOR/ES: Francisco Javier Gimeno Blanes
Margarita Rodríguez Ibáñez

‹Lo que sabemos es una gota de agua;
lo que ignoramos es el océano›
Isaac Newton

‹Nada en la vida debe ser temido, solo comprendido.
Ahora es el momento de comprender más, para que
podamos temer menos›
Marie Curie

Agradecimientos

En primer lugar, quiero expresar mi más profundo agradecimiento a mi familia que siempre ha estado a mi lado en cada paso que he dado a lo largo de mi vida. A mis padres, por su apoyo incondicional, por creer en mí y por inculcarme desde pequeño los valores del esfuerzo, la constancia y la dedicación. A mi hermana, por ser una fuerte fuente de motivación en mi día a día. Sin vosotros, nada de esto habría sido posible.

A mi novia, gracias por tu paciencia, por acompañarme en los momentos difíciles y por ser mi mayor soporte emocional durante este proceso. Has sido una pieza clave para que pudiera culminar este trabajo y te agradezco todo el cariño, comprensión y aliento que me has brindado a lo largo de este camino.

Por último, quiero dedicar unas palabras de agradecimientos a mis tutores y a todos los profesores y compañeros que me han acompañado durante estos años de formación. Gracias por los conocimientos, consejos y enseñanzas pues han sido fundamentales para mi crecimiento personal y profesional y por ello, estoy profundamente agradecido.

Este trabajo es el resultado del esfuerzo colectivo, de la suma de las personas que me rodean y me apoyan, y por ello, a todos y cada uno de vosotros, gracias.

Resumen

En la era de la información, el análisis de datos y la predicción de comportamientos en los mercados financieros se han vuelto herramientas esenciales para inversores y analistas. Este proyecto se enfoca en desarrollar un modelo predictivo que analice la relación entre la actividad en X (Twitter) y los valores de los Credit Default Swaps (CDs) de diversas entidades financieras.

Este estudio se centra en cuatro bancos europeos: Credit Suisse, Deutsche Bank, Commerzbank y Banca Monte dei Paschi di Siena, instituciones que han enfrentado desafíos financieros importantes en los últimos años. El análisis utiliza técnicas avanzadas de Procesamiento del Lenguaje Natural (NLP) y aprendizaje automático (ML) en Matlab, explorando cómo la actividad en redes sociales puede impactar en la percepción del riesgo financiero, reflejada en los Credit Default Swaps (CDS).

Para llevar a cabo esta investigación, se utilizaron datos históricos de los CDs extraídos de la base de datos DataStream con tweets obtenidos mediante la API de antiguamente Twitter y la plataforma Graphext. El periodo de estudio abarca desde enero de 2017 hasta mayo de 2023.

El desarrollo del modelo predictivo sigue varias etapas clave. Primero, se realizó un análisis de sentimiento de los tweets, clasificándolos en positivos, negativos o neutrales, lo que permite cuantificar las emociones expresadas en redes sociales respecto a las empresas analizadas. Posteriormente, se integraron estos datos con variables adicionales como la verificación de las cuentas de los usuarios y el número de seguidores para mejorar la relevancia del análisis. Finalmente, se utilizó información histórica de los CDS para entrenar el modelo y mejorar la precisión de las predicciones.

El uso de Matlab como herramienta central del proyecto facilita el procesamiento de grandes volúmenes de datos y la ejecución de algoritmos complejos de predicción, dada su capacidad para manejar cálculos matemáticos intensivos y su amplia gama de herramientas de análisis. El enfoque metodológico aplicado incluye la tokenización y normalización de los tweets, el análisis de frecuencia e intensidad, y la implementación de modelos predictivos basados en técnicas estadísticas y de aprendizaje automático.

Este proyecto busca proporcionar una base sólida para futuras investigaciones en el ámbito de la correlación entre redes sociales y los Credit Default Swaps, explorando cómo la comunicación en X puede tener un impacto en la percepción del riesgo crediticio de las entidades financieras. Además, el estudio ofrece una herramienta útil para entender la relación entre eventos críticos y las fluctuaciones en los CDS, con el fin de predecir comportamientos anómalos en los mercados financieros.

ABSTRACT

In the information age, data analysis and the prediction of behaviour in financial markets have become essential tools for investors and analysts. This project focuses on developing a predictive model that analyses the relationship between activity on X (Twitter) and the values of Credit Default Swaps (CDs) of various financial institutions.

This study focuses on four European banks: Credit Suisse, Deutsche Bank, Commerzbank and Banca Monte dei Paschi di Siena, institutions that have faced significant financial challenges in recent years. The analysis uses advanced Natural Language Processing (NLP) and machine learning (ML) techniques in Matlab, exploring how social media activity can impact the perception of financial risk, as reflected in Credit Default Swaps (CDS).

To conduct this research, we used historical CD data extracted from the DataStream database with tweets obtained through the former Twitter API and the Graphext platform. The study period spans from January 2017 to May 2023.

The development of the predictive model follows several key steps. First, a sentiment analysis of the tweets was performed, classifying them as positive, negative or neutral, which allows us to quantify the emotions expressed on social networks with respect to the companies analysed. Subsequently, this data was integrated with additional variables such as user account verification and number of followers to improve the relevance of the analysis. Finally, historical information from the CDS was used to train the model and improve the accuracy of the predictions.

The use of Matlab as the central tool of the project facilitates the processing of large volumes of data and the execution of complex prediction algorithms, given its capacity to handle intensive mathematical calculations and its wide range of analysis tools. The methodological approach applied includes tokenisation and normalisation of tweets, frequency and intensity analysis, and the implementation of predictive models based on statistical and machine learning techniques.

This project aims to provide a solid basis for future research in the area of the correlation between social networks and Credit Default Swaps, exploring how communication on X can have an impact on the perception of credit risk of financial institutions. In addition, the study offers a useful tool for understanding the relationship between critical events and CDS fluctuations, in order to predict abnormal behaviour in financial markets.

Índice general

Capítulo 1

1 Introducción

1.1 Contexto

En los últimos años, la relación entre la comunicación en redes sociales y el comportamiento de los instrumentos financieros ha cobrado una relevancia significativa. En particular, el uso de **Credit Default Swaps (CDS)** como indicador del riesgo crediticio ha mostrado ser un área en la que las opiniones vertidas en redes sociales, como **X** (anteriormente Twitter), pueden tener un impacto considerable. Los **CDS** son derivados financieros que permiten a los inversores protegerse frente al posible impago de una deuda, actuando como un seguro frente a la quiebra de una entidad.

Estudios recientes, como el de **Rodríguez-Ibañez (2024)**[1], han demostrado que la actividad en redes, especialmente en momentos críticos, puede influir en las fluctuaciones de los valores de los **CDS**, incrementando el riesgo percibido por los inversores.

Este proyecto se enfoca en analizar la relación entre la actividad en **X** y los valores de los **CDS** de cuatro bancos europeos: **Credit Suisse**, **Deutsche Bank**, **Commerzbank** y **Monte dei Paschi di Siena**. Estas entidades han enfrentado en los últimos años desafíos financieros importantes que han sido objeto de amplias discusiones en plataformas digitales. Para este análisis, se utilizarán datos históricos de los **CDS** obtenidos a través de **DataStream**, una herramienta avanzada de análisis financiero que proporciona datos detallados y exhaustivos sobre los mercados financieros globales. La actividad en **X** se analizará mediante técnicas de **Procesamiento de Lenguaje Natural (NLP)**, clasificando los tweets en categorías de sentimiento (positivo, negativo o neutral) para correlacionar los estados de ánimo públicos con los cambios en los **CDS**.

Este estudio pretende contribuir al creciente cuerpo de investigación que explora cómo las redes sociales pueden influir en los mercados financieros, proporcionando una herramienta que permita a los inversores y analistas prever momentos de alto riesgo crediticio basados en la actividad en **X**.

1.2 Motivación

En un contexto financiero cada vez más volátil, los **Credit Default Swaps (CDS)** se han convertido en un indicador esencial del riesgo crediticio percibido de las entidades financieras. Los **CDS** permiten a los inversores cubrirse frente a un posible impago de deuda, por lo que su valor está directamente relacionado con la percepción de riesgo asociada a una institución. Dado que el valor de los **CDS** es sensible a los cambios en las percepciones de riesgo, es fundamental explorar nuevas fuentes de información que puedan influir en estas percepciones.

Una de estas fuentes es **X**, una plataforma que ha ganado relevancia en los últimos años como un canal de comunicación que puede moldear la opinión pública. A través de la rápida difusión de noticias y opiniones, **X** se ha transformado en un espacio donde la percepción del riesgo de las instituciones financieras puede amplificarse de manera significativa. Estudios recientes, como el de **Rodríguez-Ibañez**[2], han demostrado que los eventos críticos en bancos europeos y la comunicación en **X** sobre estos eventos tienen un impacto directo en el comportamiento de los **CDS**. El análisis de estas dinámicas ofrece un enfoque innovador para la gestión del riesgo crediticio.

Este proyecto busca aprovechar esta dinámica para desarrollar un modelo predictivo que integre el análisis de sentimiento de los tweets con los valores históricos de los **CDS**, con el objetivo de identificar patrones que anticipen incrementos en el riesgo crediticio percibido. La capacidad de prever estos momentos puede ser clave para que los inversores y analistas tomen decisiones informadas en un entorno financiero cada vez más interconectado y dependiente de la percepción pública.

Al centrarse en bancos que han enfrentado situaciones críticas, como **Credit Suisse**, **Monte dei Paschi di Siena**, **Deutsche Bank** y **Commerzbank**, este estudio no solo tiene el potencial de mejorar la comprensión del impacto de los medios sociales en los mercados de **CDS**, sino que también puede ofrecer una herramienta práctica para prever días de alto riesgo crediticio, lo que es crucial para minimizar riesgos y optimizar estrategias de inversión.

1.3 Objetivos e Hipótesis

Objetivos

1. **Desarrollar un modelo predictivo basado en el análisis de redes sociales:** El objetivo principal del proyecto es diseñar un modelo que permita predecir los movimientos en los valores de los Credit Default Swaps (CDS), utilizando para ello el análisis de sentimiento de los tweets en X y los datos históricos de los CDS.

2. **Identificar días de riesgo elevado para las entidades financieras:** Utilizando los datos históricos de CDS y los patrones de actividad en X, se identificarán días clave en los que el riesgo crediticio percibido se incrementa significativamente. Este objetivo se logrará al correlacionar los picos de actividad en redes (especialmente los tweets con sentimiento

negativo) con los aumentos en los CDS, permitiendo anticipar posibles fluctuaciones o crisis.

3. **Evaluar el impacto de la comunicación en redes sociales sobre los valores de los CDS:** Se medirá el impacto de la actividad en X en los movimientos de los CDS utilizando análisis de correlación y modelos predictivos. Este objetivo evaluará en qué medida los tweets, especialmente aquellos con sentimiento negativo o emitidos por usuarios influyentes, amplifican el riesgo percibido y, por ende, los valores de los CDS.

4. **Integrar variables adicionales en el análisis de sentimiento:** Con el fin de aumentar la precisión del modelo predictivo, se incorporarán variables como la verificación de cuentas, el número de seguidores, y la repercusión (retweets y likes) de los tweets. Estas variables permitirán entender mejor el impacto de los actores más influyentes en la red social sobre los valores de los CDS.

5. **Desarrollar una herramienta práctica para la toma de decisiones en gestión de riesgos:** Finalmente, el objetivo es proporcionar una herramienta útil que permita anticipar aumentos en el riesgo crediticio de las entidades estudiadas.

Hipótesis

- **Hipótesis 1:** Los tweets con sentimiento negativo están correlacionados con aumentos en los valores de los CDS, lo que sugiere un incremento en la percepción de riesgo crediticio. Esta hipótesis será evaluada mediante un análisis de correlación entre los resultados del análisis de sentimiento en plataformas digitales y las fluctuaciones históricas de los CDS de las entidades financieras.

- **Hipótesis 2:** Los tweets publicados por usuarios verificados o con una gran cantidad de seguidores tienen un impacto más significativo en los valores de los CDS que aquellos publicados por usuarios con menor influencia en X.

- **Hipótesis 3:** Los eventos críticos de los bancos estudiados, que generan picos de actividad en X, están correlacionados con aumentos en la volatilidad de los CDS, lo que sugiere un incremento en la percepción de riesgo crediticio. Para probar esta hipótesis, se identificarán momentos clave en la historia reciente de las entidades, se analizarán los picos de actividad en redes sociales y se compararán con los aumentos en la volatilidad de los CDS utilizando modelos de análisis predictivo.

Capítulo 2

2 Estado del arte

El análisis de grandes volúmenes de datos ha adquirido un rol central en la comprensión de los mercados financieros modernos. Las plataformas digitales generan cantidades masivas de datos que, cuando se procesan adecuadamente, permiten obtener información valiosa sobre el comportamiento del mercado. Este auge ha fomentado la convergencia de disciplinas como el **análisis de sentimiento**, el **aprendizaje automático (machine learning)** y las **finanzas**, ofreciendo nuevos enfoques para predecir y comprender las dinámicas de los mercados financieros, incluidos los **Credit Default Swaps (CDS)**. En particular, el análisis de redes sociales como **X** (anteriormente Twitter) ha demostrado ser un recurso útil para captar, en tiempo real, las reacciones de los inversores y otros actores financieros ante eventos críticos, impactando de forma directa los valores de activos financieros.

2.1 Estudios Clásicos en el Análisis de Sentimiento Financiero

El uso de las redes para predecir movimientos financieros ha sido explorado desde hace más de una década. Uno de los estudios más influyentes en este campo fue realizado por **Bollen, Mao y Zeng (2011)**[3], quienes analizaron los sentimientos expresados en **X** para predecir cambios en el índice **Dow Jones Industrial Average (DJIA)**. Su investigación demostró que el análisis de las emociones reflejadas en los tweets puede proporcionar pistas sobre la dirección del mercado bursátil. Los resultados de su estudio mostraron que las emociones públicas, como el optimismo o el miedo, tenían una correlación significativa con los movimientos del mercado. Este estudio fue pionero al establecer que el análisis de sentimiento en redes sociales podía ser una herramienta valiosa para anticipar comportamientos financieros.

Poco después, **Mittal y Goel (2012)**[4] continuaron esta línea de investigación utilizando técnicas de **machine learning** para desarrollar un modelo predictivo de precios de acciones basándose en los sentimientos extraídos de **X**. Su estudio clasificó los tweets en tres categorías: positivos, negativos y neutrales, y correlacionó estas clasificaciones con los movimientos de las acciones. El estudio reveló que la inclusión del análisis de sentimiento en los modelos predictivos mejoraba significativamente la precisión de las predicciones de precios de acciones. Esto reforzó la idea de que el análisis de medios sociales, más allá de ser solo un indicador secundario, puede actuar como una fuente directa de información relevante para predecir movimientos del mercado.

Estos trabajos sentaron las bases para la aplicación del análisis de sentimiento en el ámbito financiero, demostrando que las emociones expresadas en las redes sociales pueden ser un indicador clave del comportamiento de los mercados. Aunque estos estudios se centraron principalmente en el mercado bursátil, sus métodos y hallazgos crearon un marco sólido para explorar el impacto del análisis de sentimiento en otros instrumentos financieros[5], como los **Credit Default Swaps (CDS)**.

2.2 Integración de Datos Sociales y Financieros

La combinación de datos sociales y financieros ha demostrado ser una estrategia poderosa para mejorar la precisión en la predicción de comportamientos del mercado. Los primeros estudios en esta área se centraron en el análisis exclusivo de datos financieros tradicionales, como precios históricos, índices bursátiles o informes económicos. Sin embargo, la creciente disponibilidad de datos provenientes de las redes sociales, especialmente **X,** ha permitido a los investigadores integrar estas dos fuentes de información para obtener predicciones más precisas y relevantes.

Un ejemplo destacado es el estudio de **Oliveira, Cortez y Areal (2017)**[6], que demostró que la inclusión de datos sociales en modelos predictivos financieros mejoraba significativamente su rendimiento. Su investigación combinó datos de redes sociales con datos financieros tradicionales para predecir movimientos en los mercados de acciones. Los resultados indicaron que los **modelos híbridos**, que incorporan tanto datos sociales como financieros, lograban un mejor rendimiento en comparación con los modelos que se basan únicamente en datos históricos del

mercado. Este estudio subrayó la importancia de las redes sociales como fuente de información complementaria en el análisis financiero.

Por otro lado, **Pagolu (2016)**[7] también exploraron cómo el análisis de sentimientos extraído de **X** podía integrarse con datos financieros tradicionales para mejorar la predicción de los precios de las acciones. Su estudio concluyó que los modelos que incluían datos de redes sociales eran más efectivos para capturar tendencias de corto plazo, ya que las emociones y opiniones expresadas en tiempo real en plataformas como **X** influían de manera inmediata en la percepción de los inversores y en el comportamiento del mercado.

Estos estudios destacan la creciente relevancia de las plataformas de comunicación digital en el ámbito financiero[8], demostrando que la integración de datos sociales con datos tradicionales permite captar mejor las dinámicas del mercado. Esta integración no solo mejora la precisión de las predicciones, sino que también ofrece una visión más completa de cómo factores externos, como las emociones y percepciones públicas, pueden influir en los precios de activos financieros.

2.3 Impacto de las Redes Sociales en los Credit Default Swaps

El análisis del impacto de las redes sociales en los **Credit Default Swaps (CDS)** es un campo emergente que ha cobrado relevancia en los últimos años, ya que los **CDS** son utilizados para medir el riesgo crediticio percibido de una entidad. Aunque los valores de los CDS han sido tradicionalmente afectados por factores económicos y políticos, recientes estudios como el de **Rodríguez-Ibañez**[2], que investigó la correlación entre la actividad en **X** y el comportamiento de los **CDS** de bancos europeos clave como **Credit Suisse**, **Deutsche Bank**, **Monte dei Paschi di Siena** y **Commerzbank**. El estudio analizó cómo ciertos eventos críticos, como reestructuraciones, sanciones regulatorias o ventas de activos, se reflejaban en la actividad en **X** y cómo dicha actividad afectaba los valores de los **CDS**.

Los hallazgos de este estudio fueron significativos. Se demostró que un aumento en la actividad de **X**, especialmente en torno a eventos negativos, estaba asociado con un incremento en los valores de los **CDS**, lo que indica un aumento en el riesgo crediticio percibido. Más allá de la cantidad de tweets, el estudio también subrayó la importancia de quién emitía estos mensajes[9]. Los tweets publicados por usuarios influyentes, como periodistas financieros, analistas o cuentas verificadas, tuvieron un impacto más significativo en los valores de los **CDS** que los tweets de usuarios con menor relevancia en la plataforma.

Este estudio es uno de los primeros en establecer una relación directa entre la comunicación en redes sociales y el comportamiento de los **CDS**. Sus resultados sugieren que las redes pueden no solo reflejar la situación financiera de una empresa, sino también amplificar la percepción de riesgo en momentos críticos. Este enfoque introduce una nueva dimensión en el análisis de los **CDS**, indicando que el análisis de medios sociales, combinado con datos financieros tradicionales, puede ser clave para anticipar movimientos en los **CDS** y gestionar mejor el riesgo crediticio.

2.4 Redes Sociales y Crisis Financieras

En situaciones de incertidumbre y volatilidad, la rápida difusión de rumores, noticias y opiniones en redes sociales puede exacerbar los problemas financieros de una entidad y desencadenar reacciones inmediatas en los mercados [10], incluidas corridas bancarias y fluctuaciones en los valores de los **Credit Default Swaps (CDS)**.

Un estudio emblemático en esta área es el de **Anthony Cookson (2023)**[11], que investigó cómo la crisis del **Silicon Valley Bank (SVB)** fue agravada por la actividad en **X**. A medida que circulaban rumores sobre la salud financiera del banco, los inversores y depositantes comenzaron a retirar fondos rápidamente, lo que desembocó en una corrida bancaria. El análisis de **Cookson** mostró que los rumores difundidos en **X** no solo reflejaron la realidad del problema financiero, sino que también contribuyeron a agravar la crisis al acelerar el pánico entre los inversores. El estudio utilizó el algoritmo **VADER** para analizar el sentimiento de los tweets, lo que permitió evaluar la emocionalidad de las publicaciones relacionadas con el SVB y su impacto en el comportamiento de los inversores.

El trabajo de **Cookson**[11] puso de manifiesto que las redes no solo actúan como un reflejo de los eventos financieros, sino que también pueden moldearlos e intensificarlos. En este contexto, **X** se convierte en un canal que influye directamente en la toma de decisiones de los actores del mercado, a veces exacerbando las crisis financieras. Pueden amplificar de manera exponencial la volatilidad del mercado en cuestión de minutos, lo que subraya la importancia de monitorear activamente estas plataformas durante períodos críticos.

Este fenómeno de amplificación es comparable a lo encontrado por **Rodríguez-Ibañez**[2], que subrayó cómo los eventos negativos reflejados en **X** pueden afectar significativamente los valores de los **CDS** de grandes bancos europeos. Ambos estudios resaltan que las RRSS son un factor que no debe ser subestimado en el análisis de riesgo financiero, especialmente en escenarios de crisis.

La capacidad de **X** para amplificar las crisis financieras plantea nuevas preguntas sobre cómo gestionar la comunicación en estas plataformas y sobre cómo los inversores pueden utilizar la información disponible en redes sociales para anticipar eventos críticos. A medida que las estas continúan evolucionando como una herramienta clave para la diseminación de información financiera, es fundamental entender su impacto en la percepción del riesgo y su influencia en las decisiones del mercado.

2.5 Impacto del Sentimiento en la Predicción Financiera

El análisis de sentimiento ha demostrado ser una herramienta clave para predecir movimientos financieros, especialmente cuando se trata de mercados altamente sensibles como los de acciones y **Credit Default Swaps (CDS)**. Las emociones expresadas en los medios sociales, como **X**, pueden generar percepciones inmediatas sobre el riesgo o la estabilidad financiera de una entidad, influyendo directamente en el comportamiento de los inversores.

Un estudio importante en este ámbito es el de **Bales (2023)**[12], que analizó el impacto de la incertidumbre generada en las redes sociales sobre el rendimiento de las acciones bancarias. Utilizando técnicas de **análisis de componentes principales** y **transformaciones wavelet**, los autores encontraron que la incertidumbre captada en **X** tenía un impacto significativo en los rendimientos de las acciones bancarias a corto plazo. Este análisis sugiere que los mercados

financieros son altamente sensibles a los cambios en el sentimiento, especialmente cuando estos reflejan temores o preocupaciones sobre la estabilidad de las instituciones financieras.

A largo plazo, el estudio de **Stephan Bales**[12] también señalaron que la incertidumbre política y económica tiene un mayor impacto en los precios de las acciones que la incertidumbre generada por plataformas digitales. Sin embargo, en contextos de alta volatilidad, pueden generar reacciones inmediatas que amplifican los movimientos del mercado a corto plazo. Este hallazgo refuerza la idea de que las redes sociales son una fuente clave de información en situaciones de crisis, aunque su impacto tiende a ser más transitorio en comparación con otros factores macroeconómicos.

Por otro lado, el estudio de **Nofer y Hinz (2015)**[13] profundizó en cómo los tweets relacionados con el mercado financiero pueden ser utilizados para predecir movimientos bursátiles. Sus resultados revelaron que los tweets que expresaban un sentimiento negativo estaban fuertemente correlacionados con caídas en los precios de las acciones, mientras que el sentimiento positivo tenía un efecto menos pronunciado. Este hallazgo es consistente con otros estudios que sugieren que las emociones negativas, como el miedo o la incertidumbre, tienen un mayor impacto en la percepción del riesgo que las emociones positivas.

Estos estudios subrayan el poder de las redes sociales para moldear la percepción del riesgo financiero en tiempo real, lo que las convierte en una herramienta valiosa para predecir fluctuaciones en los mercados. En el contexto de los **CDS**, donde el riesgo crediticio es el principal factor que observar, el análisis de sentimiento en **X** puede ofrecer una ventaja competitiva a los inversores al anticipar cambios en el comportamiento de los **CDS** antes de que se reflejen en los datos financieros tradicionales.

Capítulo 3

3 Materiales y métodos

Este capítulo detalla las técnicas, herramientas y metodologías empleadas para desarrollar y ejecutar el sistema de análisis basado en la correlación entre los datos del mercado financiero (específicamente **Credit Default Swaps - CDS**) y la actividad en **X**. Se comienza describiendo el proceso de obtención de los datos y se finaliza con las técnicas de visualización y análisis que permiten obtener los resultados pertinentes. Aunque el proyecto ha sido diseñado con fines investigativos, sus resultados pueden extrapolarse a un entorno profesional para su uso por inversores o empresas.

A lo largo del capítulo, se presenta una especificación de los distintos enfoques metodológicos, buscando que el sistema sea escalable y ágil para su utilización en futuras aplicaciones, siguiendo una estructura organizada en fases.

3.1 Conjuntos de Datos

En este apartado, se describen las fuentes de datos utilizadas en el estudio y los diferentes tipos de información empleados para analizar la correlación entre la actividad en redes sociales y los **Credit Default Swaps (CDS)**. La correcta selección y tratamiento de los conjuntos de datos es fundamental para garantizar la precisión y relevancia de los resultados obtenidos. Los datos utilizados en este proyecto provienen de múltiples fuentes, que incluyen datos financieros estructurados, como los CDS de los bancos seleccionados, y datos no estructurados, como los tweets relacionados con las entidades financieras estudiadas.

A continuación, se describen en detalle los tipos de datos utilizados, su relevancia en el análisis y las herramientas empleadas para su recolección y procesamiento.

3.1.1 Big Data y su relevancia en el proyecto

El concepto de **Big Data**[14] es clave para comprender la magnitud de la información que se maneja en este proyecto, ya que permite gestionar grandes cantidades de datos generados en tiempo real. Las arquitecturas de **Big Data** permiten el análisis y almacenamiento eficiente de estos datos, facilitando la predicción de eventos financieros. Para entender su complejidad, se destacan tres características fundamentales:

1. **Volumen**: La capacidad para manejar grandes cantidades de datos, como los generados por millones de usuarios en redes sociales. En nuestro caso, esto se refleja en la enorme cantidad de **tweets** relacionados con bancos.

2. **Velocidad**: Los datos fluyen de manera continua y en tiempo real, lo que obliga a los sistemas a procesarlos rápidamente. Para un análisis efectivo, es crucial gestionar el **flujo continuo** de tweets sobre las entidades bancarias, para identificar patrones en los **CDS** en tiempo casi real.

3. **Variedad**: Los datos provienen de múltiples fuentes y en diferentes formatos. En este estudio, los tweets son ejemplos de datos **no estructurados**, lo que requiere técnicas avanzadas de procesamiento para convertirlos en datos utilizables en análisis financiero.

3.1.2 Diferentes tipos de conjuntos de datos

Los datos con los que trabajamos en este proyecto pueden clasificarse en tres grandes grupos:

- **Datos estructurados**: Son aquellos que tienen una estructura definida y se almacenan en bases de datos relacionales, como los valores diarios de los **CDS** obtenidos de fuentes financieras como **Thomson Reuters DataStream**.

- **Datos semi-estructurados**: Presentan una organización interna, pero no siguen un esquema rígido. Los archivos **JSON** y **XML** que pueden contener tweets son ejemplos de este tipo de datos.

- **Datos no estructurados**: Son aquellos que no tienen una estructura definida, como los tweets, que pueden incluir texto, imágenes, hashtags y menciones sin una organización clara.

3.1.3 CDs y eventos financieros

En esta sección, se aborda el concepto de los Credit Default Swaps (CDS), su relevancia en el contexto del estudio, las herramientas utilizadas para obtener los datos y un análisis de las fuentes de información financieras empleadas.

3.1.3.1 Que son los CDS

Como se mencionó en la Introducción, los **Credit Default Swaps (CDS)**[15] son derivados financieros utilizados para protegerse frente a impagos de deuda. En esta sección, nos centraremos en los datos financieros asociados a estos instrumentos y su relevancia en el análisis del riesgo crediticio.

Los **CDS** se utilizan principalmente en los mercados financieros para dos fines principales:

1. **Cobertura (Hedging)**: Las instituciones compran CDS para protegerse contra posibles impagos de bonos o créditos de empresas. Por ejemplo, un banco que posea una cantidad significativa de bonos de una empresa puede comprar CDS para mitigar el riesgo de que la empresa incumpla con sus pagos.

2. **Especulación**: Los inversores también utilizan los CDS para especular sobre la solvencia de una empresa. Si un inversor cree que una empresa corre el riesgo de incumplir con su deuda, puede comprar un CDS como una apuesta a que el valor del CDS aumentará a medida que aumente el riesgo de default.

El **valor del CDS**[16] se determina principalmente por la probabilidad percibida de que el emisor de la deuda no pueda cumplir con sus obligaciones de pago. Factores como las condiciones macroeconómicas, las tasas de interés y la salud financiera del emisor son elementos cruciales en esta valoración. Un aumento en la prima del CDS sugiere que el riesgo percibido de incumplimiento está aumentando, mientras que una disminución indica que se percibe una mejora en la solvencia del emisor.

3.1.3.2 Herramientas comerciales para obtener información sobre CDS

Los datos sobre los **CDS** fueron obtenidos a través de **Thomson Reuters DataStream**, una plataforma comercial ampliamente utilizada por analistas financieros, que proporciona datos históricos y en tiempo real de diferentes instrumentos financieros. **DataStream**[17] es una herramienta esencial para el análisis financiero y económico debido a su gran cobertura y acceso a datos de múltiples mercados.

Además de DataStream, otras herramientas comerciales relevantes para obtener datos

de CDS y otros activos financieros incluyen:

- **Bloomberg:** Considerada una de las principales plataformas para la recolección de datos financieros en tiempo real[18], Bloomberg proporciona una amplia gama de herramientas de análisis y es utilizada globalmente por inversores y profesionales.

- **Yahoo! Finance**: Es una plataforma reconocida que proporciona datos financieros detallados tanto en tiempo real como históricos[19]. Ofrece herramientas como gráficos interactivos, estadísticas financieras, análisis de beneficios y flujo de caja, entre otros. Yahoo! Finance también permite el uso de su API para extraer datos de manera programática y realizar análisis personalizados.

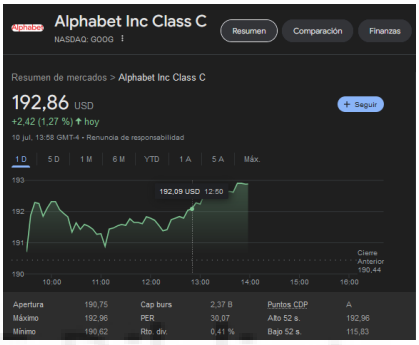


Figura 2.2: Resultado bursátil de Google para GOOG.

- **Google Finance**: Es otro motor de búsqueda que permite obtener rápidamente cotizaciones de acciones, información sobre empresas y datos históricos.[20] Aunque no ofrece tantas funciones avanzadas como Yahoo! Finance, es una herramienta útil para obtener datos bursátiles básicos de manera rápida.

Figura 2.3: Resultado bursátil de Yahoo! Finance para GOOG.

3.1.3.3 Detalles sobre DataStream

El principal recurso utilizado en este análisis fue **DataStream**, un servicio proporcionado por **Refinitiv Eikon**, que ofrece acceso a una amplia base de datos financieros con información histórica y en tiempo real de los mercados de valores globales. DataStream es una herramienta líder en el análisis financiero y económico, utilizada ampliamente por profesionales y académicos para obtener datos filtrados y relevantes de empresas y mercados en todo el mundo.[17]

Los datos de CDS recopilados a través de DataStream incluyen los valores de cierre diario, la evolución histórica de los precios, y otros factores que influyen en la valoración del riesgo financiero de las empresas. Esta información es fundamental para el análisis del riesgo crediticio y para evaluar cómo las fluctuaciones del mercado afectan la percepción de la estabilidad financiera de las empresas.

Esta herramienta destaca por mantener un buen equilibrio entre coste y producto, siendo más económico que Bloomberg, pero con prestaciones similares. Algunas de las características clave que justificaron su elección incluyen:

1. **Acceso a datos históricos exhaustivos**: DataStream permite obtener datos históricos detallados de **Credit Default Swaps** (CDS), lo que es esencial para realizar análisis longitudinales y detectar patrones a lo largo del tiempo.

2. **Amplia cobertura de mercados financieros**: La plataforma proporciona acceso a datos de múltiples mercados y tipos de activos financieros, desde acciones hasta bonos y derivados como los CDS, lo que lo convierte en una herramienta versátil.

3. **Filtrado avanzado de datos**: Permite la aplicación de filtros avanzados para seleccionar los datos relevantes para el análisis, facilitando el procesamiento y la interpretación de grandes volúmenes de información.

4. **Integración con herramientas de análisis**: DataStream se puede integrar con otras plataformas de análisis de datos y software estadístico, lo que permite realizar un análisis más profundo de las tendencias del mercado y los factores de riesgo.

3.1.4 Base de datos de Tweets

En esta sección se describe el conjunto de datos relacionados con los tweets utilizados en el proyecto, así como las técnicas empleadas para su extracción, almacenamiento y análisis.

Los datos de redes sociales, específicamente los tweets, juegan un papel crucial en este estudio al proporcionar información sobre el sentimiento y la actividad relacionada con los eventos de las entidades financieras seleccionadas.

3.1.4.1 Estructura y contenido de los datos

Los tweets fueron recolectados utilizando la plataforma **Graphext**, una herramienta que facilita la búsqueda y análisis de datos no estructurados provenientes de diversas fuentes [21]. El conjunto de datos incluye todas las interacciones en X que mencionan a cuatro bancos: **Credit Suisse, Deutsche Bank, Commerzbank y Monte dei Paschi di Siena**, entre el **1 de enero de 2017** y el **23 de mayo de 2023**.

Los tweets se recolectaron a través de las cuentas oficiales de las entidades **(@CreditSuisse**, **@BancaMPS**, **@DeutscheBank**, y **@commerzbank**), permitiendo un análisis detallado de la frecuencia de las publicaciones, las interacciones y el sentimiento que se genera en torno a eventos clave.

El conjunto de datos incluye las siguientes variables:

1. **Texto del tweet:** El contenido del tweet, que se somete a análisis de sentimiento para clasificarlo como positivo, negativo o neutral.

2. **Fecha y hora:** Información temporal que permite correlacionar los tweets con eventos específicos y fluctuaciones de los CDS.

3. **Número de retweets, likes, y respuestas:** Estas métricas indican el alcance e impacto de cada tweet, proporcionando una medida de su relevancia en el contexto de redes sociales.

4. **Hashtags y menciones:** Los hashtags y las menciones (@usuario) permiten identificar temas y actores clave en la conversación.

5. **Usuario y verificación:** Información sobre el usuario que publicó el tweet, incluyendo si la cuenta está verificada o no, lo que añade un peso adicional a los tweets emitidos por fuentes de mayor autoridad.

3.1.4.2 Procesamiento de los datos

Los tweets recolectados representan datos no estructurados, por lo que fue necesario aplicar técnicas de **Procesamiento del Lenguaje Natural (NLP)**[22] para transformarlos en un formato que pudiera ser utilizado en el análisis. Se aplicaron los siguientes pasos de procesamiento:

- **Tokenización:** Proceso mediante el cual se dividen los textos en unidades más pequeñas llamadas tokens (palabras o frases).

- **Normalización:** Incluye la conversión de texto a minúsculas, eliminación de caracteres especiales, URLs, y stopwords para reducir el ruido en los datos.

- **Análisis de sentimiento:** Mediante el uso de herramientas de NLP, se asignó un valor de sentimiento a cada tweet (positivo, negativo o neutral) para medir el tono de las publicaciones y correlacionarlo con los movimientos de los CDS.

3.1.4.3 Relevancia de los tweets en el análisis

El análisis de los tweets se centró en correlacionar la actividad en X con eventos críticos en la historia reciente de los bancos seleccionados. Algunos de estos eventos incluyeron **reestructuraciones, sanciones regulatorias, y recortes de personal**, que a su vez mostraron fluctuaciones significativas en los valores de los CDS. La frecuencia y el tono de los tweets se analizaron en relación con estos eventos, buscando patrones que pudieran ayudar a predecir variaciones en el riesgo financiero de las entidades.

La relevancia de esta base de datos radica en la capacidad de observar cómo los datos no estructurados de redes sociales pueden complementar los datos financieros tradicionales (como los CDS) para ofrecer una visión más integral del comportamiento de las entidades financieras ante eventos adversos.

3.1.5 Base de datos conjunta

En este apartado se describe cómo se integraron las diversas fuentes de datos para realizar un análisis conjunto de los eventos financieros y la actividad en redes sociales. La integración de estos conjuntos de datos permite correlacionar la información estructurada y no estructurada para obtener una visión más completa de la relación entre los **Credit Default Swaps** (CDS) y el sentimiento expresado en **X** respecto a las entidades financieras estudiadas.

3.1.5.1 Fuentes de Datos Integradas

Para el análisis, se combinaron dos fuentes de datos principales:

1. **Datos Financieros (CDS)**: Los valores diarios de los **Credit Default Swaps** de las cuatro entidades financieras (**Credit Suisse**, **Deutsche Bank**, **Commerzbank** y **Monte dei Paschi di Siena**) que fueron obtenidos de DataStream.

2. **Datos de Redes Sociales (Tweets)**: Estos datos incluyen tanto el contenido textual de los tweets como las métricas de interacción, proporcionando una medición del sentimiento y la actividad social en torno a los eventos clave de cada banco.

3.1.5.2 Integración de Datos

La integración de los **datos estructurados (CDS)** y **no estructurados (tweets)** se realizó siguiendo un enfoque temporal, correlacionando los picos de actividad en X con los cambios en los valores de los CDS. Para ello, fueron organizados de manera uniforme, considerando todo el rango de estudio.

Este proceso se llevó a cabo de la siguiente manera:

1. **Sincronización Temporal**: Se armonizaron las diferentes frecuencias de los conjuntos de datos. Los valores diarios de los CDS se emparejaron con la actividad diaria de X, permitiendo un análisis conjunto continuo a lo largo del periodo de estudio.

2. **Análisis de Sentimiento y Fluctuaciones de CDS**: Se utilizó el análisis de sentimiento de los tweets para correlacionar el tono (positivo, negativo o neutral) con los aumentos o disminuciones en los valores de los CDS. El objetivo era identificar si una emoción negativa o positiva en redes sociales precedía o seguía a cambios significativos en los CDS.

3. **Visualización Conjunta**: Los datos integrados fueron representados mediante gráficos y visualizaciones avanzadas, facilitando la identificación de patrones de correlación entre la actividad en X y las variaciones en los CDS.

3.1.5.3 Utilidad de la Base de Datos

La visualización de los datos conjuntos se realizó mediante gráficos que mostraban la evolución diaria tanto de los valores de los CDS como de la frecuencia y el **sentimiento de los tweets**. Esta representación permitió analizar posibles correlaciones entre la actividad en redes sociales y las **fluctuaciones en los CDS.**

3.2 Métodos

En este proyecto se emplearon diversas técnicas avanzadas para abordar el análisis de la relación entre la actividad en X y los Credit Default Swaps. Estas técnicas pueden dividirse en tres grandes categorías: **Procesamiento del Lenguaje Natural (NLP)**[22], que permite extraer información valiosa del texto no estructurado de las redes sociales; **Técnicas de aprendizaje estadístico o Machine Learning (ML)**[23], empleadas para construir modelos predictivos y de clasificación que analizan y correlacionan los datos.

Cada una de estas técnicas tiene un papel crucial en el análisis del proyecto y en la obtención de resultados consistentes y precisos. A continuación, se describen en detalle los enfoques y métodos utilizados en cada una de estas áreas.

3.2.1 Procesamiento del Lenguaje Natural (NLP)

El **Procesamiento del Lenguaje Natural (NLP)**[22] es una técnica fundamental en este proyecto, ya que permite extraer información relevante de grandes volúmenes de texto no estructurado provenientes de las redes sociales, en este caso, X. A través de diversas técnicas de análisis de texto, es posible interpretar la emoción de los usuarios y su relación con los valores de los CDS. En este contexto, el NLP es clave para comprender cómo la percepción pública, expresada a través de los tweets, puede tener un impacto en los mercados financieros.

3.2.1.1 Análisis Léxico

El análisis léxico se utilizó para medir el sentimiento expresado en los tweets relacionados con los bancos seleccionados. Se empleó **VADER** (Valence Aware Dictionary and Sentiment Reasoner) una herramienta ampliamente utilizada para el análisis de sentimiento en inglés, la cual asigna valores de sentimiento a palabras individuales y clasifica los textos en positivos, negativos o neutros.

Para medir el sentimiento expresado en los tweets sobre los bancos seleccionados, se utilizó **VADER (Valence Aware Dictionary and Sentiment Reasoner)**[24], una herramienta diseñada para analizar el tono de los mensajes. VADER clasifica las palabras de un texto en tres categorías: positivas, negativas o neutrales, sumando los valores para obtener el sentimiento general de cada tweet. Por ejemplo, un tweet que diga "El banco X está en problemas" sería clasificado como negativo, mientras que un mensaje como "El banco Y está creciendo rápidamente" sería positivo.

Cada palabra del tweet es evaluada en función de su polaridad (positiva o negativa), y los valores de las palabras se suman para calcular el sentimiento general del tweet. Este enfoque permite obtener una estimación del sentimiento expresado por los usuarios en relación con los eventos financieros de los bancos analizados, sin requerir grandes volúmenes de datos etiquetados para el entrenamiento.

3.2.1.2 Modelos Basados en Reglas

Los **modelos basados en reglas** fueron otro enfoque empleado para mejorar la precisión del análisis. Estos modelos utilizan reglas lingüísticas para interpretar el contexto de las palabras dentro de un tweet. Por ejemplo, las negaciones como "no" o "nunca" pueden alterar significativamente el significado de una frase. Este enfoque también considera la estructura gramatical para capturar mejor las emociones expresadas.

3.2.1.3 Preprocesamiento del Texto

Antes de aplicar los modelos de análisis léxico y basados en reglas, es necesario realizar una serie de preprocesamientos que permiten normalizar el texto y transformarlo en un formato adecuado para el análisis computacional. Las técnicas utilizadas para el preprocesamiento son las siguientes:

1. **Tokenización**: El texto se divide en palabras o tokens individuales. Esta segmentación es crucial para analizar cada palabra de forma separada. Para los tweets, se emplearon delimitadores como espacios, signos de puntuación y menciones (@usuario).

2. **Normalización del Texto**: En esta etapa, el texto se transforma a una forma estándar. Esto incluye convertir el texto a minúsculas, eliminar caracteres especiales, enlaces y hashtags irrelevantes. Los procesos más comunes de normalización utilizados en este proyecto fueron:

    - **Lematización**: Se reduce cada palabra a su forma base o "lema". Esto permite analizar las palabras en su forma más simple, evitando redundancias.

    - **Stemming**: Similar a la lematización, este proceso recorta las palabras a sus raíces, lo que acelera el análisis, aunque con menor precisión en algunos casos.

3. **Etiquetado de Partes del Discurso (PoS Tagging)**[25]: Esta técnica asigna etiquetas gramaticales a cada token, como sustantivo, verbo o adjetivo. Esto ayuda a entender el papel de cada palabra en una oración y mejora la comprensión del texto. En este proyecto, se emplearon tanto modelos basados en reglas como modelos estadísticos para realizar el etiquetado.

4. **Reconocimiento de Entidades Nombradas (NER)**[26]: Esta técnica identifica nombres propios en el texto, tales como personas, organizaciones, fechas y cantidades. El NER es especialmente útil para extraer menciones relevantes de los bancos y otros actores financieros importantes en los tweets.

3.2.1.4 Parsing Sintáctico y Dependencias

El **parsing sintáctico**[27] analiza la estructura gramatical de las oraciones, identificando las relaciones entre las palabras. Esta técnica permite entender cómo las palabras interactúan entre sí y cómo se agrupan para formar significados complejos. Existen dos tipos principales de análisis que se utilizaron en este proyecto:

- **Dependencias entre palabras**: Identifica la relación entre el sujeto, verbo y objeto en cada

tweet, lo que permite comprender cómo las palabras se relacionan entre sí.

- **Estructura jerárquica**: Examina la organización de las oraciones para detectar patrones más complejos de comunicación, ayudando a interpretar el significado más profundo del texto.

3.2.1.5 Limitaciones y Desafíos del NLP

Aunque el procesamiento del lenguaje natural ha avanzado significativamente, presenta ciertos desafíos. El principal problema es la identificación de matices sutiles como el sarcasmo o la ironía, que pueden alterar el verdadero significado de un tweet. Estos elementos dependen en gran medida del contexto y son difíciles de detectar para los algoritmos actuales. Sin embargo, el uso de técnicas más avanzadas como los **Grandes Modelos de Lenguaje (LLMs)**[28], como **BERT** (Bidirectional Encoder Representations from Transformers)[29], ha mejorado la capacidad para interpretar el contexto de manera más precisa.

3.2.2 Técnicas de aprendizaje estadístico o Machine Learning (ML).

El **aprendizaje automático (Machine Learning)**[23] es otra de las herramientas fundamentales en este proyecto. Se utilizó para construir modelos predictivos y de clasificación que permitieron analizar y correlacionar los datos de las redes sociales con las fluctuaciones en los Credit Default Swaps (CDS). El uso de **ML** en este proyecto se centró en aplicar algoritmos que permitieran identificar patrones complejos entre los datos financieros y los textos no estructurados provenientes de X.

3.2.2.1 Modelos de Predicción y Clasificación

En el proyecto se utilizaron dos tipos principales de modelos de **Machine Learning**: **modelos de regresión** y **modelos de clasificación**, cada uno de ellos diseñado para abordar distintos aspectos del análisis.

1. **Regresores**[30]: Estos modelos predicen valores continuos, como los precios de los CDS a lo largo del tiempo. Se emplearon para modelar la relación entre los datos de redes sociales y los valores de los CDS, buscando identificar cómo las fluctuaciones en el sentimiento de X influyen en las variaciones de los CDS.

    - **Aplicaciones comunes**: Predicción de precios de activos financieros, como los CDS, y análisis de tendencias en series temporales.

    - **Modelos empleados**: Regresión lineal y regresión polinómica, ajustados para capturar las tendencias lineales y no lineales presentes en los datos.

2. **Clasificadores**[31]: Los clasificadores asignan categorías a los datos. En este caso, los clasificadores se utilizaron para predecir el sentimiento de los tweets, clasificándolos como positivos, negativos o neutrales. Estos modelos fueron clave para correlacionar

el sentimiento social con los datos financieros.

- **Aplicaciones comunes**: Clasificación de textos, como el análisis de sentimiento en redes sociales, y sistemas de recomendación.

- **Modelos empleados**: Redes neuronales, Support Vector Machines (SVM) y regresión logística, los cuales permiten clasificar de manera precisa el sentimiento de grandes volúmenes de datos.

3.2.2.2 Algoritmos de Machine Learning Utilizados

Se utilizaron varios algoritmos de **Machine Learning** para ajustar los modelos predictivos y de clasificación, los más destacados son:

La **Regresión Lineal**[32] es una técnica común en Machine Learning utilizada para predecir valores futuros basándose en una relación entre variables. Por ejemplo, si queremos predecir el valor de los CDS (variable dependiente Y) basándonos en la cantidad de tweets negativos (variable independiente X), podemos aplicar la regresión lineal para identificar la relación entre ambas variables. Si los tweets negativos aumentan, podríamos esperar que el valor de los CDS también aumente.

Ecuación de la Regresión Lineal: $\boldsymbol{Y = a + bX}$

	REGRESIÓN LINEAL
Funcionamiento	Donde *"a"* es la intersección y *"b"* es la pendiente.
Aplicaciones	Utilizada para identificar tendencias y relaciones lineales en datos históricos.
Ventajas y Desventajas	Fácil de interpretar. Rápido de entrenar, incluso con grandes conjuntos de datos. Asume relación estrictamente lineal entre las variables.

Cuadro 3.6: Algoritmo Regresión Lineal

La **Regresión Gaussiana**[33] es una extensión de la regresión lineal que permite modelar relaciones no lineales mediante el uso de funciones de base. En lugar de suponer una relación lineal entre las variables, introduce **funciones kernel** que permiten capturar relaciones más complejas.

	REGRESIÓN GAUSSIANA
Funcionamiento	Extensión de la Regresión Lineal con incorporaciones de términos no lineales mediante una fusión de funciones.
Aplicaciones	Modelar relaciones más complejas en datos que no siguen las tendencias lineales simples
Ventajas y Desventajas	Capacidad para modelar relaciones complejas. Flexible ante no linealidad. Requiere ajuste complejo de parámetros. Mayor costo computacional.

Cuadro 3.7: Algoritmo Regresión Gaussiana

El algoritmo **SVM**[34] es ampliamente utilizado para problemas de clasificación y regresión. La idea principal es encontrar el **hiperplano** que mejor separa las clases en un espacio de alta dimensión. SVM se basa en la maximización del margen entre las diferentes clases, utilizando **funciones kernel** para permitir la clasificación no lineal.

Ecuación del modelo:

La función de decisión para el clasificador SVM puede expresarse como: $f(x) = w^{Tx} + b$
Donde **w** es el vector normal del hiperplano y **b** es el término independiente.

	Support Vector Machine (SVM)
Funcionamiento	Utiliza funciones para transformar los datos y encontrar un límite de decisión óptimo.
Aplicaciones	Se basa en la clasificación del texto por categorías de sentimiento obteniendo una predicción del impacto en el mercado.
Ventajas y Desventajas	Excelente rendimiento en alta dimensión. Robusto en problemas no lineales. Requiere ajuste fino de hiperparámetros. Propenso al sobreajuste.

Cuadro 3.8: Algoritmo SVM

Los **Árboles de Decisión**[35] son modelos de predicción que utilizan una estructura jerárquica de decisiones para clasificar datos o predecir valores continuos. Se basa en la partición recursiva de los datos en subconjuntos más pequeños, utilizando reglas basadas en las **características** de los datos.

Ecuación de decisión:

La construcción de un árbol de decisión se basa en la minimización de una métrica de impureza como el **índice Gini** o la **entropía**: $G_{ini} = 1 - \sum_{i=1}^{n} p_i^2$

Donde $\boldsymbol{p_i}$ es la proporción de una clase específica.

	ARBOLES DE DECISIÓN
Funcionamiento	Se dividen los datos en subconjuntos más pequeños hasta alcanzar una decisión final y óptima.
Aplicaciones	Es un modelo idóneo para predecir resultados categóricos o continuos.
Ventajas y Desventajas	Fácil de interpretar. Funciona bien con datos mixtos. Propenso al sobreajuste. Sensible a las variaciones de los datos.

Cuadro 3.9: Algoritmo Arboles de Decisión

3.2.2.3 Prevención del Sobreajuste (Overfitting)

Uno de los desafíos más importantes al utilizar modelos de **Machine Learning** es evitar el **sobreajuste**[36], que ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento, capturando tanto las tendencias generales como el ruido específico de esos datos. Esto impide que el modelo generalice bien en datos nuevos.

Para prevenir el sobreajuste, se aplicaron técnicas como:

1. **Validación Cruzada (Cross-validation)**[37]: Se dividieron los datos en subconjuntos (folds), entrenando el modelo en diferentes combinaciones de estos subconjuntos para asegurarse de que el modelo generalizara bien en todos los datos.

2. **Regularización**: Se aplicaron técnicas de regularización para penalizar modelos demasiado complejos, lo que ayuda a reducir el riesgo de sobreajuste.

3. **Reducción de Dimensionalidad con PCA**[38]: El **Análisis de Componentes Principales (PCA)** se utilizó para reducir la dimensionalidad del conjunto de datos, eliminando las variables irrelevantes o redundantes. Esto permitió mejorar la capacidad de generalización de los modelos al enfocarse solo en las variables más importantes.

3.2.3 Media de riesgo financiero (volatilidad)

En este apartado, abordaremos conceptos financieros clave utilizados en este trabajo que son fundamentales para entender y gestionar el riesgo en los mercados financieros, especialmente en el contexto del análisis de **Credit Default Swaps (CDS)**. Las estrategias y herramientas financieras aplicadas para el **trading** y la **gestión de riesgos** desempeñan un papel esencial en la maximización de ganancias y la minimización de pérdidas, al igual que en la mejora de la precisión de los modelos predictivos, como es el caso de este estudio. A continuación, se detallan dos conceptos clave que fueron utilizados: la **Desviación Estándar**[39] y el **Índice de Volatilidad (VIX)**[40].

En este apartado, se profundiza en conceptos financieros clave que son fundamentales para entender y gestionar el riesgo en los mercados, particularmente en el análisis de los **CDS**. La volatilidad es un indicador crítico en la gestión de riesgos financieros, ya que proporciona una visión cuantitativa del nivel de incertidumbre en torno a los activos. Las métricas como la **desviación estándar** y su relación indirecta con el **Índice de Volatilidad (VIX)** se utilizan para evaluar y medir este riesgo.

3.2.3.1 Desviación Estándar

La **desviación estándar**[39] es una medida estadística que cuantifica la dispersión de los datos en relación con su media. En el ámbito financiero, esta métrica es esencial para medir la volatilidad, entendida como el grado de fluctuación de un activo financiero en un periodo determinado. En el contexto de este proyecto, la desviación estándar aplicada a los precios de los CDS es una herramienta crucial para evaluar el riesgo percibido por los inversores sobre la estabilidad financiera de una empresa.

1. **Interpretación en nuestro contexto**: Un **valor alto de desviación estándar** en los CDS indica que los precios han experimentado **grandes fluctuaciones**, lo que sugiere una **mayor incertidumbre** en el mercado sobre la solvencia de la entidad financiera. Este aumento en la volatilidad puede estar relacionado con diversos factores como la publicación de noticias financieras adversas, cambios en las calificaciones crediticias, o rumores de insolvencia.

2. **Relación con el análisis de sentimiento**: La desviación estándar, al medir la dispersión de los precios, se puede correlacionar con el sentimiento general del mercado. Un aumento en la desviación estándar puede reflejar un incremento en la preocupación de los inversores, que puede estar influido por rumores, noticias negativas, o eventos desfavorables amplificados en plataformas como X. Al combinar la desviación estándar con el análisis de sentimiento, es posible mejorar la capacidad predictiva de un modelo en cuanto a las variaciones futuras en los precios de los CDS.

3. **Aplicación práctica**: La desviación estándar no solo permite estimar el riesgo, sino también identificar periodos de alta volatilidad o comportamientos inusuales en el mercado, lo que permite a los inversores ajustar sus estrategias de manera proactiva. En este estudio, esta métrica actúa como un indicador clave para prever cambios en el comportamiento de los CDS basados en la información extraída de las redes sociales.

4 3.2.3.2 Índice de Volatilidad (VIX)

Aunque en este estudio no se utilizó directamente el **Índice de Volatilidad (VIX)**[41], es importante mencionarlo como una referencia complementaria, ya que la desviación estándar está relacionada con el concepto de volatilidad que este índice mide. El VIX es un indicador que mide la volatilidad implícita en el mercado de acciones para los próximos 30 días, y se usa comúnmente como una medida del "miedo" o la incertidumbre en los mercados financieros.

1. **Impacto en los precios de los CDS**: Un aumento en el VIX indica que los inversores esperan una mayor volatilidad en el mercado, lo que a menudo conlleva un aumento en los precios de los CDS. Cuando el mercado se vuelve más volátil, los inversores buscan mayor protección contra el riesgo de impago, incrementando la demanda de CDS y, por ende, su precio.

2. **Conexión con el análisis de sentimiento en X**: El VIX puede ser influido por eventos que generan incertidumbre en los mercados, y que son amplificados en redes sociales como X. Un tweet negativo sobre la estabilidad de una entidad financiera que se viraliza puede desencadenar una reacción en cadena en los mercados, aumentando el VIX y, en consecuencia, los precios de los CDS. Este fenómeno puede reflejar un aumento en el riesgo percibido debido a la difusión de información a través de las redes sociales.

3. **Integración en el modelo predictivo**: Aunque el VIX no fue empleado explícitamente en este proyecto, su concepto está relacionado con la desviación estándar y la volatilidad general del mercado, lo que permite obtener una visión más completa del riesgo. Incluir una métrica como el VIX, junto con la desviación estándar y el análisis de sentimiento en redes sociales, podría enriquecer futuros modelos predictivos para anticipar mejor los movimientos de los CDS en situaciones de alta volatilidad.

Capítulo 4

4 Experimentos

En este capítulo se detalla el conjunto de experimentos realizados para explorar y validar las hipótesis planteadas sobre la relación entre la actividad en redes sociales, específicamente en X, y las fluctuaciones en los Credit Default Swaps de las entidades financieras en riesgo. A través de una serie de análisis basados en técnicas avanzadas de Procesamiento del Lenguaje Natural (NLP) y aprendizaje automático, se busca identificar patrones y correlaciones clave que puedan predecir el comportamiento de los CDS a partir de la actividad social.

El principal objetivo de estos experimentos es proporcionar una evaluación empírica sólida de cómo los datos no estructurados provenientes de redes sociales pueden influir en la percepción del riesgo financiero. Para ello, se han implementado diversas metodologías que abarcan desde el análisis semántico y de sentimiento hasta la construcción de modelos predictivos, utilizando tanto datos históricos como técnicas modernas de análisis estadístico.

Los experimentos se han organizado en diferentes fases, que incluyen la recolección de datos, la implementación de modelos predictivos, y la validación de resultados. Cada una de estas fases está diseñada para responder preguntas específicas, tales como: ¿qué impacto tienen las noticias negativas en los precios de los CDS? ¿Cómo afecta el volumen de comunicación en redes sociales al riesgo percibido por los inversores?

A lo largo del capítulo, se desglosarán las distintas técnicas aplicadas en cada fase experimental, se discutirá la selección de las entidades financieras bajo estudio y se presentará el flujo algorítmico implementado para el procesamiento de grandes volúmenes de datos. Este análisis no solo busca identificar relaciones directas entre las variables, sino también prever cómo eventos críticos pueden alterar la estabilidad financiera de una empresa, lo que lo convierte en una herramienta útil tanto para investigadores como para profesionales del sector financiero.

4.1 Técnicas implementadas en los experimentos

Durante el desarrollo, se aplicaron diversas técnicas y enfoques analíticos para abordar el comportamiento de los CDS, en relación con la actividad en X. Las técnicas utilizadas y mencionadas en capítulos anteriores han sido seleccionadas y adaptadas para extraer el máximo valor de los datos no estructurados, así como para capturar la complejidad del riesgo financiero percibido por los inversores.

4.1.1 Enfoque metodológico

El enfoque metodológico para los experimentos se centró en maximizar la integración de diferentes fuentes de datos en un modelo de análisis unificado, que permitiera correlacionar la información textual de las redes sociales con los datos financieros cuantitativos. Para ello, se dividió el proceso en tres fases principales:

1. **Preprocesamiento de Datos:** A través de técnicas de tokenización y normalización, los datos no estructurados (tweets) fueron procesados para convertirlos en unidades manejables y uniformes. Se establecieron reglas específicas para filtrar el ruido generado por los usuarios y optimizar la calidad de la información obtenida, asegurando que los datos procesados fueran representativos de la opinión pública sobre las entidades financieras seleccionadas.

2. **Análisis Semántico y de Sentimientos:** Para medir el sentimiento general en torno a los bancos, se aplicaron herramientas avanzadas de análisis semántico que permitieron una categorización precisa de los tweets en función de su polaridad (positiva, negativa o neutra). Esto se complementó con un enfoque semántico que fue más allá de la simple categorización, permitiendo identificar patrones contextuales que afectan la percepción del riesgo.

3. **Modelos Predictivos y Correlacionales:** La última fase implicó la implementación de modelos de regresión y clasificación, entrenados tanto con los datos históricos de los CDS como con las variables sentimentales extraídas de los tweets. El uso de algoritmos de aprendizaje automático permitió establecer relaciones entre la actividad en X y las fluctuaciones de los CDS, evaluando cómo los patrones identificados pueden predecir movimientos futuros en los precios de los CDS.

4.1.2 Análisis semántico y sentimental

En este apartado, el foco se encuentra en la aplicación de técnicas avanzadas de procesamiento del lenguaje natural para extraer información relevante del gran volumen de datos no estructurados provenientes de las redes sociales. Estas técnicas fueron clave para capturar el sentimiento y el contexto detrás de las menciones a las entidades financieras analizadas en la plataforma X.

Técnicas implementadas:

1. **Tokenización y normalización de textos**[42]**:** Para poder procesar los datos textuales obtenidos de los tweets, se llevó a cabo un preprocesamiento de estos, que consistió en la tokenización (dividir los tweets en palabras o frases más pequeñas) y la normalización (estandarización de las palabras, eliminando variaciones ortográficas y léxicas, como convertir todas las palabras a minúsculas y eliminar signos de puntuación). Este paso fue crucial para garantizar que los tweets pudieran ser analizados de manera uniforme y precisa en las siguientes etapas.

2. **Análisis de sentimiento con VADER**[24]**:** El análisis del sentimiento se realizó utilizando la herramienta VADER, que está optimizada para el análisis de texto corto, como los tweets. Cada palabra o grupo de palabras fue clasificada como positiva, negativa o neutra, y los tweets se agruparon en estas categorías para obtener una visión general del sentimiento expresado hacia cada entidad bancaria. Este proceso permitió cuantificar el tono emocional de la conversación en torno a cada banco y su relación con las fluctuaciones observadas en sus CDS.

3. **Ponderación del sentimiento por influencia de los usuarios:** Para mejorar la precisión del análisis, se introdujo un sistema de ponderación que asignó mayor relevancia a los tweets de usuarios con más seguidores o con un mayor número de interacciones (retweets, likes y respuestas). Esto permitió ajustar el análisis para que reflejara no solo el sentimiento general, sino también la influencia potencial de ciertos usuarios clave en la propagación de la percepción sobre las entidades financieras.

4. **Frecuencia de menciones y su relación con los CDS:** Se examinó la frecuencia de menciones de los bancos en X, particularmente en los momentos de mayor fluctuación en sus CDS. El número de menciones se comparó con los cambios en los valores de CDS en diferentes periodos de tiempo, observando si un aumento en la actividad social coincidía con variaciones significativas en los CDS.

5. **Correlación entre el sentimiento y las fluctuaciones en los CDS**[43]**:** Para determinar si había una relación entre el sentimiento expresado en los tweets y las fluctuaciones en los CDS, se realizó un análisis de correlación. Este análisis examinó si los cambios en el sentimiento general (positivo, negativo o neutro) coincidían temporalmente con cambios en los valores de CDS de los bancos. Los resultados permitieron evaluar si había indicios de que los movimientos del mercado financiero estuvieran influidos por la actividad en redes sociales.

4.1.3 Hipótesis de trabajo

En el contexto de este estudio, las hipótesis de trabajo surgieron a partir de la necesidad de explorar la relación entre la actividad en X y los Credit Default Swaps de los bancos seleccionados. Con el fin de estructurar los análisis y guiar el enfoque metodológico, se formularon las siguientes hipótesis:

4.1 – TÉCNICAS IMPLEMENTADAS EN LOS EXPERIMENTOS

1. **H1: Las menciones en X relacionadas con los bancos no siempre reflejan de manera inmediata el comportamiento de los CDS**.

   La hipótesis plantea que, aunque puede existir una relación entre la frecuencia o el volumen de menciones sobre un banco y su situación financiera, esta relación no es directa ni inmediata. Se presume que algunas menciones, especialmente aquellas sin contenido relevante o sin correlación con eventos financieros significativos, no tendrían un impacto directo en los CDS.

2. **H2: El sentimiento negativo en las redes sociales está asociado con un aumento en el valor de los CDS**.

   Dado que los CDS representan el riesgo percibido sobre una entidad financiera, la hipótesis asume que el contenido negativo en redes sociales, que refleje preocupaciones sobre la estabilidad de un banco, podría estar relacionado con aumentos en los valores de los CDS. Esta relación se espera particularmente en momentos de crisis o eventos de incertidumbre, donde los inversores responden a las percepciones públicas sobre el banco.

3. **H3: El volumen de actividad en X influye en el comportamiento de los CDS, actuando como un indicador anticipado**.

   Se postula que, en algunos casos, un alto volumen de actividad en redes sociales puede preceder a movimientos en los CDS. En particular, un incremento significativo en la actividad ya sea en términos de menciones o interacciones, podría actuar como un indicador de posibles fluctuaciones futuras en los valores de los CDS.

4. **H4: La interacción entre los movimientos en los CDS y la actividad en X genera un ciclo de retroalimentación**.

   Esta hipótesis sugiere que el comportamiento en los mercados financieros y la actividad en redes sociales pueden influirse mutuamente. Un aumento en los CDS puede generar más actividad y menciones en X, lo que a su vez intensifica la percepción pública y podría llevar a una mayor fluctuación en los CDS, creando un ciclo de retroalimentación entre ambas esferas.

4.1.4 Flujo algorítmico implementado

Esta sección detalla los diferentes usos y funcionalidades que se pueden llevar a cabo en el programa desarrollado. Se denominan versiones a los diversos procesos que el usuario puede utilizar durante la ejecución del programa.

En la versión actual del desarrollo, el programa se ha dividido en bloques denominados **procesos**, organizados por versiones. Esto permite al usuario comenzar en distintos puntos del sistema según convenga a su estudio. Estos procesos funcionan de la siguiente manera:

El **Proceso V1** se encarga de **transformar datos no estructurados en datos estructurados**, priorizando la escalabilidad de estos datos para su posterior uso en los otros procesos.

PROCESO V1	DESCRIPCIÓN
Extracción Bolsa	Es el dedicado a extraer los datos sin procesar de los CDS y convertirlo en una variable utilizable con los datos a utilizar.
Extracción Tweets	Dedicado a extraer los datos sin procesar de los tweets, filtrándolos y limpiándolos para su posterior uso. Se encarga de agregar las variables importantes como el Agregado por Sentimiento o variables más específicas como el Sentimiento ponderado por seguidores.

Cuadro 3.1: Proceso V1

El **Proceso V2**, por otro lado, es responsable de **filtrar los datos según las fechas** deseadas. Además, este proceso desempeña la función crucial de relacionar los datos de los tweets obtenidos con los datos de los CDs.

PROCESO V2	DESCRIPCIÓN
Fechas Específicas	Proceso opcional, su finalidad es filtrar por fechas todo el contenido obtenido, ya sea entre un margen amplio de los datos o reducido, para obtener y estudiar únicamente por ejemplo un mes específico a seleccionar mediante una interfaz o manualmente.
Unificación de Datos	Es uno de los procesos más importantes, dedicado a la correcta unión en una misma variable de los datos de los tweets y CDs permitiendo comparar posteriormente estos datos o directamente o relacionar los datos de tweets de un día con días posteriores o anteriores de los CDS.

Cuadro 3.2: Proceso V2

Una vez obtenidos los datos de una manera estructurada y según las características que buscábamos y con las variables que queremos utilizar, procedemos a encontrarnos con el **Proceso V3**, este se encarga de la tarea de **introducirnos en un análisis de los datos básicos**.

PROCESO V3	DESCRIPCIÓN
Análisis Exploratorio Básico	Realiza la función de representar los datos en procesos básicos como representación lineal, correlación mutua o histograma de variables.
Análisis Por Secuencia	Es el encargado de realizar los primeros estudios complejos. Se abordarán puntos clave como la implementación de un SVM mediante Regresión Gaussiana, la creación de un Clasificador de Árbol y su respectiva representación.

Cuadro 3.3: Proceso V3

Al proceder a ejecutar el **Proceso V4**, nos encontramos con un **análisis más sofisticado** que incluye una división entre Regresor y Clasificador.

PROCESO V4	DESCRIPCIÓN
Clasificador	Predice los valores de manera discreta, utilizando la técnica con mejor resultado, como pueden ser SVM o Árboles de decisión.
Regresor	Este modelo recibe directamente todos los datos originales sin transformación, aunque es propenso al sobreajuste en el caso de grandes cantidades de datos
Regresor con PCA	Regresor similar al anterior, pero con una transformación mediante PCA para reducir la dimensionalidad del modelo y entrenar el regresor con los componentes principales capturando así la información más relevante, pero perdiendo datos.

Cuadro 3.4: Proceso V4

Como agregado final se optó realizar el **Proceso V5**, el cual es un sistema de **predicción mediante el uso de datos de días conocidos**, se estimasen los días que los CDS siguiesen comportamientos erráticos.

PROCESO V5	DESCRIPCIÓN
Análisis Días Clave	Clasificador donde se realiza una estimación por precisión de días con comportamientos en el mercado anómalos.

Cuadro 3.5: Proceso V5

4.2 Proceso v1 - Modelo de datos

Este capítulo busca realizar un seguimiento completo de la información que se obtuvo en la extracción, hasta su posterior tratado hasta el resultado final. Para llevar a cabo este cometido, se analizarán y describirán en detalle las diversas fuentes de información a partir de las cuales se han obtenido los datos que se utilizarán durante todo el proceso de este proyecto.

Una vez se haya realizado un análisis de dichas fuentes de información se procederá al desarrollo del modelo con el cual se pretende derivar a las funciones que representen la previsualización de los datos que previamente se han analizado.

Finalizando este proceso, se llevará a cabo un análisis de los diversos cambios que sufrirán los datos con el fin de facilitar su posterior uso en las funciones finales que los utilicen, utilizando para ello los apartados bien delimitados que permite el Live Script de Matlab indicándose en todo momento y dividido por bloques un historial de los cambios que van sufriendo los datos, esto con el fin de mantener un orden en el tratado de estos.

4.2.1 Datos en Crudo (Raw Data)

Comenzando con el primer punto de esta sección, este pretende describir las diversas fuentes de información que se obtienen en su formato original (Raw Data), indicándonos esto, que se podría traducir como (Datos en Crudo) que estaremos observando la extracción de los datos vírgenes y sin adulterar de los cuales se van a obtener datos que son susceptibles de ser utilizados para aportar de contenido la arquitectura que se pretende desarrollar. Una vez analizada la información con la que se dispone, se procederá a seleccionar los campos cuya información será utilizada durante el proyecto y justificando debidamente su elección.

Las fuentes de información con las que es posible dotar de contenido a la arquitectura como es entendible y debido a la naturaleza de dos distintos tipos de datos son divididas en dos secciones, la información obtenida y asociadas a los datos históricos de las empresas y todos sus valores de CDs, y por otro lado la información asociada al sentimiento y tweets de las empresas seleccionadas.

4.2.2 Información asociada a los CDs

Los datos históricos asociados a las empresas son obtenidos a través de la aplicación DataStream de la empresa Refinitiv Eikon, una compañía ahora parte del London Stock Exchange Group, este se trata de un programa de pago que gracias a la participación de la Universidad Rey Juan Carlos se han podido obtener los datos, este servicio destaca por su amplia base de datos de series temporales financieras, que abarca más de 70 años de datos históricos, siendo una base de datos altamente utilizada para realizar análisis macroeconómicos, crear escenarios de mercado y probar estrategias de inversión.

Como es de esperar ante una aplicación tan sofisticada se conoce que esta más que por un usuario medio es mayormente utilizado por profesionales financieros e investigadores académicos, por ello el uso de la herramienta es esencial para aquellos que necesitan un análisis profundo de datos financieros y macroeconómicos, lo cual la convierte en la herramienta perfecta para llevar a cabo nuestro proyecto.

El archivo que descargamos de la base de datos se descargó en un archivo .xlsx debido a que de esta manera esos datos podían ser utilizados por la propia Universidad Rey Juan Carlos para sus propios estudios. El archivo está diferenciado en este caso por las empresas extraídas en una especie de columnas como podemos observar:

Updated at 19:16:39		Updated at 19:13:17		Updated at 19:06:01		Updated at 19:11:41	
CREDIT SUISSE	CSGNSYEUAM=	DEUTSCHE BANK	DBSYEUAL=	MONTE DEI PASCHI DI SIEN	BMPSSYEUAM=	COMMERZBANK	CBKGSYEUAM=
Timestamp	Mid Spread Close	Timestamp	Mid Spread Close	Timestamp	Mid Spread Close	Timestamp	Mid Spread Close
22/05/2023	160.05	22/05/2023	168.35	22/05/2023	298.46	22/05/2023	85.1
19/05/2023	162.58	19/05/2023	157.41	19/05/2023	298.31	19/05/2023	86.08
18/05/2023	186.525	18/05/2023	201.195	18/05/2023	349.085	18/05/2023	100.365
17/05/2023	186.245	17/05/2023	200.915	17/05/2023	348.805	17/05/2023	100.085
16/05/2023	186.228	16/05/2023	200.898	16/05/2023	348.788	16/05/2023	100.068
15/05/2023	186.216	15/05/2023	200.886	15/05/2023	348.776	15/05/2023	100.056
12/05/2023	186.202	12/05/2023	200.872	12/05/2023	348.762	12/05/2023	100.042
11/05/2023	186.195	11/05/2023	200.865	11/05/2023	348.755	11/05/2023	100.035
10/05/2023	186.172	10/05/2023	200.842	10/05/2023	348.732	10/05/2023	100.012
09/05/2023	186.16	09/05/2023	200.83	09/05/2023	348.72	09/05/2023	100
08/05/2023	177.16	08/05/2023	196.33	08/05/2023	348.65	08/05/2023	97.01
05/05/2023	177.15	05/05/2023	196.16	05/05/2023	348.68	05/05/2023	97.01
04/05/2023	181	04/05/2023	197.53	04/05/2023	349.22	04/05/2023	100
03/05/2023	175.06	03/05/2023	191.07	03/05/2023	348.98	03/05/2023	95.04
02/05/2023	175.08	02/05/2023	191.33	02/05/2023	348.93	02/05/2023	96.03
01/05/2023	170.09	01/05/2023	185.1	01/05/2023	348.87	01/05/2023	90.07
28/04/2023	170.04	28/04/2023	185.2	28/04/2023	349.06	28/04/2023	90.09
27/04/2023	172.94	27/04/2023	189.91	27/04/2023	348.47	27/04/2023	90.04
26/04/2023	175.09	26/04/2023	192.52	26/04/2023	348.87	26/04/2023	92.06
25/04/2023	172.77	25/04/2023	182.3	25/04/2023	364.23	25/04/2023	88.11

Figura 4.1: Ejemplo .xlsx de los datos de tres empresas

En la figura podemos observar cómo se diferencian esos datos, en este caso como observamos, los datos de los CDs para cada uno de los días, el programa del cual se han extraído los datos, DataStream ofrece los datos directamente así, pudiendo posteriormente coger estos datos mediante Matlab y utilizarlos y dividir los datos según la empresa que queramos utilizar y estudiar en ese momento.

4.2.3 Información de Tweets

La información que se obtuvo de los tweets, como hemos mencionado en esta memoria reiteradamente ya no es accesible debido al cierre de las APIs de X, aun así, y con la esperanza de que la actual empresa X vuelva en un futuro algún tipo de acceso al estudio como de la API de la cual se obtuvieron los datos se va a proceder a una explicación exhaustiva de la extracción realizada.

4.2.- PROCESO V1 - MODELO DE DATOS

Para obtener los datos, seleccionamos el lenguaje de programación adecuado y definimos qué datos eran más relevantes para nuestro estudio. Decidimos guardar los datos en formato .csv, lo que facilita su procesamiento posterior y permite identificar rápidamente los formatos de cada campo.

Para poner en contexto en el siguiente extracto de código se puede observar cómo es el proceso de extracción de una manera rudimentaria y como este permite la selección de todos los datos de los tweets pudiendo enfocarse la obtención de esos tweets sobre una empresa en específica o sobre una mención la cual nos pueda interesar.

En nuestro caso se eligió realizar la extracción de estos datos mediante el lenguaje de programación Python, este es muy potente para la conectividad general de las APIs pues tiene una gran versatilidad y muchas bibliotecas para poder conectarse más fácilmente a las APIs que se requieran. Para poder extraer los datos como se ha comentado utilizaremos pues justamente la biblioteca llamada *"tweepy"*, además necesitaremos una cuenta de desarrollador de X para poder obtener las credenciales de acceso a la (API key, API secret key, Access token y Access token secret).

El caso del uso de estas APIs es un poco especial, mientras plataformas como OpenAI simplemente te dan una API key y te permiten obtener todos los datos, en el caso de X, nos encontramos los pasos excepcionales de la API secret key y los Access token, pues estos tokens son necesarios para realizar la solicitud en nombre del usuario en específico y autentificar que la solicitud proviene de una aplicación legítima y registrada. No se pretende ahondar más en este tema pues se aleja de la finalidad del proyecto.

```
# Importamos la librería necesaria que hemos comentado
import tweepy

# Configuramos de las credenciales de acceso a la API de X
api_key = "TU_API_KEY"
api_secret_key = "TU_API_SECRET_KEY"
access_token = "TU_ACCESS_TOKEN"
access_token_secret = "TU_ACCESS_TOKEN_SECRET"

# Nos auteticamos con la API de X
auth = tweepy.OAuthHandler(api_key, api_secret_key)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)


# Buscamos los tweets que queramos según la palabra clave
keyword = "Banco_A_Elegir"
tweets = api.search_tweets(q=keyword, lang="es", count=10)

# Datos obtenidos y guardados
for tweet in tweets:
    user_name = tweet.user.screen_name
    created_at = tweet.created_at
    tweet_text = tweet.text
    user_followers_count = tweet.user.followers_count
    user_location = tweet.user.location
```

En este caso simple de ejemplo del código podemos observar cómo simplemente estamos obteniendo los datos de nombre, fecha de creación, tweet, followers y localización, pero podemos obtener muchísimos más campos como los siguientes:

CAMPO	TIPO	DESCRIPCIÓN
tweet_id	String	ID del tweet
tweet_text	String	Texto del tweet
tweet_created_at	DateTIme	Fecha y hora de creación del tweet
tweet_geo	String	Información geográfica del tweet
tweet_place	String	Lugar asociado con el tweet
tweet_retweet_count	String	Número de retweets
tweet_favourite_count	String	Número de favoritos (likes)
tweet_lang	String	Idioma del tweet
tweet_user_id	String	ID del usuario
tweet_user_name	String	Nombre del usuario
tweet_user_location	String	Ubicación del usuario
tweet_user_description	String	Descripción del perfil del usuario
tweet_user_followers_count	String	Número de seguidores
tweet_user_friends_count	String	Número de amigo (seguidos)
tweet_user_created_at	String	Fecha de creación del perfil del usuario
tweet_favourites_count	String	Número de tweets favoritos del usuario
tweet_user_verified	String	Indicador de si el usuario está verificado
tweet_user_statuses_count	String	Número de tweets del usuario
tweet_user_profile_image	URL	URL de la imagen de perfil del usuario

Cuadro 4.2: Campos obtenidos por la API de X

Como podemos observar, acabamos recibiendo una gran cantidad de datos los cuales son susceptibles a poder utilizarse posteriormente si así lo deseamos. Estos datos como se ha comentado acaban guardándose en un archivo .csv que tiene el siguiente aspecto:

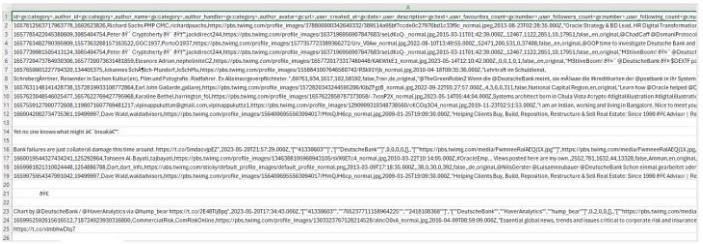


Figura 4.2: Ejemplo .csv de los datos de tres empresas

Aunque tiene un aspecto visual difícil de comprender debido a las características del archivo, es muy amigable con el tratado pues todos los datos se dividen en diferentes columnas, que en estos casos se encuentran estas diferenciadas por comas, mientras que en cada fila encontramos un tweet diferente relacionado con los datos que hemos obtenido.

Una vez analizados todos los campos que pueden sernos de utilidad, se ha llegado a la conclusión que, debido a futuras incorporaciones y a la posibilidad de no obtener los resultados requeridos, guardar la mayor cantidad de datos posibles para posteriormente poder usarlo como una nueva variable en caso de necesitarlo, y debido a que este aumento de datos no supone un gran peso en la volumetría.

4.2.4 Datos Procesados de los CDS (Processed Data CDS)

Esta sección pretende esclarecer de manera efectiva el diseño de las funciones realizadas, a partir del cual se pretende describir como van afectando estos datos refinados al cómputo global del proyecto. Como ha sucedido anteriormente el procesamiento de los datos se realiza de manera independiente para el proceso de los datos de los CDs en este apartado, y los datos de los Tweets que se explicarán siguiendo el mismo formato en el apartado siguiente.

Procederemos pues a describir las transformaciones que afectarán a los datos y como se realizan para pasar de los Raw Data a los Processed Data que queremos utilizar en partes más avanzadas del proyecto. Para una mayor comprensión, durante los siguientes apartados se realizará una aproximación detallada acerca de cada una de las transformaciones realizadas sobre el conjunto de datos originales, disponiendo de esta forma un conjunto de fases explicativas de cada una de estas transformaciones.

4.2.4.1 Fase 1

Para la fase inicial del procesamiento de estos datos, se busca principalmente obtener los datos los cuales obtuvimos de las plataformas pertinentes, para el caso que nos acontece en este apartado, tenemos la obtención de los archivos de los CDs, los cuales se obtendrán mediante el programa DataStream ya comentado donde conseguiremos los archivos correspondientes a cada empresa.

4.2.- PROCESO V1 - MODELO DE DATOS

Una vez obtenidos los datos procederemos a introducirlos en Matlab mediante el uso de funciones y ayudándonos de su bibliografía para poder obtener los datos directamente de los .xlsx o .csv. Aunque el código creado soporta ambos tipos de archivos, como se ha mencionado con anterioridad nos centraremos en su obtención mediante el archivo .xlsx mediante la siguiente parte de código:

```
[nombreArchivo, rutaArchivo] = uigetfile(nombre, 'Selecciona datos de bolsa');
archivoExcel = fullfile(rutaArchivo, nombreArchivo);

% Leer datos del archivo de Excel
[M_Data, M_Cell, M_raw] = xlsread(archivoExcel);
```

En este podemos observar cómo empezamos utilizando la función de Matlab "*uigetfile*", este se encarga de abrir un cuadro de diálogo que muestra el explorador de archivo del sistema en la carpeta actual, lo que permite al usuario seleccionar el archivo correspondiente, o en caso de existir el nombre que tiene por parámetro lo selecciona automáticamente. Y posteriormente obtenemos los parámetros del archivo, dónde se encuentra y el nombre que tiene.

Una vez obtenidos esos datos, directamente llamamos a la función de Matlab "*xlsread*" que se encarga automáticamente de leer el documento y extraerlo en filas y columnas en Matlab a través del formato *cell* en tres variables diferentes que usaremos dependiendo de nuestras necesidades, para el caso que nos interesa M_Cell donde nos da una visualización muy parecida a la del documento de Excel.

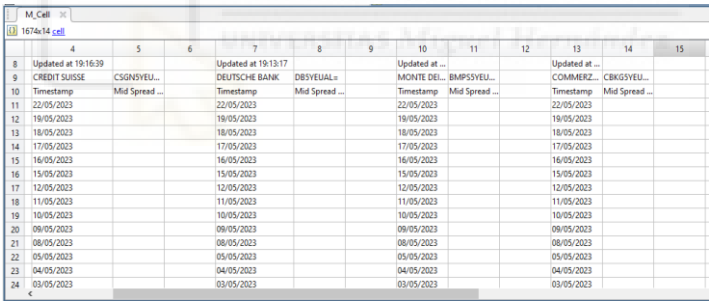


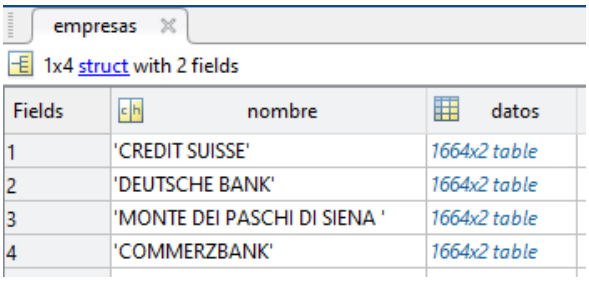
Figura 4.3: Visualización variable M_Cell

Por motivos de legibilidad en el documento no se agregarán grandes extensiones de código pues ese menester pertenece al apartado de los Anexos de este documento donde se puede vislumbrar en su plenitud el código utilizado durante todo el proceso, por ello se van a detallar los siguientes procesos realizados de manera lo más comprensible posible.

4.2.4.2 Fase 2

En esta segunda fase del procesamiento se busca obtener los datos correspondientes de la empresa a estudiar extrayendo los valores de las empresas y posteriormente seleccionando la que nos interesa entre todas las variables para obtener finalmente su resultado.

Una vez pues obtenidas las variables en Matlab y mediante la selección de la empresa comparándola con el archivo seleccionado, se realiza un bucle que recorre todas las variables separando de manera eficaz todas las empresas, dividiéndola en filas y columnas correspondientes a cada empresa, obteniendo de este proceso una variable de tipo "*struct*" donde se divide en dos columnas, la columna "*nombre*" que indica el nombre correspondiente de la empresa, y la columna "*datos*" que tiene los valores de CDs de cada empresa, teniendo de esta manera una función flexible y adaptable en el caso de que el documento introducido posea mayores empresas que quieran ser estudiadas posteriormente pudiendo seleccionar la que mejor nos convenga.


empresas

1x4 struct with 2 fields

Fields	nombre	datos
1	'CREDIT SUISSE'	1664x2 table
2	'DEUTSCHE BANK'	1664x2 table
3	'MONTE DEI PASCHI DI SIENA '	1664x2 table
4	'COMMERZBANK'	1664x2 table

Figura 4.4: Resultado obtención de datos CDs

Para esta selección, mediante el uso del listado de empresas que hemos obtenido en el "*struct*" se muestra un menú, o poniendo directamente un nombre similar a la empresa que deseaos seleccionar elegimos la empresa a estudiar y guardamos su valor en una variable llamada "*bolsa*" para continuar posteriormente con el proceso.

4.2.4.3 Fase 3

En esta última fase, se persigue realizar una limpieza de los valores de la variable bolsa, eliminando de esta manera posibles conflictos futuros derivado de falta de información o posibles errores en la obtención de los datos mediante el programa DataStream, además y debido a la naturaleza del uso de Matlab, se modificarán los campos de las variables de la "*bolsa*" para transformarlos en los datos más apropiados para su correcto uso a lo largo de todo el proyecto.

Comenzando con el proceso, y mediante indexación procederemos a sustraer cualquier valor que pueda tener un valor nulo, o como se muestra en Matlab NaN:

```
% Buscamos NaNs en todas las celdas
index = cellfun(@isnan,bolsa.Date,'uni',false);
% Ponemos "1" booleano en filas donde hayan NaN
index = cellfun(@any,index);

% Nos quedamos con las filas que no son NaNs
bolsa.Date = bolsa.Date(~index);
bolsa.Data = bolsa.Data(~index);
```

Limpiando de esta manera los datos para poder ser tratados sin dificultades. Una vez tenemos todos los datos filtrados y posicionados donde queremos únicamente queda transformar la columna de las fechas y pasarla a ser de tipo "*String*" a tipo "*DateTime*" pudiendo de esta manera posteriormente unificar los datos obtenidos durante este proceso con el posterior de los Tweets, permitiendo una gran versatilidad a la hora de eliminar datos de fechas, agregarlos u ordenarlos.

```
timezone = 'Europe/Madrid';
bolsa.Date = datetime(bolsa.Date,"Format","dd/MM/yyyy",'TimeZone',timezone);
```

4.2.5 Datos Procesados de los Tweets (Processed Data Tweets)

Como se ha comentado en el apartado anterior, en este procederemos a explicar igualmente en sus fases correspondiente el proceso desde la obtención de los datos en Raw hasta obtener los datos Processed. Debido a que el apartado de la obtención de los tweets y al ser el enfoque principal del proyecto, llevará un mayor trabajo de procesamiento, pues además de extraer los datos correspondientes, deberemos agregar el sentimiento y especificar las variables pertinentes dependientes de todas las variables que hemos podido ver en la extracción completa de todos los parámetros posibles que se pueden obtener mediante la API.

4.2.5.1 Fase 1

En esta fase inicial del procesamiento, similar a lo realizado en el apartado anterior, procedemos a obtener los datos los cuales hemos obtenido del archivo .csv como hemos podido observar en la Figura 4.2, y como se ha comentado se ha realizado mediante el uso de un programa de extracción de datos en Python mediante una API.

Una vez obtenidos los datos procederemos a cargarlos en el programa de Matlab mediante el uso de funciones personalizadas. Como podemos observar, este proceso también soporta archivos de tipo .xlsx aunque el uso común es el de los .csv como hemos comentado.

```
% Solicitar al usuario el archivo de Excel o CSV
[filename, path] = uigetfile(nombre, 'Seleccionar archivo de Tweets');
filepath = fullfile(path, filename);

% Realizar la sustitución de caracteres inválidos en los nombres de las columnas
opts = detectImportOptions(filepath);
opts.VariableNames = matlab.lang.makeValidName(opts.VariableNames);
```

En este trozo de código observamos como iniciamos el proceso con la función de Matlab "*uigetfile*", donde como ya se ha mencionado es el encargado de mostrar un cuadro de diálogo en el cual mostrar y de esta manera poder seleccionar el archivo correspondiente que se quiere utilizar.

Posteriormente a esto realizaremos una sustitución de los caracteres inválidos que puedan existir como nombre de columna, esto nos facilitará posteriormente cambiarles el nombre a las columnas dependiendo de su característica y que valores tenga ayudándonos así a poder definirlas de una manera fija para identificar cada una de manera más sencilla facilitándonos su procesamiento más adelante.

```
    processedData = struct();
    processedData.filename = filename;
    processedData.path = path;
    processedData.data = readtable(filepath,opts);

    processedData.data.Properties.VariableNames{'date_gx_date_'} = 'created_at';
    processedData.data.created_at       =       datetime(processedData.data.created_at,
'InputFormat', 'yyyy-MM-dd''T''HH:mm:ss.SSSX', 'TimeZone', 'local');
    processedData.data.Properties.VariableNames{'text_gx_text_'} = 'text';
    processedData.Data = processedData.data;
    processedData = rmfield(processedData, 'data');
```

Como observamos en este código hacemos una transformación completa de como recibimos los datos, devolviéndolos tras la ejecución de estar función llamada "*procesarTweetSentNuevos*" de una manera completamente diferente. Para empezar, creamos un struct donde guardaremos toda la información de interés con las variables necesarias, además modificaremos el nombre de algunas columnas haciendo que estas sean genéricas tanto para los datos de los CDs como para los datos de los Tweets.

Finalizando este proceso le daremos un formato de "*datetime*" a la variable del tiempo pues de esta manera conseguimos podemos organizar todas las variables de una manera más eficaz y escalable, esto permite representar los puntos en el tiempo especificándolo por año, mes, día, hora, minuto y segundo, permitiendo una flexibilidad para el procesamiento mayor que la que teníamos antes con un formato plano.

4.2.5.2 Fase 2

Una vez que los datos han sido obtenidos en Matlab, habiendo de esta manera extraído directamente todas las variables relevantes. Donde además se incluye el nombre del archivo, la fecha de cada tweet y varias métricas importantes como el número de seguidores, verificación, retweets, citas, respuestas, etc. Así es como podemos observarlo en la siguiente Figura.

processedTweets.Data

	1 id_gx_category_	2 author_id_gx_category_	3 author_name_gx_category_	4 author_handler_gx_category_	5 author_avatar_gx_url_	6 user_created_at_gx_date_	7 user_descripti
1	1.6578e+18	1.6926e+09	'Richard Sachs PMP CMC'	'richardpsachs'	'https://pbs.twimg.com/...	'2013-08-23T02:28:35.000Z'	'Oracle Strategy
2	1.6578e+18	3.0854e+09	'Peter □ Cryptoherty □ □'	'jackdirect244'	'https://pbs.twimg.com/...	'2015-03-11T01:42:39.000Z'	''
3	1.6578e+18	1.5574e+18	'GGC1937'	'PortoG1937'	'https://pbs.twimg.com/...	'2022-08-10T13:49:55.000Z'	''
4	1.6577e+18	3.0854e+09	'Peter □ Cryptoherty □ □'	'jackdirect244'	'https://pbs.twimg.com/...	'2015-03-11T01:42:39.000Z'	''
5	1.6577e+18	1.6577e+18	'Eleanore Adrian'	'nepheliniteCZ'	'https://pbs.twimg.com/...	'2023-05-14T12:10:42.000Z'	''
6	1.6577e+18	134405375	'Johannes Schölch-Mundorf'	'JoSchMu'	'https://pbs.twimg.com/...	'2010-04-18T09:35:36.000Z'	'Lehrkraft im Sch
7	1.6576e+18	1.5728e+18	'Earl John Gallarde'	'gallarej'	'https://pbs.twimg.com/...	'2022-09-22T05:27:57.000Z'	''
8	1.6576e+18	1.6576e+18	'Karaline Bethel'	'harrington_fol'	'https://pbs.twimg.com/...	'2023-05-14T05:44:04.000Z'	'Systems archite
9	1.6576e+18	1.1981e+18	'vipinappukuttan@gmail.com'	'vipinappukutta1'	'https://pbs.twimg.com/...	'2019-11-23T02:51:53.000Z'	'I am an Indian,
10	1.6600e+18	19499997	'Dave Wald'	'waldadvisors'	'https://pbs.twimg.com/...	'2009-01-25T19:09:30.000Z'	'Helping Clients
11	1.6600e+18	125292864	'Tahseen Al-Bayati'	'tajbayati'	'https://pbs.twimg.com/...	'2010-03-22T10:14:05.000Z'	'#OracleEmp... V
12	1.6600e+18	1.2549e+09	'Dart'	'dart_info'	'https://pbs.twimg.com/...	'2013-03-09T17:18:35.000Z'	''
13	1.6600e+18	19499997	'Dave Wald'	'waldadvisors'	'https://pbs.twimg.com/...	'2009-01-25T19:09:30.000Z'	'Helping Clients
14	1.6600e+18	7.1872e+17	'CommercialRisk'	'ComRiskOnline'	'https://pbs.twimg.com/...	'2016-04-09T08:59:09.000Z'	'Essential global
15	1.6599e+18	1.1493e+18	'Principe D''Italia Della TERRA ...	'principe_dei'	'https://pbs.twimg.com/...	'2019-07-11T10:18:05.000Z'	'Emperor of the
16	1.6599e+18	1.1493e+18	'Principe D''Italia Della TERRA ...	'principe_dei'	'https://pbs.twimg.com/...	'2019-07-11T10:18:05.000Z'	'Emperor of the

Figura 4.5: Visualización variables processedTweets

Con los datos importados, el siguiente paso crucial es extraer y organizar las variables de interés, siendo este proceso fundamental para transformar los datos en bruto en información estructurada y útil para el análisis posterior.

```
    % Extracción de variables de interés
    twt.Name = strrep(processedTweets.filename, '.csv', '');
    twt.Date_by_tweet                                                    =
datetime(dateshift(processedTweets.Data.created_at,'start','day'),
"Format","dd/MM/yyyy",'TimeZone','Europe/Madrid');
    twt.Followers_by_tweet = processedTweets.Data.user_followers_count_gx_number_;
    twt.Verified_by_tweet = processedTweets.Data.user_verified_gx_boolean_;
    twt.Retweets_by_tweet = processedTweets.Data.retweets_gx_number_;
    twt.Quotes_by_tweet = processedTweets.Data.quotes_gx_number_;
    twt.Replies_by_tweet = processedTweets.Data.replies_gx_number_;
    twt.Number_tweet_by_user = processedTweets.Data.user_tweets_count_gx_number_;
```

En este código, comenzamos con la obtención del nombre del archivo de donde se han importado los tweets. Esto es útil para rastrear el origen de los datos. Posteriormente usamos la función "*strrep*" de Matlab para eliminar la extensión .csv del nombre del archivo. Cabe destacar que este proceso está detallando el caso del uso de la extensión .csv como se había comentado.

Una vez esto, obtenemos los datos y los distribuimos en diversas variables dentro del struct que en este caso hemos llamado "*twt*" donde guardamos todas las variables de interés que utilizaremos más adelante, como "*Followers_by_tweet*" o "*Verifies_by_tweet*". Cada una de estas líneas de código se encargan de extraer una métrica usada posteriormente.

4.2.5.3 Fase 3

En esta última fase, se realiza uno de los aspectos más importantes de todo el proceso de extracción de los tweets que es el Tokenizado y la generación del sentimiento.

```
    % Tokenización y cálculo del sentimiento
    twt.Documents_by_tweet                                               =
tokenizedDocument(processedTweets.Data.text,'Language','en');
    twt.Sentiment_by_tweet = vaderSentimentScores(twt.Documents_by_tweet);
```

Como se puede observar en el código, antes de realizar el análisis de sentimiento, es fundamental preparar los textos de los tweets, en este caso, usamos para llevar este menester a cabo, la función "*tokenizedDocument*" de Matlab para convertir todos los textos de los tweets en documentos tokenizados, este proceso como se ha explicado en los inicios de esta memoria divide el texto en unidades manejables (tokens) y se ha configurado el idioma inglés pues los tweets estaban escritos en este idioma.

Una vez tokenizados los textos, podemos proceder con el análisis de sentimientos, esto se realiza con una herramienta popular para este análisis llamada VADER (Valence Aware Dictionary and Sentiment Reasoner), que es usada especialmente en textos cortos como

tweets. VADER es un léxico y un algoritmo de reglas que se ajustan bien a los sentimientos expresados en redes sociales. Con esta función se le da un puntaje de sentimiento a cada tweet tokenizado donde se le asigna un puntaje desde -1 (muy negativo), con un puntaje de 0 (neutro) hasta un puntaje de 1 (muy positivo).

Finalmente, los datos de sentimiento procesados se almacenan junto con las otras variables de interés y una vez completado este paso se han transformado los datos en bruto de los tweets en un conjunto organizado y estructurado de variables listas para análisis avanzados. Esta organización garantiza que todos los datos relevantes estén fácilmente accesibles y preparados para su integración con otras partes del proyecto.

4.3 Proceso v2 - Unificación de datos

En este capítulo se profundizará en la unificación de datos provenientes de diferentes fuentes, con el objetivo de integrar y analizar de manera conjunta la información previamente procesada. Este proceso es crucial para obtener una visión completa de los datos, permitiendo de esta manera realizar a futuro un análisis más complejo y detallado.

El principal objetivo de este proceso es combinar y alinear temporalmente los datos históricos de las empresas, obtenidas a través de la herramienta DataStream y ya filtrada en el proceso anterior de CDs, con los datos de sentimiento derivados de los tweets asociados a estas mismas empresas. Este proceso nos permitirá evaluar cómo las variaciones en los datos financieros están correlacionadas con las percepciones y sentimientos expresados en las redes sociales. Además, se busca identificar y analizar las fechas clave que puedan tener un impacto significativo en los datos, proporcionando un contexto adicional para la interpretación de resultados.

Para llevar a cabo la unificación de datos, este capítulo se dividirá en dos secciones principales. Una primera que será las Fechas Clave que en esa sección se identificará y seleccionará un conjunto de fechas relevantes para el análisis, donde estas fechas pueden coincidir con eventos significativos de la empresa como anuncios importantes o cambios en la logística. Y por otro lado la propia Unificación de datos, esta segunda sección se centrará en la integración de los datos históricos de la empresa con los datos del sentimiento, asegurando una correcta alineación temporal y la creación de tablas unificadas que permitan un análisis conjunto.

4.3.1 Fechas clave

La identificación de fechas clave es un paso crucial en el análisis de los datos financieros y del sentimiento, ya que permite contextualizar los eventos y su impacto en el comportamiento del mercado. El primer paso en la identificación de fechas clave consiste en determinar los eventos que podrían tener un impacto significativo en los datos, estos datos pueden ser de una naturaleza variada pero la manera más eficiente de identificarlos es mediante el análisis de datos históricos o noticas de gran peso que hayan sucedido.

El análisis de fechas específicas implica procesar los datos en segmentos temporales alrededor de cada fecha clave. Este enfoque permite observar cómo los eventos afectan a los datos antes y después de su ocurrencia. Para ello se ha utilizado un intervalo temporal definido de 15 días antes y después de cada fecha clave. Estas listas se almacenan en una lista para su posterior procesamiento.

```
num_interval = 15;
fechas = {'07/07/2019', '15/09/2020', '17/03/2019', '03/07/2020', '30/04/2021'};
fechas = datetime(fechas, 'InputFormat', 'dd/MM/yyyy');
Table_Datos = cell(size(fechas)); % Inicializa prueba como una matriz de celdas
```

Una vez obtenidas estas fechas, por cada fecha clave, se utiliza la función "*filtradoFecha*" para obtener los datos correspondientes al intervalo definido. Esta permite seleccionar un intervalo o seleccionar manualmente dicho intervalo.

```
for i = 1:numel(fechas)
    fechaInicio = fechas(i) - days(num_interval);
    fechaFin = fechas(i) + days(num_interval);
    Table_Datos{i} = filtradoFecha(FullTable_Bolsa, fechaInicio, fechaFin);
```

La función realiza el filtrado de datos en función de un intercalo de tiempo proporcionando flexibilidad, en un inicio esta convierte la columna Date en datetime para facilitar el manejo y comparación de fechas en caso de que este proceso no se hubiese hecho con anterioridad. Una vez realizada la conversión se obtienen las fechas máximas y mínima y se asegura de que todas las fechas coincidan en zona horaria.

Posterior a esto la función entra en un bucle que realizar una verificación de los argumentos y datos introducidos y se asegura que las fechas sean validas y estén dentro del rango permitido por las variables de los CDs y tweets. Cuando se obtiene este resultado se limpian y filtran y devuelve los resultados correspondientes.

Una vez obtenidos los datos del filtrado se utilizan para entrenar y validar los modelos de regresión y clasificación para evaluar el comportamiento de las variables de interés alrededor de las fechas que se han seleccionado como clave.

4.3.2 Unificación de datos

La unificación de datos es el paso más importante y crucial pues se usa para combinar diferentes fuentes de información y obtener una visión integrada. Este proceso permite analizar de manera conjunta los datos de los CDS y sentimiento, facilitando la identificación de patrones y la realización de análisis complejos.

En primera instancia, se procederá a una carga de datos previamente procesados en el Proceso v1 donde se estructurará para su unificación.

```
load(fullfile(fullfile(pwd, 'Variables/Proceso_v1_Bolsa.mat')));
load(fullfile(fullfile(pwd, 'Variables/Proceso_v1_Tweets.mat')));
```

Una vez obtenidos estos datos, se crea una tabla unificada con las variables temporales de los datos de los CDS y del sentimiento, permitiendo así obtener los datos que coinciden en ambos y no obtener errores en la unificación, una vez obtenidos se procede a ordenarlos de manera ascendente por fecha cada una de las variables.

Posteriormente ocurre uno de los procesos más cuidadosos para que todos los datos coincidan y no perder información relevante. Las variables de bolsa y sentimiento se alinean temporalmente y se crean tablas unificadas que contienen ambos tipos de datos. Se calculan las fechas mínimas y máximas comunes entre los datos de la variable bolsa y sentimiento asegurando que solo se consideren los periodos en los que ambos conjuntos de datos están disponibles. Y se procede a generar un listado que abarque estas fechas.

Cuando ya se han calculado todos los procesos para tener un conjunto de fechas correspondiente y ordenadas, se crea una tabla unificada que contiene tanto los datos de “bolsa” como de “sentimiento” ya alineados temporalmente.

```
    % Crear la tabla unificada
    FullTable = table('size', [length(Date_List), width(SentTable) + 1],
'VariableTypes', [{'datetime'}, repmat({'double'}, 1, width(SentTable) - 1),
{'double'}]);
    FullTable.Properties.VariableNames = [SentTable.Properties.VariableNames,
'Bolsa'];
    FullTable.Date = Date_List;
```

Para el caso, FullTable que es nuestra tabla unificada se crea con el tamaño adecuado para contener todas las fechas y datos correspondientes de sentimiento y bolsa, donde se especifica el tamaño de la table con el número de fechas de "*DateList*" y el número de columna, siendo este igual a las de "*SentimentTable*" más una siendo esta última el valor de los datos de la bolsa.

Finalmente, se filtran las fechas sin datos de bolsa y se crean variables adicionales para facilitar el análisis, este es el caso de por ejemplo las variables de "*sube*", "*baja*", "*estable*", "*CLASE*". Donde indican el valor de cada acción y "*CLASE*" siendo una variable categórica que toma el valor de 1 si sube, 0 si es estable o -1 si baja.

4.4 Proceso v3 - Análisis de Datos

En esta sección se abordará el análisis de los datos obtenidos y procesados en etapas anteriores. El objetivo principal es aplicar técnicas analíticas y de procesamiento que permitan extraer la información valiosa, identificar patrones y relaciones entre las variables, y evaluar la precisión de los modelos predictivos desarrollados.

El Proceso v3 y como se mencionó anteriormente, se dividirá en dos secciones fundamentales, siendo la primera "Análisis Exploratorio Básico" y una segunda parte que es "Análisis y Procesado por Secuencia". Estas secciones fueron diseñadas para ofrecer una comprensión profunda de los datos y proporcionar una base sólida para la implementación de los modelos predictivos y clasificadores que veremos en los siguientes puntos.

En el primer proceso se realizarán diversas técnicas de análisis exploratorio para comprender la estructura y las características de los datos, se realiza una exploración inicial de las variables mediante la generación de mapas de calor, la relación mutua de las varia de las variables a través de diagramas de dispersión XY, se generarán histogramas para visualizar la distribución de las variables y se realizará una representación cruzada para seleccionar las variables por clases.

En cuanto al segundo proceso, profundiza en el análisis de los datos mediante la implementación de modelos de regresión y clasificadores. Se comenzará con la aplicación de una regresión gaussiana para analizar la relación entre los datos de los CDS y todas las demás variables. Además, se empleará un clasificador de árbol para categorizar los datos según las variables que se seleccionen. Obteniendo finalmente los resultados para evaluar la precisión de los modelos.

4.4.1 Análisis Exploratorio Básico

Es una etapa importante del proyecto pues nos permite conocer de antemano antes de iniciar con estudios más complejos como se estructuran nuestros datos y cuál es el posible camino que seguir conociendo las características disponibles. En esta sección se realizan diversos análisis y visualizaciones que facilitan la identificación de patrones, relaciones y distribución de las variables involucradas.

Procediendo con la exploración, se procede a carga los datos y su estructuración previamente procesados en el Proceso v2. Se utilizarán las variables resultantes de este proceso "*FullTable*" que tiene la unificación de los datos y la variable "*info*" que contienen datos de interés de las variables e información como por ejemplo el uso horario utilizado anteriormente.

4.4.1.1 Correlación Lineal

El siguiente paso una vez asegurado nuestro entorno de trabajo y seleccionadas las variables necesarias, procederemos a llevar a cabo el análisis exploratorio, comenzando con la correlación lineal entre las diferentes variables. Esto se logra mediante la generación de un mapa de calor que visualiza las correlaciones entre todas las variables de la variable "*FullTable*".

```
clf; figure;
info.heatMap = heatmap(corrcoef(Tabla_Datos{:,2:end}));
info.heatMap.XData = Tabla_Datos.Properties.VariableNames(2:end);
info.heatMap.YData = Tabla_Datos.Properties.VariableNames(2:end);
```

Esta correlación nos permite identificar relaciones directas entre pares de variables proporcionando una visión inicial de posibles dependencias. Podemos observar en el código de arriba como se realiza dicho mapa de calor.

4.4.1.2 Relación Mutua de Variables

Una vez realizado esa prueba procedemos a utilizar la Relación Mutua de Variables, para explorar más a fondo las relaciones entre las variables. Para ello se crea un diagrama de dispersión XY (scatter plots) que muestran las relaciones mutuas entre todas las variables.

```
% Crear un layout de subplots en formato de tabla
num_variables = length(Tabla_Datos.Properties.VariableNames) - 1;
figure;

for a = 1:num_variables
    for b = 1:num_variables
        % Determinar posición del subplot
        position = (b-1) * num_variables + a;
        subplot(num_variables, num_variables, position);

        % Crear scatter plot
        scatter(Tabla_Datos{:, a+1}, Tabla_Datos{:, b+1}, '.b');
        set(gca, 'XTick', []);
        set(gca, 'YTick', []);

        % Establecer etiquetas solo en los bordes izquierdo e inferior
        if b == num_variables
            xlabel(Tabla_Datos.Properties.VariableNames{a+1});
        end
        if a == 1
            ylabel(Tabla_Datos.Properties.VariableNames{b+1});
        end
    end
end
```

Estos diagramas permiten visualizar posibles patrones o tendencias entre pares de variables, proporcionando una visión más detallada de sus interacciones.

4.4.1.3 Histograma de Variables

Para completar la distribución de variables individuales, se generan unos histogramas que muestras la frecuencia de los valores de cada variable. Estos histogramas permiten visualizar la distribución de los datos, identificando posibles sesgos, asimetrías o valores atípicos.

```
% Obtener el número total de variables
numVariables = length(Tabla_Datos.Properties.VariableNames) - 1;

% Calcular el número de filas y columnas necesarias
numFilas = ceil(sqrt(numVariables));
numColumnas = ceil(numVariables / numFilas);

% Crear la figura
figure;
k = 1;
for a = 1:numVariables
    subplot(numFilas, numColumnas, k)
    histogram(Tabla_Datos{:, a+1}, 20);
    title(Tabla_Datos.Properties.VariableNames(a+1))
    k = k + 1;
end
```

4.4.1.4 Representación Cruzada

Finalmente, se realizará una representación cruzada de selección de variables dividida por clases, permitiendo de esta manera analizar cómo se distribuyen las variables en función de diferentes categorías. Esto se logrará mediante la creación de diagramas de dispersión que muestren la relación entre dos variables específicas puesta en cada uno de los ejes diferenciadas por la variable "*CLASE*".

```
Eje_X = 2;
Eje_Y = 4;

figure();hold on;
scatter(FullTable_Bolsa{Tabla_Datos.CLASE==-
1,Eje_X},FullTable_Bolsa{Tabla_Datos.CLASE==-1,Eje_Y},'.r')
scatter(FullTable_Bolsa{Tabla_Datos.CLASE==0,Eje_X},FullTable_Bolsa{Tabla_Datos.CL
ASE==0,Eje_Y},'.k')
scatter(FullTable_Bolsa{Tabla_Datos.CLASE==1,Eje_X},FullTable_Bolsa{Tabla_Datos.CL
ASE==1,Eje_Y},'.b')
xlabel(FullTable_Bolsa.Properties.VariableNames{Eje_X});ylabel(FullTable_Bolsa.Pro
perties.VariableNames{Eje_Y})
legend('Clase -1', "clase 0", "Clase 1");
```

Esta representación cruzada permitirá visualizar cómo se distribuyen las variables seleccionadas en función de las clases, proporcionando una comprensión más profunda de las diferencias y similitudes entre las categorías.

En cuanto al segundo proceso, profundiza en el análisis de los datos mediante la implementación de modelos de regresión y clasificadores. Se comenzará con la aplicación de una regresión gaussiana para analizar la relación entre los datos de los CDS y todas las demás variables. Además, se empleará un clasificador de árbol para categorizar los datos según las variables que se seleccionen. Obteniendo finalmente los resultados para evaluar la precisión de los modelos.

4.4.2 Análisis y Procesado por Secuencia

Este apartado nos permite llevar a cabo un análisis más profundo y detallado de los datos, aplicando técnicas de modelado y clasificación que permitan extraer información valiosa y realizar predicciones precisas. Se han utilizado modelos avanzados de regresión y clasificación, y se representan de manera que se pueda evaluar la eficacia de estos modelos y su capacidad de predecir valores futuros.

Al igual que en el proceso anterior, se procede a hacer una carga de los datos procesados en las etapas anteriores, por lo tanto, se guardarán igualmente las variables correspondientes de Proceso v2. Esta preparación asegura que todos los datos necesarios estén disponibles para un análisis más exhaustivo y que el entorno de trabajo sea limpio y claro para el procesamiento de los modelos.

4.4.2.1 Regresión Gaussiana

El primer análisis realizado será una regresión gaussiana que permita evaluar la relación entre los datos de la variable bolsa y todas las demás variables. Este tipo de regresión nos permite modelar las relaciones no lineales y capturar patrones complejos que puedan darse en los datos.

Este aspecto tiene mayor complejidad y se decidió realizar dos funciones para realizar un análisis más completo y que sea escalable para diversos casos, una primera llamada "*casoEstudioSVMGauss*" que se centra en dar las directrices para calcular el SVM de Gauss y liego otra función que hay dentro de esta que es una llamada "*RegresorLinealSVM*", que realiza un análisis de regresión utilizando "*Support Vector Machine (SVM)*" para diferentes conjuntos de variables y diferentes días de historial. Además, se encarga de calcular y comparar el "*Root Mean Squared Error (RMSE)*" para cada modelo y se selecciona el modelo RMSE más bajo y que por tanto ha tenido mejores resultados.

Por lo tanto, una vez inicializadas las variables de los datos se procede a preparar las variables de entrenamiento para diferentes días de historial.

```
VariableEstudio = {'Intensity', 'CLASE'};

VariableEntrenamiento{1} = {'Intensity', 'Sentimiento'};



for n = 2:diasAtras
    % Crear una celda para la iteración actual
    celda_iteracion = VariableEstudio;

    % Agregar los componentes
    for i = 1:n-1
        celda_iteracion    =    [celda_iteracion,    strcat(VariableEstudio,    '-',
num2str(i))];
    end
```

```
        % Eliminar la columna específica si coincide
        if any(strcmp(celda_iteracion, columna))
            celda_iteracion(strcmp(celda_iteracion, columna)) = [];
        end

        % Guardar la celda de la iteración actual en VariableEntrenamiento
        VariableEntrenamiento{n} = celda_iteracion;
    end
```

Como observamos en el código, en cada iteración se crean conjuntos de variables de entrenamiento que incluyen la variable objetivo y otras variables relevantes, excluyendo la columna específica donde se coincide.

Una vez realizado esto, se procede con el entrenamiento de Regresión, donde se entrena un modelo para diferentes días de historial utilizando la función que habíamos comentado.

```
    [trainedModel,    validationRMSE(n)]    =    RegresorLienalSVM(FulltableBolsa_nxD,
VariableEntrenamiento{n}, columna);
```

Dentro de esta que estará adjuntada junto con todas las funciones y código al final de esta memoria, se calcula el RMSE de validación para cada modelo y se guarda el modelo entrenado junto con su RMSE.

Estos procesos están dentro de la función completa divididos en tres, un caso de estudio de Solo Bolsa, otro caso de estudio de Solo Sentimientos y un último caso de estudio de Sentimiento y Bolsa. De esta manera se realizan procesos similares a los comentados para cada una de las variables correspondientes en cada apartado. Una vez realizados, se procede a la identificación del mejor modelo obteniendo y guardando los resultados correspondientes.

4.4.2.2 Clasificador de Árbol

El segundo análisis es aplicar un clasificador de árbol para categorizar los datos según las variables seleccionadas. Los clasificadores de árbol son una herramienta muy eficaz para clasificar los datos, ya que pueden manejar variables tanto numéricas como categóricas y son fáciles de interpretar.

Al igual que en el regresor, se siguió una estrategia similar para su desarrollo, se generó una función que poseía tres diversos casos, el estudio con Solo Bolsa, que únicamente utiliza las variables de los CDS, el Solo Sentimiento, que realiza lo mismo, pero únicamente con dichas variables de sentimiento y un último caso conjunto de Bolsa y Sentimiento. En el calificador las funciones fueron "*casoEstudioClassifierTree*" y la función que realizaba el clasificador "*trainClassifierTree*".

Una vez comprendido esto y llamando a la función del caso de estudio específico, se procede a la preparación de las variables de entrenamiento, estas se preparan para diferentes días del historial.

```
    VariableEstudio = {'Intensity', 'CLASE'};
    VariableEntrenamiento{1} = {'Intensity', 'Sentimiento'};

    for n = 2:diasAtras
        % Crear una celda para la iteración actual
        celda_iteracion = VariableEstudio;

        % Agregar los componentes
        for i = 1:n-1
            celda_iteracion    =    [celda_iteracion,    strcat(VariableEstudio,    '-',
num2str(i))];
        end

        % Eliminar la columna específica si coincide
        if any(strcmp(celda_iteracion, columna))
            celda_iteracion(strcmp(celda_iteracion, columna)) = [];
        end

        % Guardar la celda de la iteración actual en VariableEntrenamiento
        VariableEntrenamiento{n} = celda_iteracion;
    end
```

Donde en cada iteración se crean conjuntos de variables de entrenamiento que incluyen las variables objetivo y otras variables relevantes para el estudio.

Como hemos comentado los procesos siguientes se calculan cada vez para los diferentes casos que hemos citado. Se comienza pues con el entrenamiento del Clasificador.

```
    for n = 1:diasAtras
        FulltableBolsa_nxD = calculoBolsaMemoria(n, trainingData);
        [trainedClassifier,                  validationAccuracy(n)]                  =
trainClassifierTree(FulltableBolsa_nxD, VariableEntrenamiento{n}, columna);
        trainTreeClass.SoloBolsa{n} = trainedClassifier;
        trainTreeClass.SoloBolsa{n}.validacionAcc = validationAccuracy(n);
    end

    validacionAcc.SoloBolsa = validationAccuracy;
```

Se entrena para diversos días utilizando la función "*trainClassifierTree*". Donde en esta al contrario que el Regresor, se calcula el Accuracy de validación para cada modelo y se guarda el modelo junto con su Accuracy correspondiente.

Como podemos observar en el código este caso que es el SoloBolsa se guarda en la variable correspondiente. Una vez obtenidos todos los procesos que se debían calcular, se procede a realizar una identificación del mejor modelo posible.

```
% Inicializar variables para el máximo y su posición
max_valor = -Inf;
max_estructura = '';
posicion_max = NaN;

% Iterar sobre las estructuras dentro del struct
campos = fieldnames(validacionAcc);
for i = 1:numel(campos)
    % Obtener los valores de validación de la estructura actual
    valores = validacionAcc.(campos{i});

    % Encontrar el valor máximo de esta estructura
    [max_valor_actual, max_posicion_actual] = max(valores);

    % Actualizar si se encuentra un nuevo máximo
    if max_valor_actual > max_valor
        max_valor = max_valor_actual;
        max_estructura = campos{i};
        posicion_max = max_posicion_actual;
    end
end

trainTreeClass.datosAcc.max_valor = max_valor;
trainTreeClass.datosAcc.max_estructura = max_estructura;
trainTreeClass.datosAcc.posicion_max = posicion_max;
trainTreeClass.datosAcc.columna = columna;
```

Encontrando de esta manera el modelo con el Accuracy más alto entre los modelos entrenados, guardando la información.

4.4.2.3 Representación de resultados

Finalmente, se procede a la representación de los resultados obtenidos de los modelos de regresión y clasificador que mejores resultados han ofrecido. Esto incluye la comparación de las predicciones con los datos reales, lo que permitirá evaluar la precisión y eficacia de los modelos actuales desarrollados y el rendimiento que tienen cada uno de ellos.

4.5 Proceso v4 – Análisis Exhaustivo

En esta sección se abordará el análisis de los datos utilizando diferentes técnicas de clasificación y regresión. El objetivo principal es aplicar estos métodos para extraer información valiosa, identificar patrones y evaluar la precisión de los modelos predictivos desarrollados.

El Proceso v4, se dividirá en tres partes importantes, una primera siendo "Análisis de Clasificador", una segunda parte "Análisis de Regresión", y por último una tercera siendo esta, el "*Análisis de Regresión con PCA*". Estas secciones han sido desarrolladas con el fin de obtener unos resultados concluyentes, profundizando en el desarrollo de cada función y como exprimir todo su potencial para tener posteriormente una sólida implementación de los modelos predictivos y clasificadores que veremos en los siguientes puntos.

4.5.1 Análisis de Clasificador

En esta primera sección se detalla el análisis de un clasificador mediante un enfoque paso a paso. Este análisis tiene como objetivo identificar patrones y evaluar la precisión del modelo usando diversas variables de entrenamiento y técnicas de equilibrio de datos.

El primer caso y como ya va siendo recurrente en estos procesos, es realizar la carga de datos que se han procesado en etapas anteriores del proyecto. Esto asegura que todos los datos necesarios estén disponibles y no haya errores a la hora de su uso, siendo los correctos para realizar el análisis y tener un entorno de trabajo limpio y listo para el procesamiento.

4.5.1.1 Estructura de datos

Iniciando con los procesos, se realiza la etapa de estructuración de datos, en esta etapa, se genera una matriz de datos históricos que incluye 30 días de historial para cada observación. Este conjunto de datos posteriormente se divida en dos subconjuntos, uno de entrenamiento y otro para pruebas. Esta división es crucial hacerla de manera correcta para evaluar un rendimiento real del modelo de datos, obteniendo unos resultados sobre datos no vistos durante el entrenamiento.

```
    n = 30;
    [BD.FulltableBolsa_nxD] = CrearTablaDatosHistoricos(n, FullTable_Bolsa);
    BD.Split = 0.8;
    BD.Train_Set  =  BD.FulltableBolsa_nxD(1:floor(length(BD.FulltableBolsa_nxD.Date)  *
BD.Split), :);
    BD.Test_Set   =   BD.FulltableBolsa_nxD(floor(length(BD.FulltableBolsa_nxD.Date)   *
BD.Split) + n:end, :);
```

La generación de esta matriz permite crear una base sólida para el análisis, considerándose un historial suficiente de datos para capturar posibles tendencias y patrones relevantes.

Seleccionar las variables adecuadas es fundamental para el rendimiento del modelo, es por ello, que lo siguiente que se realiza es la elección de las variables objeto y las variables de entrada que se utilizarán en el entrenamiento del modelo. Se diferencian entre variables de sentimiento y variables de bolsa, ajustando la selección según la relevancia de cada una.

```
BD.Variables_Seleccionadas = [BD.Var.Intensity, BD.Var.Sentimiento,
BD.Var.meanSent, BD.Var.medianSent, BD.Var.Std, BD.Var.xNumTweets, BD.Var.xRetweets,
BD.Var.xFollowers, BD.Var.xVerified, BD.Var.bolsa, BD.Var.bolsaVariacion,
BD.Var.bolsaNorm, BD.Var.sube, BD.Var.baja, BD.Var.estable, BD.Var.CLASE];
```

Esta selección de variables permite ajustar el modelo para que se enfoque en las características más relevantes, consiguiendo de esta forma mejorar así su capacidad predictiva.

4.5.1.2 Equilibrado de muestras

El equilibrio de clases es crucial en problemas en clasificación, especialmente cuando hay una desproporción significativa entre clases. En este paso, se equilibran las clases mayoritarias y minoritarias en el conjunto de entrenamiento para asegurar que el modelo no esté sesgado hacia la clase que pueda ser más frecuente.

```
Equilibrado = true;

if Equilibrado
    POS = Train_Set(Train_respueta == 1, :);
    NEG = Train_Set(Train_respueta ~= 1, :);
    if sum(Train_respueta == 1) > sum(Train_respueta ~= 1)
        Invert = 1;
        H = POS;
        L = NEG;
    else
        Invert = 0;
        H = NEG;
        L = POS;
    end
    indic = randperm(length(H{:, 1}));
    H = H(indic, :);
    HL = [H(1:length(L{:, 1}), :); L];
    clase = ones(2*length(L{:, 1}), 1); clase(1:length(L{:, 1})) = 0;

    if Invert; clase = ~clase; end

    indic = randperm(length(clase));
    Train_Set = HL(indic, :);
    Train_respueta = clase(indic);
end
```

Este proceso asegura que el modelo tenga una representación equilibrada de las clases, lo que deriva en una mejora en la capacidad para generalizar nuevos datos.

4.5.1.3 Entrenamiento del Clasificador

Finalmente, se entrena el modelo del clasificador y se evalúa su rendimiento. Durante el entrenamiento, se realiza una selección automática de variables y se aplica la técnica de "*Análisis de Componentes Principales (PCA)*", para mejorar la eficiencia del modelo y así intentar evitar un sobreajuste de los datos.

```
   for d = 1:n_dias_evaluar
       inicio = tic;
       txt = ['Calculando día ' num2str(d) ' de ' num2str(n_dias_evaluar, 0) ' ...'];
       disp(txt);
       waitbar(d / n_dias_evaluar, f, txt);

       dias_muetra = d;
       id = zeros(n + 1, 16);
       id(1 + n - dias_muetra:end - 1, :) = ones(dias_muetra, 1) *
Variables_Seleccionadas;
       id = id'; id = [id(:); 0];

       predictores = Train_Set(:, id == 1);

       X = predictores;
       [X_coef{d}, X_pca{d}, X_var{d}, X_ts{d}, X_exp{d}, X_mu{d}] =
pca(table2array(predictores));

       if gpuDeviceCount > 0
           In_Data = gpuArray(X_pca{d});
           Out_Data = gpuArray(Train_respueta);
       else
           In_Data = (X_pca{d});
           Out_Data = Train_respueta;
       end

       if n_pca > length(In_Data(1, :)); n_pca = length(In_Data(1, :)); end

       tic
       MDl{d} = fitcsvm(In_Data(:, 1:n_pca), Out_Data, 'KernelFunction', 'gaussian');
       toc

       Xn = BD.Test_Set(:, id == 1);
       Xn_norm = Xn - X_mu{d};
       Xn_PCA = table2array(Xn_norm) * X_coef{d};

       testeadores = Xn_PCA;
       ACC_tain(d) = sum(MDl{d}.predict(X_pca{d}(:, 1:n_pca)) == Out_Data) /
length(Out_Data);
       ACC_test(d) = sum(MDl{d}.predict(testeadores(:, 1:n_pca)) == Test_respuesta) /
length(Test_respuesta);
   end
   close(f)
```

Una vez se ha obtenido un entrenamiento del modelo se procede finalmente a la evaluación del rendimiento, esta evaluación se realiza tanto al conjunto de entrenamiento como al de prueba, utilizando métricas como la precisión Accuracy para medir el rendimiento del clasificador.

4.5.2 Análisis de Regresión (Paso a Paso)

En esta segunda sección se describe el análisis de regresión utilizando "*Support Vector Machines (SVM)*", paso a paso. El objetivo de este análisis es evaluar la capacidad del modelo para predecir valores continuos a partir de un conjunto de datos y variables seleccionadas, tanto de bolsa como de sentimiento.

Lo primero que debemos de hacer es proceder a la carga de los datos ya procesados con anterioridad.

4.5.2.1 Estructura de datos

En este apartado de la estructuración de datos, se va a realizar un proceso similar al Clasificador, generando una matriz de datos históricos con 30 días de historial y dividiendo los datos en dos conjuntos, siendo estos entrenamiento y prueba.

```
% Selección de variable objetivo
BD.respuesta = BD.FulltableBolsa_nxD.bolsaNorm0;
BD.Train_respuesta = BD.respuesta(1:BD.split_index, 1);
BD.Test_respuesta = BD.respuesta(BD.split_index + n:end, 1);

% Variables de Sentimiento y de Bolsa
BD.Var.Intensity = true; BD.Var.Agregated = true; BD.Var.xNumTweets = true;
BD.Var.xRetweets = true; BD.Var.xFollowers = true; BD.Var.xVerified = true;
BD.Var.bolsa = true; BD.Var.bolsaVariacion = true; BD.Var.bolsaNorm = true;
BD.Var.sube = true; BD.Var.baja = true; BD.Var.estable = true; BD.Var.CLASE = true;

% Vector de variables seleccionadas
BD.Variables_Seleccionadas = [BD.Var.Intensity, BD.Var.Agregated,
BD.Var.xNumTweets, BD.Var.xRetweets, BD.Var.xFollowers, BD.Var.xVerified,
BD.Var.bolsa, BD.Var.bolsaVariacion, BD.Var.bolsaNorm, BD.Var.sube, BD.Var.baja,
BD.Var.estable, BD.Var.CLASE];
```

Una vez divididos estos conjuntos de datos, se seleccionan en variables objetivo y variables de entrada que se utilizarán en el entrenamiento del modelo. Todo este proceso es completamente similar al caso del Clasificador, a diferencia de que las variables que se manejan en este caso pueden ser variables con cantidades continuas y con el objetivo de predecir valores numéricos, mientras que el clasificador se enfoca en la categorización de los datos de entrada y salida.

4.5.2.2 Combinaciones de Variables

Diferenciándonos pues del Clasificador, en este apartado se procederá a la evaluación de diversas combinaciones de variables para determinar cuáles proporcionan el mejor rendimiento del modelo. Esto incluye un bucle que recorre todas las combinaciones posibles de las variables que anteriormente se han seleccionado en la estructuración de los datos.

```
    % Generación de combinaciones de variables
    n_variables = sum(BD.Variables_Seleccionadas);
    variable_combinations = dec2bin(0:2^n_variables - 1) - '0';
    if n_variables < 13
        num_zeros_to_add = 13 - n_variables;
        variable_combinations                =                [variable_combinations,
zeros(size(variable_combinations, 1), num_zeros_to_add)];
    end
    variable_combinations = variable_combinations(2:end, :);

    variable_indices = find(BD.Variables_Seleccionadas);
    num_combinations = 2^n_variables;
```

Este proceso asegura que se exploren todas las posibles variable seleccionando así las que mejor rendimiento tengan, evaluándola en términos de rendimiento predictivo en los siguientes apartados.

4.5.2.3 Entrenamiento del Modelo

El modelo de regresión se entrena para cada combinación de variables y se evalúa utilizando el conjunto de prueba. La métrica utilizada para evaluar el rendimiento del modelo es el "*Error Cuadrático Medio (RMSE)*". Durante el entrenamiento, se optimizan los hiperparámetros del modelo para mejorar su rendimiento.

```
    % Bucle para evaluar diferentes combinaciones de variables
    for i = 1:num_combinations - 1
        selected_vars = variable_combinations(i, :);

        for d = 1:n_dias_evaluar
            % Selección de variables para el entrenamiento
            id = zeros(n + 1, length(BD.Variables_Seleccionadas));
            id(1 + n - d:end - 1, :) = ones(d, 1) * selected_vars;
            id = id';  id = [id(:); 0];

            % Entrenamiento del modelo SVM de regresión
            BD.predictores = BD.Train_Set(:, logical(id));
            BD.testeadores = BD.Test_Set(:, logical(id));
            BD.regressionSVM           =           fitrsvm(table2array(BD.predictores),
BD.Train_respuesta, 'OptimizeHyperparameters', 'all');

            % Evaluación del modelo
            ResultadoSVM{i}.regressionSVM{d} = BD.regressionSVM;
            ResultadoRMSE(i).RMSE_Entrenamiento(d)       =       loss(BD.regressionSVM,
table2array(BD.predictores), BD.Train_respuesta);
            ResultadoRMSE(i).RMSE_Test(d)           =           loss(BD.regressionSVM,
table2array(BD.testeadores), BD.Test_respuesta);
        end
    end
```

Este proceso permite evaluar el rendimiento del modelo en diferentes combinaciones de variables, guardando finalmente el resultado que ha tenido una mejor configuración para obtener el mejor rendimiento posible.

4.5.3 Análisis de Regresor Con PCA

Finalmente, en este proceso se lleva a cabo la última sección que destaca por ser un análisis de regresión, pero en este caso con la utilización de "*Análisis de Componentes Principales (PCA)*" para mejorar el rendimiento del modelo de regresión y no toparnos con el sobreajuste.

El objetivo principal de esta sección es evaluar como la reducción de dimensionalidad mediante PCA puede afectar a la precisión del modelo predictivo. Comparándose de esta manera los resultados obtenidos con el proceso sin PCA y con PCA, utilizando diversas configuraciones de variables.

El primer paso, y obviamente esencial es la carga de los datos que se han procesado en etapas anteriores.

4.5.3.1 Estructura de datos

Para esta estructuración de datos y como venimos viendo en las secciones anteriores se genera una matriz de datos históricos con 30 días de historial y seguimos un esquema muy parecido al anterior, el regresor, sobre todo debido a la misma naturaleza y el trato y procesamiento de los datos. Ya hemos visto que posteriormente entonces, se procede a una división de los datos en dos subconjuntos, obteniendo entrenamiento y otro de pruebas.

4.5.3.2 Selección de Variables

Una vez en este punto y con una estructuración predefinida, procederemos a la etapa crítica de la construcción del modelo predictivo, siendo esta la selección de sus variables objetivo y de sus variables de entrada, utilizadas posteriormente en el entrenamiento de este modelo.

Estas configuraciones permiten comparar el rendimiento del modelo con la utilización de las diversas variables, los que nos permitirá también observar cómo se visualizan en los resultados las diferencias de la utilización o no utilización de PCA dependiendo de las variables seleccionadas.

4.5.3.3 Entrenamiento del Modelo

Este modelo de regresión, por lo tanto, se entrena y evalúa para cada configuración de variables utilizando PCA para reducir la dimensionalidad. La métrica igual que en la sección anterior utilizada para la evaluación del rendimiento del modelo es el "*Error Cuadrático Medio (RMSE)*". Y durante el periodo de entrenamiento se optimizan los hiperparámetros del modelo para mejorar de esta manera su rendimiento.

El uso de PCA en este contexto permite reducir la dimensionalidad de los datos, eliminando de esta forma la multicolinealidad y capturando las variaciones más significativas

de las variables con menos componentes. Esto no solo mejorará la eficiencia computacional, sino que también puede mejorar la precisión del modelo al enfocarse en todas las características más relevantes.

```
   for d = 1:n_dias_evaluar
           inicio = tic;
           txt = ['Calculando día ' num2str(d) ' de ' num2str(n_dias_evaluar, 0) '
...'];
           disp(txt);
           waitbar(d / n_dias_evaluar, f, txt);

           dias_muetra = d;
           id = zeros(n + 1, length(Variables_Seleccionadas));
           id(1 + n - dias_muetra:end - 1, :) = ones(dias_muetra, 1) *
Variables_Seleccionadas;
           id = id'; id = [id(:); 0];

           % Selección de predictores
           predictores = Train_Set(:, id == 1);

           if usarPCA
               % PCA
               X = predictores;
               [X_coef{d},      X_pca{d},      ~,      ~,      ~,      X_mu{d}]      =
pca(table2array(predictores));
               In_Data = X_pca{d};
           else
               In_Data = table2array(predictores);
           end

           if gpuDeviceCount > 0
               In_Data = gpuArray(In_Data);
               Out_Data = gpuArray(Train_respuesta);
           else
               Out_Data = Train_respuesta;
           end

           if usarPCA && n_pca > length(In_Data(1, :))
               n_pca = length(In_Data(1, :));
           end

           % Regresión SVM con diferentes kernels con validación cruzada y
regularización
           if usarPCA
               In_Data_Train = In_Data(:, 1:n_pca);
           else
               In_Data_Train = In_Data;
           end

           regressionSVM_Gaussian{d}      =      fitrsvm(In_Data_Train,      Out_Data,
'KernelFunction', 'gaussian', 'Standardize', true, 'BoxConstraint', 1);
           if calcularLinear
               regressionSVM_Linear{d}     =     fitrsvm(In_Data_Train,     Out_Data,
'KernelFunction', 'linear', 'Standardize', true, 'BoxConstraint', 1);
           end
           if calcularPoly
               regressionSVM_Poly{d}      =      fitrsvm(In_Data_Train,      Out_Data,
'KernelFunction', 'polynomial', 'Standardize', true, 'BoxConstraint', 1);
           end

           % Transformación de datos de test
           Xn = BD.Test_Set(:, id == 1);

           if usarPCA
```

```
                Xn_PCA = (table2array(Xn) - X_mu{d}) * X_coef{d}; % Centrar los datos
de prueba con la media de los datos de entrenamiento
                testeadores = Xn_PCA(:, 1:n_pca);
            else
                testeadores = table2array(Xn);
            end

            % Evaluación
            pred_train_gauss = predict(regressionSVM_Gaussian{d}, In_Data_Train);
            pred_test_gauss = predict(regressionSVM_Gaussian{d}, testeadores);

            if calcularLinear
                pred_train_linear = predict(regressionSVM_Linear{d}, In_Data_Train);
                pred_test_linear = predict(regressionSVM_Linear{d}, testeadores);
            end

            if calcularPoly
                pred_train_poly = predict(regressionSVM_Poly{d}, In_Data_Train);
                pred_test_poly = predict(regressionSVM_Poly{d}, testeadores);
            end

            % Calcular RMSE
            RMSE_Gaussian(d) = sqrt(mean((pred_train_gauss - Out_Data).^2));
            RMSE_Gaussian_Test(d) = sqrt(mean((pred_test_gauss - Test_respuesta).^2));

            if calcularLinear
                RMSE_Linear(d) = sqrt(mean((pred_train_linear - Out_Data).^2));
                RMSE_Linear_Test(d)       =       sqrt(mean((pred_test_linear       -
Test_respuesta).^2));
            end

            if calcularPoly
                RMSE_Poly(d) = sqrt(mean((pred_train_poly - Out_Data).^2));
                RMSE_Poly_Test(d) = sqrt(mean((pred_test_poly - Test_respuesta).^2));
            end
        end
```

Los resultados obtenidos pues de los modelos de regresión se visualizan para cada configuración de variables. Esta visualización permite comparar el impacto de usar o no PCA y evaluar la precisión del modelo ante diferentes configuraciones.

4.6 Proceso v5 – Análisis de Días Clave

En esta última sección se abordará un análisis muy diferente a todo lo anterior comentado, en este se abordará el análisis de los días clave, específicamente aquellos días que puedan considerarse críticos o de crisis para una empresa.

El objetivo principal de este proceso es identificar aquellos días clave utilizando modelos de clasificación que se entrenan con características derivadas de los datos históricos de los CDs y de los sentimientos. Este análisis puede ayudar a comprender cómo ciertos eventos pueden influir significativamente en el comportamiento del mercado y en la percepción pública.

Comenzando pues con el proceso y como es recurrente en procesos anteriores, procederemos a realizar la carga de datos previamente procesados. Garantizando que toda la información necesaria para el proceso se encuentre lista para el análisis en el entorno de trabajo.

4.6.1 Generación de Características Derivadas

Una vez cargados los datos, se procede a generar características derivadas. Esto implica extraer y crear nuevas variables a partir de los datos originales ya cargados, las cuales pueden ser más representativas y útiles para el entrenamiento de clasificación que se quiere llevar a cabo.

Antes de nada, se han de identificar en el histórico de los datos las fechas clave junto son sus descripciones. Estas fechas representan eventos importantes que se espera que tengan un impacto significativo en la propia empresa.

Por lo tanto, una vez se han extraído las variables relevantes de "*FullTable*", se crean nuevas características derivadas, como la relación de tweets por seguidor y la relación de retweets por tweet. Estas nuevas características pueden ofrecer información adicional que no es directamente evidente en las variables originales.

```
% Seleccionar las variables correctas de FullTable_Bolsa
dates = FullTable_Bolsa.Date;
dates.TimeZone = '';
intensity = FullTable_Bolsa.Intensity;
sentiment = FullTable_Bolsa.Sentimiento;
mean_sentiment = FullTable_Bolsa.Mean_Sentiment;
median_sentiment = FullTable_Bolsa.Median_Sentiment;
std_sentiment = FullTable_Bolsa.Std_Sentiment;
num_tweets = FullTable_Bolsa.xNumTweets;
num_retweets = FullTable_Bolsa.xRetweets;
num_followers = FullTable_Bolsa.xFollowers;
num_verified = FullTable_Bolsa.xVerified;
bolsa = FullTable_Bolsa.bolsa;
bolsa_variacion = FullTable_Bolsa.bolsaVariacion;
bolsa_norm = FullTable_Bolsa.bolsaNorm;
sube = FullTable_Bolsa.sube;
```

```
    baja = FullTable_Bolsa.baja;
    estable = FullTable_Bolsa.estable;
    clase = FullTable_Bolsa.CLASE;

    % Crear una tabla de datos para el aprendizaje automático
    data = table(dates, intensity, sentiment, mean_sentiment, median_sentiment,
std_sentiment, ...
                 num_tweets, num_retweets, num_followers, num_verified, ...
                 bolsa, bolsa_variacion, bolsa_norm, sube, baja, estable, clase);
    data.is_key_date = ismember(dates, fechasDeutscheBank);

    % Generar características derivadas
    data.tweets_per_follower = data.num_tweets ./ data.num_followers;
    data.retweets_per_tweet = data.num_retweets ./ data.num_tweets;
    data.verified_ratio = data.num_verified ./ data.num_followers;
```

Estas características derivadas pueden ser cruciales ya que pueden revelar en alguna ocasión relaciones ocultas y patrones en los datos que pueden derivar a ser esenciales para una clasificación efectiva.

4.6.2 Entrenamiento del Modelo

Una vez hemos obtenido todos los datos correspondientes que esperamos, se procede al entrenamiento y la evaluación de varios modelos de clasificación. El objetivo de utilizar diversos modelos es identificar al mejor clasificador para predecir los días clave. Utilizándose técnicas de validación cruzada para asegurar que los resultados sean generalizables y no específicos del conjunto de datos de entrenamiento.

Primero de todo, se procede a balancear las clases, para que el conjunto de datos no se encuentre sesgado, este comportamiento ya lo hemos visto en apartados anteriores de esta memoria. Este equilibrio pues, es fundamental cuando se trabaja con estos datos desbalanceados como es nuestro caso con la “bolsa” y los “sentimientos”, para obtener el mejor resultado posible en la identificación de los días de crisis.

```
    % Balancear el conjunto de datos
    idx_key_dates = find(data.is_key_date);
    idx_non_key_dates = find(~data.is_key_date);
    num_key_dates = length(idx_key_dates);
    num_non_key_dates = length(idx_non_key_dates);

    if num_non_key_dates > num_key_dates
        idx_non_key_dates = idx_non_key_dates(randperm(num_non_key_dates,
num_key_dates));
    elseif num_key_dates > num_non_key_dates
        idx_key_dates = idx_key_dates(randperm(num_key_dates, num_non_key_dates));
    end

    balanced_data = data([idx_key_dates; idx_non_key_dates], :);
```

Tras esto, se procede a una evaluación de los clasificadores utilizando como se ha comentado la validación cruzada. Observando su precisión y seleccionando el modelo que tenga un mejor rendimiento.

4.6.- PROCESO V5 – ANÁLISIS DE DÍAS CLAVE

Y finalmente, y finalizando el proceso, se evalúa el mejor modelo en el conjunto completo de datos, para posteriormente identificar las fechas predichas como días clave. Comparando de esta manera las predicciones obtenidas con las fechas clave conocidas e identificadas en un inicio. Esta evaluación se realiza mediante diferentes métricas para aumentar el contexto, siendo estas la "*Precisión*", "*Exhaustividad*" y "*Tasa de acierto del modelo*".

```
predictions = predict(bestModel, data{:, 2:end-1});
predicted_key_dates = dates(logical(predictions));

disp('Fechas predichas como clave:');
disp(predicted_key_dates);
disp('Fechas conocidas como clave:');
disp(fechasDeutscheBank);
[common_dates, ia, ib] = intersect(predicted_key_dates, fechasDeutscheBank);
disp('Fechas comunes:');
disp(common_dates);

num_true_positives = sum(ismember(predicted_key_dates, fechasDeutscheBank));
num_false_positives = sum(~ismember(predicted_key_dates, fechasDeutscheBank));
num_false_negatives = sum(~ismember(fechasDeutscheBank, predicted_key_dates));

precision = num_true_positives / (num_true_positives + num_false_positives) * 100;
recall = num_true_positives / (num_true_positives + num_false_negatives) * 100;
accuracy = (num_true_positives + (sum(~predictions & ~data.is_key_date))) /
length(data.is_key_date) * 100;

disp(['Precisión (Precision): ', num2str(precision), '%']);
disp(['Exhaustividad (Recall): ', num2str(recall), '%']);
disp(['Tasa de acierto (Accuracy): ', num2str(accuracy), '%']);
```

Este análisis detallado intenta predecir de esta manera los días clave, proporcionando una información valiosa para la toma de decisión ante este conocimiento y para la gestión de riesgos que pueda tomar una empresa bajo estos conocimientos.

Capítulo 5

5 Resultados

En este capítulo se van a presentar los resultados obtenidos del análisis de los datos históricos de las empresas y los sentimientos expresados en X, observando los resultados de los modelos de clasificación y regresión que hemos desarrollado en los puntos anteriores. Se evaluará la precisión de estos modelos y se analizará cómo los datos de los sentimientos influyen en los comportamientos del mercado bursátil, en particular, en el caso de estudio que han sido los bonos de Credit Default Swaps. Además, observaremos los resultados obtenidos del caso particular de la obtención de los días clave partiendo de los valores históricos.

5.1 Análisis Exploratorio Básico

En esta primera sección se presentan los resultados obtenidos de este análisis básico donde se permite identificar patrones y relaciones básicas entre las variables, proporcionando de esta manera comprensión exhaustiva de los datos antes de aplicar modelos más complejos.

5.1.1 Correlación Lineal

Para visualizar correctamente la correlación entre las diferentes variables, se generaron mapas de calor (heatmaps), que muestran la correlación lineal entre las variables. Los resultados obtenidos fueron:

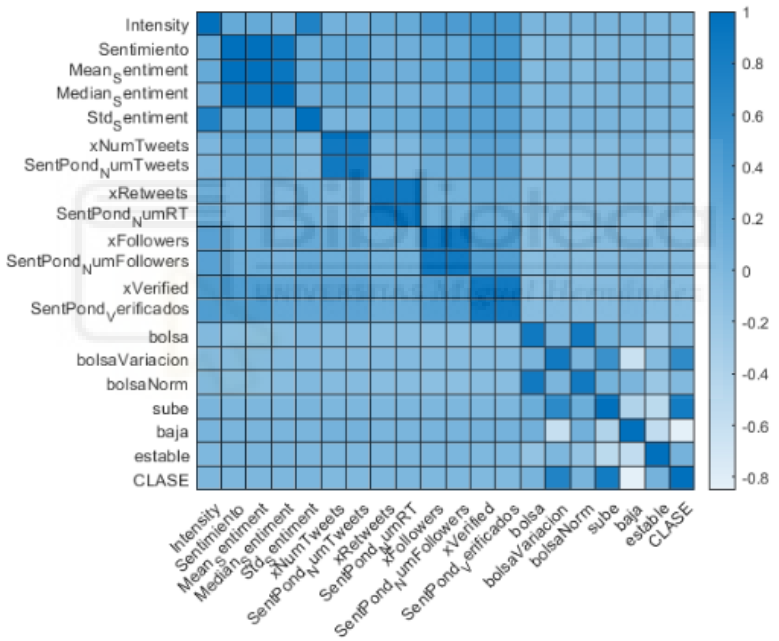


Figura 5.1: Mapa de calor (HeatMap)

Para observar esta interpretación, se dividió en una coloración azul oscura (valor cercano a 1), que indica una alta correlación positiva entre las variables, la coloración azul clara (valor cercano al -1) que indica una alta correlación negativa, y por último una coloración intermedia (valor cercano al 0) que indica una correlación baja.

Observando los datos podemos concluir, que podemos encontrar una ligera correlación positiva, obviamente despreciando las variables entre sí, entre la cantidad de tweets que hay, es decir, la intensidad, con el sentimiento ponderado por verificado, o la cantidad de usuarios verificados en sí.

Sin embargo, y con este estudio no encontramos ninguna correlación mínima entre los datos de los tweets con los datos de los Credit Default Swaps.

5.1.2 Relación Mutua de Variables

La relación mutua de variables es un análisis que permite entender cómo interactúan entre sí las diferentes variables del conjunto de datos. Para observar esta relación se ha utilizado un diagrama de dispersión XY (scatter plots), el cual muestra la relación entre pares de variables.

El diagrama en sí representa gráficamente con cada punto la observación de los datos con una variable en el eje X y otras en el eje Y. Este tipo de gráfico es útil para identificar posibles patrones, tendencias y correlaciones entre variables.

Para poder interpretarlo correctamente se debe saber el significado de los patrones y las tendencias que se obtienen. Para empezar, encontramos la dirección (positiva o negativa), que indica cual es la pendiente y la correlación entre las variables. También se ha de tener en cuenta la fuerza que podemos definir como la dispersión generada entre los puntos, conociendo que a mayor dispersión menor correlación tendrán dichas variables.

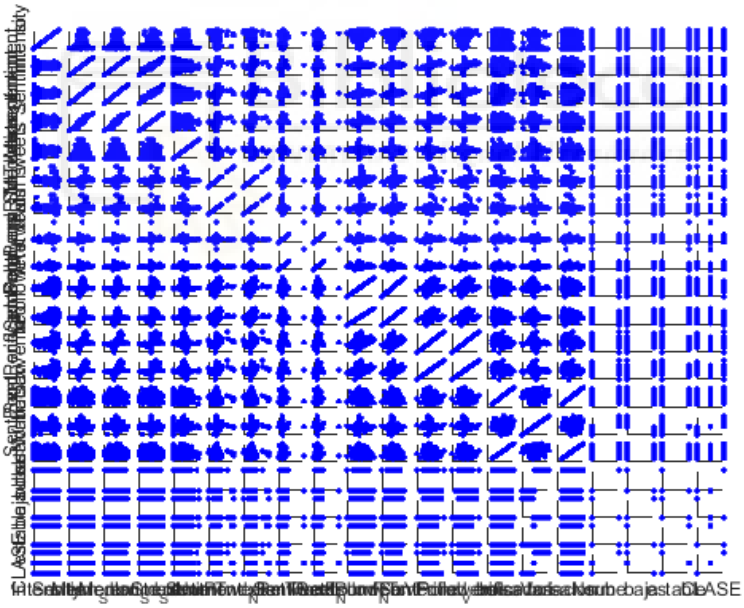


Figura 5.2: Relación Mutua de Variables (HeatMap)

Como se puede observar en el resultado, comprendemos una relación clara entre las variables que son iguales, es decir, Intensidad con Intensidad, conociendo de esta manera que los que aparecen como una línea ascendente indican este valor de correlación.

Fijándonos con estos conocimientos en los demás resultados observamos como ninguna de las variables tratadas tiene realmente un peso considerable en cuanto a la correlación,

por ejemplo, podemos observar la intensidad de los Tweets con cualquier variable de la bolsa, y observar cómo los puntos no tienen ningún patrón perceptible y únicamente se forman puntos aleatorios que no influyen entre sí.

5.1.3 Histograma de Variables

El análisis de histogramas es fundamental para comprender la distribución de las variables en el conjunto de datos. El histograma nos va a mostrar la frecuencia con la que ocurren diferentes valores de una variable, lo que nos brinda una visión de la distribución, sesgo y posibles valores atípicos que pueda haber.

Al igual que en el caso anterior, definiremos diferentes patrones para tener en cuenta para comprender los datos obtenidos. Para empezar, nos encontramos con la Distribución de los datos, los histogramas permiten ver la distribución de estos, identificando si las variables siguen distribuciones normales, sesgadas o bimodales. Continuando nos encontramos la Identificación de Sesgos, estos se muestran con un sesgo positivo, negativo o neutro. Y, por último, podemos observar la detección de valores atípico donde se pueda externalizar un valor que no concuerde con lo esperado y común pudiendo así suprimirlo para que no altere los datos.

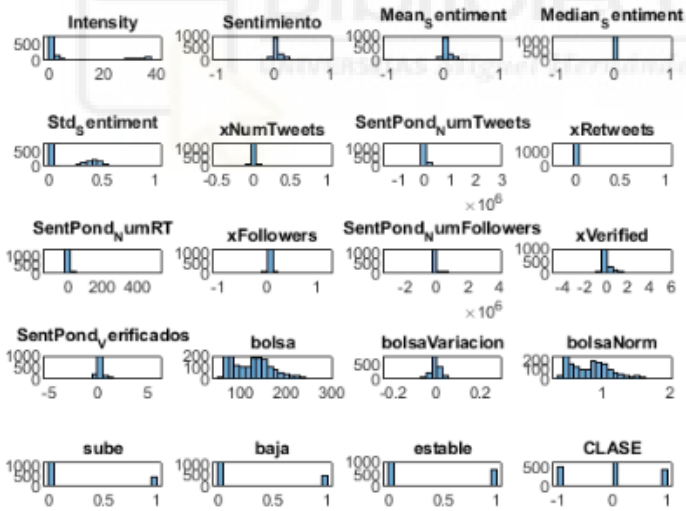


Figura 5.3: Mapa de calor (HeatMap)

Observando el resultado obtenido comprendemos que mayoritariamente los valores de sentimiento son valores neutros que no influyen en una decisión que podamos definir como drástica que luego derive en una variación en el valor de los CDS.

5.1.4 Representación Cruzada

La representación cruzada de variables es una técnica utilizada para analizar la relación entre dos variables específicas, diferenciadas por categorías o clases. En este proyecto se ha utilizado para poder visualizar como se distribuyen las variables seleccionadas en función de la variable "CLASE", que es un indicativo normalizado se si la bolsa sube (valor 1), se mantiene estable (valor 0) o la bolsa baja (valor -1).

A la hora de realizar un análisis de resultado de dos variables se usó mediante código para el Eje X la intensidad del sentimiento y para el Eje Y en este caso el número de retweets, pero se puede realizar con cualquier otro tipo de variables.

Como hemos comentado de los indicativos existen tres casos (valor 1) la mayoría de los puntos se concentran en valores más altos de intensidad de sentimiento y número de retweets. Esto sugiere que los tweets con alta intensidad de sentimiento y alta actividad en términos de retweets están asociados con días en los que el valor de los CDS suba. La alta intensidad del sentimiento positivo y retweets podría estar relacionada con el interés de los inversores.

El siguiente caso sería el caso de (valor 0) donde los puntos están dispersos a lo largo de todo el gráfico, sin una concentración clara en un área específica. Sugiriendo esto que los días en los que el valor de los CDS se mantiene estable no tienen una fuente de relación con la intensidad del sentimiento o número de retweets.

Y por último (valor -1) donde los puntos se concentran en valores bajos tanto de intensidad como de número de retweets. Esto indica que la baja intensidad del sentimiento y la baja actividad de retweets podrían estar correlacionados con el pésimo interés de los inversores viéndolo reflejado también en el riesgo de estas empresas.

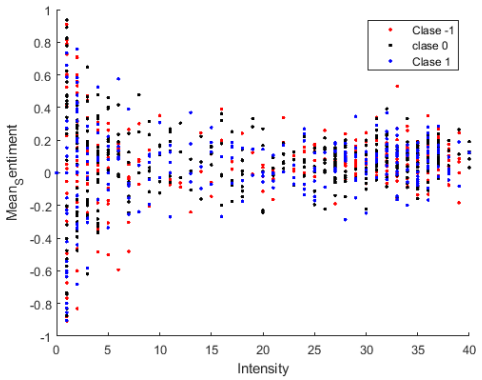


Figura 5.4: Representación Cruzada

Observando el resultado obtenido comprendemos que no hay ningún tipo de correlación directa entre las variables pues en el caso del Sentimiento Medio frente a la Intensidad, el valor de los CDS no se ven afectados pues no siguen ningún patrón perceptible en los casos de que el riesgo (la variable CLASE), suba se mantenga estable o baje.

5.2 Análisis y Procesado por Secuencia

Esta sección es una etapa crítica del proyecto pues es donde se aplican las técnicas avanzadas para profundizar en la relación entre las variables que hemos obtenido y hemos observado en el apartado anterior. En este caso, se utilizan modelos de regresión y clasificadores para hacer predicciones basadas en los datos procesados.

5.2.1 Regresión Gaussiana

Es una técnica de aprendizaje automático que modela las relaciones no lineales entre las variables. En este proyecto, se utilizó para capturar la relación entre variables de sentimiento y los valores de la bolsa, proporcionando un modelo que predice el valor futuro de la bolsa con base en las características de los tweets.

A la hora de observar su resultado los datos utilizados son las variables relacionadas con el sentimiento como (número de tweets, seguidores, retweets, entre otras) y los datos de bolsa se utilizaron como predictores.

Para su evaluación se utilizó una métrica del error cuadrático medio (RMSE) para evaluar el rendimiento del modelo en los conjuntos de entrenamiento y pruebas.

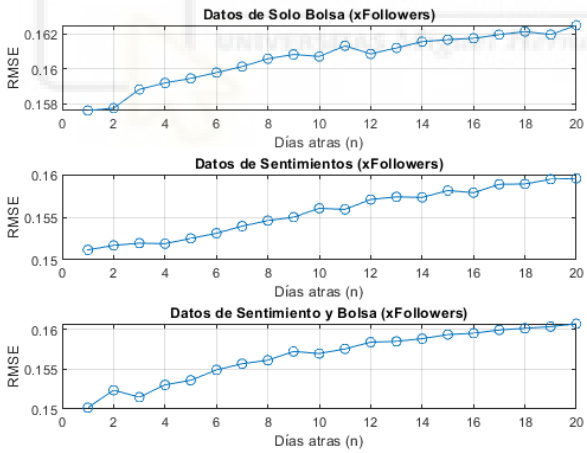


Figura 5.5: Regresión Gaussiana

Si nos fijamos en el resultado obtenido, tenemos tres subplot con valores diferentes, aunque muy similares en resultados. Para estas representaciones estamos cogiendo como podemos observar en el título de cada subplot los Datos de Solo los valores de CDS, únicamente con los Sentimientos o teniendo en cuenta ambos datos, esto frente a una variable preseleccionada que en este caso específico es el Sentimiento ponderado por el número de seguidores que tiene esa persona, es decir, a mayor número de seguidores mayor importancia le damos a su tweet.

Una vez comprendido esto, vemos como curiosamente el RMSE aumenta a lo largo de los días, con una tendencia a estabilizarse alrededor del día 15. El hecho de que el modelo no aumente drásticamente podría indicar que podría ser razonablemente preciso dentro del conocimiento de que las variables no tienen una gran correlación como hemos observado en el punto anterior.

Aun con estas, fijándonos en este caso en el valor directo que obtenemos del RMSE podemos conocer que el valor de en torno a 0,16 y con un rango de fluctuación muy estrecho nos está indicando que el modelo no está prediciendo bien. Esto podría estar indicándonos que, aunque si hay cierto valor que puede estar capturándose hay mucho ruido y podría deberse a que puede haber otros factores importantes que el modelo no está capturando como por ejemplo otros indicadores macroeconómicos.

5.2.2 Clasificador de Árbol

El clasificador es una técnica de clasificación que utiliza una estructura jerárquica para dividir los datos en subconjuntos basados en reglas de decisión. En este proyecto se utilizó para predecir si la bolsa subiese, bajase o se mantuviese estable en función de las variables seleccionadas.

En nuestro caso los datos utilizados fueron las variables del sentimiento y de la bolsa, como predictores para clasificar los datos, por último, para su evaluación se utilizó la precisión (Accuracy) para evaluar la capacidad del clasificador para predecir correctamente las clases de los datos.

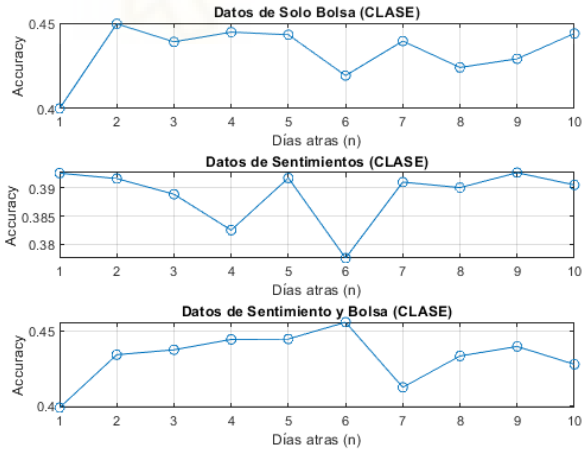


Figura 5.6: Clasificador de Árbol

Observando ellos resultados y parecido al Regresor, nos encontramos con los mismos tres subplots divididos en las variables de Solo valores de CDS, Sentimiento o un conjunto de ambas. Para este caso, y al ser un clasificador utilizamos la variable CLASE que nos indica si el valor de los CDS aumenta, se mantiene estable o disminuye, obteniendo así un valor de Accuracy que observamos que se mantiene bastante estable en los tres casos, teniendo

mayor valor en los dos casos en los que interfiere los valores de los CDS. Este valor es bastante bajo, aun así, podemos obtener como conclusión, que utilizar únicamente los valores de los tweets tiene un mayor porcentaje de fallos en la obtención de los resultados que utilizar los valores de la bolsa.

Por lo tanto, podemos concluir que, en el caso del Clasificador, el modelo tiene problemas de precisión y estabilidad, lo que indica que posiblemente no está capturando correctamente la relación entre las características de las clases. Por otro lado, la fluctuación constante que obtenemos puede ser representativo de que los datos o las características no son suficientemente representativas y el modelo no puede captar patrones significativos.

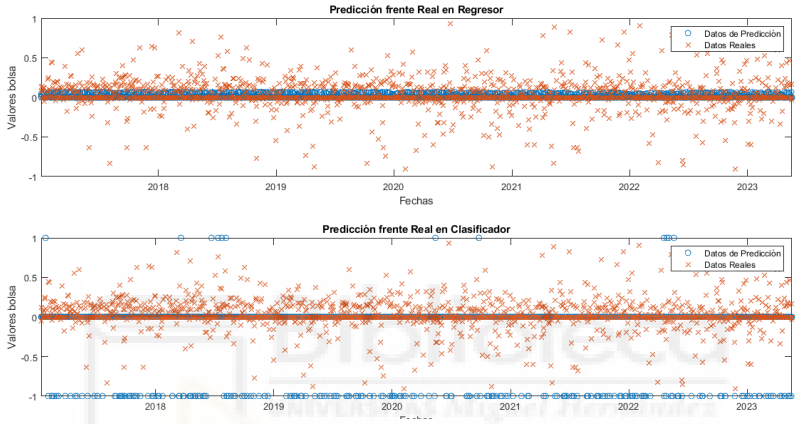


Figura 5.7: Resultados Análisis y Procesado Por Secuencia

Observando esta figura como resultado final, nos encontramos como un primer subplot el Regresor y en una segunda fila el Clasificador. Ambas gráficas representan la predicción que hemos obtenido de los valores frente al valor real.

Como podemos observar y aunque el Regresor acierta algunos valores se mantiene todo el tiempo en un valor intermedio similar al 0 por lo que muchos de los aciertos que tiene se deben a que la variación que tiene es mínima y se dan positivos ciertos valores de predicción que son irreales puesto que se debe a que los valores de la bolsa acaban tocando este valor céntrico.

Por otro lado, en el Clasificador observamos justo lo opuesto, este nos está intentando indicar si el valor sube, se mantiene o baja. Y aunque hay algunos casos que coincide al igual que sucedía en el Regresor, la mayor parte son de casualidad debido a que en este ejemplo realizado la mayor parte de las predicciones están destinadas a que el valor baja o se mantiene estable.

5.3 Análisis Exhaustivo

El análisis exhaustivo es la fase final del procesamiento de datos y modelado, donde se aplican diferentes técnicas de regresión y clasificación con un mayor grado de detalle. En este punto, se exploran las capacidades predictivas de los modelos desarrollados, buscando optimizar su rendimiento y precisión en la predicción del comportamiento del mercado

5.3.1 Análisis de Clasificador

Para este caso y a diferencia del modelo de Clasificador anterior, teniendo en cuenta las dificultades que encontraba el anterior modelo, se obtuvo el conjunto de datos inicial y se dividió en un 80% de datos para el entrenamiento del modelo y un 20% para el test.

Debido a las complicaciones que teníamos y que parecía que los datos no eran suficientemente representativos se realizó un equilibrado de las muestras tomando un número de muestras igual a la mayoritaria y minoritaria para intentar evitar que el modelo vire hacia la variable que más muestras tiene. Además de esto se utiliza un análisis con PCA y se calculan los valores de Accuracy tanto para los valores de Test o para los valores de Entrenamiento.

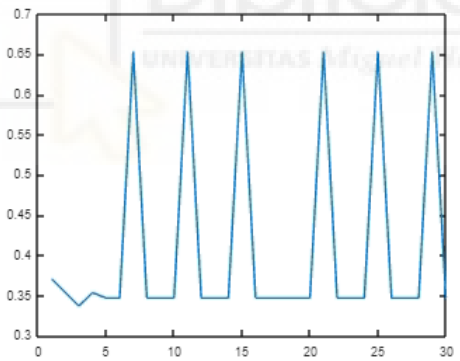


Figura 5.8: Accuracy Test

Observando el resultado que obtenemos primeramente de los valores de Test, vemos claramente que el modelo no es consistente pues el conjunto fluctúa en aproximadamente un periodo de 6 días, entre los valores de 0,35 hasta 0,65. Esto indica claramente que el modelo tiene dificultades para tener una predicción consistente del sistema, indicando que está obteniendo solo parcialmente las relaciones entre las variables.

Y la fluctuación constante y periódica puede deberse a un indicativo de que el modelo está siendo influenciado por algún fenómeno temporal que no se ha tenido en cuenta en el estudio.

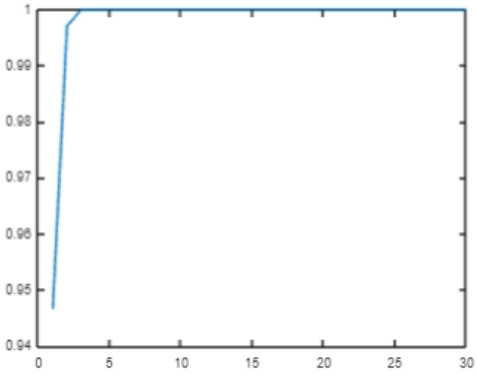


Figura 5.9: Accuracy Train

Por otro lado, el hecho de que en el Entrenamiento el Accuracy tienda rápidamente a 1 está indicando claramente que hay un sobreajuste en los datos, pues asume que aproximadamente al tercer día ya está memorizando los datos de entrenamiento lo que le permite obtener una precisión perfecta, pero a costa de generar correctamente nuevos datos como observamos en el Accuracy Test.

5.3.2 Análisis de Regresión

Para este proceso de Regresión se utilizan múltiples combinaciones de variables obteniendo mediante bucles la mejor combinación posible evaluando su desempeño mediante el uso del RMSE. De esta manera obtenemos para estos casos el mejor modelo posible con sus hiperparámetros ajustados para luego poder obtener los mejores resultados y estimaciones posibles.

Para este caso se toma un 70% de datos para entrenamiento y un 30% para realizar los test, teniendo en cuenta los últimos 30 días.

Iter	Eval result	Objective: log(1+loss)	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	BoxConstraint	KernelScale	Epsilon	KernelFunction	PolynomialOrder	Standardize
1	Best	0.080182	1.0022	0.080182	0.080182	0.027112	-	0.00082289	linear	-	true
2	Best	0.079902	0.19761	0.079902	0.080842	0.77874	-	2.6754	linear	-	false
3	Accept	0.079902	0.2146	0.079902	0.079995	0.11689	3.4993	2.0009	gaussian	-	true
4	Accept	0.079902	0.23478	0.079902	0.079971	823.29	-	17.279	polynomial	3	false
5	Accept	0.079902	0.20412	0.079902	0.079902	2.7869	-	1.5488	polynomial	2	false
6	Accept	0.1232	0.20893	0.079902	0.079903	109.02	1.1317	0.68141	gaussian	-	true
7	Accept	0.079902	0.11407	0.079902	0.079902	3.4907	-	9.2635	linear	-	false
8	Accept	0.079902	0.10522	0.079902	0.079973	93.64	1.2661	4.5442	gaussian	-	true
9	Accept	0.079902	0.099039	0.079902	0.079902	365.86	-	3.5504	linear	-	false
10	Best	0.079836	0.20583	0.079836	0.079902	0.0016706	1.2518	0.0087378	gaussian	-	true
11	Accept	0.080053	0.22294	0.079836	0.079902	0.0076138	3.2942	0.059876	gaussian	-	true
12	Accept	0.080014	0.18348	0.079836	0.079836	0.0010963	3.7691	0.06069	gaussian	-	true
13	Accept	0.079902	0.097866	0.079836	0.079836	0.015716	-	27.947	linear	-	false
14	Accept	0.13613	0.10387	0.079836	0.079836	110.98	-	0.74566	linear	-	false
15	Accept	0.080417	1.2748	0.079836	0.079836	421.98	-	0.013974	linear	-	false
16	Accept	0.079902	0.11562	0.079836	0.079836	17.774	-	16.271	linear	-	false
17	Accept	0.079902	0.094124	0.079836	0.079836	328.88	408.57	12.98	gaussian	-	true
18	Accept	0.079902	0.11802	0.079836	0.079836	365.45	0.0024342	30.649	gaussian	-	true
19	Accept	0.079902	0.11006	0.079836	0.079836	514.19	107.92	2.9577	gaussian	-	true

Figura 5.10: Tabla Optimización Hiperparámetros

Este es el resultado del proceso de la optimización de los hiperparámetros para el modelo de regresión. Observando levemente la tabla vemos a lo largo de cada fila cada una de las iteraciones y cómo podemos observar en la columna de KernelFunction, utilizando un modelo u otro a parte de los propios hiperparámetros. Una vez obtenidos todos nos guardamos la información del mejor y procedemos a realizarle las pruebas

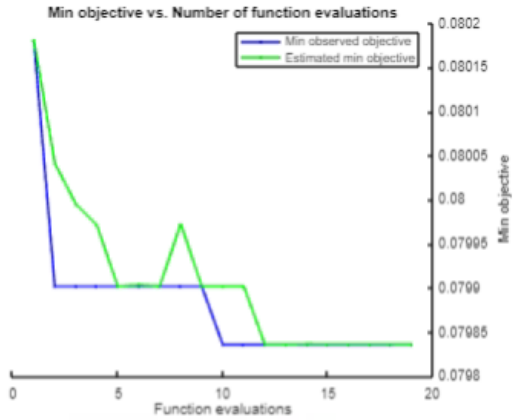


Figura 5.11: Convergencia de la Optimización de Hiperparámetros

Por último, este gráfico final muestra la convergencia de la optimización de los hiperparámetros para el modelo de Regresión, donde se grafica el mínimo objetivo es decir el error que se tiene, frente al número de evaluaciones de función.

Por lo tanto, la línea azul indica el valor mínimo observado del objetivo (el error) tras cada evaluación, es decir, el mejor rendimiento que se ha lograd en cada etapa de la optimización y, por otro lado, la línea verde representa la estimación del objetivo mínimo en función de la información obtenida hasta el momento.

Observamos que al inicio tiene un rápido descenso, lo que nos indica una caída rápida tanto en la línea verde como en la azul, lo que nos indica que se ha encontrado una combinación previa de hiperparámetros que reducen rápidamente el error. Luego observamos que se estabiliza a partir de la evaluación 10, cerca del valor de 0,079 lo que indica que se ha alcanzado un mínimo local y los hiperparámetros están cerca de un rendimiento óptimo.

Una vez obtenido la mejor configuración se procede a guardar el RMSE de Train y Test en una variable para poder comparar los datos en el proceso siguiente.

5.3.3 Análisis de Regresor con PCA

El Análisis de Componentes Principales (PCA) es una técnica utilizada para reducir la dimensionalidad de los datos, lo que permite eliminar la colinealidad entre las variables y capturar solo la información más relevante. En este análisis, se aplicó PCA al modelo de regresión para mejorar la precisión y evitar el sobreajuste de los datos. La reducción de dimensionalidad mediante PCA mejoró significativamente la eficiencia computacional del modelo y redujo el riesgo de sobreajuste.

Conociendo el ajuste de hiperparámetros óptimo para los diferentes modelos, y comparando las diferentes configuraciones “Con Sentimiento”, “Con valores CDS” y “Ambas”, se procede a realizar el estudio de que valor es el que produce un mejor resultado intentando obtener el mayor porcentaje de acierto.

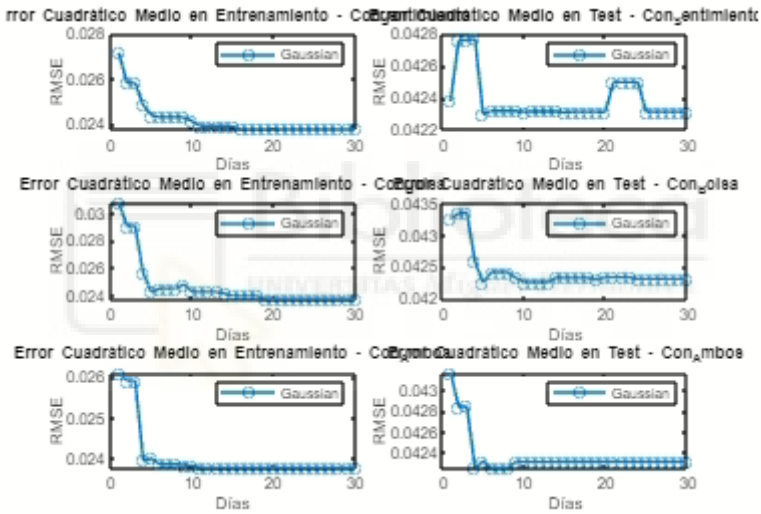


Figura 5.12: Convergencia de la Optimización de Hiperparámetros

El resultado final obtenido se divide en varios plots, en el caso presentado todos utilizan el modelo Gaussiano que en el proceso realizado fue el que mejores resultados obtuvo. Observando la imagen vemos dividido por filas el mismo modelo para valores de entrenamiento diferentes, ya sea únicamente valores de Sentimiento, solo valores de CDS o con la unificación de ambas, por otro lado, dividido por columnas obtenemos a la izquierda el modelo de entrenamiento realizado y a la derecha el modelo de Prueba obtenido.

En los tres experimentos en el conjunto de entrenamiento el RMSE tiende a disminuir rápidamente durante los primeros días evaluados, estabilizándose con valores cercanos al 0,024. Esto indica que el modelo está aprendiendo bien a ajustar los datos de entrenamiento y el error en este conjunto, lo cual es un buen indicador de que el modelo está capturando las relaciones entre las variables y la variable objetivo.

Para el caso del modelo de Test, la fila situada a la derecha, el conjunto es mayor al de entrenamiento, lo cual es lo esperado pues los datos de prueba no han sido vistos por el modelo, sin embargo, lo interesante es que el RMSE también disminuye de manera rápida y se estabiliza en torno a valores de entre 0,042 lo que sugiere que el modelo tiene una buena capacidad de generalización.

Observando en su totalidad los resultados vemos que se realiza un buen ajuste de entrenamiento. En todos los casos el RMSE baja rápidamente y se estabiliza en torno al mismo valor por lo que sabemos que el modelo no cae en sobreajuste excesivo y sabemos que la utilización del PCA funciona correctamente.

Por otro lado, vemos que la capacidad del modelo para generalizar es moderada, puesto que el conjunto del Test se estabiliza en torno a 0,042 para este caso, lo cual es un a diferencia aceptable respecto al entrenamiento.

Sin una clara ventaja al combinar Sentimiento y valores de CDS el experimento no muestra una mejora significativa en el RMSE del conjunto de test lo que sugiere que una de las fuentes es menor en su poder predictivo y por lo que hemos observado a lo largo de los experimentos ocurre con el caso de los datos de los Sentimientos y añadirla no necesariamente añade valor adicional.

Por último y centrándonos en la efectividad de las variables, el RMSE en el Test para el experimento parece dar una buena predicción de los valores.

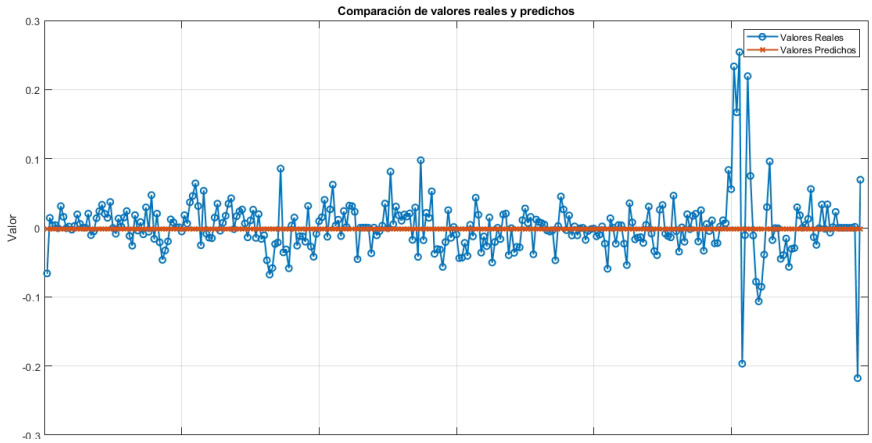


Figura 5.13: Valores Predichos frente a Reales

5.3.- ANÁLISIS EXHAUSTIVO

Una vez observados los resultados reales vemos que los valores predichos están casi alineados en cero, mientras que los valores reales muestran mucha más variabilidad. Esto indica que realmente lo que sucede es que el modelo no está capturando las variaciones dado que los modelos predichos están casi todos en torno a cero.

Las posibles razones en torno a esto es que al transformar los datos a la medida de entrenamiento puede hacer que el modelo no sea capaz de generalizar. La mejor solución a esto es realizar un entrenamiento con más datos pues es probable que los datos de entrenamiento no sean suficientes o no contengan suficiente variabilidad, por lo que el modelo no puede aprender patrones específicos importantes.

5.4 Análisis de Días Clave

El Análisis de Días Clave tiene como objetivo identificar y predecir los días críticos o de crisis en el mercado financiero. Este análisis es fundamental para evaluar cómo ciertos eventos específicos pueden influir de manera significativa en el comportamiento de la bolsa y en la percepción pública, generalmente capturada a través de datos de sentimiento en redes sociales, como los tweets. Para llevar a cabo este análisis, se entrenaron y probaron diferentes modelos de clasificación que pudieran identificar estos días clave basándose en datos históricos de mercado y sentimientos.

Para identificar los días clave en el mercado, se seleccionaron manualmente ciertos eventos importantes de empresas como Deutsche Bank, tales como:

- Anuncios financieros importantes.
- Cambios en la estructura interna de la empresa.
- Publicaciones de resultados trimestrales.
- Crisis económicas o noticias que puedan afectar la percepción pública y los mercados.

Estas fechas clave se utilizaron como punto de referencia para evaluar el impacto de los datos de sentimiento y los valores del mercado.

Como el número de días clave es considerablemente menor que los días no clave, fue necesario equilibrar las clases en el conjunto de datos para evitar que el modelo aprendiera un sesgo hacia la clase mayoritaria (días no clave). Esto se logró seleccionando aleatoriamente un subconjunto equilibrado de días no clave.

Se aplicó validación cruzada para asegurar que el modelo fuera generalizable y no se ajustara únicamente a los datos de entrenamiento. Esto permitió obtener resultados más robustos y confiables.

El modelo entrenado se evaluó en términos de su capacidad para predecir los días clave. Para esto, se utilizaron diferentes métricas de evaluación:

1. **Precisión (Precision)**: Proporción de días predichos como clave que realmente lo eran.
2. **Exhaustividad (Recall)**: Proporción de días clave reales que fueron correctamente identificados por el modelo.
3. **Exactitud (Accuracy)**: Proporción de predicciones correctas, tanto de días clave como de días no clave, sobre el total de días evaluados.

Una vez comprendidos todos los puntos observamos que al realizar la precisión obtenemos estos datos:

```
Clasificador: tree
Tasa de acierto promedio (Accuracy): 30%
Clasificador: svm
Tasa de acierto promedio (Accuracy): 40%
Clasificador: ensemble
Tasa de acierto promedio (Accuracy): 60%
El mejor clasificador es: ensemble
La mejor tasa de acierto promedio es: 60%
```

El mejor Clasificador obtenido ha sido Ensemble con un 60% en los pliegues de la validación cruzada, esto es razonable debido a que estos modelos suelen ser más robustos al combinar varios modelos individuales para mejorar el rendimiento.

```
Precisión (Precision): 1.1527%
Exhaustividad (Recall): 57.1429%
Tasa de acierto (Accuracy): 79.3249%
```

Una vez se realiza la prueba y se observan los días que se consideran como claves obtenemos una precisión de un 1,15287% lo cual es extremadamente baja, lo que significa que la mayoría de las fechas que el modelo ha predicho como clave no lo son, es decir, hay muchos falsos positivos.

Fijándonos en la Exhaustividad que es significativamente mayor, indica que el modelo logra detectar más de la mitad de los días clave reales, lo que indica que el modelo está haciendo un buen trabajo encontrado algunos días clave, pero a costa de hacer muchas predicciones incorrectas.

Por último, el Accuracy es relativamente alto, pero esto es debido a que la mayoría de los días no son clave y al tener muy pocos días clave y el modelo predice correctamente la clase mayoritaria de los días que no son clave.

El mayor problema encontrado es debido a que hay muy pocos días clave en comparación con los días no clave que son el resto de los días de los años estudiados. Aunque se realizó un balance de clases para minimizar esto, claramente el problema persiste en algunos pliegues de la validación cruzada y el modelo sobreestima la cantidad de días clave que hay por lo que genera muchos falsos positivos.

Capítulo 6

6 Conclusiones

6.1 Limitaciones en la implementación

Durante el desarrollo de este proyecto, una de las principales limitaciones encontradas fue la cantidad insuficiente de tweets disponibles relacionados con los Credit Default Swaps (CDS) de las entidades financieras estudiadas. Aunque los datos obtenidos ofrecieron una base inicial para el análisis, no fueron suficientes para obtener conclusiones más concluyentes o extrapolables a otros sectores fuera del ámbito financiero. El volumen de datos, específicamente los tweets asociados a los CDS, resultó ser limitado para realizar un análisis más profundo y significativo, afectando la capacidad del modelo para identificar patrones claros entre los sentimientos expresados en redes sociales y las fluctuaciones en los CDS.

Este hecho refleja que la tecnología utilizada para el análisis, basada en el procesamiento de lenguaje natural y técnicas de aprendizaje automático, es factible y adecuada para este tipo de estudios. Sin embargo, el problema radica en la cantidad y calidad de los datos disponibles. La escasez de tweets relevantes restringió el potencial del análisis predictivo, lo que hace necesario disponer de un mayor volumen de datos para obtener mejores resultados. Así, el modelo mostró un rendimiento limitado no por su metodología o enfoque técnico, sino por la falta de datos que representen con mayor precisión la relación entre la actividad en redes sociales y los riesgos financieros.

Otro desafío fue la necesidad de ajustar la obtención de datos de la plataforma X debido a las restricciones impuestas en su API. Esto obligó a replantear el enfoque, limitando el acceso a tweets de la manera que se había realizado. Para abordar este problema, se tuvieron que buscar alternativas y métodos complementarios para enriquecer el análisis, lo que supuso una modificación del enfoque inicial planteado para la extracción de datos.

Finalmente, un aspecto técnico importante que afectó la implementación fue el rendimiento limitado del equipo de hardware disponible. Este factor no permitió realizar análisis a gran escala o de manera más eficiente, lo que resultó en tiempos de procesamiento prolongados y, en algunos casos, en la necesidad de simplificar o dividir los procesos para que pudieran ejecutarse correctamente. Aunque la tecnología de análisis utilizada es adecuada, la falta de recursos computacionales adecuados ralentizó el progreso y afectó la capacidad para probar simulaciones más complejas y obtener resultados más rápidos y precisos. Con un sistema más potente, el análisis podría haberse desarrollado de manera más fluida y con un mayor alcance experimental.

6.2 Conclusiones de Trabajo

Tras el desarrollo de este proyecto, podemos concluir que, a pesar de las limitaciones mencionadas, las técnicas de análisis basadas en Procesamiento de Lenguaje Natural (NLP) y aprendizaje automático son factibles para investigar la relación entre la actividad en redes sociales, en particular en la plataforma X, y las fluctuaciones en los CDS. Sin embargo, el volumen de datos disponible no fue suficiente para obtener una correlación clara y generalizable entre los sentimientos expresados en los tweets y el comportamiento de los CDS.

El análisis ha demostrado que existe una ligera correlación entre algunas métricas de los tweets, como la intensidad del sentimiento y el número de interacciones, con ciertos movimientos de los CDS. No obstante, estas correlaciones no fueron lo suficientemente consistentes ni significativas para afirmar que los sentimientos en redes sociales, tal como se obtuvieron en este estudio, tienen un impacto directo o predecible en las fluctuaciones de riesgo financiero representadas por los CDS.

El proyecto ha evidenciado que, si se disponen de más datos relevantes y de mejor calidad, sería posible afinar los modelos predictivos y lograr una relación más precisa entre las redes sociales y los indicadores financieros. En particular, si se amplía el campo de estudio a otras empresas o sectores fuera del ámbito financiero, se podrían obtener conclusiones más extrapolables y robustas. El hecho de que el análisis se haya limitado a un grupo reducido de bancos y a una cantidad limitada de tweets ha condicionado la profundidad del estudio, y es por ello por lo que futuras investigaciones deben considerar una expansión en el volumen y diversidad de los datos.

A nivel técnico, la metodología utilizada ha sido sólida. Las herramientas implementadas, como el análisis semántico y de sentimiento, así como los modelos de regresión y clasificación, han demostrado ser apropiadas para este tipo de estudios. El problema radica en la escasez de datos de entrada, lo que impidió explotar el verdadero potencial de estas técnicas. Esto refuerza la conclusión de que, con datos más completos y representativos, los modelos podrían alcanzar un nivel de precisión significativamente mayor.

Se ha demostrado que el análisis de sentimiento en redes sociales y su relación con los CDS es una línea de investigación prometedora, pero requiere un mayor conjunto de datos para ser concluyente. Los resultados obtenidos han sido útiles para comprender mejor las limitaciones y potenciales de la tecnología aplicada en este contexto, y abren la puerta a futuras mejoras y expansiones en este campo de estudio.

6.3 Trabajo futuro

El trabajo realizado en este proyecto ha evidenciado áreas clave que pueden mejorarse y ampliarse en estudios futuros. Las limitaciones observadas, especialmente en lo que respecta al volumen y la calidad de los datos disponibles, sugieren varias oportunidades para optimizar y fortalecer el enfoque de análisis en investigaciones posteriores.

1. Ampliación del conjunto de datos: Como ya se ha comentado reiteradas veces uno de los objetivos futuros debe ser la ampliación del conjunto de datos. Para mejorar la precisión de los modelos predictivos y aumentar la relevancia de los resultados, es esencial obtener un mayor volumen de datos. Esto podría lograrse mediante la expansión del análisis a empresas de otros sectores fuera del ámbito financiero, lo que permitiría verificar si la relación entre la actividad en redes sociales y los CDS es un fenómeno específico del sector financiero o si también se presenta en otros sectores económicos.

2. Utilización de otras fuentes de datos de redes sociales: Dado que la plataforma X (anteriormente Twitter) limitó el acceso a su API durante el transcurso del proyecto, es recomendable explorar otras fuentes de datos de redes sociales para complementar el análisis. Plataformas emergentes como Threads, Reddit, o incluso análisis de comentarios en plataformas de inversión como StockTwits, pueden proporcionar nuevas perspectivas y enriquecer los modelos de análisis. La integración de estas fuentes diversificadas de datos sociales permitiría obtener un panorama más completo de cómo se percibe el riesgo financiero en diferentes medios.

3. Optimización de los modelos predictivos: Si bien los modelos de regresión y clasificación implementados fueron efectivos dentro de las limitaciones del proyecto, es posible mejorar aún más su precisión mediante la incorporación de técnicas de optimización más avanzadas. Modelos más complejos, como redes neuronales profundas (deep learning), o la implementación de modelos híbridos que combinen análisis de redes sociales con otros indicadores macroeconómicos, podrían ofrecer una visión más precisa y detallada de los patrones financieros.

4. Mejora en la infraestructura tecnológica: Durante el desarrollo de este proyecto, las limitaciones del hardware disponible ralentizaron el proceso de análisis y modelado. Para estudios futuros, se recomienda disponer de una infraestructura tecnológica más avanzada, capaz de procesar grandes volúmenes de datos de manera eficiente. El uso de servidores en la nube, soluciones de computación distribuida o el acceso a plataformas de alto rendimiento serían alternativas válidas para acelerar la experimentación y mejorar el rendimiento de los modelos.

5. Exploración de nuevos métodos de análisis de sentimiento: Si bien el análisis de sentimiento implementado en este proyecto se basó en herramientas robustas como VADER, existen otros enfoques más avanzados, como modelos de lenguaje preentrenados tipo BERT o GPT, que podrían capturar de manera más matizada el contexto y la polaridad de los tweets. Estos modelos tienen la capacidad de detectar sutiles cambios de tono y significado, lo que podría mejorar significativamente la precisión en la identificación de sentimientos vinculados a eventos críticos que afectan a los CDS.

6.4 Apreciaciones personales finales

A lo largo del desarrollo de este proyecto, he adquirido una serie de conocimientos y habilidades que han sido fundamentales tanto en el ámbito académico como en el profesional. El proceso me ha brindado la oportunidad de profundizar en áreas clave del análisis de datos y el uso de herramientas avanzadas como MATLAB, así como de comprender la relación entre la actividad en redes sociales y el comportamiento financiero de los mercados.

En primer lugar, uno de los aprendizajes más importantes ha sido la capacidad de gestionar grandes volúmenes de datos no estructurados, como los provenientes de redes sociales, y aplicar técnicas avanzadas de procesamiento de lenguaje natural (NLP) para extraer información relevante. Esto me ha proporcionado una base sólida para abordar problemas complejos en el ámbito de los datos, que es de creciente importancia en el entorno empresarial actual. Además, me ha permitido adquirir una mayor familiaridad con las técnicas de análisis de sentimientos, un área que será cada vez más relevante para entender la percepción pública y su impacto en las decisiones financieras.

Otro aspecto que destaco es la importancia de la flexibilidad y la capacidad de adaptación durante el desarrollo de un proyecto. Los contratiempos, como la limitación en el acceso a la API de X, me obligaron a buscar soluciones alternativas y a replantear el enfoque del estudio. Este tipo de desafíos, si bien complejos, fueron una oportunidad para desarrollar habilidades de resolución de problemas y tomar decisiones informadas en tiempo real, algo fundamental en cualquier contexto profesional.

Asimismo, el proyecto me ha permitido experimentar de primera mano la relevancia del análisis predictivo en el sector financiero, un campo que ha captado mi interés desde hace tiempo. Poder implementar modelos de regresión y clasificación para prever cambios en los CDS en función de los datos de redes sociales fue una experiencia enriquecedora, ya que me permitió vincular dos áreas aparentemente dispares —las finanzas y las redes sociales— en un solo marco de análisis.

Además, este proyecto me ha brindado una perspectiva más clara sobre la importancia de la infraestructura tecnológica en el desarrollo de proyectos de análisis de datos a gran escala. Las limitaciones del hardware con las que me encontré durante el procesamiento de datos fueron un recordatorio de lo esencial que es disponer de recursos adecuados para gestionar grandes volúmenes de información de manera eficiente.

Por último, el proyecto también ha contribuido a mejorar mis habilidades organizativas y de gestión del tiempo. Dado el alcance y la duración del estudio, fue crucial establecer plazos claros y mantener una estructura coherente en cada una de las etapas del proyecto. Este aprendizaje será valioso para futuros proyectos y desafíos profesionales que requieran un alto nivel de planificación y ejecución.

7 Bibliografía

[1] M. Rodríguez-Ibáñez *et al.*, «On the Statistical and Temporal Dynamics of Sentiment Analysis», doi: 10.1109/ACCESS.2020.2987207.

[2] M. Rodriguez-Ibañez, M. Angel Garcia Martínez, F.-J. Gimeno-Blanes, R. Juan Carlos, y J. Doe, «An empirical analysis of the dynamic relation between Twitter communication and credit default swap Article Information», *Journal: Journal of XYZ*, 2024, doi: 10.1234/xyz.

[3] J. Bollen, H. Mao, y X. Zeng, «Twitter mood predicts the stock market», *J Comput Sci*, vol. 2, n.º 1, pp. 1-8, mar. 2011, doi: 10.1016/J.JOCS.2010.12.007.

[4] A. Mittal y A. Goel, «Stock Prediction Using Twitter Sentiment Analysis».

[5] A. Liu, «Data Science and Data Scientist», 2015.

[6] N. Oliveira, P. Cortez, y N. Areal, «The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices», *Expert Syst Appl*, vol. 73, pp. 125-144, may 2017, doi: 10.1016/J.ESWA.2016.12.036.

[7] V. S. Pagolu, K. Nayan, R. Challa, G. Panda, y B. Majhi, «Sentiment Analysis of Twitter Data for Predicting Stock Market Movements», SCOPES.

[8] E. Marsh y D. Perzanowski, «MUC-7 EVALUATION OF IE TECHNOLOGY: Overview of Results», 1998.

[9] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, y P. M. Cuenca-Jiménez, «A review on sentiment analysis from social media platforms», *Expert Syst Appl*, vol. 223, p. 119862, ago. 2023, doi: 10.1016/J.ESWA.2023.119862.

[10] M. Rodríguez-Ibáñez *et al.*, «Sentiment Analysis of Political Tweets From the 2019 Spanish Elections», doi: 10.1109/ACCESS.2021.3097492.

[11] J. Anthony Cookson Colorado -Boulder *et al.*, «Social Media as a Bank Run Catalyst *», 2024.

[12] S. Bales y H. P. Burghof, «Public attention, sentiment and the default of Silicon Valley Bank», *The North American Journal of Economics and Finance*, vol. 69, p. 102026, ene. 2024, doi: 10.1016/J.NAJEF.2023.102026.

[13] M. Nofer y O. Hinz, «Using Twitter to Predict the Stock Market: Where is the Mood Effect?», *Business and Information Systems Engineering*, vol. 57, n.º 4, pp. 229-242, ago. 2015, doi: 10.1007/s12599-015-0390-4.

[14] A. & A. O. Sherry Tiao | Senior Manager, «What Is Big Data?», https://www.oracle.com/es/big-data/what-is-big-data/.

[15] Alba Megías | INEAF, «Credit default swap, ¿qué es y cómo funciona?», https://www.ineaf.es/tribuna/credit-default-swap/.
[16] D. Tori, «Credit default swaps», en *Elgar Encyclopedia of Post-Keynesian Economics*, Edward Elgar Publishing Ltd., 2009, pp. 77-78. doi: 10.17016/feds.2022.023.
[17] «Refinitiv EIKON (Datastream)». [En línea]. Disponible en: http://eikon.thomsonreuters.com/index.html
[18] About us, «Bloomerg», https://www.bloomberg.com/latam/acerca-de-bloomberg/.
[19] BolsaZone, «Yahoo Finance, el portal de información financiera completamente gratuito», https://bolsazone.com/news/yahoo-finance-portal-informacion-financiera-gratuito/.
[20] ámbito, «Google Finance: así funciona la web ideal para inversores », https://www.ambito.com/tecnologia/google-finance-asi-funciona-la-web-ideal-inversores-n5910401.
[21] Ministerio para la tramitación digital y de la función pública, «Graphext », https://datos.gob.es/es/casos-exito/graphext.
[22] Amazon Web Service, «¿Qué es el Procesamiento de lenguaje natural (NLP)?», https://aws.amazon.com/es/what-is/nlp/.
[23] IBM, «¿Qué es el machine learning (ML)? », https://www.ibm.com/es-es/topics/machine-learning.
[24] C.J.Hutto, «vaderSentiment 3.3.2 », https://pypi.org/project/vaderSentiment/.
[25] Prompts para IA, «Guía completa del etiquetado POS en español: Procesamiento del lenguaje natural en España», https://prompt.uno/procesamiento-del-lenguaje-natural/etiquetado-pos-part-of-speech/.
[26] keepcoding, «Reconocimiento de entidades nombradas (NER)», https://keepcoding.io/blog/reconocimiento-de-entidades-nombradas-ner/.
[27] Arimetrics, «Analizador Sintáctico – Parser », https://www.arimetrics.com/glosario-digital/analizador-sintactico-parser.
[28] IBM, «¿Qué son los grandes modelos de lenguaje (LLM)? », https://www.ibm.com/es-es/topics/large-language-models.
[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, y Kristina Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», https://arxiv.org/abs/1810.04805.
[30] A. Novales, «Análisis de Regresión», 2010.
[31] R. Salman y V. Kecman, «Regression as classification», en *2012 Proceedings of IEEE Southeastcon*, 2012, pp. 1-6. doi: 10.1109/SECon.2012.6196887.
[32] MATLAB, «¿Qué es la regresión lineal?», https://es.mathworks.com/discovery/linear-regression.html.
[33] MATLAB, «Regresión de procesos gaussianos», https://es.mathworks.com/help/stats/gaussian-process-regression.html.
[34] MATLAB, «Introducción a Support Vector Machine (SVM)», https://es.mathworks.com/discovery/support-vector-machine.html.
[35] IBM, «¿Qué es un árbol de decisión? », https://www.ibm.com/es-es/topics/decision-trees.
[36] IBM, «¿Qué es el sobreajuste? », https://www.ibm.com/es-es/topics/overfitting.

[37] Miguel Sotaquirá, «Validación cruzada y k-fold cross-validation», https://www.codificandobits.com/blog/validacion-cruzada-k-fold-cross-validation/.
[38] Cristina Gil Martínez, «Análisis de Componentes Principales (PCA)», https://rpubs.com/Cristina_Gil/PCA.
[39] Economipedia, «Desviación estándar o típica: Qué es y ejemplos», https://economipedia.com/definiciones/desviacion-tipica.html.
[40] «Volatilidad».
[41] Plus500, «¿Qué es Índice de Volatilidad VIX?», https://www.plus500.com/es-es/instruments/vix.
[42] Geofferey Pagel, «Tokenización: Todo lo que necesita saber», https://www.weareplanet.com/es/blog/que-es-la-tokenizacion.
[43] MATLAB, «Correlación lineal», https://es.mathworks.com/help/matlab/data_analysis/linear-correlation.html.