UNIVERSIDAD CATÓLICA DEL MAULE

FACULTAD DE CIENCIAS DE LA INGENIERÍA
ESCUELA INGENIERÍA CIVIL INFORMÁTICA

CIENCIA DE DATOS APLICADA A PRÁCTICAS DE GESTIÓN DE RIESGO EN SOCIEDADES CHILENAS QUE REPORTAN A LA COMISIÓN DE MERCADOS FINANCIEROS Y LA AUTOMATIZACIÓN DE PROCESOS REPETITIVOS MEDIANTE UNA APLICACIÓN WEB

Armin Herrera Medel

Ricardo J. Barrientos Rojel

Cristian Martínez

Proyecto de título para optar al Título Profesional de Ingeniero Civil Informático

TALCA, ENERO 2025

UNIVERSIDAD CATÓLICA DEL MAULE

FACULTAD DE CIENCIAS DE LA INGENIERÍA
ESCUELA INGENIERÍA CIVIL INFORMÁTICA

Proyecto de título para optar al Título Profesional de Ingeniero Civil Informático

Ciencia de datos aplicada a prácticas de gestión de riesgos en sociedades chilenas que reportan a la comisión de mercados financieros y la automatización de procesos repetitivos mediante una aplicación web

Armin Brahiam Herrera Medel

COMISIÓN EXAMINADORA **FIRMA**

PROFESOR GUÍA
Mg./Dr. Ricardo Barrientos _____

PROFESOR COMISIÓN
Mg./Dr. Nombre Apellidos _____

PROFESOR COMISIÓN
Mg./Dr. Nombre Apellidos _____

NOTA FINAL EXAMEN DE TÍTULO: _____

TALCA, ENERO 2025

AGRADECIMIENTOS

Dado el culmine del desarrollo de este proyecto para la finalización de este ciclo como estudiante de Ingeniería Civil Informática en la Universidad Católica del Maule, quisiera expresar mi más sincera gratitud a todas las personas e instituciones que hicieron posible este logro.

En primer lugar, a mis padres, José Herrera y Ruth Medel, quienes fueron el pilar fundamental de esta etapa. Su apoyo incondicional, tanto económico como emocional, fue crucial para que pudiera alcanzar este sueño. También quiero destacar a mi abuela Edelmira Cuevas y a mi tío Jonathan Medel, quienes, más allá de mi hogar, me brindaron siempre su respaldo y confianza durante todos estos años de estudio.

Agradezco profundamente a la Universidad Católica del Maule por la oportunidad de formarme profesionalmente en sus aulas y por ser un espacio donde no solo adquirí conocimientos y amistades, sino también valores que llevaré conmigo toda la vida. Quiero extender este agradecimiento a mis profesores, quienes marcaron mi camino académico con su paciencia y guía. De manera especial, expreso mi reconocimiento al profesor Cristian Martínez, mi profesor guía y principal gestor para el desarrollo de esta tesis, cuya orientación y compromiso fueron esenciales para este proyecto.

Asimismo, agradezco al señor Richard Poblete, mi jefe de práctica en Oriencoop, por brindarme la oportunidad de aprender y crecer en el ámbito profesional.

A Dios, mi gratitud infinita por darme la fuerza, la suerte y las circunstancias necesarias para desarrollarme como persona en un contexto lleno de oportunidades.

A todos ustedes, mi más sincero agradecimiento por ser parte de este logro, que no habría sido posible sin su apoyo y confianza. Este es tanto mi triunfo como el de todos ustedes.

RESUMEN

En la presente tesis se llevará adelante un proyecto de Ciencia de los datos con el objetivo de adquirir conocimiento relacionado con las sociedades anónimas chilenas. El trabajo se enfocará en el análisis de las prácticas de gestión de riesgos en sociedades chilenas que reportan a la Comisión para el Mercado Financiero (CMF). En particular, se evaluará el nivel de cumplimiento de las prácticas de gobierno corporativo relacionadas con la gestión de riesgos y sus respectivos desempeños financieros.

El proyecto incluirá análisis exploratorio de datos, técnicas de segmentación mediante algoritmos de clusterización como K-Means y BIRCH, y el entrenamiento de modelos predictivos utilizando estrategias avanzadas de Aprendizaje Automático. Estas herramientas permitirán generar información relevante, antes desconocida.

Adicionalmente, se desarrollará una aplicación web basada en Python, diseñada para automatizar el proceso completo. Esta permitirá la carga de datos, la transformación y estandarización de estos, la selección de parámetros y la visualización interactiva de los resultados. Su propósito es ofrecer una solución práctica e integral para facilitar el análisis y la toma de decisiones en el contexto de análisis corporativo chileno.

Este enfoque integral no sólo busca mejorar la comprensión de las prácticas de gestión de riesgos, sino también promover la Ciencia de datos como un recurso clave para un mejor análisis y entendimiento del objeto de estudio, en este caso empresas.

ABSTRACT

This thesis presents a Data Science project aimed at gaining insights related to Chilean corporations. The research focuses on analyzing risk management practices in Chilean companies that report to the Financial Market Commission (CMF). Specifically, it assessed the compliance level of corporate governance practices related to risk management and their corresponding financial performance.

The project involves exploratory data analysis, segmentation techniques using clustering methods such as K-Means and BIRCH, and the training of predictive models through advanced Machine Learning strategies. These tasks are organized using a work methodology in order to uncover relevant and previously unknown information.

Additionally, a web application is developed with Python to streamline the entire process. This application allows users to upload data, transform and standardize it, select parameters, and interactively visualize the results. Its goal is to offer a practical and comprehensive solution to support analysis and decision-making.

This holistic approach seeks not only to improve the understanding of risk management practices but also to promote Data Science as a vital resource for enhancing the analysis and comprehension of the study subject. In this case, corporations.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Capítulo I.

Introducción y Objetivos.

1. Introducción y Objetivos

En las organizaciones disminuir la incertidumbre asociada a los riesgos se ha convertido en una actividad esencial y al mismo tiempo se busca aumentar el valor de la empresa (Robles et al., 2019). La gestión de riesgos proporciona herramientas para afrontar los diversos riesgos a los que están expuestas las organizaciones. Nasteckienė (2021) plantea que en este proceso coexisten las prácticas formales y no formales de gestión de riesgos, y es así como en las empresas se configura un proceso no lineal, por lo que, no existe una claridad respecto del inicio y final del mismo.

Los avances en TICs desde fines de los 90 a la fecha permitieron una generación exponencial de datos. La gestión y análisis de este gran volumen de datos fue posible gracias al trabajo conjunto de disciplinas como la estadística, matemática, programación, inteligencia artificial, entre otras, dando lugar a un campo conocido como "Ciencia de Datos" (Barbaglia et al., 2021). Su principal objetivo es la obtención de información y conocimiento a partir del análisis de datos estructurados y no estructurados mediante diferentes técnicas y herramientas. Si bien se ha aplicado de manera exitosa en Salud, Educación, Agricultura, Cadena de Suministro y Deportes (Liu y Huang 2017, Cravero y Sepúlveda 2021, Subrahmanya et al. 2022, Jahani et al. 2023, D'Urso et al. 2023), existen limitados aportes en Economía y Finanzas.

En Chile, el avance en la adopción de prácticas de gobierno corporativo ha sido limitado (Moraga y Ropero, 2018). A nivel general de prácticas de gobierno corporativo, Flores y Undurraga (2019) analizan la NCG 385. En su trabajo señalan que no existe una relación entre el sector industrial de la empresa y el grado de adopción de prácticas, así como que no existe una relación entre el grado de adopción de prácticas con el grado de solvencia. De la misma manera, Torres et al. (2021) abordan la NCG 385 concluyendo que la adopción de prácticas de gobierno corporativo no tiene efecto sobre las rentabilidades de las empresas chilenas estudiadas.

En este contexto, las investigaciones referentes a la gestión de riesgos y desempeños financieros han dado resultados contradictorios y se concentran principalmente en empresas financieras (Horvey y Ankamah, 2020), por lo que, el presente trabajo consiste en aplicar

diferentes técnicas de Ciencia de Datos a través del proceso del descubrimiento de conocimiento en base de datos (Fayyad y Stolorz, 1997) a datos financieros y de cumplimiento de prácticas de gobierno corporativo de gestión de riesgos que las sociedades chilenas informan a la Comisión de Mercado Financiero con el propósito de tener un mejor conocimiento de las mismas.

1.1. Problema u Oportunidad

En la actualidad, la necesidad de analizar y comprender datos relacionados con la gestión de riesgos y el desempeño financiero se ha vuelto esencial en el ámbito académico, particularmente para apoyar investigaciones y vinculaciones al medio que contribuyan al entendimiento de prácticas de gobierno corporativo y sus impactos. En este contexto, académicas de la Facultad de Economía de nuestra Universidad han requerido de la aplicación de herramientas avanzadas de análisis para procesar datos y generar conocimiento que respalden futuras acciones.

Aunque existen datos relevantes disponibles, no se han aplicado técnicas avanzadas de Ciencia de Datos para analizar de manera eficaz las relaciones entre indicadores clave relacionados con gestión, estructura organizativa y desempeño financiero. Esta falta de análisis limita la posibilidad de identificar patrones valiosos que puedan contribuir a la toma de decisiones estratégicas y al conocimiento de valor.

En este contexto, esta tesis busca abordar estas limitaciones mediante la aplicación de diferentes técnicas de Ciencia de Datos, incluyendo agrupamiento de datos y predicción. La aplicación de las mismas facilitarán el procesamiento de los datos, la identificación de patrones clave y la clasificación de empresas. La propuesta también incluye el desarrollo de una aplicación web que permita automatizar este proceso dando la facilidad al usuario de cargar su conjunto de datos, para luego visualizarlos, agrupar dinámicamente con diferentes cantidades de grupos, además de entrenar un modelo predictivo con datos previamente etiquetados.

El objetivo general de este trabajo es aplicar técnicas avanzadas de Ciencia de Datos para analizar el cumplimiento de prácticas de gobierno corporativo y su relación con indicadores financieros clave, con el propósito de generar información que facilite la toma de decisiones estratégicas y el entendimiento de los datos.

1.2. Hipótesis

La aplicación de técnicas avanzadas de Ciencia de Datos, como el agrupamiento de datos y las técnicas predictivas, permitirá identificar patrones entre el cumplimiento de prácticas de gobierno corporativo relacionadas con la gestión de riesgos y los indicadores financieros de las empresas chilenas, generando información útil para la toma de decisiones.

1.3. Objetivo General

Llevar adelante un proyecto de Ciencia de Datos para analizar una base de datos de sociedades anónimas chilenas, con el fin de generar conocimiento relacionado con el desempeño de las mismas a nivel financiero y de cumplimiento de prácticas de gobierno corporativo y de gestión de riesgos.

1.3.1. Objetivos Específicos

Los objetivos específicos de esta tesis son los siguientes:

- Limpiar los datos recopilados de manera que sean de calidad y confiables para la aplicación de técnicas posteriores.
- Realizar un análisis exploratorio de datos para alcanzar un conocimiento preliminar de los mismos.
- Desarrollar modelos de agrupamiento basados en técnicas de Aprendizaje Automático, para segmentar las empresas en grupos homogéneos basados en sus prácticas de gobierno corporativo y desempeño financiero.
- Analizar grupos de empresas de manera de caracterizarlas apropiadamente.

- Desarrollar modelos predictivos basados en técnicas de Aprendizaje Automático que permitan clasificar una empresa de manera precisa según características relacionadas con prácticas de gobierno corporativo y desempeño financiero.
- Desarrollar una aplicación web que permita visualizar el procedimiento llevado adelante en este Proyecto de forma amigable y sencilla.

1.4. Alcance del Proyecto

El alcance del presente proyecto incluye el desarrollo de modelos basados en técnicas de Ciencia de Datos y una aplicación web complementaria.

1. **Para el Análisis de Datos**:
    a. Recopilar, pre-procesar, seleccionar y analizar los datos relacionados con las sociedades chilenas, de manera de tener un conocimiento preliminar de las mismas para definir el uso de técnicas precisas del Aprendizaje Automático.
    b. Aplicar técnicas de agrupamiento de datos y de predicción, para identificar patrones y relaciones claves y predecir características relevantes de las empresas.
    c. Generar información útil a partir del análisis de indicadores financieros, como rentabilidad sobre el patrimonio, rentabilidad sobre los activos y razones de endeudamiento, entre otros.
2. **Para la Aplicación Web**:
    a. Diseñar una plataforma web interactiva que permita a los usuarios cargar datos financieros y de gobierno corporativo en formato CSV de manera sencilla.
    b. Facilitar la visualización de resultados mediante gráficos claros e intuitivos.
    c. Incluir funcionalidades que permitan usar y mostrar los resultados de aplicar técnicas de agrupamiento y de predicción sobre los datos cargados.

El proyecto está limitado al análisis de los datos entregados por académicas de la Facultad de Ciencias Sociales y Económicas de la UCM. A su vez, la aplicación web no incorpora la generación de datos sintéticos ya que su ejecución requiere un entendimiento previo de los clústers generados.

1.5. Contribución Esperada

Este proyecto tiene como objetivo generar aportes significativos en diversos ámbitos relacionados con el análisis y la visualización de datos, integrando tecnologías avanzadas y metodologías innovadoras. En particular, se espera contribuir en las siguientes áreas:

- **Metodológica**
  El proyecto establece una metodología replicable para analizar y visualizar datos de diversa índole, combinando técnicas de estadística, aprendizaje automático y visualización interactiva. Este enfoque permite explorar y comunicar características claves de manera clara y efectiva.

- **Tecnológica**
  El proyecto contempla el desarrollo de una solución tecnológica accesible que integra herramientas de Aprendizaje Automático con un enfoque en la comprensión y la interacción con los datos. A través de un entorno web práctico, se ofrece una plataforma que facilita tanto el análisis como la interpretación de los datos de forma visual e intuitiva.

- **Educativa**
  El proyecto promueve el aprendizaje y la aplicación de técnicas avanzadas de análisis y visualización, fomentando el desarrollo de competencias en el uso integrado de estas herramientas junto con tecnologías web. Esto busca capacitar a los usuarios para entender, comunicar y trabajar con datos de manera eficiente y acercar el uso de modelos de Inteligencia Artificial a problemas reales.

Capítulo II.

Marco Teórico.

2. Marco Teórico

A continuación, se presenta el marco teórico que sustenta este proyecto. En esta sección, se analizarán conceptos necesarios para abordar la problemática planteada, de manera de establecer las bases conceptuales de la investigación, facilitando la comprensión de las relaciones entre los datos, las técnicas de análisis y los objetivos del proyecto.

2.1 Gestión de riesgos y desempeño financiero

El riesgo es la combinación de la probabilidad de que ocurra un evento negativo y el impacto o tamaño de ese evento; a mayor probabilidad y mayor pérdida potencial, mayor es el riesgo (Tocabens, 2011) dado que su control es crucial para cualquier organización. La gestión de riesgos se ha consolidado como un elemento clave en la gobernanza empresarial, especialmente en un entorno globalizado. Su enfoque integral, conocido como Enterprise Risk Management (ERM), conocido también como Gestión de riesgo empresarial como muestra la Figura 2.1, busca identificar y mitigar riesgos, integrándolos en la cultura y los procesos operativos de las empresas (Shad et al., 2019). Este enfoque holístico aborda riesgos financieros, operativos y estratégicos, optimizando la toma de decisiones y garantizando la sostenibilidad. Además, la adopción de ERM refleja una gerencia sólida, gestionando eficazmente los recursos y mitigando incertidumbres críticas, lo que los inversionistas valoran como un indicador de estabilidad financiera y crecimiento a largo plazo (Ahmad et al., 2021; Slamet et al., 2023).

Figura 2.1. Gestión de riesgo empresarial (ERM)

Un aspecto particularmente relevante dentro de la gestión de riesgos es el rol del Comité de Gestión de Riesgos (CGR), que actúa como un agente del Consejo de Administración y desempeña un papel decisivo en la implementación y supervisión de estrategias de mitigación. La existencia de este comité no sólo mejora la divulgación sobre riesgos, sino que también ha mostrado un impacto positivo en el desempeño financiero, particularmente en empresas no financieras, donde su aplicación se alinea con las mejores prácticas de gobernanza (Ramlee y Ahmad, 2020). Investigaciones recientes destacan cómo la presencia y características del CGR pueden influir directamente en la percepción y confianza de los inversionistas, fortaleciendo la posición competitiva de las empresas en el mercado (Ayuningtyas y Harymawan, 2022). Además, desde una perspectiva teórica, la teoría de agencia establece que los accionistas delegan en los altos ejecutivos la responsabilidad de administrar la empresa, lo que puede generar conflictos de interés cuando las prioridades de los gerentes difieren de las de los accionistas. En este contexto, el CGR juega un papel crucial al alinear estos intereses y garantizar que las decisiones relacionadas con la gestión de riesgos sean implementadas y monitoreadas adecuadamente (Bui y Krajcsák, 2024).

Por último, la gestión de riesgos no se limita a una función operativa dentro de las organizaciones, sino que representa un elemento estratégico fundamental para enfrentar los desafíos del entorno económico, político y social actual. En países como Chile, iniciativas regulatorias, como la Norma de Carácter General N° 461, han promovido la integración de la sostenibilidad y el gobierno corporativo en las memorias anuales de las empresas supervisadas, fortaleciendo la transparencia y el enfoque en políticas ambientales, sociales y de gobernanza (ASG) (CMF, 2021). Este enfoque regulatorio refleja una tendencia global hacia prácticas de gobernanza más responsables y adaptadas a las expectativas sociales y de mercado (Kot y Dragon, 2015). Por consiguiente, la capacidad de las organizaciones para identificar y gestionar riesgos no tradicionales, como los relacionados con sostenibilidad, se convierte en un diferenciador estratégico que fortalece la resiliencia corporativa y contribuye a su estabilidad a largo plazo. Esto evidencia cómo las prácticas avanzadas de gestión de riesgos, en conjunto con enfoques integrados como ERM, pueden redefinir la forma en que las empresas enfrentan sus retos y garantizan su viabilidad futura (Shad et al., 2019).

2.2 Ciencia de Datos aplicados a los negocios

La Ciencia de Datos ha emergido como una disciplina esencial para la transformación empresarial en la era digital. Su proceso central, conocido como Descubrimiento de Conocimiento en Bases de Datos (KDD), permite extraer patrones válidos, útiles y comprensibles a partir de grandes volúmenes de datos (Fayyad y Stolorz, 1997). Este proceso abarca las siguientes etapas:

- Integración y recopilación de datos
- Selección, limpieza y transformación de datos
- Aplicación de técnicas
- Evaluación e interpretación de resultados

En la Figura 2.2 se muestran las etapas del proceso KDD, el cual es un proceso iterativo y con interacción continua con el interesado del proyecto. Las mismas no sólo garantizan la calidad y relevancia de los datos utilizados, sino que también facilitan una interacción continua con los interesados, asegurando que las soluciones generadas estén alineadas con los objetivos estratégicos de las organizaciones.

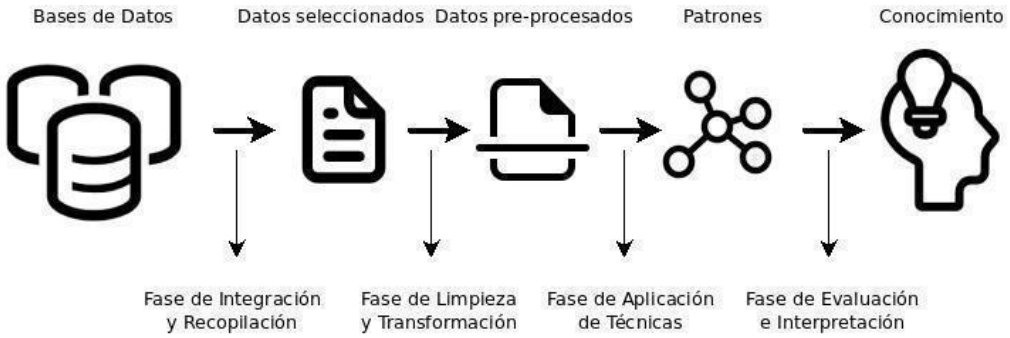

Figura 2.2. Etapas en el proceso KDD (Ibarra, 2020)

Entre las técnicas más utilizadas en Ciencia de Datos, la Estadística ha desempeñado un rol tradicional al ofrecer herramientas como el análisis multivariado, la correlación y métodos de estadística descriptiva. Sin embargo, en años recientes, la Inteligencia Artificial (IA) ha tenido una gran participación. La IA es un área de la Ciencia de la Computación dedicada a construir máquinas capaces de realizar tareas que típicamente requieren inteligencia humana. Es un

sentido más estricto, se relaciona con el desarrollo de soluciones basadas en software, hardware y datos para realizar tareas complejas en tiempo mínimo. Se subdivide en las siguientes áreas:

- Aprendizaje Automático
- Robótica
- Reconocimiento de Patrones
- Sistemas Expertos
- Sistemas de Soporte para la Toma de Decisión

Áreas como los Sistemas Expertos, el Reconocimiento de Patrones y el Aprendizaje Automático (ML), permiten no sólo analizar grandes conjuntos de datos, sino también generar predicciones y automatizar decisiones clave (Plathottam et al., 2023).

El Aprendizaje Automático consiste en interpretar, procesar y analizar datos para resolver problemas complejos. Dentro de esta área podemos encontrar diferentes paradigmas como el Aprendizaje Supervisado, No Supervisado, Semi-Supervisado y por Refuerzo. El Aprendizaje no Supervisado posibilita el descubrimiento de patrones ocultos y la reducción de la dimensionalidad en datos no etiquetados (Shobha y Rangaswamy, 2018). Para ello, se analizan y agrupan datos no etiquetados previamente, encontrando patrones ocultos sin intervención humana. Dentro del Aprendizaje no Supervisado, las técnicas más conocidas son el agrupamiento, la reducción de dimensión y la asociación. K-Means, BIRCH y Componentes Principales permiten identificar relaciones latentes en los datos, aportando valor en tareas como la segmentación de clientes y la detección de anomalías (Zhang et al., 1997). El aprendizaje Supervisado agrupa a algoritmos que son entrenados con datos que incluyen un conjunto de atributos y sus correspondientes etiquetas por cada caso; durante la fase de entrenamiento, éstos aprenden a predecir la etiqueta asociada con los atributos de entrada. Una vez alcanzado un nivel de performance adecuado, pueden ser usados para predecir las salidas (etiquetas) de nuevos casos. En la literatura se han aplicado a problemas de clasificación, regresión y forecasting. Algunas de las técnicas más conocidas son Random Forests, Support Vector Machines y Extreme Learning Machine.

Estos avances tecnológicos, respaldados por marcos teóricos sólidos y un enfoque iterativo, han convertido a la Ciencia de Datos en una herramienta indispensable para empresas que buscan tomar decisiones basadas en evidencia y optimizar sus procesos en un entorno competitivo.

2.3 Técnicas de Aprendizaje Automático

El Aprendizaje Automático es un campo de estudio que se enfoca en desarrollar modelos computacionales capaces de identificar patrones en los datos para realizar predicciones o clasificaciones. En este contexto, técnicas como K-Means, BIRCH, Random Forests (RF) y Extreme Learning Machine (ELM), han demostrado ser herramientas claves debido a su capacidad para abordar problemas supervisados como no supervisados, dependiendo de las características de los datos y los objetivos del análisis.

En el ámbito del Aprendizaje no Supervisado, **K-Means** propuesto por James MacQueen (1967), se destaca como una de las técnicas más utilizadas para el agrupamiento de datos. Este método particiona los datos en un número predefinido de grupos (K), donde cada cluster (o grupo) está representado por un centroide. Su objetivo es minimizar la variación dentro de los clusters, logrando una separación clara entre ellos. Este método es eficiente y escalable, lo que lo hace adecuado para grandes conjuntos de datos. Sin embargo, presenta limitaciones como su sensibilidad a la inicialización de los centroides y su bajo desempeño cuando los clusters no tienen formas esféricas o contienen valores atípicos (Lloyd, 1982). En la Figura 2.3 se puede apreciar un agrupamiento de datos mediante esta técnica.

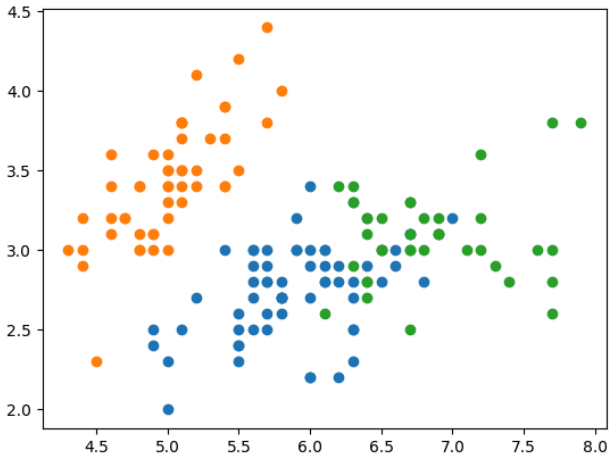

Figura 2.3. Clustering usando K-Means

Para superar algunas de las limitaciones de K-Means, el algoritmo **BIRCH** (Balanced Iterative Reducing and Clustering using Hierarchies), desarrollado por Zhang et al. (1996), ofrece una alternativa eficiente para el agrupamiento de grandes volúmenes de datos. Este método organiza los datos en estructuras jerárquicas llamadas "características de clúster" (CF), que representan resúmenes de los datos originales. A través de estos resúmenes, BIRCH puede realizar agrupamientos preliminares rápidos como podemos ver en la Figura 2.4, los cuales pueden refinarse posteriormente mediante otros métodos como K-Means. Aunque su capacidad para manejar grandes conjuntos de datos es notable, su uso está restringido a datos numéricos, y su rendimiento depende de la configuración adecuada de parámetros como el umbral y el factor de ramificación (Rasmussen et al., 2001).

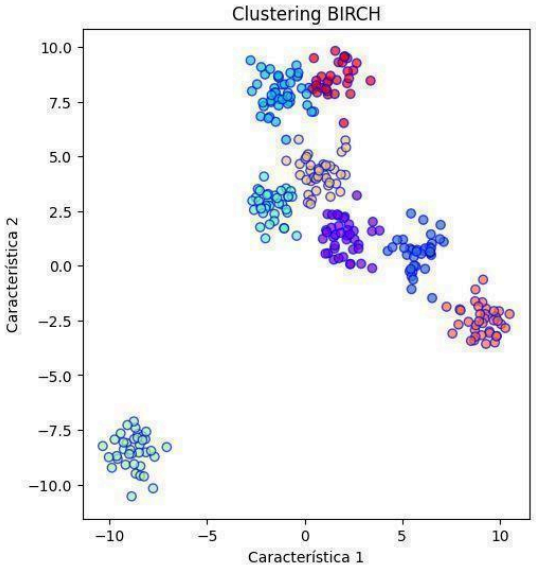

Figura 2.4. Clustering usando BIRCH

Estos algoritmos son fundamentales para proyectos de análisis de datos, ya que permiten abordar diferentes tareas según las características de los datos y los objetivos del análisis. Se han aplicado de manera exitosa a problemas de detección de fraudes, networking, urbanización, entre otros. Más detalles en Hilas y Mastorocostas (2008), Usama et al. (2019) y Wang y Biljecki (2022).

Random Forests (RF) es una técnica de aprendizaje supervisado basado en árboles de decisión propuesto por Breiman (2001). Su funcionamiento se basa en la construcción de múltiples árboles de decisión a partir de subconjuntos aleatorios de datos y características. Las predicciones de estos árboles son combinadas para obtener resultados robustos y precisos, lo que reduce la varianza y el riesgo de sobreajuste. Es especialmente útil en problemas de clasificación y regresión debido a su capacidad para manejar grandes volúmenes de datos, variables tanto categóricas como numéricas, y su capacidad para evaluar la importancia de las características (Liaw y Wiener, 2002). Sin embargo, su principal desventaja radica en el consumo de recursos computacionales cuando se trabaja con conjuntos de datos muy grandes. En la Figura 2.5 se presenta el funcionamiento general de Random Forests.

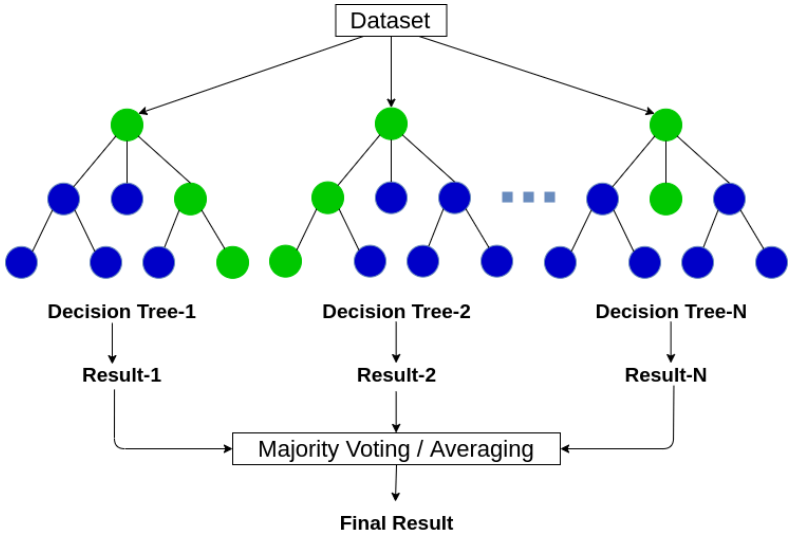

Figura 2.5. Representación de árboles de decisión de Random Forests.

Extreme Learning Machine (ELM) es un método supervisado perteneciente a la familia de redes neuronales, introducido por Guang-Bin Huang, Qin-Yu Zhu y Chee-Kheong (2006). Se caracteriza por su rapidez y simplicidad, ya que las conexiones entre la capa de entrada y la capa oculta son asignadas de forma aleatoria, eliminando la necesidad de ajustes iterativos de parámetros. Los pesos entre la capa oculta y la capa de salida se calculan mediante una solución cerrada basada en mínimos cuadrados. Aunque ELM ha demostrado ser eficaz en problemas de clasificación y regresión, su desempeño puede verse afectado en presencia de datos ruidosos o con una alta correlación entre variables (Kasun et al., 2013). En la Figura 2.6 se presenta la versión original de ELM.

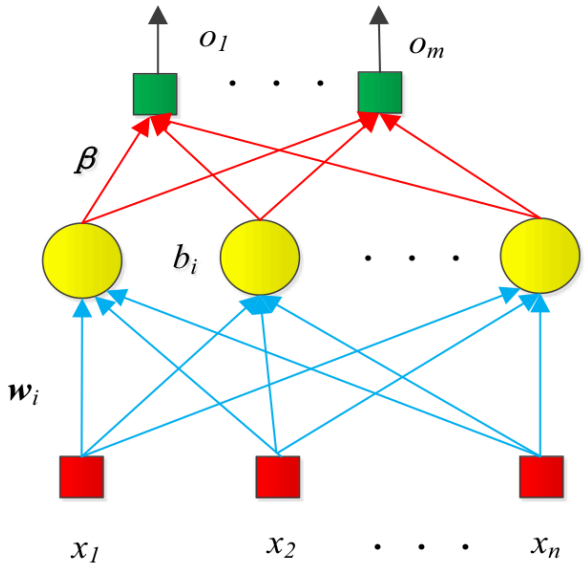

Figura 2.6. Arquitectura básica de una ELM

En la literatura se pueden encontrar numerosas aplicaciones de técnicas de Aprendizaje Supervisado. Entre ellas, aplicadas a problemas de Ecología, imágenes médicas, descubrimiento de drogas, entre otros. Más detalles en Crisci et al. (2012), Aljuaid y Anwar (2022), y Obaido et al. (2024).

Mientras Random Forests y ELM destacan en el ámbito supervisado por su capacidad predictiva, K-Means y BIRCH son esenciales para descubrir patrones y estructuras en los datos sin etiquetas. La combinación de estos enfoques puede ofrecer una visión integral y enriquecedora de los datos, maximizando su valor en aplicaciones prácticas como el análisis financiero, la detección de fraudes y la segmentación de clientes.

2.4 Python y Django: Lenguaje y Framework para Aplicaciones Web

Python es un lenguaje de programación interpretado, de alto nivel y de propósito general, desarrollado por Guido van Rossum en 1991. Conocido por su sintaxis clara y sencilla, ha ganado popularidad en diversas áreas como desarrollo web, análisis de datos, inteligencia artificial y automatización. Su flexibilidad y la capacidad de adaptarse a múltiples paradigmas de programación como la programación orientada a objetos, estructurada y funcional, lo

convierten en una opción versátil para proyectos de cualquier magnitud. Además, cuenta con un ecosistema rico en bibliotecas como NumPy, Pandas y Scikit-learn, que permiten llevar a cabo tareas complejas de manera eficiente. Este enfoque modular y la capacidad de ejecutar código en diversas plataformas sin necesidad de modificaciones, son algunas de las ventajas que ofrece este lenguaje. Su logo de representación se presenta en la Figura 2.7.

Figura 2.7. Logo de Python

Dentro del ámbito del desarrollo web, existen varios frameworks populares que facilitan la creación de aplicaciones robustas y escalables. Flask y Django son dos de los más conocidos, cada uno con sus particularidades y ventajas. Mientras que Flask es más ligero y flexible, Django se destaca por ser un framework más estructurado y de alto nivel. Dado que esta tesis se enfoca en el uso de Django, se profundizará en las características y ventajas que éste ofrece.

Es un framework de desarrollo web basado en Python que se distingue por su enfoque en la rapidez, la seguridad y la escalabilidad. Desde su lanzamiento en 2005, ha sido reconocido por simplificar el proceso de desarrollo de aplicaciones web, permitiendo que los desarrolladores se concentren en la lógica de negocio en lugar de en tareas repetitivas. Sigue la arquitectura Model-View-Template (MVT) como se muestra en la Figura 2.8, lo que facilita la separación de la lógica de negocio, la presentación y la interacción con los datos. Además, Django integra un sistema de ORM (Object-Relational Mapping), lo que elimina la necesidad de escribir consultas SQL manualmente, simplificando la interacción con bases de datos, además de utilizar la gestión de los datos a través del uso de objetos de Python.

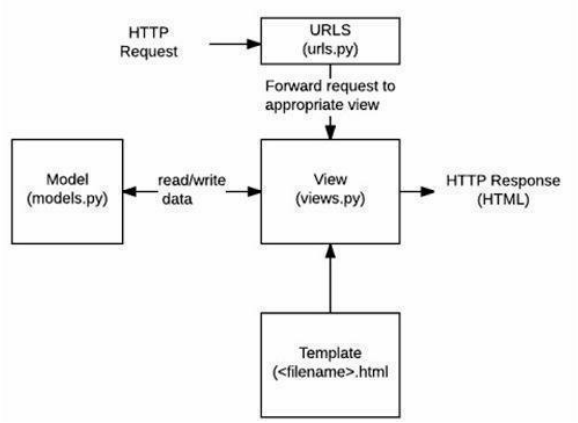

Figura 2.8. Estructura utilizada para el desarrollo en Django (MVT)

La arquitectura Modelo-Vista-Plantilla (MVT) seguida por Django se forma de:

1. Modelo: Representa los datos y lógica de negocio. Esta sección define la estructura de los datos y cómo se interactúan con ellos; además se encarga de recibir las consultas de la vista para obtener los datos de la base de datos y también de enviar los datos procesados para ser mostrados por una plantilla HTML (template).
2. Vista: Contiene la lógica de la aplicación. Esta sección actúa como intermediario entre el modelo y las plantillas utilizando la recepción de solicitudes HTTP desde el cliente (frontend) para devolver lo solicitado.
3. Template: Presenta los datos ante el usuario. Se encarga de renderizar la información enviada desde la Vista mediante contenido en formato HTML.

El sistema de plantillas (Template) de Django permite renderizar HTML dinámico, facilitando la creación de páginas web interactivas y actualizadas en tiempo real. Su amplia documentación y comunidad activa proporcionan un soporte constante, lo que hace que sea una herramienta robusta y confiable para el desarrollo web. Además, incluye mecanismos de seguridad integrados, protegiendo las aplicaciones contra amenazas comunes como CSRF, XSS y SQL Injection. Esta combinación de características hace de Django una opción ideal para proyectos web que requieren una base sólida, seguridad avanzada y facilidad de escalabilidad.

En conclusión, las herramientas de desarrollo utilizadas para llevar a cabo este proyecto se indican en la Tabla 2.1.

Tabla 2.1. Tecnologías utilizadas en el proyecto

Nº	Nombre	Descripción
1	Python	Lenguaje de programación
2	Django	Framework de desarrollo web
3	Chart.js	Librería de JavaScript para la visualización de gráficos
4	SQLite	Sistema de gestión de bases de datos
5	Pandas	Biblioteca para la manipulación y análisis de datos estructurados
6	Sklearn	Conjunto de herramientas para Aprendizaje Automático, que incluye modelos, preprocesamiento y evaluación
7	Matplotlib	Biblioteca para crear gráficos estáticos, animados e interactivos en Python
8	Seaborn	Biblioteca para visualización de datos estadísticos, basada en matplotlib
9	Json	Módulo para trabajar con datos en formato JSON
10	Numpy	Biblioteca para el cálculo numérico eficiente, especialmente con arreglos multidimensionales
11	Imblearn	Biblioteca para el manejo de datos desbalanceados. Incluye técnicas de submuestreo y sobre muestreo
12	Sdv	Biblioteca para generar datos sintéticos

En conclusión, luego de haber presentado los diferentes temas abordados, podemos afirmar que la integración de los mismos ofrece una comprensión más profunda sobre cómo la gestión de riesgos, el análisis de datos y las tecnologías emergentes se complementan para mejorar la toma de decisión. La cohesión entre técnicas avanzadas de análisis, como el Aprendizaje Automático y la Ciencia de Datos, con el desarrollo de soluciones tecnológicas mediante Python y Django, permite a las organizaciones no solo enfrentar los desafíos actuales, sino

también anticiparse a riesgos futuros. Así, la combinación de estas herramientas y enfoques proporciona un marco robusto para optimizar el desempeño organizacional, fomentar la innovación y mejorar la resiliencia en un entorno cada vez más competitivo y dinámico.

Capítulo III.

Estado del arte.

3. Introducción

En esta sección se discutirán artículos relacionados con la gestión de riesgos, las prácticas de gobierno y el desempeño financiero de empresas que en general, cotizan en Bolsa. El análisis de los mismos tiene por objetivo mostrar el abordaje que han tratado diferentes autores al respecto y cuál es la posición adoptada en este Proyecto.
Finalmente, se hará una mención respecto al uso de tecnologías sobre el problema de estudio.

3.1. Harvey y Ankamah (2020): Enterprise risk management and firm performance: empirical evidence from Ghana equity market

Los autores analizan la relación lineal y no lineal entre la gestión de riesgos empresarial y el desempeño de las empresas en el mercado de valores de Ghana. Para ello, analizaron empresas durante el período 2010-2016 mediante técnicas de regresión. Algunos puntos salientes del trabajo son:

- El conjunto de datos se forma de 30 empresas participantes en la Bolsa de Valores de Ghana. Los atributos de entrada son el índice de ERM (entre 0 y 1), tamaño de la empresa, eficiencia (gastos operativos/activos totales), propiedad (diferencia entre dueños locales y extranjeros), edad (antigüedad en el mercado), y apalancamiento. Los atributos dependientes son ROA, ROE y el índice Q de Tobin.
- Propusieron 3 modelos lineales de regresión para predecir valores de ROA, ROE y el índice de Tobin.
- Los autores muestran que las empresas grandes y de capitales extranjeros tienen mejores valores de ROE y Q de Tobin.

Si bien el trabajo es de interés por el desarrollo de 3 modelos predictivos, existen algunos aspectos que podrían mejorarse. Primero, el tamaño del conjunto de datos es reducido por lo que los modelos pueden no ser buenos para predecir la performance de nuevas empresas. Por otra parte, no se analiza en profundidad el conjunto de datos: el análisis es de tipo univariado. Finalmente, no proponen la comparación de resultados de los modelos desarrollados con otras técnicas provenientes del Aprendizaje Automático.

3.2. Gouiaa (2018): Analysis of the effect of corporate governance attributes on risk management practices

El autor examina cómo los atributos de la gobernanza corporativa, en particular los relacionados con el Consejo de Administración, afectan a las prácticas de gestión de riesgos en empresas de Canadá que cotizan en la Bolsa. Entre otros:

- El conjunto de datos corresponde a 162 empresas no financieras de Canadá, durante el período 2012-2013. Entre los atributos de entrada se encuentran el tamaño del Consejo de Administración, porcentaje de directores independientes en el Consejo, promedio en años de los directores, frecuencia de reuniones, si están separados los comités de riesgos y el comité auditor, ROA, tipo de industria, entre otros. El atributo de salida es un índice de manejo del riesgo.
- Proponen un modelo de regresión para predecir el índice de manejo del riesgo en función de atributos fuertemente ligados al Consejo de Administración de las empresas.
- El autor destaca que las empresas con consejos más grandes, independientes y con más reuniones, presentan menores niveles de riesgo.

Aunque el trabajo es interesante porque mide el manejo del riesgo, existen aspectos que merecen un mejor tratamiento. No se realiza un análisis exploratorio de datos profundo, limitándose a estadística descriptiva univariada. Por otra parte, ciertos atributos del conjunto de datos son subjetivos como el porcentaje de directores independientes o la frecuencia de reuniones del Consejo. A su vez, la falta de aplicación de técnicas de Aprendizaje Automático no permiten analizar diferentes tipos de relaciones entre atributos y empresas como también predecir diferentes comportamientos de las mismas.

3.3. Moraga Flores y Ropero Moriones (2017): Gobierno corporativo y desempeño financiero de las empresas chilenas

En este trabajo se analiza la adopción de las prácticas de gobierno corporativo en empresas chilenas inscritas en la Superintendencia de Valores y Seguros, durante el período 2015-2016. A continuación, se indican algunos puntos destacables:

- El conjunto de datos está formado por 197 empresas chilenas, de las cuales 159 cotizan en la Bolsa de Comercio de Santiago.
- Proponen un indicador de adopción de gobierno corporativo (IAGCEx) que depende de una relación lineal del promedio de las prácticas de gobierno. A continuación calculan la correlación de este indicador y la rentabilidad neta, el retorno sobre el patrimonio (ROE) y el índice Q de Tobin.
- Los autores muestran que la adaptación de prácticas de gobierno corporativo no influye directamente en la rentabilidad neta, retorno sobre patrimonio y el índice Q de Tobin.

La propuesta presenta algunas oportunidades de mejora. En primera instancia, los datos se limitan a un único período. Por otra parte, el análisis de los mismos necesita un mayor desarrollo y análisis. Finalmente, los autores proponen un único modelo de regresión lineal de las prácticas corporativas para luego calcular correlación con indicadores de desempeño financiero. En este sentido, carece del uso de técnicas más avanzadas de Aprendizaje Automático como así también la posibilidad de predecir la performance de las empresas en función de diferentes atributos de las mismas.

3.4. Uso de Software

Si bien se disponen de herramientas como Google Colab, R, SPSS, entre otras, las mismas no realizan un estudio automático sobre un conjunto de datos con desarrollo e interpretación de diversos modelos basados en Estadística y Aprendizaje Automático. Se necesita que el conjunto de datos sea tratado por un equipo de Ciencia de Datos a efectos de definir modelos que generen conocimiento de valor. Por otra parte, la aplicación web que se pretende alcanzar requiere de conocimientos de tecnologías web como también de Ciencia de Datos para vincular los modelos desarrollados.

Teniendo en cuenta la discusión de artículos relacionados, la falta de un análisis profundo de datos, el uso de diversos modelos basados en técnicas de Aprendizaje Automático como así también una herramientas (en este caso) web que facilite el análisis y presentación de resultados, se considera que las actividades a desarrollar en el presente Proyecto son pertinentes y relevantes.

En la próxima sección se brindarán detalles de cómo se llevará adelante el proyecto de Ciencia de Datos.

Capítulo IV.

Metodología y Propuesta de Solución.

4. Metodología y Propuesta de Solución

Este capítulo describe las estrategias metodológicas y técnicas adoptadas para abordar la problemática planteada. Se explica el producto final propuesto, la metodología seleccionada para guiar el proceso de desarrollo, y los estudios de factibilidad que respaldan la viabilidad del proyecto.

4.1. Producto Final

El producto principal de este proyecto es una aplicación web desarrollada en Python utilizando el framework Django, diseñada para facilitar la automatización de procesos analíticos y predictivos. Esta herramienta está orientada a usuarios no técnicos y proporciona una interfaz visual intuitiva, dentro de las principales funciones del producto tenemos:

- Carga de conjunto de datos,
- Visualización gráfica del conjunto de datos,
- Preparación y transformación de datos,
- Entrenamiento de modelos basados en Aprendizaje Supervisado y No Supervisado,
- Visualización de resultados de los modelos basados en Aprendizaje Supervisado y No Supervisado.

La arquitectura de la aplicación se basa en un modelo **cliente-servidor**:

1. **Backend:** Implementado en Django junto a Python, maneja la lógica del negocio, la carga de datos, el preprocesamiento, el entrenamiento de modelos y el almacenamiento en bases de datos relacionales (SQLite). Es importante destacar que gran parte del desarrollo será utilizando el paradigma orientado a objetos.
2. **Frontend:** desarrollado con HTML, CSS (Bootstrap), y JavaScript (Chart.js y otras bibliotecas). Principalmente, está orientado y propuesto para ofrecer visualizaciones sencillas y amigables con formularios interactivos.
3. **Integración con Aprendizaje Automático:** a través de bibliotecas como scikit-learn, pandas y numpy, junto con la integración que ofrece Python y Django, las funcionalidades antes descritas serán implementadas de manera sencilla y rápida.

La aplicación permitirá que los usuarios interactúen con modelos de IA sin necesidad de escribir código, promoviendo el acceso a tecnologías avanzadas mediante un enfoque visual y amigable.

4.2. Adaptación Metodológica

Para el presente proyecto, se emplea la metodología **CRISP-DM** (Cross Industry Standard Process for Data Mining), CRISP-DM (Cross Industry Standard Process for Data Mining), lanzada en 1997 bajo el financiamiento del Programa ESPRIT de la Unión Europea y liderada por empresas como SPSS, Teradata, Daimler AG, NCR y Ohra, es una de las más utilizadas en Minería de Datos/Ciencia de Datos. Su primera versión se presentó en 1999 y se publicó como una guía paso a paso para proyectos de minería de datos (Gallardo, 2018). Esta metodología consta de varias etapas que guían el proceso de análisis de los datos.

1. **Comprensión de los datos:**
   Una vez obtenidos los datos del interesado del proyecto, estos son analizados para una comprensión de los mismos de manera de definir/determinar nuevos datos y próximas tareas a realizar.
2. **Preparación de los datos:**
   A efectos de desarrollar modelos específicos y obtener resultados confiables, previamente es necesario llevar adelante tareas de limpieza sobre los datos recolectados.
3. **Modelado:**
   En esta etapa se desarrollan modelos y se aplican técnicas (principalmente de Estadística e Inteligencia Artificial) de manera de alcanzar conocimiento sobre los datos.
4. **Evaluación:**
   Los resultados obtenidos en la etapa anterior son evaluados de manera de realizar ajustes en caso de ser necesario como también para el desarrollo de nuevas propuestas.
5. **Despliegue:**
   Se pone en producción la solución obtenida durante el proyecto. En este punto también

es importante mencionar que el entregable puede ser un informe, una aplicación, una API, entre otros.

La adaptación considera los siguientes puntos:

- **Tiempo:** El proyecto se llevará a cabo en 4 meses, considerando etapas iterativas e incrementales.
- **Equipo:** Desarrollo a cargo del estudiante, con retroalimentación de un profesor experto en el área de Inteligencia Artificial y asesoría de dos académicas de la Facultad de Ciencias Sociales y Económicas de nuestra Universidad.
- **Entregables:**
    - Análisis exploratorio de datos del conjunto de datos inicial
    - Análisis exploratorio final y modelos de agrupamiento de datos y predictivos
    - Prototipo funcional de la aplicación web.
    - Informe general de la problemática abordada.

4.3. Carta Gantt

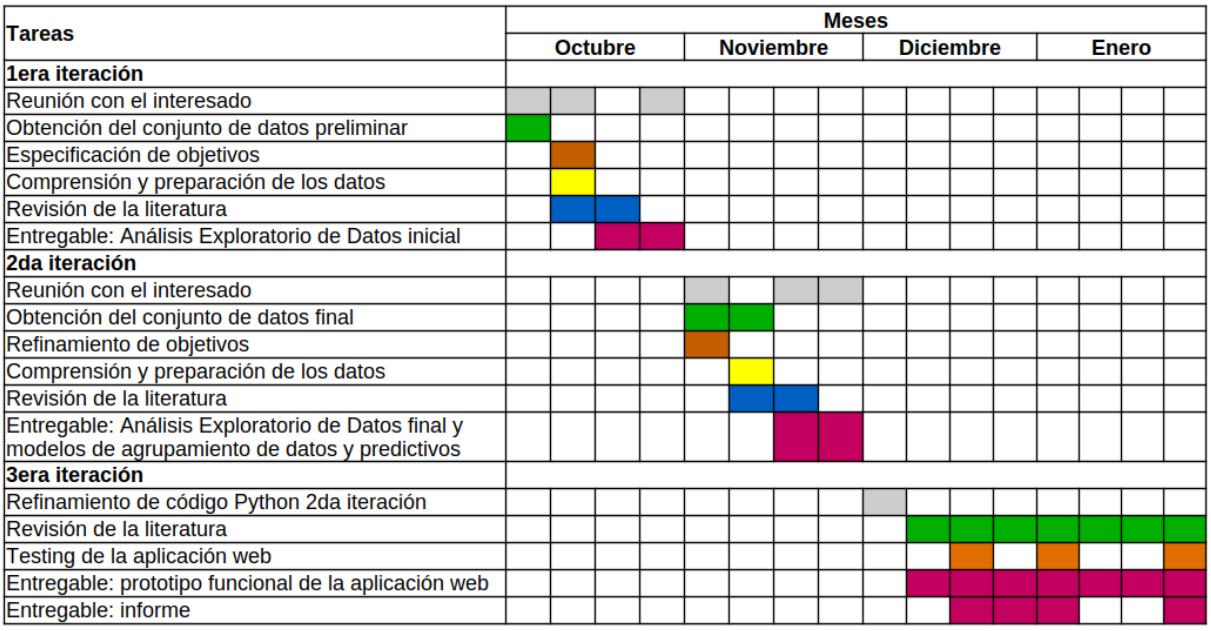

Figura 4.1. Carta Gantt

En la Figura 4.1 se presenta el cronograma de actividades, estructurado en 3 iteraciones. Como se puede observar, en cada una de ellas hay al menos un entregable. En particular, las primeras 2 iteraciones están relacionadas fuertemente con la aplicación de CRISP-DM al problema de estudio. En ambas iteraciones, se pueden apreciar las actividades de comprensión del negocio, comprensión y preparación de los datos, modelado, evaluado y despliegue. En base a esto último y como ya se mencionó, al final de cada iteración se presenta un entregable y al final del proyecto, se espera el despliegue de uno de los entregables de la tercera iteración. Por otra parte, la tercera iteración está relacionada con el desarrollo de aplicación web y del informe final. Gracias a las prestaciones de Python y Django, el código desarrollado en la 2da iteración es usado fuertemente de base para la tercera y última iteración. Es importante destacar que en esta última iteración, la revisión de la literatura se relaciona con el manejo de las tecnologías necesarias para la aplicación web y para la redacción del informe.

Finalmente, para este caso, los interesados del Proyecto son las académicas de la UCM junto con la Comisión Examinadora.

4.3. Estudio de Factibilidad

4.3.1. Factibilidad técnica

El desarrollo del presente Proyecto requiere de tecnologías, servicios y hardware que permitan un normal desarrollo de las actividades y finalmente, la presentación de los entregables en tiempo y forma.

En cuanto a tecnologías, mediante Google Colab, se tiene acceso a Python y otras tecnologías indicadas en la Tabla 2.1. Con las tecnologías listadas previamente, es posible llevar adelante las tareas relacionadas con Ciencia de Datos. Para el desarrollo de la aplicación web, es necesario disponer de Python, Django y otras tecnologías también indicadas en dicha tabla. Las mismas son accesibles en este caso, a través de Visual Studio Code.

Respecto a hardware, para las tareas relacionadas con Ciencia de Datos y de desarrollo web, se requiere al menos de un procesador I5 con 8GB de RAM y al menos 250GB de disco.

Por otra parte, el uso de Google Colab y el uso de la aplicación web, requieren al menos una conexión a Internet de 100Mb. de ancho de banda.

Debido a que se dispone de las tecnologías, el hardware mínimo y la conexión a Internet, es proyecto es técnicamente viable.

4.3.2. Factibilidad operativa

Para que el proyecto sea exitoso es necesario que el entregable sea desplegado y usado. En ese sentido, la propuesta del uso de CRISP-DM con un enfoque iterativo incluyendo reuniones semanales y quincenales con el interesado del proyecto, permite no sólo una evaluación constante de lo realizado sino también una familiaridad con las tecnologías involucradas.

La propuesta de una aplicación web sencilla, amigable y familiar a los interesados del proyecto refuerza su futura usabilidad. Por otra parte, la decisión de desarrollar una aplicación web en vez de un informe técnico, motiva a que los interesados usen el prototipo y su vez, adquieran nuevos conocimientos de manera autónoma.

Finalmente, el desarrollo del proyecto está a cargo de un estudiante con supervisión directa de un profesor guía experto en inteligencia artificial, asegurando la orientación técnica adecuada y el cumplimiento de los objetivos.

Por lo antes expuesto el proyecto es operativamente factible.

4.3.3. Factibilidad económica

Análisis de Costos

En la Tabla 4.1 se detallan los costos relacionados con el Proyecto. En particular, están relacionados con honorarios y servicios mínimos para asegurar la concreción del mismo. Por otra parte, no se indican los costos asociados para que la aplicación web esté en producción, costos de mantenimiento de la solución como tampoco operativos actuales puesto que el producto final no reemplaza a alguno que esté actualmente en funcionamiento. Finalmente, los honorarios indicados fueron obtenidos del Colegio de Ingenieros de Chile.

Tabla 4.1. Desglose de costos

Item	Costo Mensual CLP	Costo Total CLP
Honorarios Desarrollador Full Stack (full time)	$3.000.000	$12.000.000
Honorarios Asesoría de Experto en Inteligencia Artificial (10 hs semanales)	$800.000	$3.200.000
Licencias de Software	$0	$0
Conectividad	$40.000	$160.000
Gastos operativos (impresión, traslados, etc)	$200.000	$800.000
Total del proyecto		$16.160.000

La información proporcionada en relación a los honorarios de los profesionales fue obtenida en publicaciones oficiales del Colegio de Ingenieros de Chile agregada a los anexos

Análisis de Beneficios

El producto final de este proyecto presenta grandes oportunidades de retorno de inversión. Entre ellas:

- Servicio de consultoría, por parte de los interesados a empresas y organismos del Estado en temas relacionados principalmente con la gestión de riesgos empresarial,
- Licenciamiento a terceros interesados en utilizar el producto final.

Sólo a modo de ejemplo, una asesoría sobre el mismo problema de estudio, a cargo de una empresa tecnológica con sede en Chile, tiene un costo aproximado de 10.000.000 CLP.

Considerando el desglose de costos y los beneficios posibles, el presente Proyecto es factible económicamente.

4.3.4. Factibilidad legal

El desarrollo del Proyecto respeta los derechos de propiedad intelectual y de licencias asociadas con las tecnologías utilizadas.

Django y otras tecnologías indicadas en la Tabla 2.1 están bajo licencias permisivas, permitiendo su uso comercial y académico sin restricciones.

Por otra parte, el conjunto de datos usado es público y de libre acceso. Asimismo, se cumple con normativas de protección de datos, ya que no se almacena información sensible ni requiere datos personales que infrinjan leyes de privacidad como la Ley de Protección de Datos Personales de Chile (Ley N°19.628).

Por lo antes indicado, el proyecto es viable legalmente.

4.3.5. Factibilidad Social

El Proyecto además de buscar dos entregables de interés, tiene un impacto social positivo.

La aplicación web permite a usuarios no provenientes de las Ciencias de la Computación, la posibilidad de acceder a herramientas avanzadas de análisis y predicción sin necesidad de conocimientos técnicos. A su vez, se promueve el aprendizaje y la democratización de tecnologías de Inteligencia Artificial, beneficiando a comunidades académicas y empresariales interesadas en la gestión de datos.

De esta manera, se considera que el proyecto es factible socialmente.

4.3.6. Factibilidad Ambiental

En el marco del Proyecto se buscó que las técnicas aplicadas sean lo más eficientes posible. Es decir, que hagan uso del menor cómputo posible y los resultados sean altamente satisfactorios. Por otra parte, para el entrenamiento y uso de los modelos se requiere únicamente de CPU. En cuanto a las tecnologías usadas para la aplicación web, no requieren de un alto cómputo o almacenamiento de datos.

Todo esto refuerza la viabilidad ambiental y compromiso con prácticas sostenibles, siendo este proyecto ambientalmente viable.

Capítulo V.
Desarrollo de la Propuesta y Resultados.

5. Desarrollo de la Propuesta y Resultados

5.1 Iteracion 1

5.1.1 Reuniones con el Interesado del Proyecto

Se organizaron 3 reuniones durante la primera iteración. La primera tuvo como objetivo conocer el problema de estudio propuesto por las académicas de la Facultad de Ciencias Sociales y Económicas de nuestra Universidad. En esta primera instancia, se revisó un artículo de la literatura para contextualizar el problema de la gestión de riesgos empresarial.

La segunda reunión estuvo dirigida a discutir sobre el primer conjunto de datos a analizar. También, a revisar objetivos del Proyecto.

La tercera estuvo enfocada en la presentación de un análisis de datos preliminar como también la propuesta de actividades posteriores.

5.1.2 Base de datos preliminar

Los datos fueron entregados por los interesados del proyecto. Estos provienen de la base de datos pública de la Comisión del Mercado Financiero (CMF), sobre prácticas de gobierno corporativo de sociedades anónimas abiertas fiscalizadas por esta entidad y que continuaban reportando bajo la Norma de Carácter General N°385, con información disponible hasta el 23 de noviembre 2023. El formulario está compuesto por 99 prácticas; sin embargo, se consideraron 22 del cuestionario que corresponden a la categoría de control y gestión de los riesgos correspondiente al periodo 2020. Las respuestas son de carácter dicotómica, indicando “SI” en el caso de que la sociedad esté adoptando la práctica y “No” si no la adopta, para efectos del análisis de los datos, Si = 1 y No = 0. Para un análisis más general sobre las preguntas, se agruparon en 4 subcategorías que miden el cumplimiento de las prácticas. En la Tabla 5.1 se presenta un detalle de los atributos.

Tabla 5.1. Descripción de la base de datos preliminar

Atributo	Tipo de dato	Operación de cálculo
Procedimiento de gestión y control de riesgos (VAR 1)	Numerico (%)	Sumatoria de prácticas adoptadas / 9 practicas
Canal de denuncias (VAR 2)	Numerico (%)	Sumatoria de practicas adoptadas / 4 practicas
Diversidad y sucesión (VAR 3)	Numerico (%)	Sumatoria de prácticas adoptadas / 5 practicas
Estructura salarial y compensación de la alta gerencia (VAR 4)	Numerico (%)	Sumatoria de prácticas adoptadas / 4 practicas
Industria	Carácter	Rubro de la empresa

5.1.3 Análisis de Datos preliminar

A los efectos de no extenderse en el informe, el análisis preliminar de datos se presentará en la siguiente iteración.

Sin embargo, es importante notar que los atributos listados en la tabla anterior no son suficientes para un estudio profundo de Ciencia de Datos. Por otra parte, según se observa en la sección 2 referida al estado del arte, se necesitan datos adicionales para un mejor estudio y posterior obtención de conocimiento sobre las empresas. La falta de un estudio profundo ausente en la literatura, motivó a los interesados del proyecto a conseguir datos adicionales de las empresas chilenas y artículos relacionados con la gestión de riesgos.

5.2 Iteración 2

5.2.1 Base de datos final

A los indicados en la Tabla 5.1, se agregaron datos financieros de las empresas chilenas. Esta se obtuvo de las estadísticas públicas de la CMF, referente a los indicadores financieros bajo estándar IFRS de emisores de valores, correspondiente al cierre de los estados financieros al 31 de diciembre de 2020.

De esta manera, la base de datos consta de respuestas a 22 preguntas sobre diferentes prácticas de gobierno corporativo de gestión de riesgos, información financiera y rubro por cada empresa. Por otra parte, considerando que los datos recolectados corresponden a cumplimiento de prácticas empresariales y datos financieros, la cantidad de empresas con datos completos asciende a 97. En la Tabla 5.2 se listan los atributos.

Tabla 5.2. Descripción de la base de datos final

Atributo	Tipo de dato	Operacion de cálculo
Procedimiento de gestión y control de riesgos (VAR 1)	Numerico (%)	Sumatoria de prácticas adoptadas / 9 practicas
Canal de denuncias (VAR 2)	Numerico (%)	Sumatoria de practicas adoptadas / 4 practicas
Diversidad y sucesión (VAR 3)	Numerico (%)	Sumatoria de prácticas adoptadas / 5 practicas
Estructura salarial y compensación de la alta gerencia (VAR 4)	Numerico (%)	Sumatoria de prácticas adoptadas / 4 practicas
Rentabilidad sobre el patrimonio (ROE)	Numerico (%)	Ganancia (pérdida) *100 / Patrimonio total - Ganancia (pérdida)
Rentabilidad sobre los activos (ROA)	Numerico (%)	Ganancia (pérdida) *100 / Total de activos
Capital (CT)	Numerico (%)	(Activos corrientes totales - Pasivos corrientes totales) / 1000
Razón de endeudamiento (RE)	Numerico (%)	Total pasivos / Patrimonio Total
Industria	Carácter	Rubro de la empresa

5.2.2 Transformación de datos

Para entrenar modelos de Aprendizaje Automático de manera eficiente, es fundamental trabajar con datos numéricos, ya que la mayoría de los algoritmos no son compatibles con

datos categóricos. Por esta razón, se realizaron transformaciones específicas para estandarizar y adaptar los datos de acuerdo con sus características. En el caso de las columnas numéricas, se aplicó un proceso de estandarización utilizando herramientas de Scikit-learn. Este enfoque garantiza que los valores de estas columnas se encuentren en una escala uniforme, facilitando su procesamiento por los algoritmos y mejorando el rendimiento del modelo.

Para las columnas categóricas, se implementó una transformación que convierte cada categoría en un valor numérico único. Este procedimiento, conocido como codificación de categorías, asegura que las variables categóricas sean interpretadas correctamente por los algoritmos sin introducir sesgos relacionados con el orden o magnitud, ya que se limita a asignar identificadores únicos a cada categoría. Ambas técnicas combinadas no sólo preparan los datos de manera adecuada para los modelos, sino que también mejoran la eficiencia del entrenamiento y la capacidad del modelo para generalizar.

5.2.3 Análisis exploratorio de datos

Se realizó un Análisis Exploratorio de datos (EDA) de manera de tener un primer contacto con los datos (Meloun y Militký, 2011). A través de estadística descriptiva, correlación, gráfica univariada, bivariada y multivariada y del análisis de los mismos, se avanzó en técnicas que describen mejor la relación entre los datos utilizados**.**

Como parte del análisis exploratorio de datos, se presenta la Tabla 5.3.

Tabla 5.3. Estadística descriptiva

Atributo	Promedio	D. Estándar	Minimo	Maximo
VAR1	0.430699	0.401661	0	1
VAR2	0.546392	0.454047	0	1
VAR3	0.257732	0.358188	0	1
VAR4	0.185567	0.240113	0	1
ROE	6.557508	27.662551	-68.015100	172.631900
ROA	1.531478	16.185785	-129.283400	44.734800

CT	3.038212e+07	1.129428e+08	-1.807588e+07	1.072535e+09
RE	1.073257	1.739010	0	14.458600

A nivel general, las empresas en promedio no superan el 55% de adopción de prácticas de gobierno corporativo, siendo el canal de denuncias la práctica con mayor nivel de adopción (54,6% de las empresas), en contraste con la estructura salarial y compensación de la alta gerencia que es la menos adoptada por las entidades (18,5% de las empresas). En relación al rendimiento financiero, se observa que en promedio son resultados positivos, aunque son empresas que presentan mayores deudas que recursos propios. En este sentido, se ha de considerar que los indicadores financieros presentan una alta dispersión de datos con respecto de la media, a diferencia, del nivel de adopción que tiene una menor dispersión de los datos.

Continuando con el análisis exploratorio de los datos también se realizó el cálculo de la correlación numérica de los datos, información de vital importancia para conocer qué variables son más parecidas entre sí. Esta relación se refleja en la Figura 5.1.

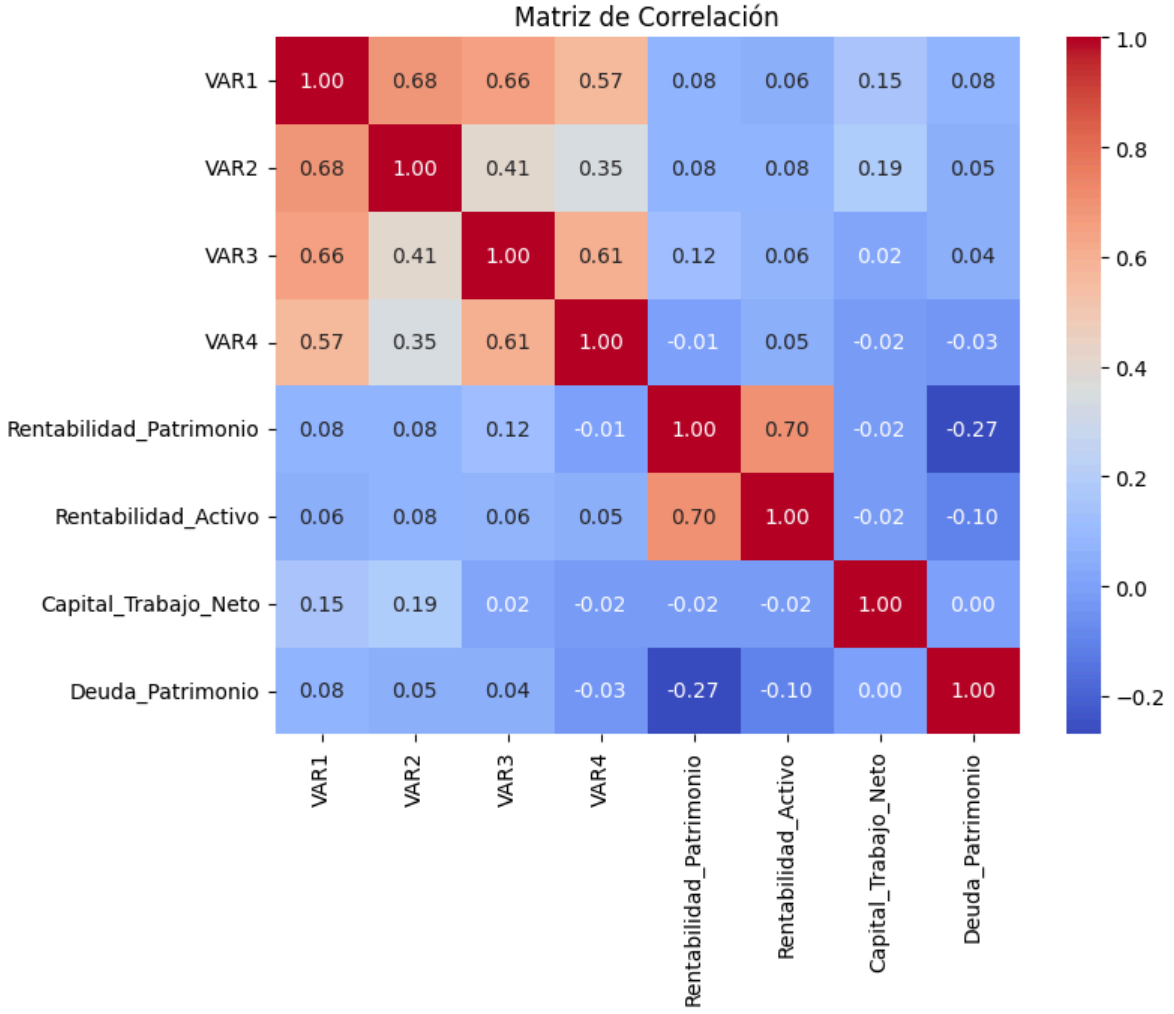

Figura 5.1. Matriz de correlación

La matriz presentada refleja una alta correlación en colores rojos, mientras que la baja correlación podemos verla en color azul. Dado esto, observamos que los atributos de las distintas políticas de gestión (VAR1, VAR2, VAR3 y VAR4) están altamente correlacionadas entre sí, mientras las variables financieras cuentan con una baja correlación entre ellas, exceptuando la correlación entre la Rentabilidad Patrimonial (ROE) y la Rentabilidad de Activos (ROA) con un valor de *0.7*, pudiendo inferir que ambas variables se comportan de manera parecida.

También es importante poder analizar de manera gráfica cómo se comportan los valores financieros dentro de nuestro conjunto de datos. A continuación, se presenta el atributo de Rentabilidad sobre el patrimonio (ROE) en la Figura 5.2, y Capital (CT) en la Figura 5.3.

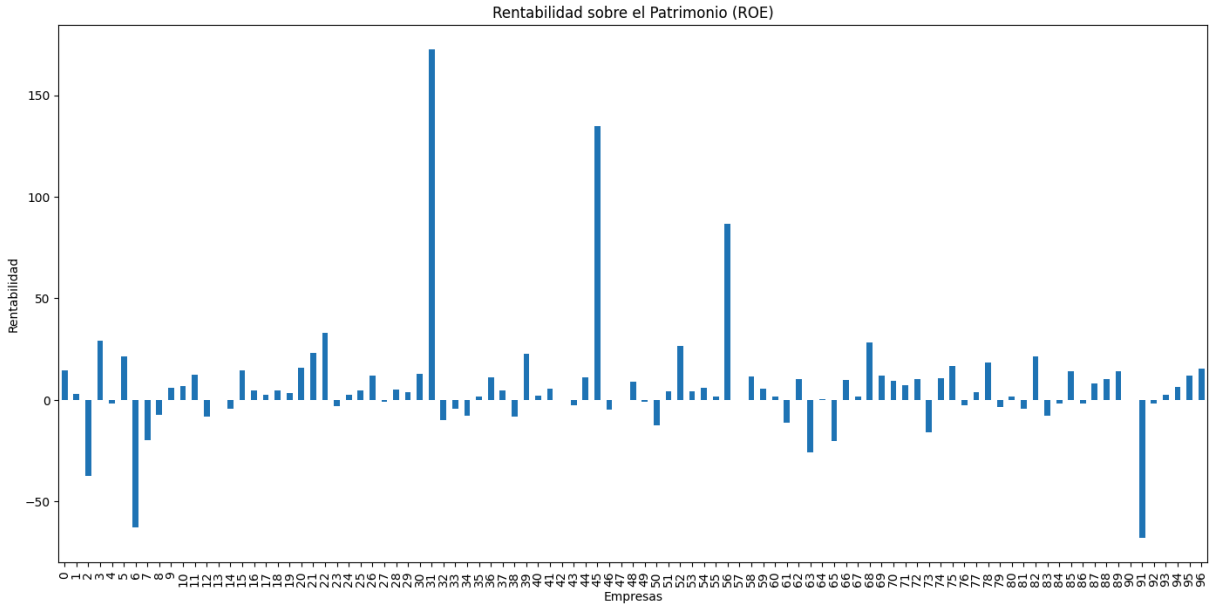

Figura 5.2. Rentabilidad sobre el patrimonio de las empresas

En el análisis de la gráfica de rentabilidad sobre el patrimonio, podemos inferir que las empresas analizadas tienen una rentabilidad muy similar entre sí, destacando sólo 3 empresas con resultados sobresalientes positivos y 3 empresas con resultados insuficientes o por debajo del promedio. Esto implica que, en general, las empresas tienen una rentabilidad bastante homogénea entre 50 y -50, con unos pocos casos que se desvían significativamente hacia los extremos positivos o negativos.

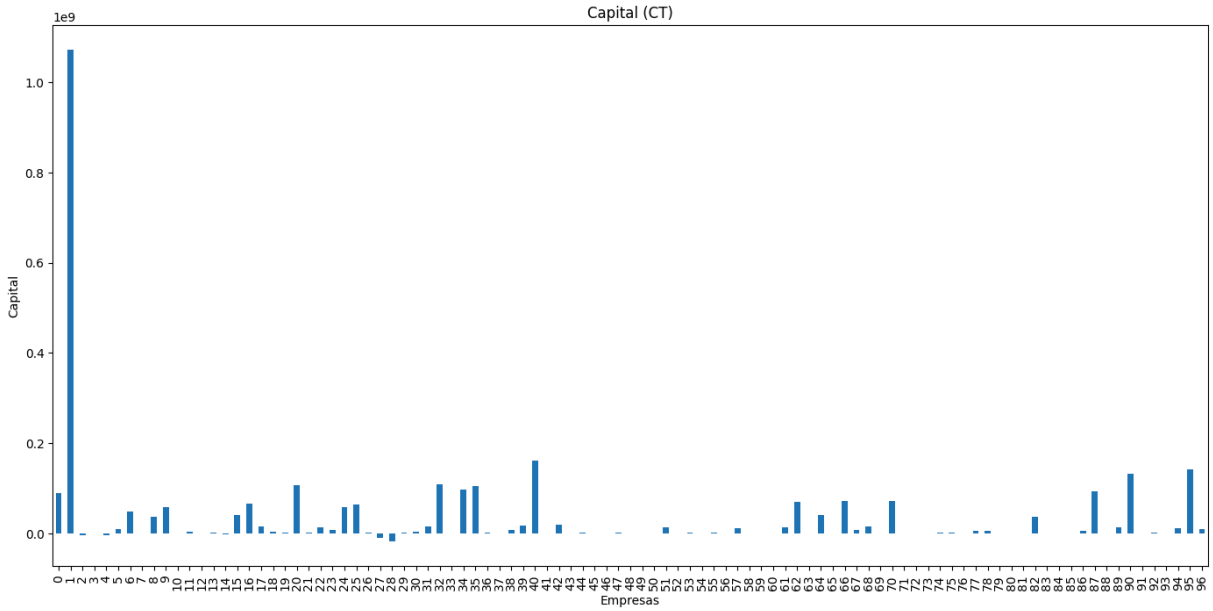

Figura 5.3. Capital de las empresas

En la gráfica de la Figura 5.3, relacionada con el capital de las empresas, se observa una clara tendencia de valores positivos que oscilan entre 0 y 0.2. A partir de esto, podemos inferir que la mayoría de las empresas cuentan con un capital positivo. Además, se destaca una sola empresa, cuyo valor de capital supera 1, diferenciándose significativamente del resto.

Tras haber analizado el comportamiento de los datos financieros recopilados, es fundamental analizar cómo se desempeñan las empresas en relación con el cumplimiento de las políticas de gestión definidas en la Tabla 5.2 como VAR1 y VAR2. A continuación, se procederá a analizarlas, ya que estos son los atributos más representativos y utilizados a lo largo del Proyecto. Estos se presentan en las Figuras 5.4 y 5.5. En este análisis se realizó una sumatoria de todas las empresas que tengan un porcentaje mayor o igual a 0.5, considerando este valor de piso como de cumplimiento aceptable y aquellas con un valor menor a 0.5, como de bajo cumplimiento. Estas cantidades se representan con los colores verde y rojo, respectivamente.

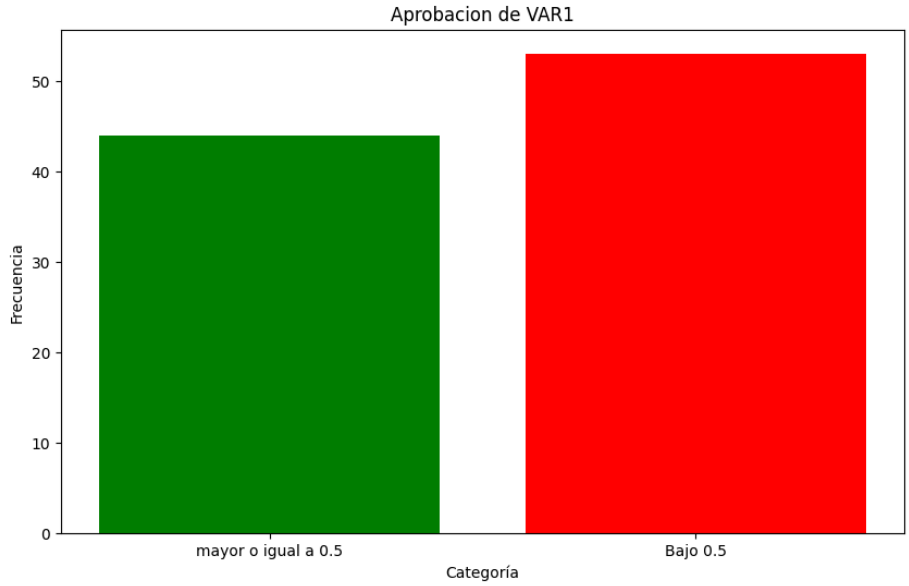

Figura 5.4. Nivel de aprobación VAR1

El atributo VAR1 relacionado con la aprobación de las políticas de gestión (Procedimiento de gestión y control de riesgos), se observa un nivel predominantemente negativo. Sin embargo, esta tendencia negativa está influenciada por una diferencia de 9 empresas que inclinan la balanza hacia un resultado negativo.

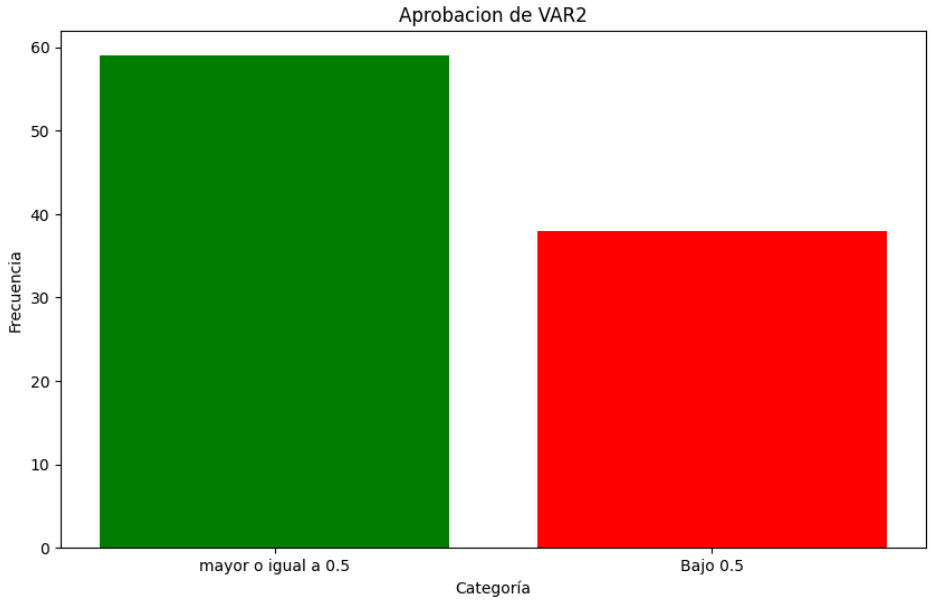

Figura 5.5. Nivel de aprobación VAR2

El atributo VAR2 referente a la aprobación de las políticas de gestión (Canal de denuncias), se aprecia una tendencia claramente positiva en los resultados. En este caso, la diferencia a favor de la aprobación positiva es considerable, con un total de 21 empresas que se inclinan hacia este tipo de resultado. Esta diferencia es significativa, dado que representa una parte importante del total de empresas analizadas, lo que sugiere que la mayoría de las empresas han cumplido de manera favorable con las políticas de gestión establecidas.

5.2.4 Aprendizaje No Supervisado

Luego de haber analizado las gráficas correspondientes a nuestro análisis exploratorio de datos (EDA), el siguiente paso consiste en adentrarnos en el proceso de clustering, una técnica de Aprendizaje no Supervisado. El clustering nos permite agrupar los datos en grupos o clusters, de tal manera que las observaciones dentro de un mismo grupo sean más similares entre sí que con las de otros grupos.

Para determinar el número óptimo de clusters en un conjunto de datos, uno de los métodos más utilizados es el método del codo (Elbow). Este método se basa en el análisis de la suma de errores cuadráticos (SSE, por sus siglas en inglés), que mide la dispersión interna dentro de cada clúster. Al realizar el clustering para diferentes números de clusters, se obtiene un gráfico que muestra cómo varía la SSE a medida que aumentamos el número de clusters representado en la Figura 5.6. El punto en el que la disminución de la SSE empieza a estabilizarse, formando una curva en forma de "codo", es considerado el número óptimo de clusters. Este es el punto en el que añadir más clusters no mejora significativamente la calidad del modelo, por lo que se interpreta como el número más adecuado para dividir los datos.

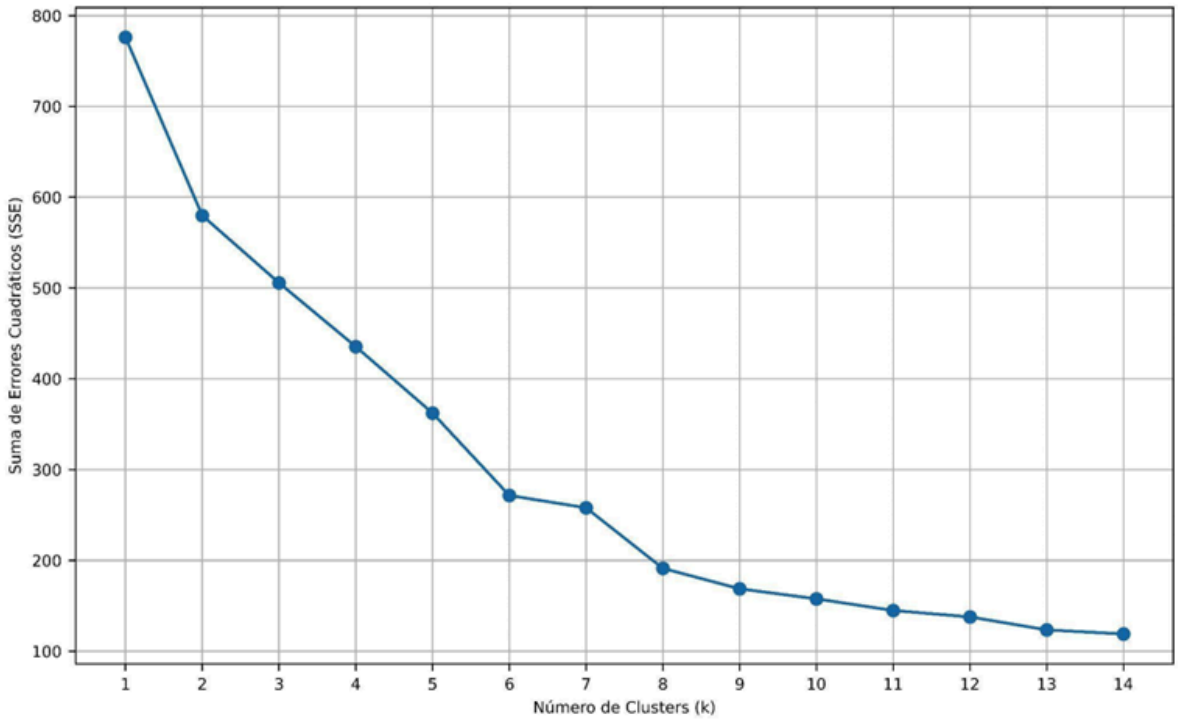

Figura 5.6. Método del codo

Dada la gráfica 5.6 podemos inferir que el número recomendado se sitúa entre 2 y 6. Considerando el tamaño de la base de datos, se optó por 4 grupos para analizar la calidad de agrupamiento de los algoritmos. En la Tabla 5.4 se presentan las métricas respecto a la calidad de los agrupamientos alcanzados por K-Means y BIRCH. En ambos se usaron todos los atributos numéricos disponibles los cuales fueron estandarizados.

Tabla 5.4. Calidad de agrupamiento para K=4

Métrica	**Métodos de agrupamiento**	
	K-Means	**BIRCH**
Silhouette	0,29	0,63
Calinski-Harabasz	24,26	12,84

Se observa que, según el índice de Silhouette, BIRCH alcanza un mejor resultado, mientras que para Calinski-Harabasz, K-Means obtiene mejores agrupaciones. Ante esta paridad, se decidió optar por los resultados de K-Means tanto para la interpretación de los grupos como para los modelos predictivos. En este contexto, se evaluará la clusterización K=4 grupos utilizando K-Means representados los grupos a continuación en la Tabla 5.5.

Tabla 5.5. Distribución de clusterización con K = 4

Grupo	Cantidad de empresas incluidas
0	36
1	38
2	22
3	1

Se presentan los informes de estadística descriptiva para los 4 grupos formados por el algoritmo K-Means en las Tablas 5.6 , 5.7 , 5.8 y 5.9 y a posterior, la interpretación de los mismos.

Tabla 5.6. Estadística descriptiva del grupo 0 (36 empresas)

Atributo	Promedio	D. Estándar	Mínimo	Máximo
VAR1	0.037037	0.102869	0.000000	0.555556
VAR2	0.055556	0.190029	0.000000	0.750000
VAR3	0.011111	0.046462	0.000000	0.200000
VAR4	0.027778	0.099602	0.000000	0.500000
ROE	6.645814	30.174533	-62.685400	134.711500
ROA	0.462658	24.598621	-129.283400	44.734800
CT	1.010364e+07	2.253181e+07	-3.910031e+06	7.142201e+07
RE	0.625158	1.075725	0.000000	5.013400

Tabla 5.7. Estadística descriptiva del grupo 1 (38 empresas)

Atributo	Promedio	D. Estándar	Mínimo	Máximo
VAR1	0.549708	0.314214	0.000000	1.000000
VAR2	0.855263	0.198232	0. 250000	1.000000
VAR3	0.178947	0.172670	0.000000	0.600000
VAR4	0.171053	0.165477	0.000000	0.500000
ROE	3.082450	10.856651	-20.442700	28.208700
ROA	1.374463	5.678745	-11.506400	18.320400
CT	2.437187e+07	4.435286e+07	-1.807588e+07	1.611571e+08
RE	1.031468	1.098344	0.010200	4.328000

Tabla 5.8. Estadística descriptiva del grupo 2 (22 empresas)

Atributo	Promedio	D. Estándar	Mínimo	Máximo
VAR1	0.888889	0.160833	0.444444	1.000000
VAR2	0.840909	0.349706	0.000000	1.000000
VAR3	0.809091	0.305363	0.000000	1.000000
VAR4	0.477273	0.254824	0.000000	1.000000
ROE	15.805045	37.372342	-25.752900	172.631900
ROA	4.246545	10.557543	-22.815400	38.953400
CT	7.577008e+07	2.258958e+08	-5.764780e+05	1.072535e+09
RE	1.270264	0.922366	0.203200	4.758700

Tabla 5.9. Estadística descriptiva del grupo 3 (1 empresa)

Atributo	Promedio	D. Estándar	Mínimo	Máximo
VAR1	0	NaN	0	0
VAR2	0	NaN	0	0
VAR3	0	NaN	0	0

VAR4	0	NaN	0	0
ROE	-68.0151	NaN	-68.0151	-68.0151
ROA	-13.7559	NaN	-13.7559	-13.7559
CT	-38514.0	NaN	-38514.0	-38514.0
RE	14.4586	NaN	14.4586	14.4586

Del análisis numérico de cada cluster, se observa que:

- El Cluster 0 agrupa a empresas con cumplimiento muy bajo de las políticas, rentabilidad moderada y endeudamiento bajo,
- El Cluster 1 a empresas con cumplimiento bajo de las políticas, rentabilidad baja y endeudamiento moderado,
- El Cluster 2 a empresas con alto cumplimiento, rentabilidad alta y endeudamiento leve,
- El Cluster 3 a empresas con nulo cumplimiento, rentabilidad negativa y endeudamiento alto.

5.2.5 Aprendizaje Supervisado

En este apartado se presentan los resultados obtenidos mediante dos modelos de Aprendizaje Supervisado: Random Forest (RF) y Extreme Learning Machine (ELM) en el contexto del uso de 4 clusters, utilizados para clasificar el comportamiento de las empresas en relación con el cumplimiento de las políticas.

Para abordar el desequilibrio de clusters (en particular, sobre el último grupo), se optó por la librería **SDV** (Patki et al., 2016) para la generación de datos sintéticos. Esto permite ampliar el conjunto de datos disponibles, incorporando características específicas de este grupo.

Para garantizar clasificaciones precisas con el menor número posible de atributos, se emplearon dos técnicas de reducción de dimensión: Recursive Feature Elimination (RFE) y Feature Importance, este último podemos ver los resultados en la Tabla 5.10.

Tabla 5.10. Feature Importance: importancia de los atributos

Atributo	Puntuación
VAR1	0,21
VAR2	0,21
VAR3	0,15
ROE	0,15
CT	0,12
ROA	0,06
VAR4	0,04
RE	0,04
Rubro	0,02

Estas técnicas permitieron seleccionar los siguientes atributos como entrada para los modelos:

- VAR1 (numérica)
- VAR2 (numérica)
- INDUSTRIA (categórica)
- Capital_Trabajo_Neto (numérico)
- Deuda_Patrimonio (numérico)

En base a experimentos preliminares, los hiper-parámetros y valores considerados para el modelo basado en Random Forests son:

- número de estimadores (árboles): [250, 500, 700]
- máximo de caracteristicas: [sqrt, log2]
- máximo de profundidad: [3, 4, 5, 7, 9, 11]
- criterio: [gini, entropy]

A su vez, los usados para el modelo basado en Extreme Learning Machine son:

- número de neuronas: [1000, 2000, 3000, 4000, 5000]

- funcion de activacion: [tanh, sigm, relu, lin]
- C (sesgo): [0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 100, 100, 1000, 10000]

Ambos modelos fueron entrenados utilizando el 70% de los datos disponibles, mientras que el 30% restante se destinó a la prueba de los mismos. Para el modelo basado en Random Forests, se intensificó el entrenamiento mediante validación cruzada con 10 folds y 3 repeticiones. Por otro lado, en el modelo basado en Extreme Learning Machine, esta opción fue descartada.

A continuación, se presentan los resultados de performance de los modelos basados en Random Forests (RF) y Extreme Learning Machine (ELM). Los resultados destacan la capacidad de ambos para predecir el nivel de cumplimiento de las políticas por parte de las empresas analizadas, como se detalla en el reporte de la Tabla 5.11.

Tabla 5.11. Reporte de clasificación para los modelos predictivos

Modelo	**Métrica**	**Modelo**	**Grupo**			
			0	**1**	**2**	**3**
RF	**Accuracy (%)**	89				
	Precision (%)		90	86	100	80
	Recall (%)		82	100	88	80
	F1-Score (%)		86	92	93	80
ELM	**Accuracy (%)**	94				
	Precision (%)		100	100	100	71

Modelo	Métrica	Modelo	Grupo			
			0	1	2	3
	Recall (%)		82	100	100	100
	F1-Score (%)		90	100	100	83

Los mejores hiper-parámetros asociados al modelo RF son:

- número de árboles: 250
- máxima profundidad del árbol: 4

El tiempo total de CPU para entrenamiento y evaluación del modelo fue de **1365** segundos.

Respecto a ELM, los hiper-parámetros del modelo son:

- número de neuronas: 1000
- función de activación: relu
- C (sesgo) : 1

El tiempo de CPU fue de **834** segundos.

El modelo basado en Random Forests ha obtenido un desempeño destacado, alcanzando una precisión global (Accuracy) del 89%. Además, al analizar su rendimiento por clases, se observa que logra un F1-Score que varía entre el 80% para la clase 3, la cual representa el menor rendimiento, y un 93% para la clase 2, que muestra el mejor desempeño en este modelo. Esto indica que, aunque el modelo tiene un buen desempeño general, existen diferencias notables en la capacidad de clasificar correctamente algunas clases específicas.

Por otro lado, el modelo basado en Extreme Learning Machine (ELM) supera a Random Forests en términos de precisión global, alcanzando un Accuracy del 94%. En cuanto al rendimiento por clases, este modelo presenta valores de F1-Score más consistentes y elevados.

Para la clase 3, que es la de menor rendimiento, el F1-Score es del 83%, mientras que para las clases 1 y 2 alcanza un excelente 100%, reflejando una clasificación perfecta en estas categorías. Esta mayor consistencia y precisión sugieren que el modelo ELM podría ser más adecuado para este conjunto de datos, especialmente si se prioriza la clasificación precisa de ciertas clases.

En resumen, mientras que ambos modelos ofrecen buenos resultados, el modelo ELM muestra un desempeño superior tanto en precisión global como en la clasificación de las diferentes clases, especialmente en las clases 1 y 2, donde logra un rendimiento óptimo.

Estos resultados sobresalientes son producto de un enfoque integral que incluyó una adecuada transformación y selección de atributos, una cuidadosa configuración de los hiper-parámetros, y un entrenamiento eficiente de los modelos predictivos. Además, cabe destacar que ambos modelos presentan tiempos de procesamiento en CPU relativamente bajos, optimizando su viabilidad para aplicaciones prácticas.

5.3 Tercera iteración

5.3.1 Introduccion

El desarrollo de una herramienta específica para abordar la problemática planteada es crucial, ya que permite realizar ajustes interactivos, seleccionar grupos de datos y cambiar dinámicamente el origen de la información utilizada. Esta funcionalidad sería especialmente útil para usuarios sin conocimientos de programación, facilitándoles la interacción con modelos de Inteligencia Artificial mediante datos en formato CSV, ampliamente utilizados en software como Microsoft Excel. Esta herramienta, además de ser sencilla y accesible, aprovecharía la familiaridad que empresas y usuarios privados ya tienen con Excel, optimizando así el análisis y procesamiento de datos.

5.3.2 Diagrama de clases del Modelo de Datos

A continuación, se presenta la estructura fundamental del diagrama de clases que modela las relaciones entre las entidades principales de esta herramienta. Esta estructura se utiliza para gestionar los datos provistos por el usuario, además de implementar el uso de llaves primarias y foráneas para una óptima relación entre las columnas clasificadas y el modelo cargado como muestra la relación en la Figura 5.7.

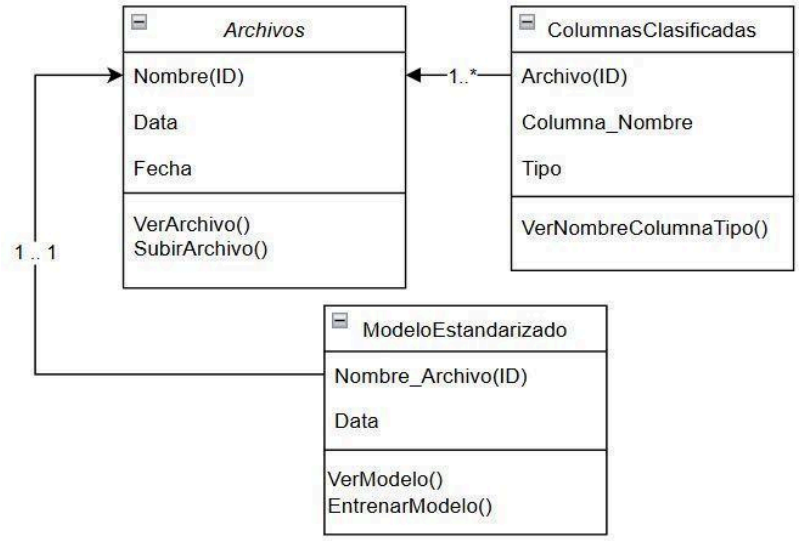

Figura 5.7. Diagrama de clases de la aplicación

5.3.3 Funcionalidades del software

Para facilitar la comprensión de las funcionalidades activas y pasivas de la aplicación web desarrollada, se presenta en la Tabla 5.12 los requisitos clave implementados y una nota del funcionamiento. Esto permitirá entender de manera más clara cómo están distribuidas y organizadas todas las funcionalidades.

Tabla 5.11. Funcionalidades del software

ID Requerimiento	Descripción	Nota
RF1: Gestión de Datos	Carga y almacenamiento de archivos en formato .csv a	Realiza una validación específica del formato y guarda la

	través de la aplicación web	información con el nombre introducido
RF2: Clasificación	Clasificación de columnas según el tipo de dato del archivo cargado	Permite seleccionar entre entero , real o categórico y guarda esta clasificación para la estandarización posterior y además permite cambiar el nombre a las columnas
RF3: EDA	Visualización de la estadística descriptiva para cada columna	Permite una navegación interactiva entre diferentes columnas y su estadística descriptiva
RF4: EDA	Visualización de gráficas interactivas para cada columna	Ofrece una navegación interactiva entre columnas y gráficas, con opciones de visualización en gráficos de barras y de torta
RF5: Preparación	Preparación de los datos por medio del uso de la clasificación anterior	Aplica técnicas específicas según el tipo de dato guardado. Valida clasificaciones incorrectas entre numéricos y categóricos, redirigiendo a la sección de clasificación para correcciones.
RF6: Método del Codo	Se entrenan los datos estandarizados con el método del Codo, generando una gráfica que muestra la relación entre el número de clústers (k) y la suma de los errores al cuadrado (SEE).	La gráfica permite poder seleccionar un punto para consultar el resultado.
RF7: Entrenamiento K-Means	Entrenamiento del modelo K-Means, permitiendo seleccionar manualmente la cantidad de grupos (entre 2 y 5) para realizar la segmentación de los datos.	Opción para ajustar *k* según la necesidad del usuario y entrenar el modelo.

RF8: Visualización de resultado 1	Permite visualizar en una tabla, la cantidad de clusters obtenidos y la cantidad de casos agrupados	Opción de volver atrás y agrupar por una nueva cantidad
RF9: Visualización de Resultados 2	Permite visualizar las características de cada cluster formado por el algoritmo K-Means	Opción de ver las diferentes estadísticas descriptivas de cada cluster y representación gráfica.
RF10: Entrenamiento Extreme Learning Machine	Permite poder entrenar el modelo basado en ELM	Opción de elegir los hiper-parámetros asociados (número de neuronas, función de activación)
RF11: Funcionalidad General	La aplicación web permite gestionar los conjuntos de datos	Se permite visualizar, usar y eliminarlos

5.3.4 Desarrollo y prácticas particulares en Django

El entorno de Django permite gestionar todas las entidades de la aplicación utilizadas para la gestión de los datos a partir de clases u objetos. A continuación, en la Figura 5.8 se presenta un ejemplo de código.

```
1  class ColumnClassification(models.Model):
2      archivo = models.ForeignKey(DataRow, on_delete=models.CASCADE, related_name='columnas')
3      columna_nombre = models.CharField(max_length=255)
4      tipo_dato = models.CharField(max_length=50, choices=[('real', 'Real'), ('categorico', 'Categórico'),('entero','Entero')])
5
6      def __str__(self):
7          return f"{self.columna_nombre} ({self.tipo_dato}) - {self.archivo.archivo_nombre}"
8
```

Figura 5.8. Creación de Modelos en Django

El uso de modelos en Django nos permite definir características particulares de las columnas de una base de datos de manera estructurada y declarativa. Por ejemplo, en la línea 3, se utiliza

max_length=255 para restringir el tamaño máximo de caracteres a 255 en un campo de texto. Además, en la línea 2 se define una clave foránea que establece una relación con el modelo DataRow, especificando también el comportamiento de eliminación en cascada a través del parámetro *on_delete=models.CASCADE*, lo que garantiza que los datos relacionados se eliminen automáticamente al borrar el registro principal.

Este enfoque, basado en el paradigma de programación orientada a objetos, no sólo simplifica la creación y gestión de tablas en la base de datos, sino que también facilita el acceso a los datos desde nuestra aplicación. Con métodos como *.objects.filter(id=id).all()*, podemos obtener todas las filas que coincidan con un identificador específico, mientras que el método *.first()* nos permite acceder al primer registro de un conjunto filtrado de manera sencilla como muestra la Figura 5.9.

```
1  def clusterizar(request, nombre_archivo):
2
3      data_row = DataRow.objects.filter(archivo_nombre=nombre_archivo).first()
4      ......
5
6
7      .....
8      return render(request, 'cargaXLS/clusterizacion_resultados.html', {
9          'nombre_archivo': nombre_archivo,
10         'cluster_counts': cluster_counts,
11         'data': data,
12     })
13
```

Figura 5.9. Acceso a datos a partir de modelos en Django

La Figura 5.10 es un ejemplo de cómo se pueden usar datos dinámicos en el frontend, a través del uso de templates. En este caso, ambos datos se envían desde la vista indicada en la Figura 5.9.

```
<div class="container mt-4">
  <h2>Describe de la Base: {{ nombre_archivo }}</h2>
  <p>Total de filas actuales: {{ total_filas }}</p>
  ...
```

Figura 5.10. Uso de variables dinámicas desde el template

Django gestiona el Frontend a través de templates. Como se puede observar en la figura anterior, el manejo de templates es simple e intuitivo.

5.3.5 Capturas de pantalla de la aplicación web

Como última sección, se presentarán capturas de pantalla según funcionalidades especificadas en la Tabla 5.11. En las Figuras 5.11 a 5.18 se presentan pantallas de la aplicación web desarrollada.

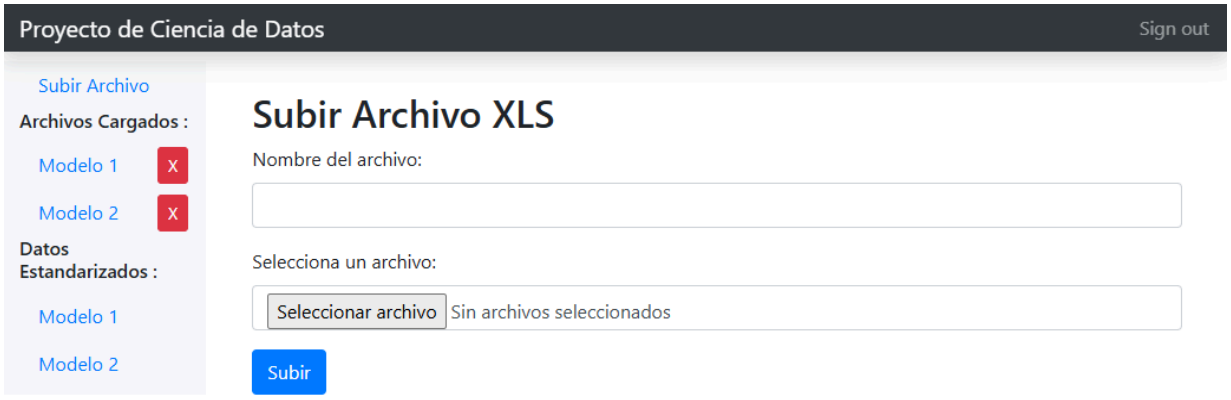

Figura 5.11. Carga de datos y visualización de cargados previos

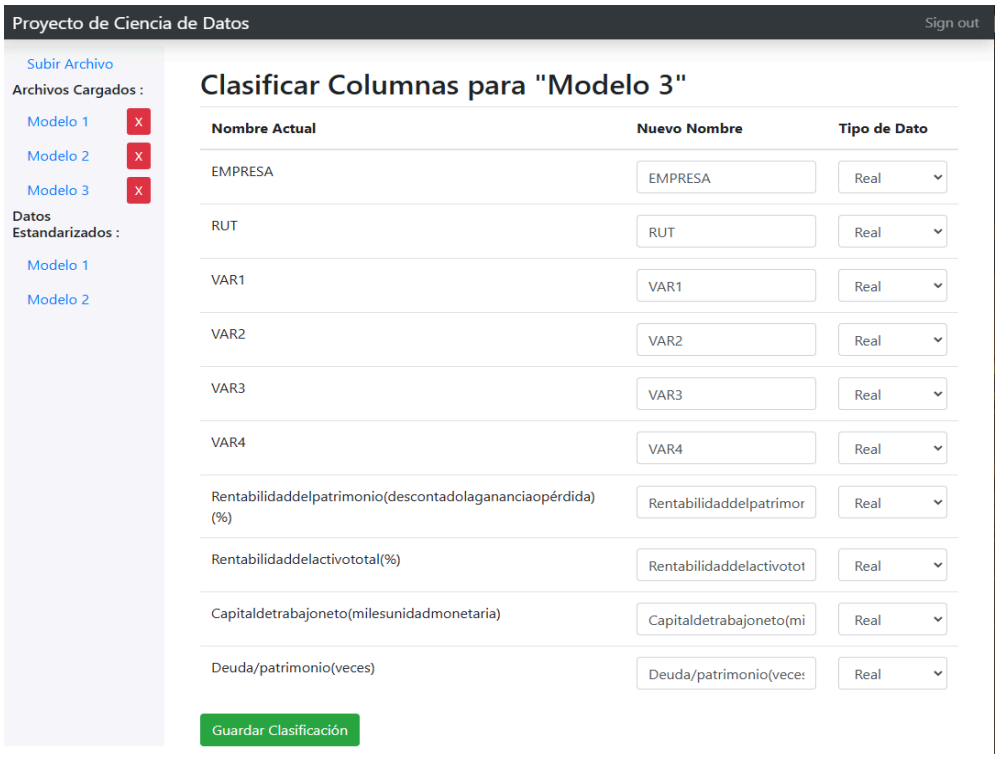

Figura 5.12. Clasificación de columnas y cambio de nombres

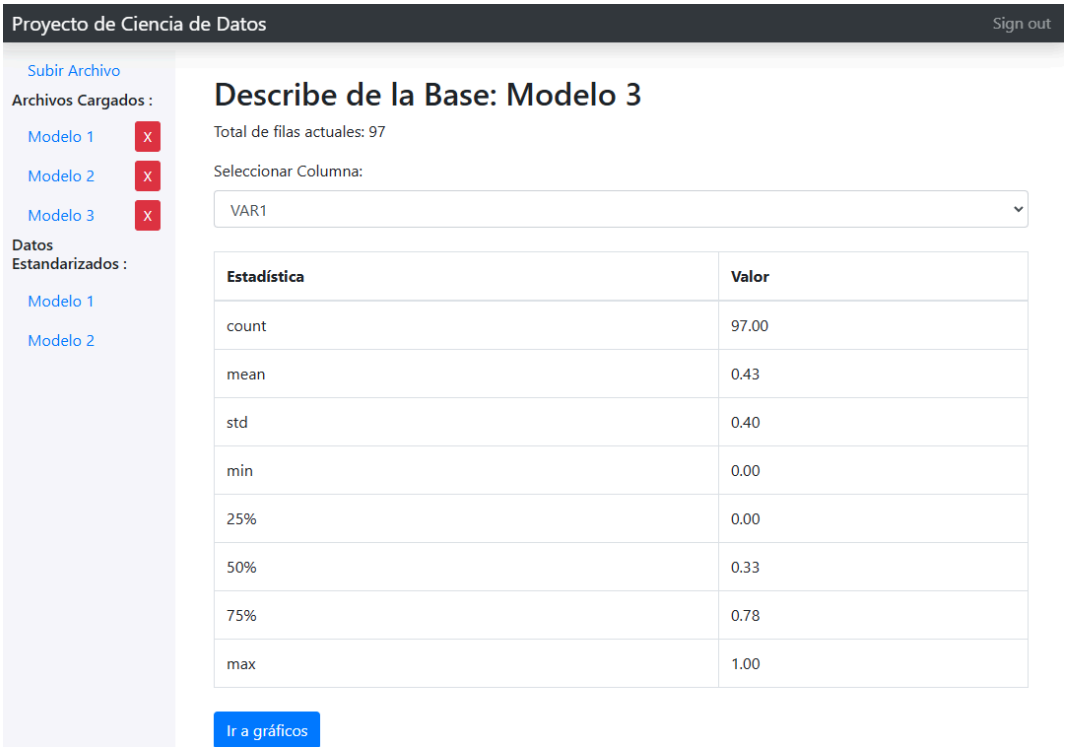

Figura 5.13. Estadística descriptiva de cada columna

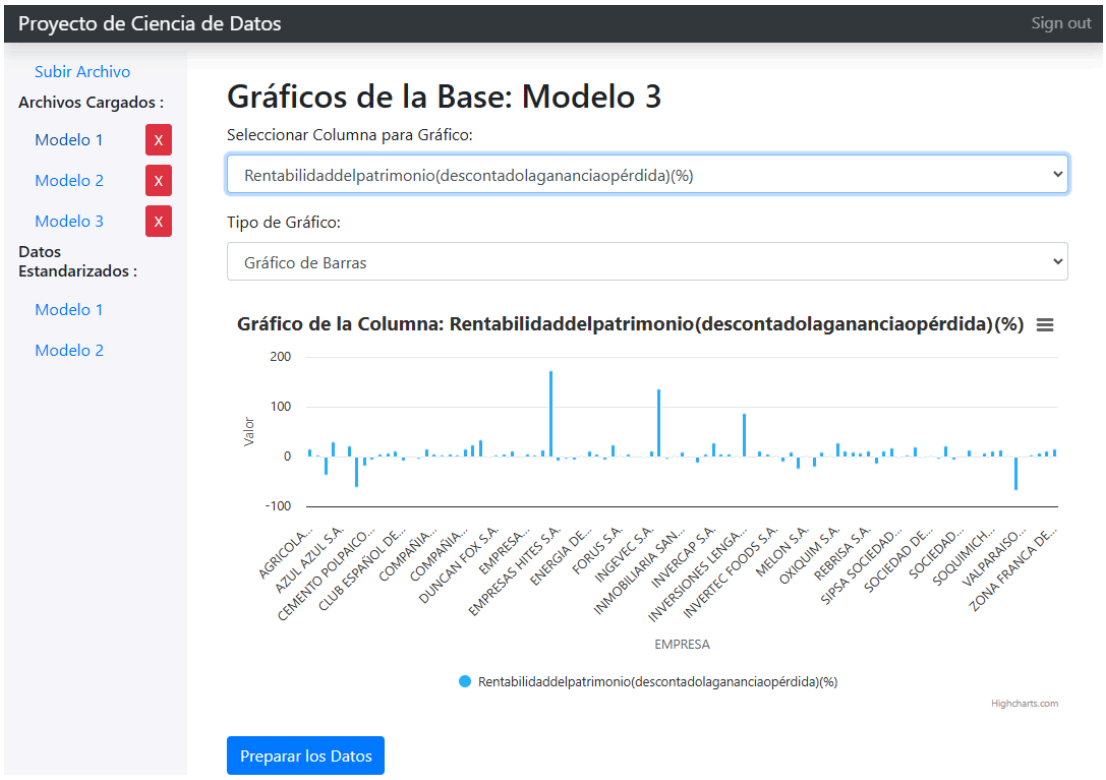

Figura 5.14. Visualización gráfica de los datos en formato barra

Figura 5.15. Visualización gráfica de los datos en formato torta

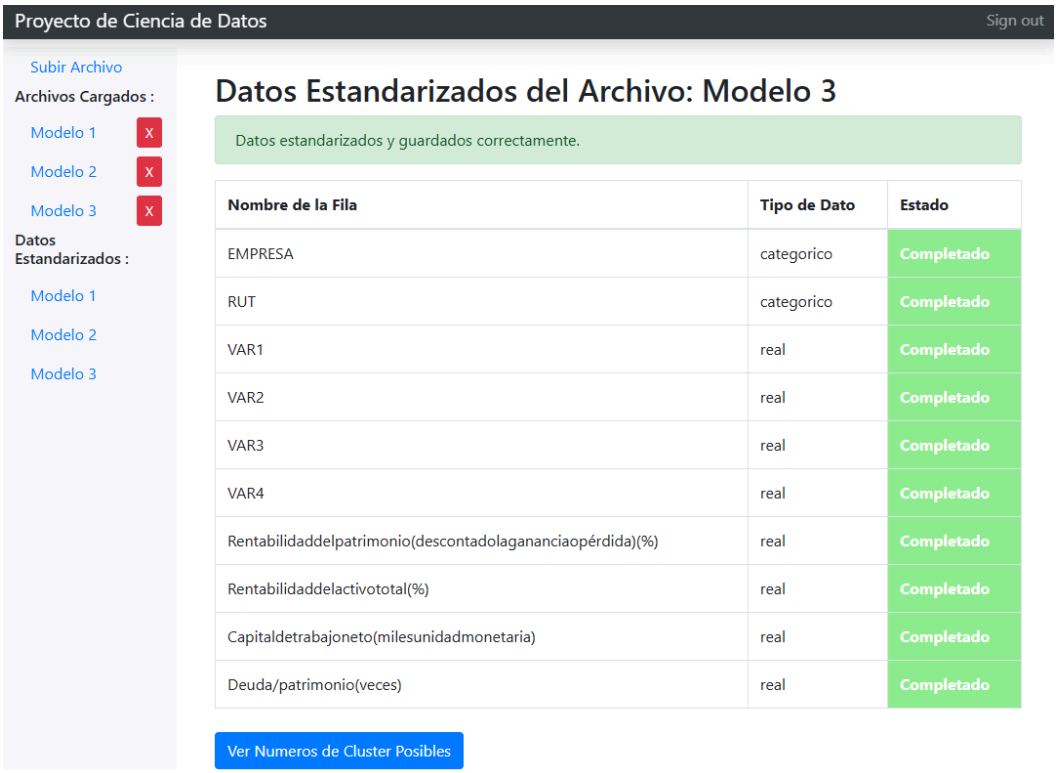

Figura 5.16. Transformación de los datos

Figura 5.17. Resultado gráfico del método del codo

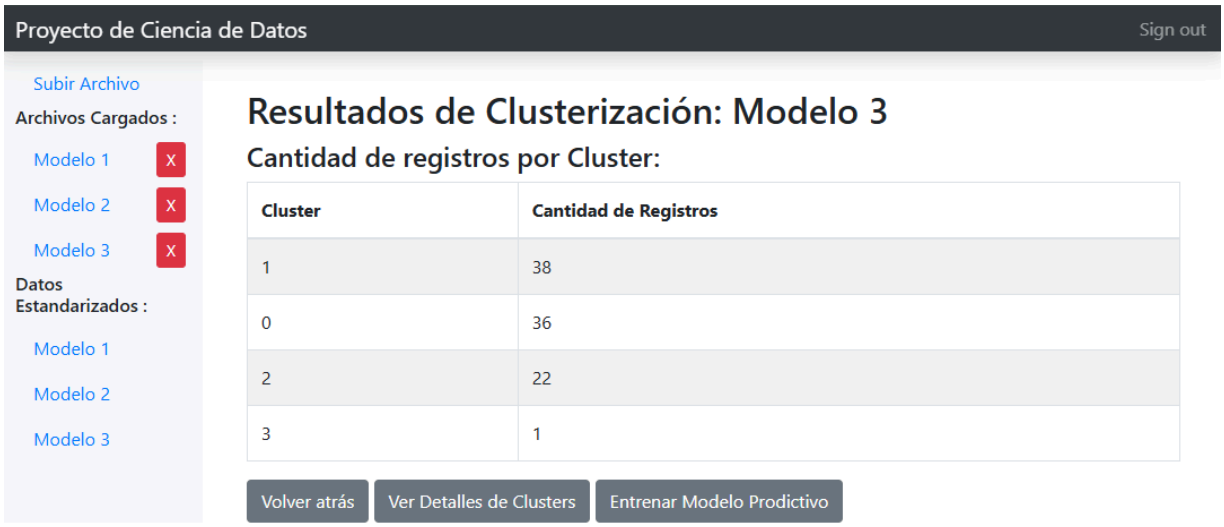

Figura 5.18. Resultado de clusterización de K-Means

Capítulo VI.
Conclusiones y Trabajos Futuros.

6. Conclusiones y Trabajos Futuros.

El haber culminado el desarrollo de este proyecto, que incluye tanto la aplicación de diferentes técnicas de Ciencia de los datos, como la implementación de una aplicación web funcional, no significa que deba darse por finalizado. Si bien se cumplieron los objetivos planteados, las posibilidades de mejora y expansión del trabajo son amplias. En este apartado se presentan las conclusiones generales obtenidas a partir de los resultados logrados, resaltando los principales aprendizajes, aportes y hallazgos que este proyecto ha generado. Además, se proponen líneas de trabajo futuro que permitirán adaptar y enriquecer la aplicación desarrollada.

6.1. Conclusiones

Luego del desarrollo de este Proyecto se amplió la experiencia respecto a Ciencia de Datos y desarrollo web. A continuación, algunas conclusiones:

- **Importancia de los datos**: A lo largo del Proyecto los datos siempre fueron una prioridad, ya que es sumamente importante conocerlos, que sean confiables y de calidad para su uso en diferentes modelos y técnicas.
- **Selección de atributos**: Se utilizó la técnica de selección de características (RFE) para reducir la dimensionalidad y seleccionar los atributos más relevantes. Esta estrategia ayudó a mejorar la precisión de los modelos y a reducir el tiempo de cómputo, lo que refuerza la idea de que una adecuada selección de características es crucial para un buen desempeño en tareas de Aprendizaje Automático.
- **Rendimiento de modelos desarrollados**: Respecto al manejo de técnicas supervisadas en este Proyecto, se puede concluir que Extreme Learning Machine se destaca respecto a Random Forests en términos de precisión y tiempo de CPU consumido.
- **Aplicación web**: La aplicación desarrollada con Django permitió crear un backend eficiente para cargar y procesar archivos XLS. Se integraron herramientas como Bootstrap para la interfaz y Chart.js para visualizar los resultados, brindando una experiencia interactiva y accesible para los usuarios.

En resumen, el desarrollo de este Proyecto implicó la integración de diversas técnicas y herramientas orientadas al análisis de datos y la implementación de una aplicación web. A

través de una metodología apropiada e iterativa, se lograron los objetivos y se verificó la hipótesis planteada, permitiendo no sólo el análisis y entrenamiento de modelos, sino también el desarrollo de una interfaz que facilita la interacción con los usuarios. Este Proyecto ha permitido adquirir un conocimiento profundo sobre la manipulación de datos, la optimización de modelos y la creación de soluciones prácticas y funcionales en el ámbito de la Ciencia de Datos.

6.2. Trabajos Futuros

Durante el desarrollo, surgieron nuevas ideas y funcionalidades que podrían enriquecer la aplicación y que no fueron abordadas por cuestiones de tiempo. A continuación, se proponen algunas mejoras basadas en la retroalimentación obtenida y el análisis del rendimiento actual. Estas propuestas permitirán mejorar la experiencia de usuario, optimizar el rendimiento y ampliar las capacidades de la aplicación.

- **Sistema de usuario**: Se propone implementar un módulo de gestión de usuarios, donde cada uno pueda tener su propio espacio para almacenar y acceder a los modelos y datos guardados. Esto ofrecerá mayor flexibilidad y usabilidad.
- **Soporte para archivos CSV con múltiples hojas**: Actualmente, la aplicación maneja archivos XLS, pero se sugiere agregar soporte para trabajar con archivos CSV que contengan múltiples hojas, lo que ampliará la versatilidad de la herramienta.
- **Nuevas técnicas de clustering**: Para mejorar la segmentación de datos, se plantea la incorporación de otras técnicas de clustering como DBSCAN y K-Modes que podrán proporcionar mejores resultados en determinados conjuntos de datos.
- **Nuevos técnicas de aprendizaje supervisado**: Se recomienda explorar la integración de otras técnicas como Support Vector Machines (SVM) para aumentar la precisión en las predicciones y ofrecer más opciones a los usuarios.
- **Mejoras en la visualización de resultados**: Se sugiere agregar más tipos de gráficos interactivos y personalizables para ofrecer una experiencia más rica y permitir una mejor interpretación de los resultados.

Bibliografía.

Referencias Bibliográficas

1. Ahmad, A., Narullia, D. y Muhammad (2021). Corporate risk disclosure: The effect of corporate governance. *Journal of Applied Managerial Accounting , Vol. 5 (101-113)*

2. Ayuningtyas, E. y Harymawan, I. (2022). Risk management committee and textual risk disclosure. *Risks*, Vol. 10 (2-30). https://doi.org/10.3390/risks10020030

3. Belete, D. y Huchaiah, M. (2021). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications. Vol. 44.* https://doi.org/10.1080/1206212X.2021.1974663.

4. Bui, H. y Krajcsák, Z. (2024). The impacts of corporate governance on firms performance: From theories and approaches to empirical findings. *Journal of Financial Regulation and Compliance*. Vol. 32 (18-46). https://doi.org/10.1108/JFRC-01-2023-0012

5. Cravero, A. y Sepúlveda, S. (2021). Use and adaptations of machine learning in big data applications in real cases in agriculture. *Electronics* Vol. 10, (5-552). https://doi.org/10.3390/electronics10050552

6. D'Urso, P., De Giovanni, L. y Swartz, T. (2023). Editorial: Big data and data science in sport. *Annals of Operational Research*. Vol. 325 (1-7)

7. Horvey, S. y Ankamah, J. (2020). Enterprise risk management and firm performance: Empirical evidence from Ghana equity market. *Cogent Economics & Finance, Vol. 8*. https://doi.org/10.1080/23322039.2020.1840102

8. Ibarra, C. (2020). *Técnicas de data mining aplicadas a la deserción de los estudiantes de la Facultad de Ciencias Exactas* [Tesis de maestría, Universidad Santo Tomás de Aquino].

9. Jahani, H., Jain, R. y Ivanov, D. (2023). Data science and big data analytics: A systematic review of methodologies used in the supply chain and logistics research. *Annals of Operations Research*, (1-58). https://doi.org/10.1007/s10479-023-05390-7

10. Musallam, S. (2020). Effects of board characteristics, audit committee and risk management on corporate performance: Evidence from Palestinian listed companies. *International Journal of Islamic and Middle Eastern Finance and Management*, Vol. 13 (691-162). https://doi.org/10.1108/IMEFM-12-2017-0347

11. Nahar, S., Azim, M. y Hossain, M. (2020). Risk disclosure and risk governance characteristics: Evidence from a developing economy. *International Journal of Accounting & Information Management. Vol. 28 (577-605).* https://doi.org/10.1108/IJAIM-07-2019-0083

12. Navarrete, J., Moraga, H. y Gallegos, J. (2023). Index to degree of adhesion to good practices of corporate governance and their effect on financial performance: Evidence for Chilean companies. *Economic Research-Ekonomska Istraživanja*, Vol. 36 (2527-706). https://doi.org/10.1080/1331677X.2022.2101016

13. Nasteckienė, V. (2021). Empirical investigation of risk management practices. *Management-Journal of Contemporary Management Issues, Vol. 26.* https://doi.org/10.30924/mjcmi.26.2.5

14. Plathottam, S., Rzonca, A., Lakhnori, R. y Iloeje, C. (2023). A review of artificial intelligence applications in manufacturing operations. *AMP Advances in Manufacturing Processes*, Vol. 25. https://doi.org/10.1002/amp2.10159

15. Ramlee, R. y Ahmad, N. (2020). Malaysian risk management committees and firms' financial performance. *Asia-Pacific Management Accounting Journal, Vol. 15 (147-167).* https://doi.org/10.24191/APMAJ.v15i2-07

16. Robles, A., Castañeda, A. y Carrizo, J. (2020). *Gestión de riesgos corporativos y la necesidad de su regulación en las empresas argentinas*. Revista Ciencias Empresariales [Universidad Blas Pascal] Vol. 4 (53 - 67)

17. Slamet, A., Christiana, A. y Kurniawati, H. (2023). Enterprise risk management and firm value: Evidence of Indonesia before and during Covid-19. *E3S Web of Conferences, Vol. 426*. https://doi.org/10.1051/e3sconf/202342602051

18. Subrahmanya, S., Shetty, D., Patil, V., Hameed, B., Paul, R., Smriti, K., Naik, N. y Somani, B. (2022). The role of data science in healthcare advancements: Applications, benefits, and future prospects. *Irish Journal of Medical Science*, Vol. 191 (1473-1483). https://doi.org/10.1007/s11845-021-02693-2

19. Zhou, T. y Zhang, Y. (2022). Risk management committee, audit committee, and firm performance: Evidence from Chinese listed companies. *Asia-Pacific Journal of Accounting & Economics*, Vol. 1 (141-182). https://doi.org/10.1080/16081625.2021.1991423

20. Horvey, S. y Ankamah, J. (2020). Enterprise risk management and firm performance: Empirical evidence from Ghana equity market. *Cogent Economics & Financ*e, Vol. 8. https://doi.org/10.1080/23322039.2020.1840102

21. Gouiaa, R. (2018). Analysis of the effect of corporate governance attributes on risk management practices. Risk Governance and Control: Financial Markets & Institutions, Vol. 8 (14-23). https://doi.org/10.22495/rgcv8i1art2

22. Flores, H. y Moriones, E. (2017). Corporate governance and financial performance of Chilean companies. *Capic Review,* Vol. 15 (31-43).

23. Espinosa, J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. Ingenieria Investigacion y Tecnologia, Vol. 21. https://doi.org/10.22201/fi.25940732e.2020.21n1.008

24. Tocabens, B. (2011). Definiciones acerca del riesgo y sus implicancias. *Revista Cubana de Higiene y Epidemiología*, Vol. 49, (470-481)

25. Patki, N., Wedge , R. y Veeramachameni, K. (2016), The Synthetic data vault. IEEE International Conference on Data Science and Advanced Analytics (DSAA), (399-410)

Anexos

A. Anexo Repositorio de la aplicación web
   https://github.com/arminhm/proyectodjango
B. Anexo Notebook Colab con clustering 3 y reportes EDA

   https://colab.research.google.com/drive/1oj4Y4OsRPf0hH1-A16HBZS6YmtYbJf6P?usp=drive_open