



UNIVERSIDAD DE PAMPLONA
FACULTAD DE CIENCIAS BÁSICAS
MAMEMÁTICA APLICADA

ESTUDIO MONOGRÁFICO SOBRE CIENCIA DE DATOS

T E S I S

Requisito para obtener el título de:
MATEMÁTICO

PRESENTA:
Erick Paulsallet Mantilla Suarez

DIRECTOR:
Luis Alberto Esteban Villamizar

Pamplona, Colombia, 2024



Índice general

Lista de Figuras	V
Lista de Tablas	VII
1. Generalidades	1
1.1. Presentación	1
1.2. Objetivos	2
1.2.1. Objetivo General	2
1.2.2. Objetivos específicos	2
1.3. Marco Conceptual	2
1.4. Estructura del documento	4
2. Marco teórico	5
2.1. Ciencia de Datos	5
2.2. Los programas académicos de ciencia de datos en Colombia	8
2.3. Historia de la Ciencia de Datos	9
2.4. El cuerpo de conocimiento de la Ciencia de datos	13
2.5. La ciencia de datos dentro del cuerpo de conocimiento de las matemáticas	16
2.6. Aprendizaje automático (Machine Learning)	18
2.6.1. Aprendizaje supervisado	20
2.6.2. Aprendizaje no supervisado	21
2.7. Tipos de problemas en ciencia de datos	21
2.7.1. Regresión	21
2.7.2. Clasificación	22
2.7.3. Agrupamiento (Clustering)	23
2.8. Técnicas para la generación de modelos	24
2.8.1. Regresión Lineal	24
2.8.2. Regresión Logística	27
2.8.3. K-means	29
2.9. El proceso de tratamiento de datos	31
2.9.1. Limpieza de datos	31
2.9.2. Normalización	33
2.9.3. Técnicas de Imputación	35

ÍNDICE GENERAL

2.10. Reducción de Dimensionalidad	38
2.10.1. Selección de Características	38
2.10.2. Extracción de Características	39
2.11. Medición de resultados (métricas)	42
2.11.1. Métricas para problemas de regresión	43
2.11.2. Métricas para problemas de clasificación	44
2.11.3. Métricas para problemas de agrupamiento	46
3. Aplicación de las técnicas seleccionadas	49
3.1. Problemas de regresión: técnica de regresión Lineal	50
3.1.1. Limpieza y Preprocesamiento del conjunto de Datos	51
3.1.2. Generación del Modelo	55
3.1.3. Evaluación del modelo	56
3.2. Problemas de clasificación: Técnica de regresión logística	57
3.2.1. Limpieza y Preprocesamiento del Conjunto de Datos	58
3.2.2. Generación del Modelo	62
3.2.3. Evaluación del modelo	65
3.3. Problemas de agrupamiento (clustering): Técnica de K-Nearest Neigh- bors (KNN)	66
3.3.1. Limpieza y Preprocesamiento del Conjunto de Datos	67
3.3.2. Generación del Modelo	67
3.3.3. Evaluación del Modelo	68
4. Conclusiones	71
Bibliografía	73

Lista de Figuras

2.1. Mapa conceptual de ciencia	6
2.2. Mapa conceptual de ciencia de datos	7
2.3. Desde los datos hasta el modelo subyacente	8
2.4. Línea de tiempo: Ciencia de Datos	11
2.5. Áreas fundamentales de las matemáticas y relación con la ciencia de datos	16
2.6. Flujo de trabajo en un modelo en general	18
2.7. Flujo de trabajo de un modelo con aprendizaje automático	19
2.8. Ejemplo de regresión	22
2.9. Ejemplo de Clasificación	22
2.10. Ejemplo de agrupamiento	23
3.1. Distribución de la variable price antes y después de las transformaciones.	52
3.2. Métricas: regresión logística	66

Lista de Tablas

2.1. Universidades que ofrecen programas en Ciencia de Datos	9
2.2. Áreas de conocimiento de la ciencia de datos según la ACM	15
2.3. Comparación entre clasificación, regresión y agrupamiento en ciencia de datos.	25
2.4. Matriz de Confusión para Clasificación Binaria	45
2.5. Matriz de Confusión para el Ejemplo	45
3.1. Resumen descriptivo de las características numéricas del conjunto de datos.	53
3.2. Descripción estadística de los datos.	59
3.3. Resultados del análisis de valores atípicos para la altura.	60
3.4. Resultados del análisis de valores atípicos para el peso.	60
3.5. Descripción estadística de las características ajustadas.	61
3.6. Métricas de evaluación del modelo K-Means.	68

Capítulo 1

Generalidades

1.1. Presentación

Este documento, presentado como una monografía ¹, tiene como propósito brindar un material sintético que contextualice la ciencia de datos dentro del cuerpo de conocimiento de las matemáticas. En un mundo cada vez más impulsado por los datos, esta disciplina se ha convertido en esencial para la toma de decisiones informadas en diversos sectores como negocios, salud y tecnología. A partir de fuentes como el **Data Science Body of Knowledge (DS-BoK)** y las recomendaciones curriculares de la **ACM**, se describen los alcances y la estructura interna de conocimientos de la ciencia de datos, incluyendo una revisión de algunos planes de estudio de programas académicos de pregrado en Colombia. Además, presenta el proceso de construcción y evaluación de modelos predictivos mediante aprendizaje automático. Se introducen tres técnicas representativas para abordar problemas clave de clasificación, agrupamiento y regresión, que serán analizadas en profundidad en los capítulos posteriores. La implementación de estas técnicas se desarrollará a lo largo de un flujo completo que incluye la preparación de los datos, la reducción de la dimensionalidad, la generación de modelos y la evaluación de su desempeño mediante métricas específicas.

¹La monografía se puede definir como una exposición por escrito sobre un tema particular de una ciencia, disciplina o área de estudio, que se desarrolla a partir de fuentes documentales.

1.2. Objetivos

1.2.1. Objetivo General

Realizar un estudio exploratorio de la ciencia de datos, enfocado en las técnicas de modelado mediante aprendizaje automático, abarcando los tres tipos principales de aplicación: clasificación, regresión y agrupamiento.

1.2.2. Objetivos específicos

- Explorar los fundamentos teóricos de la ciencia de datos en el contexto de las matemáticas aplicadas, identificando las áreas y conceptos de la matemática que sustentan las técnicas de modelado de datos.
- Seleccionar una técnica representativa para cada tipo de aplicación del aprendizaje automático: clasificación, regresión y agrupamiento, junto con un conjunto de datos adecuado para su implementación.
- Aplicar el proceso completo de generación y evaluación de modelos para las técnicas y conjuntos de datos seleccionados. Esto incluye la limpieza y preparación del conjunto de datos, particionado del conjunto de datos en datos de entrenamiento y datos de prueba, obtención del modelo y su validación mediante métricas.

1.3. Marco Conceptual

En esta sección se establecen las definiciones clave que fundamentan la comprensión del marco teórico y estructuran los temas centrales abordados en esta monografía. Estas definiciones permiten contextualizar la ciencia de datos y, en particular, el aprendizaje automático como herramienta esencial para la construcción de modelos a partir de conjuntos de datos. El marco conceptual aquí expuesto busca proporcionar una base sólida para entender los conceptos técnicos y metodológicos empleados a lo largo del trabajo de grado.

Definición Un **modelo** es una representación simplificada de un fenómeno, sistema o proceso. En ciencia de datos, un modelo se utiliza para describir, predecir o comprender datos mediante reglas matemáticas y algoritmos, buscando patrones útiles.

Definición Un **modelo estocástico** es aquel que incorpora elementos de azar o incertidumbre. Representa fenómenos que no siempre tienen un comportamiento predecible y produce resultados en términos de probabilidades. Son útiles cuando hay variabilidad inherente en los datos, como en procesos biológicos o económicos.

Definición Un **modelo determinista** es aquel donde, dadas las mismas condiciones iniciales, siempre produce el mismo resultado. Estos modelos no incluyen elementos

aleatorios y se utilizan en situaciones donde las relaciones entre variables son claras y predecibles, como en ciertas aplicaciones de física o ingeniería.

Definición El **aprendizaje automático** es una técnica que permite a las computadoras aprender de los datos sin ser programadas explícitamente para cada tarea. Utiliza modelos matemáticos para identificar patrones y hacer predicciones a partir de datos etiquetados o no etiquetados.

Definición Las **técnicas de aprendizaje automático** son métodos y algoritmos utilizados para construir modelos que analizan datos, identifican patrones y realizan predicciones o clasificaciones.

Definición Un **ingeniero de datos** o **científico de datos** es el encargado de preparar los datos para su análisis, lo que incluye tareas como limpieza, normalización, selección de características, elección de algoritmos y ajuste de modelos. Su labor es crucial para el éxito del aprendizaje automático.

Definición Una **variable** en el contexto de la ciencia de datos y el aprendizaje automático es una característica, atributo o propiedad que se mide o registra para describir un fenómeno, y cuyo valor puede variar entre observaciones dentro de un conjunto de datos.

Tipos de Variables

Definición Una **variable independiente**, en el contexto de la ciencia de datos, es aquella que se utiliza como entrada en un modelo. Su valor no depende de otras variables dentro del análisis y, a menudo, se emplea para predecir el comportamiento de una variable dependiente.

Definición Una **variable dependiente** es aquella que se modela o predice en función de las variables independientes. En aprendizaje automático, corresponde a la variable objetivo que el modelo busca estimar o clasificar.

Definición Una **variable continua** es una característica que puede asumir un número infinito de valores dentro de un rango. En ciencia de datos, estas variables se usan comúnmente en problemas de regresión, como predecir el precio de una vivienda o la temperatura de un día.

Definición Una **variable discreta** toma valores específicos y finitos. Estas variables son frecuentes en el análisis de datos categóricos o en problemas donde las salidas son enteros, como contar el número de clientes que compraron un producto.

Definición Una **variable cualitativa** (o categórica) es una característica que describe categorías o grupos no numéricos. En aprendizaje automático, estas variables son fundamentales en problemas de clasificación, como identificar el tipo de producto en un inventario.

Definición Una **variable cuantitativa** es una característica que toma valores numéricos y representa una magnitud. Estas variables pueden ser continuas o discretas y se utilizan en múltiples aplicaciones de ciencia de datos, desde regresión hasta análisis descriptivo.

Definición Una **variable aleatoria** es una característica cuyo valor resulta de un fenómeno aleatorio. En ciencia de datos, las variables aleatorias son esenciales para

modelar incertidumbre y realizar inferencias estadísticas, siendo clave en algoritmos probabilísticos y técnicas como modelos bayesianos.

1.4. Estructura del documento

Este documento se organiza en cuatro capítulos, diseñados para guiar al lector desde los fundamentos hasta los resultados obtenidos en este estudio.

En el **Capítulo 1**, se establecen los objetivos, el alcance y la relevancia del estudio, además de incluir un glosario de definiciones clave que facilita la comprensión del resto del documento.

El **Capítulo 2** está dedicado al marco teórico y comienza con una contextualización de la ciencia de datos, revisando los programas académicos de ciencia de datos en Colombia y destacando su relevancia y contribución al desarrollo de esta disciplina en el país. A continuación, se aborda la historia de la ciencia de datos, resaltando su evolución y el impacto que tiene en la actualidad. Se examina los conceptos fundamentales de este campo, situándolos dentro del cuerpo de conocimiento de las matemáticas. Además, se describe la estructura interna de la ciencia de datos, identificando sus principales áreas y subáreas de conocimiento. Por último, se presentan las técnicas de aprendizaje automático seleccionadas para su aplicación en el análisis de un conjunto de datos específico.

El **Capítulo 3** detalla el proceso de implementación de las técnicas seleccionadas. Se inicia con la descripción de los conjuntos de datos utilizados, incluyendo los procesos de depuración y preparación de los mismos. Luego, se describe la aplicación de cada técnica, junto con la evaluación de los resultados obtenidos. Para los problemas de clasificación y agrupamiento, se utiliza un conjunto de datos que registra la presencia o ausencia de enfermedades cardiovasculares. En el caso de los problemas de regresión, se trabaja con un conjunto de datos orientado a predecir el precio de venta de viviendas a partir de sus características.

Finalmente, el **Capítulo 4** expone las conclusiones derivadas del trabajo realizado, ofreciendo una visión general de los logros obtenidos y posibles líneas de investigación futura.

Capítulo 2

Marco teórico

2.1. Ciencia de Datos

Una ciencia es un conjunto de conocimientos generados mediante métodos de investigación sistemática, validados y documentados. Aunque no existe un consenso absoluto sobre los tipos de ciencias, es comúnmente aceptada la clasificación en ciencias formales y ciencias fácticas. Las ciencias formales incluyen disciplinas como las matemáticas, la lógica y la estadística, las cuales se centran en la construcción de sistemas abstractos sin necesidad de observaciones empíricas. En cambio, las ciencias fácticas (también conocidas como ciencias empíricas) abarcan tanto las ciencias naturales (como la física, la biología y las ciencias de la tierra) como las ciencias sociales (sociología, psicología, economía, entre otras), y se dedican al estudio de fenómenos observables mediante la experimentación y la observación empírica.

En este contexto, la ciencia de datos puede considerarse como una ciencia empírica, dado que se enfoca en la recolección, análisis y modelado de datos observables, generalmente derivados de fenómenos naturales o sociales. Su objetivo principal es identificar patrones y construir modelos que permitan realizar predicciones o tomar decisiones fundamentadas en los datos recopilados [Bunge \(1960\)](#).

La ciencia de datos se organiza en torno a técnicas, herramientas y procedimientos destinados a la recolección, procesamiento, análisis y visualización de datos. Este campo abarca diferentes tipos de análisis:

1. **Análisis descriptivo:** Describe las características de los datos, respondiendo preguntas como qué ocurrió y cómo es la estructura de los datos.
2. **Análisis diagnóstico:** Identifica las causas subyacentes de los fenómenos observados, determinando por qué sucedieron ciertos eventos.
3. **Análisis predictivo:** Utiliza datos históricos para hacer predicciones sobre eventos futuros, proporcionando respuestas sobre lo que podría suceder.

2. MARCO TEÓRICO

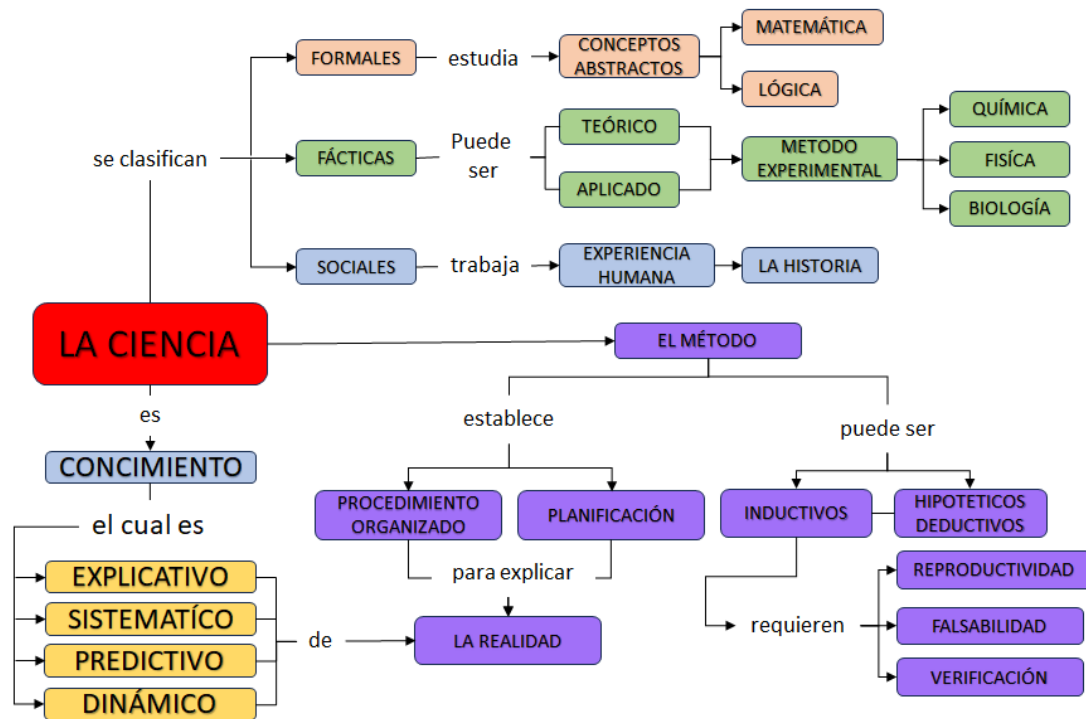


Figura 2.1: Mapa conceptual de ciencia

4. Análisis prescriptivo: Sugiere acciones o decisiones basadas en los resultados del análisis predictivo, proponiendo soluciones a problemas futuros.

Además, la ciencia de datos incluye métodos clave de predicción, como:

- Regresión: Técnica utilizada para predecir valores continuos basados en relaciones entre variables.
- Clasificación: Método para categorizar datos en diferentes grupos o clases.
- Agrupamiento (clustering): Técnica para identificar grupos naturales dentro de los datos, sin necesidad de etiquetas predefinidas.

La ciencia de datos es un campo interdisciplinario que combina conocimientos de diversas disciplinas, como la estadística, la informática y la experiencia en un dominio específico del cual se obtienen los datos a procesar. A través de sus técnicas y herramientas, es posible modelar, resumir y analizar grandes volúmenes de datos utilizando algoritmos que permiten almacenar, procesar y visualizar la información de manera eficiente. Sin embargo, el papel de expertos en los dominios específicos de aplicación es crucial, ya que proporcionan el contexto necesario para formular las preguntas correctas y hacer un uso efectivo de los modelos de datos generados [VanderPlas \(2016\)](#).

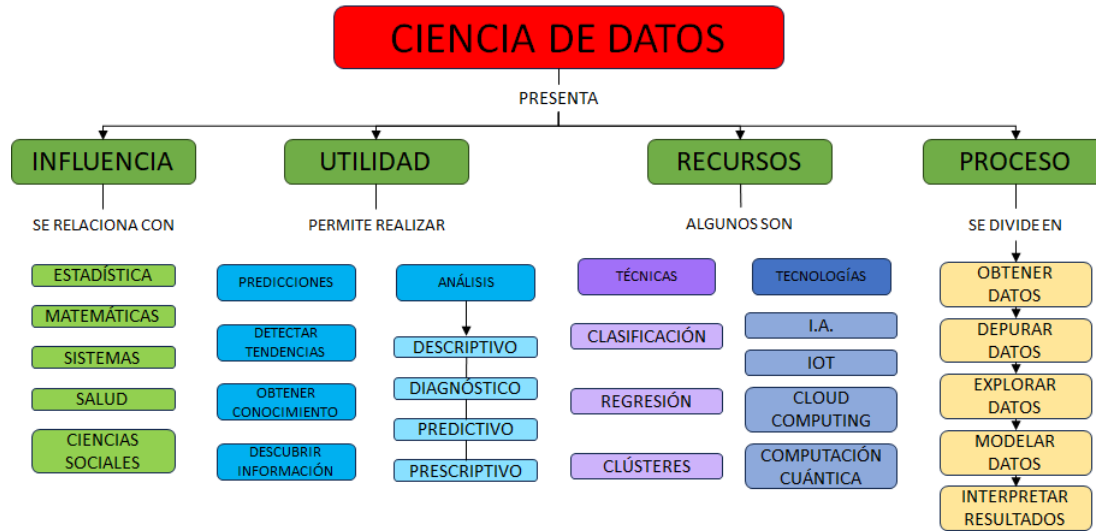


Figura 2.2: Mapa conceptual de ciencia de datos

Más que un campo autónomo, la ciencia de datos se presenta como una profesión interdisciplinaria que desarrolla habilidades aplicables a múltiples áreas y disciplinas. En este sentido, la ciencia de datos tiene como objetivo no solo generar modelos y predicciones, sino también formular y responder nuevas preguntas dentro de un ámbito de especialización elegido [Igal et al. \(2017\)](#). Estos grupos multidisciplinarios se enfocan en la creación de algoritmos, el desarrollo de técnicas computacionales y la implementación de infraestructuras para la adquisición, almacenamiento y procesamiento eficiente de datos, entre otros aspectos. [Herrero et al. \(2018\)](#)

El desarrollo de soluciones que respondan a los retos actuales del análisis de datos requiere la colaboración de equipos multidisciplinarios. Profesionales como informáticos, estadísticos, matemáticos, ingenieros, biólogos, periodistas y sociólogos, entre otros, trabajan en conjunto para generar conocimiento o descubrir modelos ocultos en los datos registrados en diversos procesos. Estos equipos multidisciplinarios se dedican a la creación de algoritmos, el desarrollo de técnicas computacionales y la implementación de infraestructuras que permitan adquirir, almacenar y procesar datos de manera eficiente [Herrero et al. \(2018\)](#).

Como disciplina, la ciencia de datos integra una amplia variedad de procesos, técnicas y habilidades, que abarcan desde la búsqueda y almacenamiento de datos hasta la depuración, selección de técnicas de modelado, configuración de algoritmos de aprendizaje y evaluación de su rendimiento. También incluye el desarrollo de interfaces de software para permitir el uso práctico de los modelos generados en entornos reales.

En este trabajo, se explora cómo el modelado de un conjunto de datos es un proceso de abstracción en el que se aplican diversas técnicas de modelado. El objetivo de este proceso es obtener un algoritmo que represente de manera eficiente las relaciones entre las variables de entrada y salida en el contexto específico de cada análisis. El

2. MARCO TEÓRICO

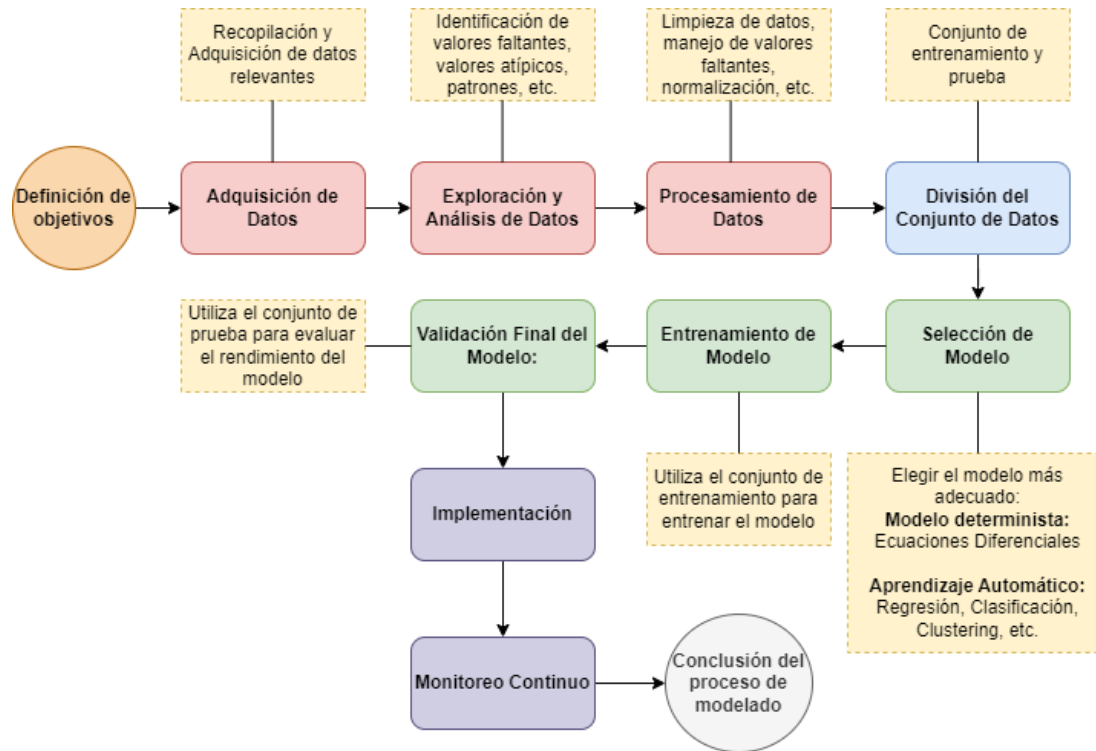


Figura 2.3: Desde los datos hasta el modelo subyacente

aprendizaje automático es una de las principales herramientas utilizadas para realizar estas abstracciones, y en este documento se compararán diferentes enfoques y técnicas, apoyándose en herramientas de software ampliamente utilizadas, como las librerías de Python. (ver figura 2.3).

2.2. Los programas académicos de ciencia de datos en Colombia

La ciencia de datos, un campo interdisciplinario que integra matemáticas, estadística, programación y conocimiento de dominio específico, ha ganado una importancia creciente en sectores como la salud, la tecnología y las finanzas. Este interés ha impulsado a muchas universidades a desarrollar programas académicos especializados que combinan teoría y práctica, con el objetivo de formar profesionales capaces de gestionar y analizar grandes volúmenes de datos.

En Colombia, esta tendencia se ha visto reflejada en la oferta académica de varias instituciones educativas, que han comenzado a ofrecer programas relacionados con la ciencia de datos. Según datos del Sistema Nacional de Información de la Educación

Superior (SNIES)¹, existen múltiples programas académicos en áreas vinculadas a la ciencia de datos, como Estadística, Ciencias de la Computación, Matemáticas con énfasis en Estadística, Ciencia de Datos, Ingeniería en Ciencia de Datos, y otros programas de analítica de datos. Esta oferta académica se ha expandido principalmente en universidades privadas, respondiendo al creciente interés en formar profesionales en este campo.

IES	PROGRAMA ACADÉMICO
Pontificia Universidad Javeriana	Ciencia de Datos
Fundación Universitaria Compensar	Ciencia de Datos
Universidad Externado de Colombia	Ciencia de Datos
Universidad Del Norte	Ciencia de Datos
Universidad Autónoma de Occidente	Ingeniería de Datos e Inteligencia Artificial
Universidad de San Buenaventura	Ingeniería de Datos y Software
Universidad de Manizales	Ingeniería en Analítica de Datos
Corporación Universitaria Iberoamericana	Ingeniería en Ciencia de Datos
Politécnico Grancolombiano	Ingeniería en Ciencia de Datos
Universidad Pontificia Bolivariana	Ingeniería en Ciencia de Datos

Tabla 2.1: Universidades que ofrecen programas en Ciencia de Datos

2.3. Historia de la Ciencia de Datos

La ciencia de datos ha experimentado un desarrollo vertiginoso, transformándose de una subdisciplina de la computación y la estadística a un campo interdisciplinario que abarca diversas áreas del conocimiento. Su historia se remonta a los años 60, cuando Peter Naur acuñó el término “ciencia de datos” para referirse al uso computacional de datos. A medida que las capacidades de procesamiento de las computadoras aumentaron, la disciplina se consolidó y se convirtió en una profesión esencial en la era de la información **Naur (1966)**.

¹Sistema Nacional de Información de Educación Superior que ofrece datos de estadísticas consolidadas en la URL <https://snies.mineducacion.gov.co/portal/ESTADISTICAS/Bases-consolidadas/>

2. MARCO TEÓRICO

Durante las décadas de 1960 y 1970, los avances en el hardware y la creación de lenguajes de programación especializados pusieron los cimientos para el análisis de datos. En los años 80 y 90, la disciplina creció significativamente con la aparición de algoritmos más sofisticados y la mejora de la capacidad de almacenamiento Muñoz (2021). La verdadera explosión se produjo a finales del siglo XX, impulsada por el crecimiento de internet y la generación masiva de datos digitales y humanos. Esto hizo que el análisis de datos dejara de ser exclusivo de académicos y estadísticos, y se convirtiera en una herramienta clave en el mundo empresarial, científico y gubernamental.

Con la llegada del siglo XXI, la ciencia de datos se estableció como una disciplina autónoma. El término “Big Data” se popularizó en la primera década del siglo, reflejando los desafíos asociados a trabajar con grandes volúmenes de datos. Además, la inteligencia artificial y la automatización de procesos impulsaron la importancia de los científicos de datos en la transformación digital de diferentes sectores Glass and Callahan (2014).

Hoy en día, la ciencia de datos es un campo que integra conocimientos de matemáticas, estadística, y computación, y se enriquece con aportes de áreas como ingeniería, ciencias sociales y ética. Su impacto es notable en sectores como la salud, finanzas, educación, entretenimiento y políticas públicas.

Historia de Python y la librería Scikit-learn

Python, concebido por Guido van Rossum a finales de la década de 1980, vio la luz con la versión 0.9.0 en febrero de 1991, enfocándose desde sus inicios en la legibilidad del código y la simplicidad. Durante la próspera década de 1990, Python experimentó un crecimiento constante, culminando con el lanzamiento de la versión 2.0 en 2000, que introdujo características clave como la recolección de basura y soporte completo para Unicode. La transición a Python 3 en 2008 marcó una etapa crucial, presentando mejoras sustanciales pero también incompatibilidades con la versión anterior. En la década de 2000, Python se consolidó como el lenguaje de elección en ciencia de datos, con bibliotecas como NumPy¹ y Matplotlib² desempeñando un papel fundamental en el análisis de datos y visualización Challenger-Pérez et al. (2014). El surgimiento de Jupyter Notebooks en 2014 revolucionó la forma en que se realiza el análisis exploratorio al permitir la combinación de código, visualizaciones y texto en un entorno interactivo Dombrowski et al. (2023). A partir de 2010, la adopción de Python en aprendizaje au-

¹NumPy es una biblioteca fundamental para la computación científica en Python. Proporciona estructuras de datos eficientes para manejar arreglos multidimensionales y funciones para realizar operaciones matemáticas en estos arreglos de manera rápida. <https://numpy.org/>

²Matplotlib es una biblioteca para la creación de gráficos 2D y 3D en Python. Proporciona una interfaz similar a MATLAB para la visualización de datos. <https://matplotlib.org/>

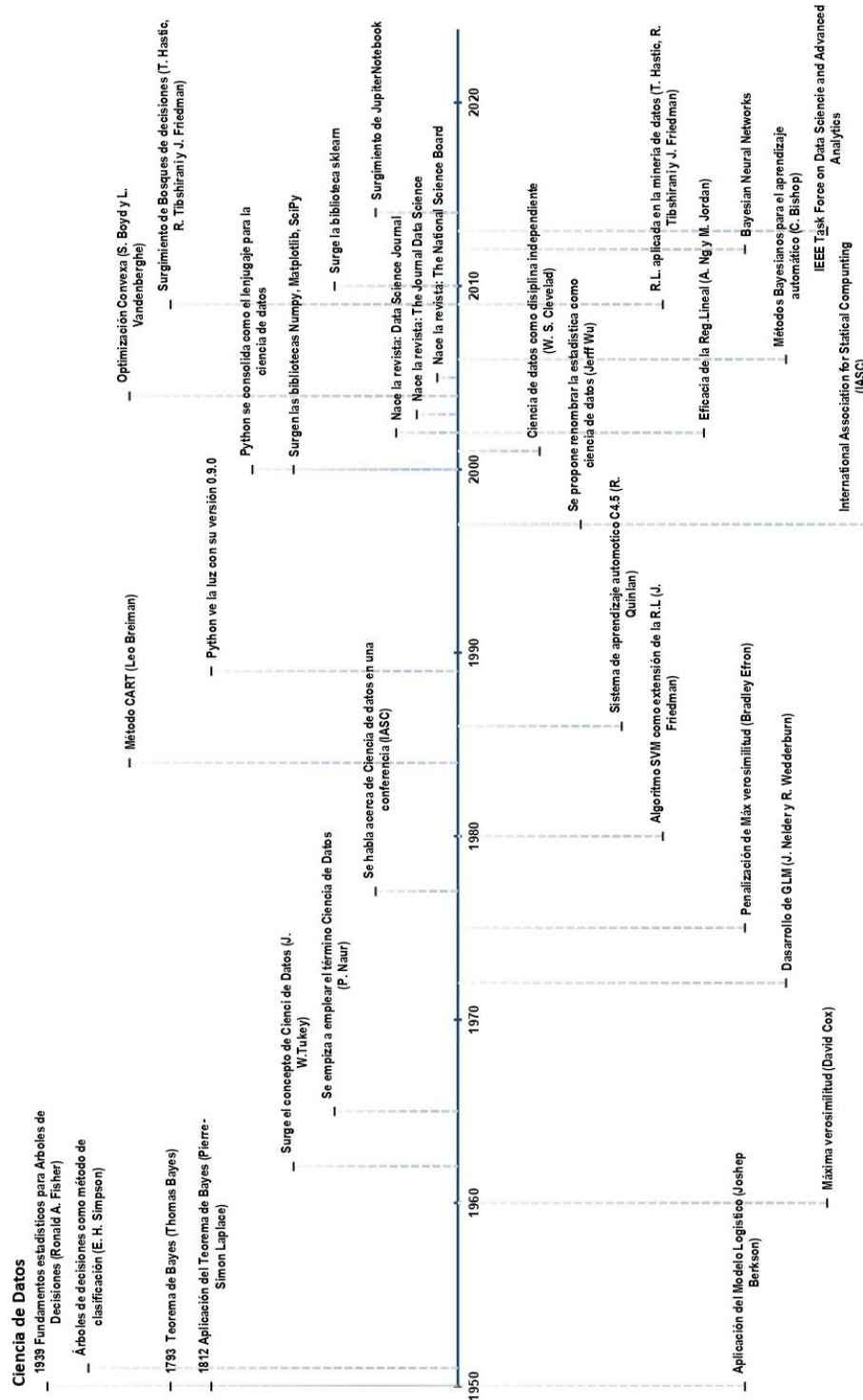


Figura 2.4: Línea de tiempo: Ciencia de Datos

2. MARCO TEÓRICO

tomático se aceleró con bibliotecas como scikit-learn¹, TensorFlow² y PyTorch³. En la actualidad, Python es omnipresente en el ámbito empresarial y académico, respaldado por una comunidad activa y el aporte constante de paquetes especializados. Su importancia en la ciencia de datos se ha cimentado gracias a su sintaxis clara, comunidad robusta y una amplia gama de bibliotecas especializadas, lo que lo convierte en la elección predilecta para proyectos de análisis de datos, aprendizaje automático y desarrollo de modelos predictivos.

Scikit-learn, una biblioteca de aprendizaje automático para Python, tuvo sus inicios en 2007 como parte del programa Google Summer of Code⁴, liderado por David Cournapeau. Concebido con la misión de proporcionar una herramienta simple y eficiente, scikit-learn se expandió gracias a la colaboración abierta y la incorporación de diversos algoritmos y funcionalidades Bisong and Bisong (2019). Desde su versión inicial, la biblioteca ha mantenido un desarrollo activo y una estrecha integración con el ecosistema de herramientas de análisis de datos en Python, incluyendo NumPy, SciPy y Matplotlib. Su interfaz simple y consistente, junto con una amplia variedad de algoritmos y utilidades para el preprocesamiento de datos, selección de modelos y evaluación, ha ganado reconocimiento en la comunidad científica y educativa. Scikit-learn sigue siendo un proyecto en evolución, con un equipo comprometido que asegura su relevancia y utilidad continua en el ámbito del aprendizaje automático.

Evolución y Relevancia Contemporánea de la Ciencia de Datos

La ciencia de datos, aunque considerada una disciplina moderna, tiene una historia que se remonta a conceptos y técnicas desarrolladas en el siglo XX. Peter Naur utilizó el término “ciencia de datos.” en la década de 1960 como un sinónimo de ciencias computacionales, y John W. Tukey lo popularizó en 1962 con su artículo sobre análisis de datos Tukey (1962). La denominación “ciencia de datos” fue formalizada en conferencias y publicaciones durante los años 90, consolidándose como un campo de estudio y profesión en la primera década del siglo XXI Press (2013).

En la actualidad, la ciencia de datos está en el corazón de la transformación digital impulsada por el Big Data y la inteligencia artificial. El procesamiento de grandes

¹scikit-learn es una biblioteca de aprendizaje automático de código abierto que proporciona herramientas simples y eficientes para análisis predictivo y minería de datos. Está construida sobre otras bibliotecas populares como NumPy, SciPy y Matplotlib <https://scikit-learn.org/>

²TensorFlow es una biblioteca de código abierto desarrollada por el equipo de Google Brain que se utiliza para implementar y entrenar modelos de aprendizaje profundo. <https://www.tensorflow.org/?hl=es-419>

³PyTorch es una biblioteca de código abierto de aprendizaje profundo desarrollada principalmente por Facebook. Se utiliza para construir y entrenar modelos de aprendizaje profundo, como redes neuronales artificiales. <https://pytorch.org/>

⁴Google Summer of Code es un programa global en línea centrado en atraer nuevos contribuyentes al desarrollo de software de código abierto.

volúmenes de datos y el desarrollo de algoritmos avanzados han permitido avances significativos en áreas como el reconocimiento de voz, la visión por computadora y el procesamiento del lenguaje natural. La integración de aprendizaje profundo y técnicas avanzadas de optimización ha abierto nuevas fronteras en el análisis predictivo y la toma de decisiones basada en datos.

El campo de la ciencia de datos es cada vez más interdisciplinario, con contribuciones de la estadística, la informática, la ingeniería y las ciencias sociales. La colaboración entre estas disciplinas potencia la capacidad para resolver problemas complejos y obtener insights valiosos a partir de los datos.

Las tendencias emergentes, como el DataOps¹ y el MLOps², están transformando la forma en que se gestionan y operan los proyectos de ciencia de datos, mientras que la adopción de tecnologías como la computación en la nube y Hadoop³ está facilitando el almacenamiento y procesamiento de datos a gran escala. Además, el enfoque en la ética de los datos y la responsabilidad en el uso de algoritmos se está convirtiendo en una prioridad crucial para abordar los desafíos de privacidad y sesgo en los modelos.

Mirando hacia el futuro, la ciencia de datos se enfrenta a nuevos retos y oportunidades con el avance de la computación cuántica y el creciente impacto del Internet de las Cosas (IoT). Estos desarrollos prometen transformar aún más el campo, ofreciendo nuevas capacidades para el análisis y la interpretación de datos.

2.4. El cuerpo de conocimiento de la Ciencia de datos

La ciencia de datos es un campo en constante crecimiento que se enfoca en la extracción de conocimiento (modelos) y la generación de ideas a partir de grandes

¹DataOps (Data Operations) es una metodología que busca mejorar la calidad y velocidad del análisis de datos mediante la integración de prácticas de desarrollo ágil y operaciones de TI. Al igual que DevOps en el desarrollo de software, DataOps se enfoca en automatizar y optimizar el flujo de trabajo de datos, desde la ingestión hasta el análisis, promoviendo una colaboración más efectiva entre equipos de datos y operaciones. <https://www.dataversity.net/what-is-dataops/>

²MLOps (Machine Learning Operations) es una práctica que combina técnicas de operaciones de TI y aprendizaje automático para gestionar el ciclo de vida de los modelos de machine learning en producción. MLOps busca automatizar y estandarizar la implementación, monitoreo y mantenimiento de modelos, asegurando su rendimiento continuo y facilitando la colaboración entre equipos de datos, desarrollo y operaciones. <https://www.microsoft.com/en-us/ai/ai-lab-mlops>

³Hadoop es un marco de software de código abierto diseñado para almacenar y procesar grandes volúmenes de datos en clústeres de computadoras. Utiliza un modelo de programación distribuido para dividir y procesar datos en paralelo, lo que permite manejar datos a gran escala de manera eficiente. Hadoop incluye componentes como el sistema de archivos distribuido Hadoop (HDFS) y el marco de procesamiento MapReduce. <https://hadoop.apache.org/>

2. MARCO TEÓRICO

volúmenes de datos. Para comprender y aplicar eficazmente los principios y métodos de la ciencia de datos, es fundamental contar con un marco de referencia sólido que defina las áreas de conocimiento y las competencias clave en este campo. En este contexto, el Cuerpo de Conocimiento de Ciencia de Datos (DS-BoK) [Provost et al. \(2023\)](#) emerge como una propuesta integral.

El DS-BoK¹ se basa en un análisis exhaustivo de diversos cuerpos de conocimiento relevantes y en el Marco de Competencias en Ciencia de Datos (CF-DS) [B.V. \(2019\)](#), con el objetivo de establecer las áreas de conocimiento y las unidades fundamentales en ciencia de datos. Este enfoque proporciona una estructura sólida para diseñar programas educativos, desarrollar materiales didácticos y evaluar las habilidades y conocimientos de los profesionales en ciencia de datos.

El DS-BoK se compone de cinco grupos de áreas de conocimiento (KAG, por sus siglas en inglés) que cubren diferentes aspectos de la ciencia de datos.

1. **Análisis de Datos (DSDA):** Incluye métodos estadísticos, como el análisis descriptivo y el análisis exploratorio de datos (EDA), así como el análisis confirmatorio de datos (CDA) para validar hipótesis formuladas [Provost et al. \(2023\)](#).
2. **Ingeniería de Ciencia de Datos (DSENG):** Engloba tecnologías, paradigmas y metodologías relacionadas con la ingeniería de software y la infraestructura tecnológica. También se incluyen en este grupo enfoques ágiles, tecnologías basadas en la nube y prácticas como DevOps [Provost et al. \(2023\)](#).
3. **Gestión de Datos (DSDM):** Se centra en la administración integral de los datos, abordando aspectos como la curación de datos, la preservación de datos y la infraestructura de datos. Además, se consideran recomendaciones de organismos como la Research Data Alliance (RDA) y modelos de gestión de datos comunitarios, como el acceso abierto y los datos abiertos. El concepto de Gestión del Ciclo de Vida de los Datos también es crucial en este grupo [Provost et al. \(2023\)](#).
4. **Métodos de Investigación y Gestión de Proyectos (DSRMP):** Este grupo se enfoca en los métodos de investigación aplicados a la ciencia de datos y en la gestión de proyectos relacionados. Aquí se exploran técnicas de investigación, métodos de experimentación y mejores prácticas para la gestión de proyectos de ciencia de datos [Provost et al. \(2023\)](#).
5. **Análisis de Negocios (DSBA):** Cubre el análisis de datos desde una perspectiva empresarial, centrándose en la agregación de datos y el análisis de múltiples fuentes de información para obtener conocimientos relevantes para la toma de decisiones en el ámbito empresarial [Provost et al. \(2023\)](#).

¹El Cuerpo de conocimientos sobre ciencia de datos (DS-BoK) está diseñado para crear una base para establecer una profesión de científico de datos para la industria en todo el mundo, dicho documento es realizado por The International Association of Business Communicators (IABC))

Abr.	Área de conocimiento	En español
AL	Algorithms and Complexity	Algoritmos y Complejidad
AR	Architecture and Organization	Arquitectura y organización
CN	Computational Science	Ciencias computacionales
DS	Discrete Structures	Estructuras Discretas
GV	Graphics and Visualization	Gráficas y visualización
HCI	Human-Computer Interaction	Interacción Humano-computador
IAS	Information Assurance and Security	Seguridad de Información
IM	Information Management	gestión de Información
IS	Intelligent Systems	Sistemas Inteligentes
NC	Networking and Communications	Redes y comunicaciones
OS	Operating Systems	sistemas operativos
PBD	Platform-based Development	Desarrollo basado en plataformas
PD	Parallel and Distributed Computing	Comp. paralela y distribuida
PL	Programming Languages	Lenguajes de Programación
SDF	Software Development Fundamentals	Fund. de desarrollo de software
SE	Software Engineering	Ingeniería de Software
SF	Systems Fundamentals	Fund. de sistemas
SP	Social Issues and Professional Practice	Asuntos sociales y prácticas Prof.

Tabla 2.2: Áreas de conocimiento de la ciencia de datos según la ACM

De igual manera, es importante resaltar el trabajo de la ACM¹ que en el año 2013 lanzó las recomendaciones curriculares para la ciencia de datos dentro de las disciplinas de la computación².

¹La ACM (Association for Computing Machinery) es una organización internacional sin ánimo de lucro dedicada a avanzar en el campo de la informática. Fundada en 1947, la ACM es una de las organizaciones más grandes y antiguas del mundo dedicadas a la informática. Su misión es fomentar la excelencia en el ámbito de la informática a través de la entrega de recursos técnicos, oportunidades de educación y foros de colaboración.

²<https://www.acm.org/education/curricula-recommendations>

2.5. La ciencia de datos dentro del cuerpo de conocimiento de las matemáticas

La matemática es una disciplina fundamental que abarca un amplio espectro de conocimientos y aplicaciones en diversas áreas. Su naturaleza abstracta y su capacidad para modelar fenómenos del mundo real hacen de las matemáticas un pilar esencial en la construcción del conocimiento científico. Antes de explorar las diversas ramas de las matemáticas y su relación con la ciencia de datos, es importante reconocer que la organización y clasificación de estas áreas pueden variar según diferentes instituciones y organizaciones. Diferentes entidades pueden tener enfoques particulares y criterios específicos para dividir y categorizar el vasto campo de las matemáticas. En este contexto, categorizamos las matemáticas en las siguientes áreas: Estadística y Probabilidad, Álgebra y Álgebra Lineal, Cálculo y Análisis Matemático, Matemáticas Computacionales, Optimización, Matemáticas Discretas, Teoría de Números y Topología, y establecemos la relación que algunas de estas áreas tienen con la ciencia de datos.

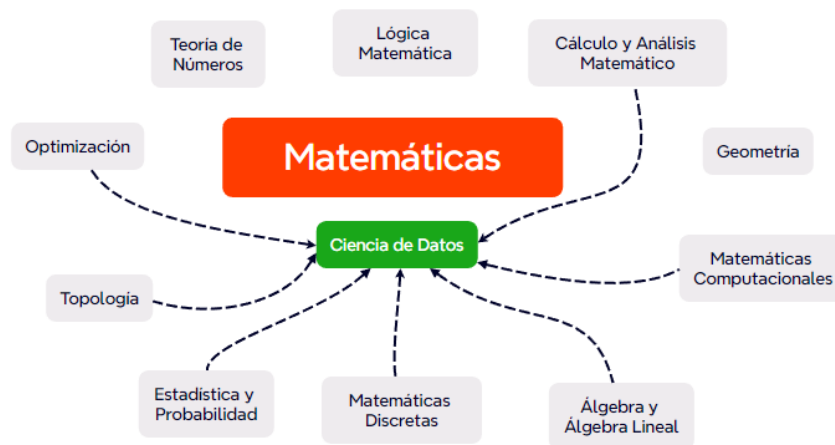


Figura 2.5: Áreas fundamentales de las matemáticas y relación con la ciencia de datos

Estadística y Probabilidad La estadística y la teoría de la probabilidad son pilares fundamentales en la ciencia de datos. La estadística descriptiva se utiliza para resumir y describir las características de los datos mediante medidas de tendencia central (como la media, mediana y moda) y de dispersión (como la varianza y desviación estándar). Estos resúmenes ayudan a entender la distribución y variabilidad de los datos antes de aplicar técnicas más avanzadas. La estadística inferencial permite hacer generalizaciones sobre una población basándose en una muestra de datos. Métodos como intervalos de confianza y pruebas de hipótesis son esenciales para validar modelos

y tomar decisiones basadas en datos. La teoría de la probabilidad subyace en la modelización predictiva y el análisis de riesgo, proporcionando un marco para manejar la incertidumbre y el comportamiento aleatorio en los datos [Leal \(2022\)](#).

Álgebra y Álgebra Lineal El álgebra lineal es fundamental en el manejo y análisis de grandes conjuntos de datos. Las operaciones con matrices y vectores son comunes en el procesamiento de datos, algoritmos de recomendación y redes neuronales en el aprendizaje profundo. Por ejemplo, la descomposición en valores singulares (SVD) y la factorización de matrices se utilizan para la reducción de dimensionalidad y el análisis de componentes principales (PCA), que ayudan a simplificar y visualizar datos complejos. El álgebra abstracta, aunque menos directamente aplicada, influye en áreas como la criptografía y la teoría de la información, que son cruciales para la seguridad y la privacidad de los datos [Strang \(2022\)](#).

Cálculo y Análisis Matemático El cálculo diferencial e integral es vital para la optimización de funciones, una tarea central en el entrenamiento de modelos de aprendizaje automático. El cálculo permite entender y modelar cambios y acumulaciones, lo cual es crucial para ajustar parámetros en modelos predictivos y encontrar valores óptimos que minimicen el error. El análisis numérico proporciona métodos para aproximar soluciones de problemas matemáticos que no pueden resolverse analíticamente. Esto es esencial en la simulación de modelos complejos y en la resolución de problemas de optimización, donde los métodos exactos son inviables debido a la gran escala de los datos.

Matemáticas Computacionales Los algoritmos y la teoría de la computación son la columna vertebral de la ciencia de datos. El diseño y análisis de algoritmos eficientes son cruciales para la minería de datos, el aprendizaje automático y el procesamiento de grandes volúmenes de datos. La criptografía asegura que los datos se mantengan confidenciales y seguros durante su análisis y almacenamiento, protegiendo la integridad y privacidad de la información [Lemus-Delgado and Pérez Navarro \(2020\)](#).

Optimización La optimización, tanto lineal como no lineal, es central en la ciencia de datos. Se utiliza para encontrar la mejor solución a problemas con restricciones, como ajustar modelos predictivos para minimizar el error de predicción o maximizar el rendimiento. Las técnicas de optimización ayudan a seleccionar los mejores parámetros y a mejorar la precisión de los modelos [Cueli et al. \(2019\)](#).

Matemáticas Discretas La teoría de grafos es crucial para el análisis de redes sociales, estructuras de datos y algoritmos de búsqueda. Permite modelar relaciones y conexiones en datos complejos, como redes de comunicación o redes sociales. La combinatoria se utiliza en problemas que involucran grandes conjuntos de combinaciones y permutaciones, importantes en la optimización y el análisis de datos [Epp \(2012\)](#).

Modelado Matemático El modelado matemático implica la construcción de mo-

delos matemáticos para representar sistemas reales, lo cual es esencial en la simulación y predicción de fenómenos basados en datos. La teoría de la información proporciona fundamentos para la compresión y transmisión de datos, esenciales en la gestión eficiente de grandes volúmenes de información. Al analizar estas relaciones, se observa cómo la ciencia de datos integra diversas áreas matemáticas para proporcionar herramientas y técnicas cruciales en el análisis y la interpretación de datos. La comprensión de estas interconexiones es fundamental para avanzar en la investigación y aplicación de la ciencia de datos en múltiples dominios [Linares et al. \(2001\)](#).

Al analizar estas relaciones, se observa cómo la ciencia de datos integra diversas áreas matemáticas para proporcionar herramientas y técnicas cruciales en el análisis y la interpretación de datos. La comprensión de estas interconexiones es fundamental para avanzar en la investigación y aplicación de la ciencia de datos en múltiples dominios.

2.6. Aprendizaje automático (Machine Learning)

El aprendizaje automático es una rama de la inteligencia artificial (IA) que dota a los sistemas de la capacidad de aprender y mejorar su rendimiento de forma automática a partir de la experiencia, sin necesidad de una programación específica para cada tarea o situación. La premisa fundamental es que estos sistemas pueden identificar patrones en los datos, ajustar sus parámetros y mejorar su desempeño conforme reciben más información, logrando así hacer predicciones y tomar decisiones más precisas con el tiempo.



Figura 2.6: Flujo de trabajo en un modelo en general

Los algoritmos de aprendizaje automático permiten crear modelos a partir de los datos iniciales y, mediante la retroalimentación y la actualización continua, optimizan sus predicciones. Esto resulta clave en escenarios donde los datos evolucionan o cuando se requiere flexibilidad en la respuesta de los sistemas. Este proceso de aprendizaje y ajuste también permite reducir los errores e incrementar la precisión del modelo conforme se le proporcionan datos adicionales [Moreno et al. \(1994\)](#).

El aprendizaje automático resulta útil en diversas situaciones clave:

- Para resolver problemas estructurados: Es ideal cuando se pueden definir reglas o patrones claros que el modelo debe aprender a reconocer a partir de ejemplos, tales como clasificación de imágenes o análisis de sentimientos.
- En problemas complejos o de difícil análisis manual: Cuando el análisis manual de la información resulta inviable por la complejidad o el volumen de los datos, el aprendizaje automático puede identificar patrones y relaciones que no son evidentes a simple vista.
- En entornos dinámicos o cambiantes: En contextos donde la información varía o se actualiza con frecuencia, el aprendizaje automático permite que el modelo se ajuste a los nuevos datos, manteniendo su capacidad de predicción y relevancia.
- Como apoyo en el análisis de grandes volúmenes de datos: En situaciones donde se dispone de grandes conjuntos de datos difíciles de interpretar manualmente, el aprendizaje automático ayuda a extraer patrones útiles y a tomar decisiones informadas en menos tiempo Norman (2019).

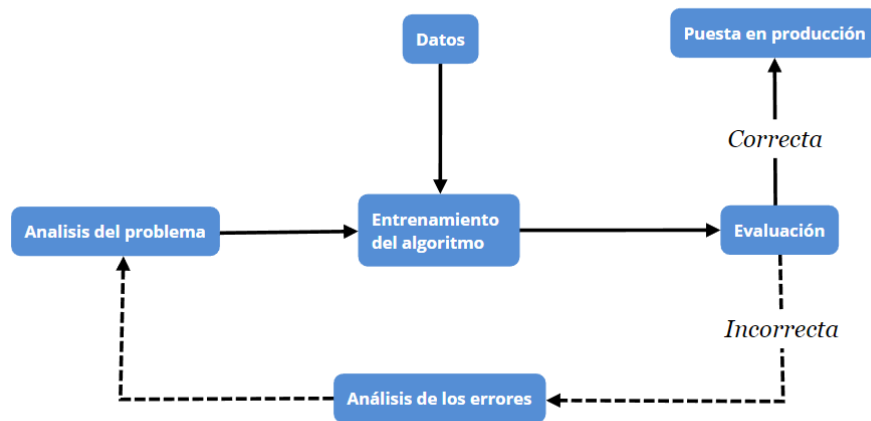


Figura 2.7: Flujo de trabajo de un modelo con aprendizaje automático

Los sistemas de aprendizaje automático se pueden clasificar de varias formas, dependiendo de los criterios que se utilicen. Aquí te presento algunas de las clasificaciones más comunes:

- En función de la manera en que se entrena:
 - Aprendizaje supervisado: El modelo se entrena con ejemplos etiquetados, es decir, datos que tienen una respuesta conocida.

2. MARCO TEÓRICO

- Aprendizaje no supervisado: El modelo se entrena con datos no etiquetados y busca patrones o estructuras inherentes en los datos.
- Aprendizaje semi-supervisado: Se utiliza cuando solo una parte de los datos está etiquetada y se busca mejorar el aprendizaje aprovechando los datos no etiquetados.
- En función de la manera en que aprenden con el tiempo:
 - Aprendizaje online: Los modelos se actualizan continuamente con nuevos datos a medida que llegan.
 - Aprendizaje batch: Los modelos se entrenan utilizando todo el conjunto de datos disponible de una vez.
- En función de la forma en la que realizan las predicciones:
 - Aprendizaje basado en instancias: Se basan en encontrar similitudes entre instancias individuales de datos.
 - Aprendizaje basado en modelos: Construyen un modelo predictivo a partir de los datos de entrenamiento.

Es importante destacar que los sistemas de aprendizaje automático no se categorizan de manera excluyente, sino que pueden aplicarse a más de una categoría según sus técnicas y algoritmos específicos. En el ámbito práctico, se enfoca principalmente en la primera clasificación, que se divide las técnicas en las categorías de aprendizaje supervisado y no supervisado [Díaz-Ramírez \(2021\)](#).

2.6.1. Aprendizaje supervisado

El aprendizaje supervisado es un enfoque central del aprendizaje automático en el que se entrena un algoritmo utilizando un conjunto de datos etiquetados. Durante el proceso de aprendizaje, el algoritmo aprende los ejemplos de entrada y las correspondientes respuestas deseadas. El objetivo principal es que el algoritmo aprenda la relación o mapeo entre estas entradas y salidas. Un concepto clave en este enfoque es un conjunto de datos etiquetados que consta de pares de ejemplos de entrada y sus respuestas correctas asociadas. Por ejemplo, en una tarea de clasificación de imágenes, las imágenes forman las entradas y las etiquetas indican la clase a la que pertenece cada imagen. En la fase de entrenamiento, el algoritmo ajusta sus parámetros utilizando estos datos etiquetados para minimizar una función de pérdida que mide la diferencia entre las predicciones del modelo y las respuestas reales. Una vez que se entrena un modelo, puede hacer predicciones basadas en datos sin etiquetar, aplicando la relación aprendida a nuevas entradas [Bobadilla \(2021\)](#). El aprendizaje supervisado se ocupa principalmente de dos tipos de problemas: clasificación, donde a cada entrada se le asigna una etiqueta basada en un conjunto predeterminado de clases, y regresión, cuyo objetivo es predecir un valor numérico.

2.6.2. Aprendizaje no supervisado

El aprendizaje no supervisado es una rama central del campo del aprendizaje automático donde los algoritmos exploran conjuntos de datos sin etiquetar sin conocer los resultados deseados. Este enfoque enfatiza varias tareas como la agrupación, que organiza automáticamente los datos en grupos en función de similitudes; reducción de dimensionalidad, que transforma los datos a un estado de dimensiones inferiores; y detección de anomalías, que busca patrones inusuales. Además, el aprendizaje no supervisado puede manejar la relación entre variables y utilizar técnicas como la codificación automática junto con redes neuronales para aprender representaciones eficientes. Este enfoque es importante para descubrir las estructuras subyacentes, los patrones ocultos y las relaciones inherentes en los datos, y es valioso en aplicaciones como la segmentación de clientes, la exploración de grandes conjuntos de datos y la organización de la información general [Bobadilla \(2021\)](#).

Estas categorías proporcionan un marco inicial para comprender cómo los diferentes sistemas de aprendizaje automático pueden aplicarse según el tipo de datos y los objetivos de modelado.

2.7. Tipos de problemas en ciencia de datos

Uno de los aspectos más destacados de la ciencia de datos es su capacidad para enfrentar una variedad de desafíos, lo que permite la toma de decisiones basadas en datos, la predicción de resultados y la revelación de patrones ocultos en datos complejos.

A continuación, se exploran los tres tipos fundamentales de problemas en ciencia de datos:

2.7.1. Regresión

El objetivo principal de los problemas de regresión es realizar predicciones de valores continuos o *floating-point numbers*, lo que se traduce en términos de programación como números reales en matemáticas [Müller and Guido \(2016\)](#). Un ejemplo típico de tarea de regresión es predecir el ingreso anual de una persona basándose en factores como su nivel educativo, edad y lugar de residencia. Cuando se realiza esta predicción de ingresos, el valor resultante es una cantidad numérica y puede tomar cualquier número dentro de un rango especificado. Otra instancia de una tarea de regresión sería predecir el rendimiento de un cultivo de maíz en función de características como los rendimientos anteriores, las condiciones climáticas y el número de empleados que trabajan en la finca. Nuevamente, el rendimiento previsto puede ser un número arbitrario en lugar de una categoría o etiqueta.

En términos más generales, la regresión se aplica en cualquier contexto donde se necesite predecir un valor numérico continuo a partir de un conjunto de variables de

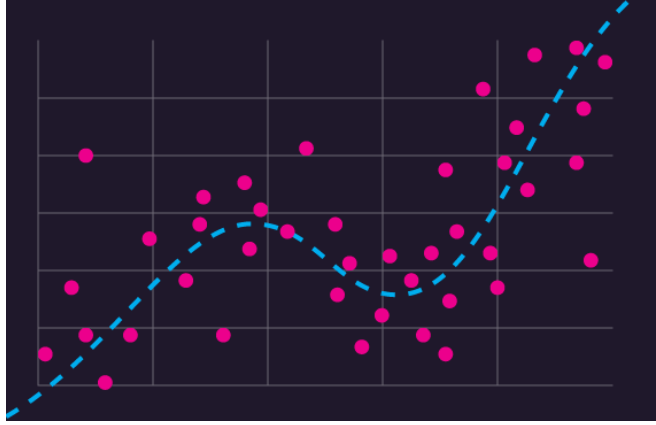


Figura 2.8: Ejemplo de regresión

entrada, y es fundamental en campos como las finanzas, la economía, la medicina e ingeniería.

2.7.2. Clasificación

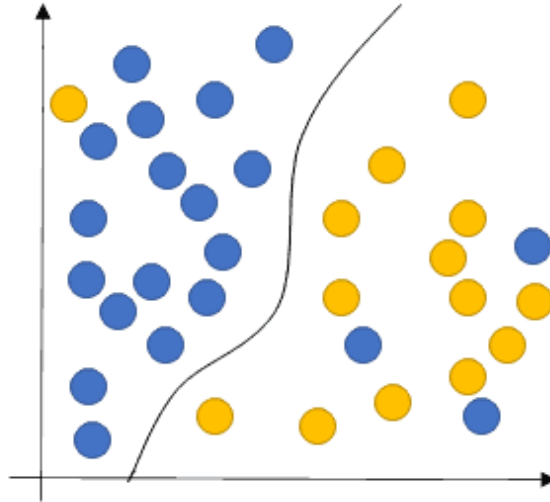


Figura 2.9: Ejemplo de Clasificación

La clasificación en ciencia de datos tiene como objetivo principal predecir con precisión a qué grupo pertenece cada muestra de datos [Bobadilla \(2021\)](#). Un ejemplo ilustrativo sería un conjunto de datos que contenga correos electrónicos etiquetados como “spam” o “no spam”. A partir de estos datos, se desarrolla un modelo que aprende

las diferencias en los patrones de correos electrónicos considerados spam y aquellos que no lo son.

Una vez entrenado el modelo, se utilizará para clasificar nuevos correos electrónicos (nuevas muestras) en una de las dos categorías: spam o no spam. Los algoritmos de clasificación trabajan con datos que suelen estar etiquetados y pueden ser de naturaleza categórica en muchos casos. A diferencia de los algoritmos de regresión, que predicen valores numéricos continuos, los algoritmos de clasificación asignan una muestra a uno de varios grupos o clases existentes.

Además de predecir a qué clase pertenece una muestra, algunos algoritmos de clasificación también pueden proporcionar la probabilidad de que una muestra dada pertenezca a cada una de las clases existentes. Es importante destacar que, aunque los algoritmos pueden devolver conjuntos numéricos de probabilidades, su objetivo final suele ser asignar una etiqueta de categoría específica a cada muestra.

2.7.3. Agrupamiento (Clustering)

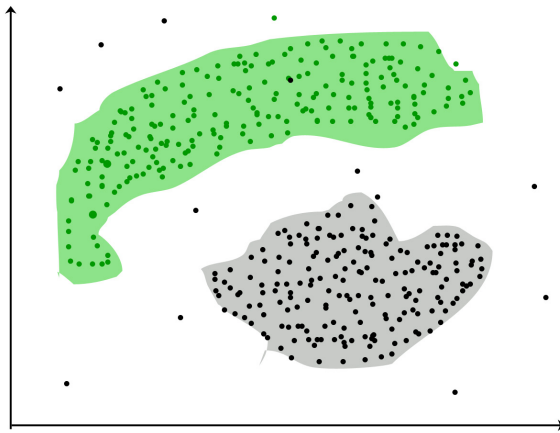


Figura 2.10: Ejemplo de agrupamiento

El agrupamiento, una técnica ampliamente reconocida en el campo de la ciencia de datos, es una forma de aprendizaje no supervisado. A diferencia de la clasificación y la regresión, el agrupamiento no trabaja con datos etiquetados; su objetivo es agrupar muestras que comparten similitudes [Bobadilla \(2021\)](#). Esta información es invaluable en diversas aplicaciones, como el marketing, donde se puede categorizar a los clientes en grupos con características similares para enviar publicidad personalizada. En las redes sociales, identificar y diferenciar los comportamientos de distintos tipos de usuarios es esencial. En el comercio electrónico, agrupar productos puede mejorar significativamente los sistemas de recomendación.

El agrupamiento permite tomar medidas personalizadas basadas en la pertenencia de una muestra a un grupo específico. Es una herramienta fundamental en la ciencia

de datos, ya que ayuda a descubrir estructuras ocultas en los datos, facilitando la segmentación y el análisis profundo en múltiples dominios.

Estas tres categorías fundamentales de problemas en ciencia de datos clasificación, regresión y agrupamiento proporcionan una base sólida para abordar una amplia variedad de desafíos analíticos, permitiendo a los científicos de datos extraer valor significativo de los datos complejos y variados con los que trabajan.

2.8. Técnicas para la generación de modelos

Una vez comprendida la importancia de los modelos estocásticos y deterministas en el aprendizaje automático, así como su impacto en la precisión y robustez de las predicciones, es fundamental explorar las técnicas específicas utilizadas para generar estos modelos. Esta sección se centrará en algunas de las principales técnicas de aprendizaje supervisado y no supervisado, esenciales para resolver diversos problemas en la ciencia de datos.

Entre los principales Modelos de aprendizaje supervisado y no supervisado se encuentran los siguientes:

2.8.1. Regresión Lineal

La regresión lineal es una técnica fundamental en el campo de la estadística y el aprendizaje automático que se utiliza para modelar la relación entre una variable dependiente y una o más variables independientes. A continuación, se proporciona una descripción detallada de la técnica, sus características, aplicaciones, fundamentos matemáticos, construcción del modelo, función de coste, optimización y conclusiones [Roque López et al. \(2021\)](#).

La regresión lineal es una técnica predictiva que busca establecer una relación lineal entre una variable dependiente (también conocida como variable de respuesta o target) y una o más variables independientes (predictoras). Esta relación se representa mediante una ecuación lineal que describe cómo los cambios en las variables independientes afectan a la variable dependiente, sus principales características son:

- **Simplicidad:** La regresión lineal es sencilla de entender e interpretar, lo que la hace ideal para explicar las relaciones entre variables.
- **Eficiencia:** Es computacionalmente eficiente y se puede aplicar a conjuntos de datos grandes con relativa facilidad.
- **Analítica:** Proporciona una base para inferencias estadísticas y permite la prueba de hipótesis sobre la relación entre variables.
- **Predictiva:** Aunque es simple, la regresión lineal puede ser una herramienta poderosa para la predicción y el pronóstico.

Características	Clasificación	Regresión	Agrupamiento
Tipo de Aprendizaje	Supervisado	Supervisado	No supervisado
Objetivo	Asignar una etiqueta de categoría a cada muestra.	Predecir un valor numérico continuo para cada muestra.	Agrupar muestras similares en grupos.
Datos de Entrada	Etiquetados (categorías)	Etiquetados (valores continuos)	No etiquetados
Salida del Modelo	Categorías o clases.	Valores numéricos.	Grupos o clusters.
Algoritmos Comunes	Regresión logística, árboles de decisión, SVM, redes neuronales.	Regresión lineal, regresión polinómica, redes neuronales.	K-means, jerárquico, DBSCAN.
Medida de Rendimiento	Precisión, recall, F1-score, ROC-AUC.	Error cuadrático medio (MSE), error absoluto medio (MAE), R^2 .	Índice de Silueta, SSE (Suma de Cuadrados Dentro del Cluster).
Ejemplos de Aplicaciones	Detección de spam, diagnóstico de enfermedades, clasificación de imágenes.	Predicción de precios de viviendas, estimación de ingresos, análisis de ventas.	Segmentación de clientes, agrupación de productos, análisis de comportamiento.

Tabla 2.3: Comparación entre clasificación, regresión y agrupamiento en ciencia de datos.

La técnica de regresión lineal se utiliza para predecir valores futuros de una variable dependiente basándose en valores conocidos de variables independientes. Además, permite identificar y cuantificar tendencias en datos históricos. También se emplea para

2. MARCO TEÓRICO

evaluar el impacto de una o más variables independientes sobre una variable dependiente. En el ámbito económico, la regresión lineal es ampliamente utilizada en econometría para modelar relaciones económicas. Finalmente, en finanzas y seguros, se utiliza para modelar el riesgo y predecir eventos adversos.

La ecuación general de la regresión lineal simple para un conjunto de datos con una única característica de entrada es:

$$y = \theta_0 + \theta_1 x + \epsilon \quad (2.1)$$

Donde:

- y es la variable dependiente.
- x es la variable independiente.
- θ_0 es la intersección con el eje y (constante).
- θ_1 es la pendiente de la línea (coeficiente).
- ϵ es el término de error o residuo.

Para la regresión lineal múltiple (múltiples características de entrada), la ecuación se generaliza a:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n + \epsilon \quad (2.2)$$

Donde

- $x_1, x_2, \dots, x_n$ son las características de entrada
- $\theta_1, \theta_2, \dots, \theta_n$ son los coeficientes asociados.
- n es el número de características de entrada

Función de Coste (Función Error)

La función de coste más común en la regresión lineal es el **Error Cuadrático Medio (MSE)**, que se define como:

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2 \quad (2.3)$$

Donde:

- m es el número de ejemplos.
- y_i es el valor observado de la característica de salida.
- $\hat{y}_i$ es el valor predicho por el modelo.

El objetivo es encontrar los coeficientes θ que minimicen esta función de coste.

Función de Optimización

Para encontrar los coeficientes θ que minimizan la función de coste, se utiliza un proceso de optimización. En regresión lineal, la optimización se realiza comúnmente mediante el método de **mínimos cuadrados ordinarios (OLS)**, que busca los coeficientes que minimicen la suma de los cuadrados de los residuos.

Matemáticamente, esto implica resolver el siguiente problema de minimización:

$$\min_{\theta_0, \theta_1, \dots, \theta_n} \sum_{i=1}^n (y_i - (\theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_n x_{in}))^2 \quad (2.4)$$

Esto puede hacerse de manera analítica o utilizando algoritmos numéricos. Una forma analítica es derivar la función de coste respecto a cada coeficiente θ , establecer estas derivadas igual a cero, y resolver el sistema resultante de ecuaciones lineales. Otra forma es utilizar algoritmos de optimización iterativa, como el descenso de gradiente, especialmente cuando se trata de grandes conjuntos de datos o modelos más complejos.

En el descenso de gradiente, los coeficientes se actualizan iterativamente en la dirección del gradiente negativo de la función de coste:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} \text{MSE} \quad (2.5)$$

Donde α es la tasa de aprendizaje, un parámetro que determina el tamaño de los pasos de actualización.

Construcción del Modelo

1. **Recopilación de Datos:** Recopilar conjunto de datos que incluyan las características de entre y la característica de salida.
2. **Análisis Exploratorio:** Analizar y visualizar (si es posible) los datos para comprender las relaciones y detectar posibles anomalías.
3. **División de Datos:** Dividir los datos en conjuntos de entrenamiento y prueba.
4. **Ajuste del Modelo:** Utilizar el conjunto de entrenamiento para ajustar el modelo de regresión, calculando los coeficientes θ que minimizan la función de coste.
5. **Evaluación del Modelo:** Evaluar el rendimiento del modelo utilizando el conjunto de prueba y métricas de evaluación como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2).

2.8.2. Regresión Logística

La regresión logística es una técnica ampliamente utilizada en estadística y aprendizaje automático para la clasificación de datos. A continuación, se ofrece una explicación

2. MARCO TEÓRICO

clara y accesible de la técnica, sus características, aplicaciones, fundamentos matemáticos, construcción del modelo, función de coste, optimización y conclusiones.

La regresión logística se utiliza para modelar y predecir resultados binarios (es decir, aquellos que tienen dos posibles valores, como “sí.” “no”, “verdadero.” “falso”). A diferencia de la regresión lineal, que predice valores continuos, la regresión logística predice probabilidades de que un evento ocurra [Norman \(2019\)](#).

La regresión logística tiene varias características importantes:

- **Clasificación Binaria:** Se usa principalmente para problemas donde el resultado puede estar en una de dos categorías.
- **Probabilística:** Predice la probabilidad de que un resultado pertenezca a una categoría específica.
- **No Lineal:** Utiliza la función logística para manejar problemas de clasificación, lo que permite modelar relaciones no lineales.

La técnica de regresión logística se utiliza en diversos casos, como el diagnóstico médico para predecir la presencia o ausencia de una enfermedad, y el análisis de crédito para determinar la probabilidad de que un cliente incumpla un préstamo. En marketing, se emplea para predecir si un cliente comprará un producto o no, mientras que en la detección de fraude, ayuda a identificar transacciones fraudulentas. Además, es utilizada en la clasificación de correos electrónicos para distinguir entre spam y no spam.

La regresión logística utiliza la función logística (sigmoide) para modelar la probabilidad de un evento. La ecuación general es:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n)}} \quad (2.6)$$

Donde:

- $P(y = 1|x)$ es la probabilidad de que el resultado sea 1 dado el vector de características x .
- θ_0 es la intersección.
- $\theta_1, \theta_2, \dots, \theta_n$ son los coeficientes de las características de entrada.

Función de Coste (Función Error)

La función de coste para la regresión logística es la **log-loss (pérdida logística)**, que mide la precisión del modelo en términos de la probabilidad predicha y el resultado real:

$$\text{Log-loss} = -\frac{1}{m} \sum_{i=1}^m [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (2.7)$$

Donde:

- m es el número de ejemplos.
- y_i es el valor real (0 o 1).
- $\hat{y}_i$ es la probabilidad predicha de que y_i sea 1.

El objetivo es encontrar los valores de θ que minimicen esta función de coste.

Función de Optimización

Para encontrar los mejores valores de θ , se utiliza un proceso de optimización, típicamente mediante el método de **máxima verosimilitud**, que maximiza la probabilidad de observar los resultados dados los valores de las variables independientes.

Otra técnica común es el **descenso de gradiente**, que ajusta iterativamente los valores de θ para minimizar la log-loss:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} \text{Log-loss} \quad (2.8)$$

Donde α es la tasa de aprendizaje, que determina el tamaño de los pasos de ajuste.

Construcción del Modelo

1. **Recopilación de Datos:** Recopilar datos históricos que incluyan la variable dependiente (resultado binario) y las variables independientes.
2. **Análisis Exploratorio:** Analizar y visualizar los datos para comprender las relaciones y detectar posibles problemas.
3. **División de Datos:** Dividir los datos en conjuntos de entrenamiento y prueba.
4. **Ajuste del Modelo:** Utilizar el conjunto de entrenamiento para ajustar el modelo de regresión logística, encontrando los coeficientes β que maximizan la función de verosimilitud.
5. **Evaluación del Modelo:** Evaluar el rendimiento del modelo utilizando el conjunto de prueba y métricas de evaluación como la precisión, la sensibilidad y la especificidad.

2.8.3. K-means

K-means es una técnica ampliamente utilizada en aprendizaje no supervisado dentro de la ciencia de datos para la agrupación de datos en clústeres. A continuación, se ofrece una explicación clara y accesible de la técnica, sus características, aplicaciones, fundamentos matemáticos, construcción del modelo, función de coste, optimización y conclusiones.

K-means busca dividir un conjunto de datos en k clústeres, donde los elementos dentro de cada clúster son similares entre sí, y diferentes de los elementos en otros

2. MARCO TEÓRICO

clústeres. A diferencia de los métodos supervisados, K-means no requiere etiquetas previas y su objetivo es descubrir patrones subyacentes en los datos.

El algoritmo K-means tiene varias características importantes:

- **Simplicidad:** Fácil de implementar y entender.
- **Eficiencia:** Escalable para grandes conjuntos de datos.
- **Iterativo:** Funciona mediante una optimización iterativa, que aunque garantiza la convergencia, no siempre encuentra la solución óptima global.

K-means se utiliza en diversos escenarios, como segmentación de clientes, compresión de imágenes, agrupación de documentos y detección de anomalías. Es una herramienta útil para encontrar patrones en datos no etiquetados, y sus aplicaciones van desde el marketing hasta la seguridad informática [Ares et al. \(2010\)](#).

Matemáticamente, K-means minimiza la suma de las distancias cuadradas entre los puntos de datos y los centroides de sus respectivos clústeres:

$$\sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (2.9)$$

Donde:

- C_i es el conjunto de puntos en el clúster i .
- μ_i es el centroide del clúster i .
- $\|x - \mu_i\|^2$ es la distancia euclidiana entre el punto x y el centroide μ_i .

Función de Coste (Función Error)

La función de coste en K-means es la **inercia**, que mide la suma de las distancias cuadradas de los puntos al centroide de su clúster:

$$\text{Inercia} = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (2.10)$$

El objetivo es minimizar esta inercia, lo que significa agrupar los puntos de manera que estén lo más cerca posible de sus respectivos centroides.

Función de Optimización

El proceso de optimización en K-means se basa en un algoritmo iterativo que reasigna puntos a clústeres y actualiza los centroides hasta que la inercia no pueda reducirse significativamente o hasta que se alcance un número máximo de iteraciones. Aunque el algoritmo garantiza convergencia, puede llegar a un mínimo local en lugar de la solución óptima global.

Construcción del Modelo

1. **Recopilación de Datos:** Obtener y limpiar los datos a ser agrupados.
2. **Preprocesamiento de Datos:** Normalizar o estandarizar los datos si es necesario.
3. **Selección del Número de Clústeres:** Determinar el valor óptimo de k mediante métodos como el método del codo (*elbow method*).
4. **Inicialización:** Seleccionar k centroides iniciales aleatoriamente.
5. **Asignación y Actualización:** Repetir el proceso de asignación de puntos a clústeres y actualización de centroides hasta la convergencia.
6. **Evaluación del Modelo:** Evaluar la calidad del agrupamiento mediante la inercia o usando métricas como la homogeneidad.

Si bien en esta sección se han explorado los algoritmos de regresión lineal, regresión logística y K-means, existen muchas otras técnicas dentro del aprendizaje supervisado y no supervisado que abordan una gran variedad de problemas y adaptan sus enfoques a diferentes tipos de datos y necesidades. Además de estos métodos de modelado, el campo del aprendizaje automático incluye técnicas adicionales, como algoritmos de limpieza y preprocesamiento de datos, que son esenciales para mejorar la calidad de los datos antes de generar un modelo. Métodos como los árboles de decisión, redes neuronales, y técnicas de reducción de dimensionalidad son ejemplos de herramientas que permiten optimizar el análisis y la interpretación de datos complejos. Este amplio repertorio, en constante expansión, no solo facilita la creación de modelos más precisos, sino que también mejora cada etapa del flujo de trabajo, enriqueciendo las posibilidades y aplicaciones del aprendizaje automático en diversas áreas.

2.9. El proceso de tratamiento de datos

2.9.1. Limpieza de datos

La ciencia de datos es un pilar fundamental de la era de la información, ya que transforma los datos sin procesar en conocimiento significativo. En este proceso, la calidad de los datos es la clave, ya que el procesamiento de un buen conjunto de datos proporciona descubrimientos valiosos y la posibilidad de predicciones precisas. En el centro de esta cualidad se encuentra la limpieza de datos, este proceso fundamental de la ciencia de datos busca preparar conjuntos de datos para su refinamiento y análisis [Medina and Galván \(2007\)](#).

Pero, ¿a qué se refiere exactamente cuando se habla de calidad de los datos? La calidad de los datos incluye la integridad, coherencia, precisión y la información actualizada del conjunto de datos. Este es el proceso de garantizar que los datos sean

2. MARCO TEÓRICO

confiables y estén libres de errores, repeticiones o valores atípicos que puedan garantizar las conclusiones sin ningún tipo de sesgo. La calidad de los datos asegura que las inferencias y decisiones resultantes estén fundamentadas en bases estadísticas sólidas, proporcionando representaciones confiables de los fenómenos analizados [Medina and Galván \(2007\)](#).

El núcleo de la ciencia de datos radica en la manipulación de datos a través de estructuras como Dataframe. Un DataFrame bien construido es la columna vertebral sobre la que se lleva a cabo cada análisis significativo. Sin embargo, este proceso de construcción es solo el primer paso; La verdadera magia radica en la capacidad de garantizar que los datos que contiene sean confiables, consistentes y libres de anomalías.

Por lo tanto, se vuelve imperativo categorizar cada dato o valor que pueda comprometer la calidad de los datos. Esta clasificación se convierte en un paso esencial para evitar que estos elementos indeseados obstaculicen la correcta aplicación de los modelos de tratamiento de datos. Identificar y abordar proactivamente estos factores contribuye significativamente a fortalecer la robustez y confiabilidad de nuestros conjuntos de datos, asegurando así que el proceso de análisis y modelado se realice sobre una base sólida y precisa. Por ello es necesario el conocer y clasificar los tipos de datos que afectan la calidad del conjunto de datos

- **Datos faltantes (Features):** Los datos faltantes son aquellos elementos que no están presentes en ciertas observaciones o registros de un conjunto de datos. Estos vacíos pueden surgir debido a diversas razones, como errores de recopilación, fallos en la transmisión de datos o simplemente la falta de respuesta en una encuesta. La presencia de datos faltantes puede afectar la capacidad de análisis y modelado. La ausencia de datos en un conjunto no siempre implica la falta completa de una observación. En ciertos contextos, como encuestas u otros escenarios similares, la carencia de información se sustituye ocasionalmente por valores específicos, ya sea números, letras o símbolos, a fin de representar la omisión de datos. Este enfoque, aunque común, introduce desafíos particulares al tratar con datos faltantes, ya que estos marcadores requieren una gestión especializada para garantizar que no distorsionen los resultados del análisis [Medina and Galván \(2007\)](#).
- **Valores atípicos (Outliers):** Los valores atípicos son observaciones inusuales que se desvían significativamente del patrón general del conjunto de datos. Estos valores pueden afectar los resultados del análisis porque alteran las medidas de tendencia central y la dispersión. los valores atípicos pueden surgir de errores humanos, mal funcionamiento de sensores IoT(Internet de las cosas)¹, o errores en programas, como la mezcla de medidas, entre sistemas métricos, en general valores que no tiene sentido con el parámetro al que está asignado [Medina and Galván \(2007\)](#).

¹Internet de las cosas son dispositivos electrónicos que recopilan datos del entorno físico o lógico y los transmite a través de la red para su procesamiento, análisis y toma de decisiones

- **Datos incorrectos e inconsistentes:** Las inconsistencias se refieren a discrepancias o contradicciones en los datos que pueden surgir debido a errores de entrada, duplicación de registros, o incluso cambios en las convenciones de codificación. Estas inconsistencias pueden manifestarse en la forma de información contradictoria entre variables o registros. Resolver las inconsistencias es esencial para mantener la coherencia y la fiabilidad de los datos, garantizando que los análisis y modelos se construyan sobre una base precisa y coherente [Medina and Galván \(2007\)](#).

Es esencial utilizar métodos efectivos para abordar y resolver los problemas relacionados con la calidad de los datos, que incluyen datos faltantes, valores atípicos e inconsistencias. Para garantizar que el análisis y la generación del modelo se realicen sobre una base sólida, es esencial la detección precisa y el manejo adecuado de estos elementos indeseados. En la siguiente sección, analizaremos una variedad de enfoques para identificar y abordar estos problemas de datos. Para optimizar la calidad y la confiabilidad de nuestros conjuntos de datos para un análisis más robusto y preciso, estas técnicas incluirán métodos para imputar datos faltantes, técnicas para gestionar valores atípicos y métodos para corregir inconsistencias.

Para garantizar que todas las características tengan escalas comparables y que los algoritmos de aprendizaje automático funcionen de manera óptima, la normalización y la estandarización son pasos cruciales en el preprocesamiento de datos. Sin embargo, aunque estos métodos abordan problemas de escala de datos, no resuelven problemas de datos faltantes o valores atípicos directamente. Es esencial utilizar técnicas de imputación específicas para mantener la integridad y la calidad de los datos. Estas técnicas permiten manejar valores faltantes y tratar con datos inusuales. Por lo tanto, examinaremos una variedad de métodos de imputación, como la imputación media, mediana y moda, así como técnicas más avanzadas como k-vecinos más cercanos (kNN) y regresión lineal. Cada una de estas estrategias ofrece una forma diferente de abordar las dificultades.

2.9.2. Normalización

Es un proceso común en el procesamiento de datos. Su objetivo es llevar todas las características del conjunto de datos al mismo rango o escala.

Esto es importante dado que los algoritmos funcionan mejor cuando todas las características tienen magnitudes similares.

Normalización mín-máx

Este método consiste en escalar cada parámetro de manera que los valores estén en el rango específico, comúnmente de 0 a 1.

$$X_{normalizado} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2.11)$$

2. MARCO TEÓRICO

donde

$X_{normalizado}$: es el valor normalizado del parámetro.

X : es el valor original del parámetro.

X_{min} : es el valor mínimo del parámetro.

X_{max} : es el valor máximo del parámetro.

Después de este proceso, todos los valores estarán normalizados en el rango 0 a 1.

La normalización puede fallar cuando hay información sesgada o existen valores atípicos (Outliers) Bobadilla (2021), al analizar la ecuación 2.11, se puede observar que para determinar el valor normalizado es necesario del valor máximo y valor mínimo, si estos valores están muy alejados de los valores pueden llegar a producir una desviación de los valores luego del proceso de normalizar.

Normalización z-score o Estandarización

La estandarización centra los datos en su media y lo escala de acuerdo con el valor de la desviación estándar.

$$X_{estandarizado} = \frac{X - \bar{X}}{\sigma} \quad (2.12)$$

donde

$X_{estandarizado}$: es el valor estandarizado del parámetro.

X : es el valor original de parámetro.

$\bar{X}$: es la media del parámetro.

σ : es la desviación estándar del parámetro.

Este enfoque es menos sensible a los valores atípicos, ya que la estandarización se basa en la media y desviación estándar que son menos influenciados por valores extremos Bobadilla (2021). En cuanto a valores vacíos en los parámetros, hay varios enfoques para manejarlos:

- **Eliminar las muestras que contienen datos faltantes** Esto se hace cuando se dispone de suficientes datos y la pérdida de algunas muestras no afecta significativamente el modelo Bobadilla (2021).
- **Insertar valores correctos cuando sea posible** Es posible insertar valores correctos si están disponibles en otras fuentes Bobadilla (2021).
- **Utilización de imputación de datos** Una técnica común es reemplazar los valores vacíos con alguna predicción. Esto podría ser simplemente la media de los parámetros, o métodos más avanzados. También se pueden agregar características adicionales para indicar si un valor está vacío o no, lo que permite al modelo aprender patrones asociados con valores faltantes Bobadilla (2021).

2.9.3. Técnicas de Imputación

La normalización o estandarización de un conjunto de datos es fundamental para facilitar la comparación entre variables cuando estas tienen escalas diferentes. Sin embargo, es importante señalar que este proceso no aborda directamente valores faltantes o atípicos en los datos. Para tratar con valores faltantes, se recurre a técnicas específicas de imputación, mientras que para manejar valores atípicos, se aplican enfoques como la identificación y eliminación de dichos valores o el uso de métodos robustos en el análisis estadístico. Ahora, se exploran algunas técnicas de imputación comunes para abordar valores faltantes y garantizar la integridad de nuestros análisis.

Imputación media-mediana-moda

La imputación con media, mediana y moda son estrategias clave para manejar valores faltantes, mejorando la integridad y consistencia de los datos en procesos de análisis y modelado.

- **Imputación con la media** Reemplaza los valores faltantes con el promedio aritmético de los datos disponibles. Es útil para variables numéricas, especialmente si los datos siguen una distribución normal, aunque puede verse afectada por valores atípicos, se debe de tener las siguientes consideraciones:
 - Aplicable a variables cuantitativas en una escala ordinal o superior.
 - Es más efectiva cuando no existen valores extremos que distorsionen el promedio.
 - Los valores faltantes deben ser aleatorios y no contener información relevante sobre el análisis.
 - Requiere un tamaño de muestra suficiente para que la media sea representativa.
- **Imputación con la mediana** Sustituye valores faltantes con la mediana, el valor central de un conjunto de datos ordenados. Este método es resistente a valores atípicos, lo que lo hace ideal en situaciones de datos sesgados o con valores extremos, se debe de tener las siguientes consideraciones:
 - Apropiaada para variables cuantitativas (ordinales o superiores).
 - Preserva la tendencia central en distribuciones no normales.
 - Los valores faltantes deben ser aleatorios y no relacionados con patrones específicos.
 - Es importante contar con un tamaño de muestra suficiente para calcular una mediana confiable.
- **Imputación con la moda** Reemplaza valores faltantes en variables categóricas con la categoría más frecuente. Es una técnica simple y efectiva para mantener la consistencia en datos cualitativos, se debe de tener las siguientes consideraciones:

- Aplicable únicamente a variables categóricas o nominales.
- Es necesario que exista una categoría dominante en la distribución.
- Los valores faltantes deben ser aleatorios y no contener información sistemática.

Imputación K-vecinos más cercanos o KNN

El método kNN o k Vecinos más cercanos es un algoritmo simple, pero eficaz que se utiliza en clasificación y regresión en el campo del aprendizaje automático. El concepto básico de kNN es muy simple. Para predecir o clasificar un nuevo punto de datos, el algoritmo determina los puntos de datos nulos existentes y encuentra k puntos en el espacio de la imagen que están más cerca del nuevo punto. Una medida de “cercanía” se basa en una medida de distancia, normalmente la distancia euclidiana en el espacio multidimensional. Cuanto más cerca esté dato faltante del dato al dato existente, mayor será la influencia en el dato resultante, ya por lo general se utiliza el promedio ponderado para determinar el nuevo valor, aunque también se puede utilizar la media. La elección entre la media ponderada y la media simple depende de la naturaleza de los datos y del supuesto de que cada vecino contribuye igualmente a determinar los valores faltantes, Sin embargo, hay algunos aspectos y consideraciones importantes a tener en cuenta al utilizar este método:

- Datos Numéricos: es más apropiado para datos numéricos, ya que implica calcular distancias entre puntos de datos. Aplicar esto a variables de categoría no es natural.
- Datos escalados: Si las variables tienen escalas muy diferentes, es posible que se deba considerar la normalización o estandarización antes de aplicar kNN.
- Distribución de Datos: Aunque el método no asume una distribución específica, puede funcionar mejor cuando los datos tienen una distribución aproximadamente normal.
- Manejo de Valores Atípicos: El método kNN puede ser sensible a los valores atípicos, ya que las distancias se calculan directamente. Es necesario procesar previamente los datos para la eliminación de estos valores.
- Elección del Parámetro k: El valor k es indispensable para obtener una buena predicción, un valor pequeño de k puede afectar la predicción debido a ruido, o, por el contrario, un valor k muy grande puede suavizar muchos los datos.
- Proporción Pequeña de Datos Faltantes: El método kNN funciona mejor cuando la proporción de datos faltantes es pequeña en comparación con el tamaño total del conjunto de datos.
- Distancia Significativa: La distancia entre puntos de datos debe tener significado en el contexto del problema. Es importante elegir características relevantes para el cálculo de distancias.

- **Computacionalmente Factible:** El método puede volverse computacionalmente intensivo a medida que aumenta el tamaño del conjunto de datos y el número de vecinos. En conjuntos de datos grandes, se pueden considerar enfoques aproximados.
- **Independencia de las Observaciones:** Se asume que los valores faltantes son aleatorios y no siguen patrones específicos relacionados con otras variables. Si existe alguna estructura de dependencia, pueden ser necesarios métodos más avanzados.
- **Validación Cruzada:** Es recomendable realizar validación cruzada u otras técnicas de evaluación para determinar la eficacia del método en el conjunto de datos específico.

Imputación mediante regresión lineal

La imputación mediante regresión lineal consiste en estimar los valores faltantes en un conjunto de datos ajustando una línea de regresión a los datos disponibles. En este método, se identifican los parámetros con valores faltantes y se seleccionan otros parámetros relacionados para predecir las faltantes. Luego, se realiza una partición del conjunto de datos, de la siguiente forma: uno completo, sin valores faltantes del parámetro dependiente y otro incompleto con valores faltantes. Se ajusta un modelo de regresión lineal con el conjunto completo y se utiliza para predecir los valores faltantes en el conjunto incompleto. Este proceso se repite para cada parámetro con valores faltantes. Sin embargo, es esencial considerar las suposiciones de la regresión lineal, para asegurar resultados precisos y válidos:

- **Linealidad:** El método asume que la relación entre las variables seleccionadas es lineal. Relaciones no lineales pueden resultar en estimaciones imprecisas.
- **Independencia de Errores:** La regresión lineal asume independencia de los errores; patrones sistemáticos pueden afectar el modelo.
- **Normalidad de Errores:** La normalidad de los errores es supuesta; desviaciones de la normalidad pueden afectar la validez de las inferencias.
- **Homocedasticidad:** Se asume igualdad de varianzas; heterocedasticidad puede afectar la precisión de los intervalos de predicción.
- **Multicolinealidad:** Alta correlación entre variables independientes puede afectar la estabilidad del modelo y debe ser evaluada.
- **Validación del Modelo:** La calidad del modelo debe evaluarse con métricas de rendimiento y validación cruzada para asegurar su generalización.
- **Conjunto de Datos Grande:** La regresión lineal puede ser más efectiva en conjuntos de datos grandes donde la variabilidad es menor.
- **Sensibilidad a Outliers:** Valores atípicos pueden tener un impacto significativo en la estimación de los coeficientes y deben ser considerados.

2.10. Reducción de Dimensionalidad

La reducción de dimensionalidad es una técnica fundamental en el ámbito de la ciencia de datos y el aprendizaje automático. Su principal objetivo es simplificar el análisis de grandes conjuntos de datos al reducir el número de variables o características, manteniendo al mismo tiempo la mayor cantidad posible de información relevante. Esta técnica es especialmente útil cuando los datos contienen una gran cantidad de variables, lo que puede generar problemas de sobreajuste, aumentar el tiempo de procesamiento y dificultar la visualización y análisis de patrones subyacentes [Jiménez et al. \(2010\)](#).

En este contexto, es importante resaltar dos enfoques principales para la reducción de dimensionalidad: técnicas de selección de características y técnicas de extracción de características.

2.10.1. Selección de Características

La selección de características, también conocida como *feature selection*, es un paso crucial en la construcción de modelos de aprendizaje automático. Consiste en elegir un subconjunto de las variables disponibles en el conjunto de datos, eliminando aquellas que no aportan valor significativo al modelo. Este proceso permite reducir la dimensionalidad del problema, mejorar la eficiencia computacional y, en muchos casos, aumentar el rendimiento del modelo. A continuación, se presentan los conceptos clave y métodos utilizados en la selección de características.

La **selección de características** es un proceso que tiene como objetivo identificar y mantener solo aquellas variables que son más relevantes para el modelo de predicción. Esto es especialmente útil cuando se trabaja con grandes volúmenes de datos, ya que no todas las variables necesariamente aportan información valiosa. Al reducir la cantidad de características, el algoritmo puede aprender de manera más eficiente, lo que también puede llevar a una mejora en el rendimiento del modelo [Vega Rojas and Hernández Martinez \(2023\)](#).

Criterios para Evaluar la Relevancia de las Características

Existen varios criterios que permiten determinar si una característica es o no relevante:

- **Impacto en el rendimiento del modelo:** Si la eliminación de una característica provoca una caída significativa en el rendimiento del modelo, esta característica se considera relevante. Por ejemplo, en un modelo de clasificación, si el puntaje F1 disminuye al eliminar una variable, esta es considerada importante.
- **Correlación con la variable objetivo:** Las características que tienen una alta correlación con la variable objetivo suelen ser más útiles para la predicción. Variables con baja o nula correlación tienden a ser eliminadas, ya que no aportan información nueva.

Métodos de Selección de Características

Varios métodos son comúnmente utilizados para realizar la selección de características. Entre los más importantes destacan:

- **Eliminación de características con muchos valores faltantes:** Variables con una gran cantidad de valores faltantes pueden ser descartadas, ya que suelen agregar ruido al modelo sin proporcionar información útil.
- **Eliminación de características con valores constantes:** Variables que tienen el mismo valor para todas las observaciones no aportan información útil al modelo, por lo que se pueden eliminar de inmediato.
- **Eliminación de características correlacionadas:** Si dos variables están altamente correlacionadas entre sí, es preferible eliminar una de ellas para evitar la redundancia de información.
- **Algoritmos basados en la importancia de las características:** Algoritmos como los bosques aleatorios pueden evaluar la importancia de cada variable y eliminar aquellas con baja importancia.

Ventajas de la Selección de Características

La correcta implementación de un proceso de selección de características aporta varias ventajas:

- **Rendimiento más rápido:** Al reducir la cantidad de variables, el modelo se entrena más rápidamente y utiliza menos recursos computacionales.
- **Mayor precisión:** Al eliminar características irrelevantes o ruidosas, se puede mejorar la precisión del modelo, permitiendo que este se enfoque en la información más útil.
- **Facilidad de interpretación:** Un modelo con menos características es más fácil de interpretar y analizar, lo que resulta especialmente valioso en contextos donde la interpretabilidad es crucial.

Este proceso puede realizarse manualmente o utilizando algoritmos automáticos, como los mencionados anteriormente, para optimizar los resultados.

2.10.2. Extracción de Características

La *reducción de dimensionalidad* es una técnica fundamental en el aprendizaje automático que busca simplificar los datos disminuyendo la cantidad de variables o características presentes en un conjunto de datos, sin perder información esencial. Una forma de lograr esta reducción es a través de la *extracción de características*, en la cual

se transforman los datos originales en un conjunto de características de menor dimensión, pero que contienen la mayor parte de la información relevante.

Diferencias entre Selección y Extracción de Características

Existen dos enfoques principales para reducir la dimensionalidad: la *selección de características* y la *extracción de características*. En la selección de características, se eliminan algunas de las columnas o atributos que se consideran irrelevantes o redundantes para el problema a resolver, manteniendo las demás características intactas. Por ejemplo, mediante el uso de algoritmos como *Random Forest*, es posible obtener una medida de relevancia de cada característica, lo que permite elegir las más importantes.

Por otro lado, la extracción de características implica no solo reducir el número de atributos, sino también transformar las características originales en nuevas características que resumen la información más relevante del conjunto de datos. Al aplicar este tipo de técnicas, las nuevas características generadas no corresponden directamente con las características originales, sino que son combinaciones de las mismas, optimizadas para representar la mayor cantidad de información posible.

Algoritmos de Extracción de Características

Uno de los enfoques más comunes para la extracción de características es el uso del *Análisis de Componentes Principales* (PCA, por sus siglas en inglés). Este algoritmo es considerado una técnica lineal de reducción de dimensionalidad y trabaja proyectando los datos a un espacio de menor dimensión, manteniendo la mayor varianza posible de los datos. Para lograr esto, el PCA emplea el método matemático conocido como *Descomposición en Valores Singulares* (SVD, por sus siglas en inglés), el cual descompone una matriz en tres matrices más simples. A continuación, se exploran ambos algoritmos en mayor detalle.

■ Análisis de Componentes Principales (PCA)

El *PCA* es un algoritmo que permite proyectar un conjunto de datos de alta dimensionalidad en un espacio de menor dimensión, manteniendo la mayor varianza posible en los datos proyectados. El proceso que sigue el PCA es el siguiente:

- **Selección de ejes:** El PCA selecciona el eje que minimiza la distancia media de los datos proyectados, utilizando el error cuadrático entre el conjunto de datos original y su proyección. Este eje es conocido como el *componente principal*.
- **Proyección:** Cada punto o muestra del conjunto de datos original es proyectado sobre el eje seleccionado, creando un nuevo conjunto de datos en el que las características originales se han transformado en una nueva dimensión, definida por el componente principal.

El objetivo de este proceso es reducir la cantidad de dimensiones manteniendo la mayor parte de la información original. Por ejemplo, si un conjunto de datos tiene

dos dimensiones X_1 y X_2 , el PCA transformará estos datos a un nuevo conjunto con una única dimensión Z_1 , en la cual los datos han sido proyectados. Esto permite trabajar con menos variables, sin perder una gran cantidad de información relevante Ringnér (2008).

■ **Descomposición en Valores Singulares (SVD)**

El algoritmo *SVD* es una técnica de descomposición de matrices que descompone una matriz en tres matrices: U , E y V^T . Este proceso es utilizado por el PCA para encontrar los *componentes principales* que definen las direcciones de mayor varianza en el conjunto de datos.

Matemáticamente, SVD descompone la matriz de datos X en tres matrices:

$$X = UEV^T$$

- U contiene los *componentes principales* del conjunto de datos.
- E es una matriz diagonal que contiene los valores singulares.
- V^T es la transpuesta de la matriz V , que define la orientación de los datos en el espacio.

El PCA utiliza las primeras columnas de la matriz U para proyectar el conjunto de datos original en un espacio de menor dimensión. Así, el número de columnas seleccionadas de U determina la cantidad de dimensiones que tendrá el conjunto de datos resultante.

Dado un conjunto de datos en dos dimensiones, con características X_1 y X_2 . El objetivo del PCA es reducir este conjunto de datos a una única dimensión, que llamaremos Z_1 . Para hacerlo, el algoritmo selecciona el eje que maximiza la varianza, y proyecta los puntos del conjunto de datos en este nuevo eje. Los nuevos datos proyectados tendrán solo una dimensión, Z_1 , que retendrá la mayor cantidad posible de la varianza original de los datos.

Finalmente, el algoritmo SVD es utilizado en PCA para descomponer la matriz original X , permitiendo seleccionar los componentes principales, que serán utilizados para reducir la dimensionalidad del conjunto de datos. De este modo, el PCA transforma los datos en un conjunto más compacto y manejable, preservando la información más relevante Abdi (2007).

La reducción de dimensionalidad no solo mejora la eficiencia computacional y la interpretabilidad de los modelos, sino que también ayuda a prevenir problemas como el sobreajuste, asegurando que los modelos capturen las relaciones fundamentales dentro de los datos sin verse afectados por el ruido o la complejidad innecesaria. Sin embargo, una vez que se ha optimizado el conjunto de características, el siguiente paso crucial es evaluar el rendimiento del modelo. Para ello, es esencial seleccionar las métricas adecuadas que permitan medir con precisión la calidad de las predicciones y el ajuste del modelo. En la siguiente sección se introducen las principales métricas utilizadas para

evaluar el desempeño de los modelos, recordando que en la próxima unidad se detallará el proceso de generación de los mismos y la selección de la métrica más apropiada en cada caso.

El éxito de cualquier modelo depende de la calidad del conjunto de datos. Las técnicas de normalización e imputación no solo garantizan la integridad y consistencia del dataframe, sino que también optimizan el rendimiento del modelo al asegurar que los datos sean aptos para el análisis. Sin embargo, en muchos casos, aun con un conjunto de datos limpio, el número de características disponibles puede ser abrumador, lo que puede llevar a modelos complejos, difíciles de interpretar y propensos a sobreajuste. Es aquí donde las técnicas de reducción de dimensionalidad juegan un papel crucial, permitiendo simplificar el espacio de características sin perder información relevante. A continuación, se explorarán los métodos más utilizados para reducir la dimensionalidad y mejorar la eficiencia de los modelos.

2.11. Medición de resultados (métricas)

La medición de la calidad de los modelos en ciencia de datos y aprendizaje automático es esencial para evaluar su rendimiento y tomar decisiones informadas. Con frecuencia, no es posible distinguir visualmente la superioridad de un resultado sobre otro, por lo que es necesario tomar medidas objetivas para comprender la precisión de los resultados. Es importante destacar que no existe una sola medida de calidad que funcione para todos los escenarios; la elección de la métrica correcta depende del tipo de problema que se aborda. Por ejemplo, hay métricas que son particularmente adecuadas para evaluar modelos de clasificación, mientras que otras son más adecuadas para resolver problemas de regresión.

La regla fundamental de la metodología de validación de modelos es clara y fundamental: no medir la calidad del modelo utilizando el mismo conjunto de datos que se utilizó para su entrenamiento. Este principio se deriva de la necesidad de evitar el sobreajuste, que es cuando el modelo se adapta demasiado a las muestras de entrenamiento y pierde su capacidad para generalizar a datos nuevos. Para reducir este riesgo, es fundamental separar el conjunto de datos en varios conjuntos: uno para el entrenamiento del modelo, otro para la validación y, finalmente, un conjunto independiente para evaluar el desempeño de los datos completamente nuevos.

Esta técnica, también conocida como validación cruzada o particionamiento de datos, ofrece una evaluación del modelo más precisa y realista, asegurando que su capacidad predictiva no esté sesgada por la familiaridad con las muestras utilizadas durante el entrenamiento.

Para que el entrenamiento y la evaluación del modelo sean exitosos, es necesario particionar correctamente el conjunto de datos. Se divide generalmente en tres grupos: entrenamiento, validación y prueba. El modelo se instruye con el conjunto de entrenamiento, que suele representar el 80 % [Bobadilla \(2021\)](#) del total de muestras. Durante la fase de ajuste fino de los hiperparámetros, la validación es importante porque permite

mejorar y optimizar el modelo sin comprometer su capacidad de generalización.

El conjunto de prueba(test), que está compuesto por muestras que no se han utilizado previamente en el entrenamiento ni en el ajuste de parámetros, es esencial para la evaluación final. Este conjunto permite medir la calidad del modelo con datos completamente nuevos, asegurando que el modelo no solo haya memorizado las muestras de entrenamiento, sino que también pueda generalizar a datos nuevos.

Para garantizar resultados confiables, los tres conjuntos de datos deben tener distribuciones de datos similares, datos representativos de la población y, lo más importante, no debe haber superposiciones entre ellos. Por decirlo de otra manera, cada muestra debe ser exclusiva de uno de los conjuntos. Esta distinción garantiza que la evaluación del modelo se lleve a cabo de manera objetiva y sin dañar la información utilizada en el proceso de entrenamiento o validación.

El conjunto de prueba pone a prueba el verdadero poder del modelo después de que ha sido entrenado con el conjunto de entrenamiento y perfeccionado con el conjunto de validación. En qué medida las predicciones coinciden con los valores objetivo no vistos previamente, se mide la calidad del modelo. Los indicadores de calidad del modelo aumentarán cuanto más se asemejen a los valores reales, validando su capacidad de generalización y predicción en situaciones del mundo real.

2.11.1. Métricas para problemas de regresión

En los problemas de regresión, el objetivo es predecir valores continuos. Las métricas de evaluación se centran en medir la discrepancia entre los valores predichos y los valores reales.

Definición Error Cuadrático Medio (MSE): El Error Cuadrático Medio mide el promedio de los cuadrados de las diferencias entre los valores predichos ($\hat{y}_i$) y los valores reales (y_i).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

El MSE es especialmente útil en contextos donde los errores grandes tienen un impacto significativo y se desea penalizarlos más severamente. Por ejemplo, en problemas donde la precisión es crucial y los errores grandes pueden llevar a consecuencias graves, como en la predicción de costos de seguros o en la estimación de riesgos financieros. La sensibilidad a los outliers puede ser beneficiosa si esos errores grandes representan situaciones inusuales pero críticas.

Definición Error Absoluto Medio (MAE): El Error Absoluto Medio mide la magnitud promedio de los errores entre los valores predichos ($\hat{y}_i$) y los valores reales (y_i) sin considerar la dirección (positivo o negativo).

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

2. MARCO TEÓRICO

El MAE es preferible en situaciones donde se requiere una medida del error que sea robusta frente a outliers y que proporcione una interpretación más intuitiva del error medio. Por ejemplo, en aplicaciones de predicción de ventas o en la estimación de tiempos de entrega donde los errores deben ser interpretables y no se desea penalizar excesivamente las desviaciones grandes. Es útil cuando los errores se distribuyen de manera más uniforme y no se desea que los errores grandes influyan desproporcionadamente en la métrica global.

Definición Raíz del Error Cuadrático Medio (RMSE): La Raíz del Error Cuadrático Medio es la raíz cuadrada del MSE.

$$RMSE = \sqrt{MSE}$$

El RMSE es útil en situaciones donde la interpretación del error debe estar en las mismas unidades que la variable de salida, facilitando la comprensión del impacto real de los errores. Esto es beneficioso en aplicaciones como la predicción de temperaturas o precios de bienes, donde una interpretación directa de las unidades es crucial. Además, el RMSE sigue penalizando los errores grandes, lo que es adecuado en casos donde se desea una mayor sensibilidad a grandes desviaciones.

Definición Coeficiente de Determinación (R^2): El Coeficiente de Determinación mide la proporción de la varianza en la variable dependiente que es explicada por el modelo. Un valor de R^2 cercano a 1 indica que el modelo explica bien la variabilidad de los datos.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

El R^2 es valioso en contextos donde se desea evaluar la capacidad explicativa del modelo en términos de variabilidad de los datos. Por ejemplo, en estudios científicos o en análisis de modelos económicos donde se desea entender cuánto del comportamiento observado es explicado por el modelo. Sin embargo, su utilidad puede ser limitada en modelos no lineales o en presencia de heterocedasticidad, donde la varianza de los errores no es constante.

2.11.2. Métricas para problemas de clasificación

La evaluación de modelos de clasificación binaria es crucial para evaluar su capacidad para hacer predicciones precisas y generalizar a datos no vistos. Es posible utilizar una variedad de métricas de calidad para analizar varios aspectos del rendimiento del modelo. En este contexto, se considera la importancia de no evaluar la calidad con el mismo conjunto de datos utilizado para el entrenamiento porque esto podría resultar en un sobreajuste, donde el modelo se adapta demasiado a las muestras de entrenamiento y no generaliza bien a los nuevos datos.

En una clasificación binaria, donde solo hay dos posibles resultados, es factible diseñar medidas de calidad más específicas que se reflejan en una matriz de confusión.

Definición La **Matriz de Confusión** es una herramienta que proporciona una visualización completa de las predicciones del modelo de clasificación. Muestra el conteo de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Estos valores se organizan en una tabla que ayuda a entender cómo se comporta el modelo en términos de aciertos y errores.

Esta matriz clasifica los resultados de la clasificación de la siguiente manera:

Predicción	Predicción Positiva	Predicción Negativa
Real Positivo(Clase 0)	TP (Verdadero Positivo)	FN (Falso Negativo)
Real Negativo(Clase 1)	FP (Falso Positivo)	TN (Verdadero Negativo)

Tabla 2.4: Matriz de Confusión para Clasificación Binaria

TP (Verdadero Positivo): Instancias positivas correctamente clasificadas como positivas.

TN (Verdadero Negativo): Instancias negativas correctamente clasificadas como negativas.

FP (Falso Positivo): Instancias negativas incorrectamente clasificadas como positivas.

FN (Falso Negativo): Instancias positivas incorrectamente clasificadas como negativas.

La matriz de confusión es útil para calcular métricas derivadas como la **precisión**, el **recall** y el **F1-Score**, y proporciona una visión detallada del rendimiento del modelo.

Para ejemplo, suponiendo que el conjunto de datos está compuesto por 100 pacientes, cuyos estudios médicos revelan dos posibles etiquetas: “Presenta Problemas Cardiacos” (Clase 0) y “No Presenta Problemas Cardiacos” (Clase 1). En el proceso de clasificación, se observa que 45 pacientes han sido correctamente predichos como “Presenta Problemas Cardiacos” (Verdaderos Positivos - TP), 35 han sido correctamente predichos como “No Presenta Problemas Cardiacos” (Verdaderos Negativos - TN), 15 han sido incorrectamente clasificados como “Presenta Problemas Cardiacos” (Falsos Positivos - FP), y 5 han sido incorrectamente clasificados como “No Presenta Problemas Cardiacos” (Falsos Negativos - FN).

	Predicción Positiva	Predicción Negativa
Real Positivo	45 (TP)	5 (FN)
Real Negativo	15 (FP)	35 (TN)

Tabla 2.5: Matriz de Confusión para el Ejemplo

La matriz de confusión da una idea de cómo funciona nuestro modelo en la clasificación binaria. Sin embargo, es fundamental considerar una variedad de medidas

de calidad para obtener una evaluación más completa y detallada de su desempeño. Estas métricas revelan más información sobre la capacidad predictiva y la precisión del modelo.

Definición Exactitud (Accuracy): La proporción de predicciones correctas respecto al total de predicciones realizadas por el modelo.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

La exactitud es útil cuando las clases están balanceadas. Sin embargo, puede ser engañosa en presencia de clases desequilibradas, ya que un modelo que predice siempre la clase mayoritaria podría tener una alta exactitud, pero no ser útil en la práctica.

Definición Precisión (Precision): La proporción de instancias positivas correctamente clasificadas entre todas las instancias clasificadas como positivas.

$$Precision = \frac{TP}{TP + FP}$$

La precisión es útil cuando el coste de los falsos positivos es alto.

Definición Recall (Sensibilidad o Tasa de Verdaderos Positivos): La proporción de instancias positivas correctamente clasificadas entre todas las instancias que son realmente positivas.

$$Recall = \frac{TP}{TP + FN}$$

El recall es útil cuando el coste de los falsos negativos es alto.

Definición F1-Score: La media armónica de precisión y recall.

$$F1 = 2 \frac{Precision \cdot Recall}{Precision + Recall}$$

El F1-Score es especialmente útil cuando hay un desequilibrio entre las clases, ya que tiene en cuenta tanto los falsos positivos como los falsos negativos y proporciona un equilibrio entre precisión y recall.

2.11.3. Métricas para problemas de agrupamiento

Evaluar la calidad de un modelo de agrupamiento puede ser más complejo, ya que no siempre hay etiquetas verdaderas disponibles para comparación. Las siguientes métricas son claves en este contexto:

Definición Índice de Silueta: El Índice de Silueta mide qué tan similares son los puntos dentro de un clúster en comparación con otros clústeres. Un valor cercano a 1 indica que los puntos están bien agrupados.

$$Silhouette = \frac{b - a}{\max(a, b)}$$

Donde a es la distancia media entre un punto y los otros puntos en su mismo clúster, y b es la distancia media entre un punto y los puntos del clúster más cercano.

Definición Inercia (Within-cluster sum of squares): La inercia mide la suma de las distancias cuadradas de los puntos dentro de un clúster al centroide del clúster. Es útil para evaluar la cohesión interna del clúster.

$$WCSS = \sum_{i=1}^n \sum_{k=1}^K \|x_i - c_k\|^2$$

Definición Índice de Davies-Bouldin: El Índice de Davies-Bouldin mide la media de las ratios de la suma de dispersión dentro del clúster y la separación entre clústeres. Un valor más bajo indica una mejor separación entre clústeres.

$$DB = \frac{1}{n} \sum_{i=1}^n \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d_{ij}} \right)$$

Definición Adjusted Rand Index (ARI): El ARI mide la similitud entre las etiquetas reales y los clústeres generados, ajustando por coincidencias aleatorias. Un valor cercano a 1 indica un alto grado de concordancia.

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{0,5 \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}$$

Definición Homogeneidad y Completitud: La homogeneidad evalúa si cada clúster contiene solo elementos de una misma clase, mientras que la completitud mide si todos los elementos de una misma clase están contenidos en un único clúster.

$$Homogeneidad = 1 - \frac{H(C|K)}{H(C)}, \quad Completitud = 1 - \frac{H(K|C)}{H(K)}$$

Donde H representa la entropía.

Definición V-measure: La V-measure combina homogeneidad y completitud en una media armónica:

$$V - measure = 2 \cdot \frac{Homogeneidad \cdot Completitud}{Homogeneidad + Completitud}$$

No existe una métrica universalmente adecuada para todos los problemas; la elección de la métrica debe estar alineada con los objetivos del modelo y las características del problema en cuestión. Evaluar un modelo de manera exhaustiva con múltiples métricas permite una comprensión más completa de su rendimiento y facilita la toma de decisiones informadas para la mejora y optimización del modelo.

Capítulo 3

Aplicación de las técnicas seleccionadas

En este capítulo se describe el proceso aplicado en la fase práctica del presente trabajo, la cual se centra en la selección y preparación de dos conjuntos de datos destinados a la generación de modelos a partir de las técnicas de aprendizaje automático seleccionadas para cada uno de los tipos de problemas típicos en la ciencia de datos: regresión lineal, clasificación mediante regresión logística y agrupamiento (clustering) utilizando la técnica de K-Nearest Neighbors (KNN).

Con el fin de mejorar la calidad de los modelos y optimizar los resultados, se realizaron procesos de limpieza y tratamiento de datos, incluyendo técnicas de reducción de dimensionalidad. Estos modelos tienen como objetivo representar de manera aproximada el comportamiento de los conjuntos de datos en casos donde no es posible obtener predicciones exactas mediante fórmulas determinísticas, permitiendo así la generación de estimaciones confiables dentro de los límites de precisión establecidos.

El desarrollo de los modelos y el análisis de datos en este trabajo de grado se llevó a cabo utilizando el lenguaje de programación Python, ampliamente reconocido en la comunidad de ciencia de datos por su sintaxis sencilla, versatilidad y la amplia gama de bibliotecas especializadas en el procesamiento de datos y el aprendizaje automático.

La elección de Python se justifica por su capacidad para gestionar tareas complejas de análisis de datos, modelado y visualización de manera eficiente, además de contar con una activa comunidad que contribuye constantemente a su evolución.

Las principales librerías de python utilizadas fueron:

- **Pandas** Esta librería fue fundamental para la manipulación y preprocesamiento de los datos. Pandas ofrece estructuras de datos eficientes, como el DataFrame, que permite manejar grandes volúmenes de datos tabulares de manera sencilla. Con esta herramienta, se gestionaron tareas de limpieza, transformación de datos y manejo de valores faltantes, tareas previas esenciales para la construcción de los modelos.
- **Matplotlib y Seaborn** Ambas librerías se utilizaron para la visualización de datos. Matplotlib permite la creación de gráficos estáticos en dos y tres dimensiones, facilitando la exploración visual de distribuciones y relaciones entre variables. Por

3. APLICACIÓN DE LAS TÉCNICAS SELECCIONADAS

su parte, Seaborn, que trabaja sobre Matplotlib, se encargó de mejorar la estética y la presentación de los gráficos, haciendo más intuitiva la interpretación de los resultados estadísticos.

- **Scikit-Learn** Esta librería proporciona una implementación eficiente de una amplia gama de algoritmos de aprendizaje automático. En este trabajo, se emplearon tres técnicas específicas: Regresión Lineal, Regresión Logística y K-Nearest Neighbors (KNN). Scikit-Learn simplifica el proceso de construcción, evaluación y comparación de modelos, permitiendo experimentar con diferentes enfoques para resolver los problemas de regresión, clasificación y agrupamiento planteados.

El entorno de desarrollo utilizado fue Jupyter Notebook, que facilita la escritura de código interactivo, la visualización de resultados y la documentación de los pasos del análisis. Jupyter permite integrar código, visualizaciones y texto explicativo en un único documento, lo cual mejora la claridad del proceso y facilita la replicación de los análisis. Además, se utilizó Google Colab para aprovechar recursos adicionales de computación en la nube, lo que permitió una mayor flexibilidad en el manejo de los datos y los modelos, especialmente en el caso de datasets grandes o complejos.

Con el entorno de desarrollo y las herramientas mencionadas, se estableció una base sólida para llevar a cabo el procesamiento, análisis y modelado de los datos. Gracias a las bibliotecas de Python, como Pandas, Matplotlib, Seaborn y Scikit-Learn, fue posible implementar de manera efectiva las técnicas de aprendizaje automático seleccionadas. A continuación, se detallan los tres modelos principales utilizados en este estudio: Regresión Lineal para el análisis de datos continuos, Regresión Logística para la clasificación binaria, y K-Nearest Neighbors (KNN) para la asignación de etiquetas en problemas de clustering. Cada modelo se implementó siguiendo el flujo de trabajo adecuado para la limpieza y optimización de los datos, garantizando que los resultados obtenidos fueran comparables y relevantes en el contexto de los problemas planteados.

3.1. Problemas de regresión: técnica de regresión Lineal

Como parte del ejercicio práctico, se seleccionó un conjunto de datos público disponible en la plataforma Kaggle, denominado **King County House Sales Dataset**, accesible a través del siguiente enlace: <http://surl.li/fetuqn>. Este conjunto de datos proporciona información detallada sobre las ventas de casas en el condado de King, Washington, durante los años 2014 y 2015.

El objetivo de este análisis es desarrollar un modelo de aprendizaje supervisado que permita predecir el precio de venta de una casa en función de diversas características descriptivas de la propiedad y su entorno. Para este propósito, se aplicará el modelo de *Regresión Lineal* como técnica principal.

El conjunto de datos contiene un total de 21 columnas que describen tanto las propiedades físicas de las viviendas como características relacionadas con su ubicación

y entorno. A continuación, se presentan las principales categorías de variables incluidas en el conjunto de datos:

- **Identificación y Temporalidad:** Este conjunto incluye variables como el *ID*, que es un número identificador único para cada casa vendida, y *Date*, que indica la fecha en la que se registró la venta de la propiedad.
- **Precio Objetivo:** La variable *Price* representa el precio de venta de la casa, expresado en dólares estadounidenses, y constituye la variable dependiente del modelo de predicción.
- **Características de la Vivienda:** Entre las características específicas de la casa se incluyen el *Bedrooms*, que indica el número de habitaciones; *Bathrooms*, que señala el número de baños, con un valor de 0.5 representando un baño con inodoro pero sin ducha; *Sqft_living*, que describe el área habitable en pies cuadrados; *Floors*, que representa la cantidad de pisos de la casa; *Waterfront*, una variable binaria que indica si la casa tiene vista al mar; *View*, un índice del 0 al 4 que califica la calidad de la vista; *Condition*, un índice del 1 al 5 que evalúa la condición general de la propiedad; *Grade*, un índice del 1 al 13 que mide la calidad de construcción y diseño de la casa; *Yr_built*, que señala el año de construcción de la casa; y *Yr_renovated*, que indica el año de la última renovación, si corresponde.
- **Ubicación y Tamaño del Terreno:** Las variables relacionadas con la ubicación y tamaño incluyen el *Zipcode*, que representa el código postal del área donde se encuentra la casa; las coordenadas *Lat* y *Long*, que indican la ubicación geográfica; *Sqft_lot*, que describe el área total del terreno en pies cuadrados; y *Sqft_lot15*, que mide el tamaño promedio de los terrenos de los 15 vecinos más cercanos.
- **Espacios Internos:** Se consideran el *Sqft_above*, que representa el área habitable sobre el nivel del suelo; el *Sqft_basement*, que mide el área habitable en el sótano; y el *Sqft_living15*, que describe el tamaño promedio del área habitable de los 15 vecinos más cercanos.

Cada registro en el conjunto de datos representa una casa y sus características asociadas al momento de la venta. Este conjunto de datos es ideal para explorar la relación entre las características físicas de las propiedades, su ubicación, y los precios de venta. El análisis permitirá evaluar el desempeño del modelo de regresión lineal y su capacidad para predecir valores en función de las variables independientes disponibles.

3.1.1. Limpieza y Preprocesamiento del conjunto de Datos

El primer paso en la generación del modelo de regresión lineal consistió en analizar la distribución de la variable objetivo, **price**, para identificar posibles sesgos que pudieran afectar la calidad del modelo. Utilizando el método `skew` de la librería `scipy.stats`, se determinó que esta variable presentaba un sesgo inicial de 4.02, lo que indica una

3. APLICACIÓN DE LAS TÉCNICAS SELECCIONADAS

distribución altamente asimétrica hacia valores altos. Este sesgo puede dificultar la interpretación de los resultados y afectar negativamente la capacidad predictiva del modelo al violar el supuesto de normalidad de los errores.

Para corregir esta asimetría, se aplicaron dos transformaciones matemáticas a la variable `price`:

1. **Transformación logarítmica:** Se aplicó el logaritmo natural a los valores de `price`, reduciendo el sesgo a 0.43.
2. **Transformación raíz cuadrada:** Se calculó la raíz cuadrada de los valores de `price`, disminuyendo el sesgo a 1.66.

Tras analizar los resultados, se optó por utilizar la transformación logarítmica debido a su capacidad para reducir significativamente el sesgo y aproximar la distribución de `price` a una forma más cercana a la normal. En la Figura 3.1 se presentan los histogramas de la variable objetivo antes y después de aplicar cada transformación. Se observa claramente cómo la transformación logarítmica produce una distribución más simétrica y adecuada para el modelado.

Esta transformación no solo facilitó la interpretación de los coeficientes del modelo en términos proporcionales, sino que también mejoró la linealidad entre la variable objetivo transformada (`price_log`) y las variables predictoras, lo que resulta esencial para cumplir con los supuestos de la regresión lineal.

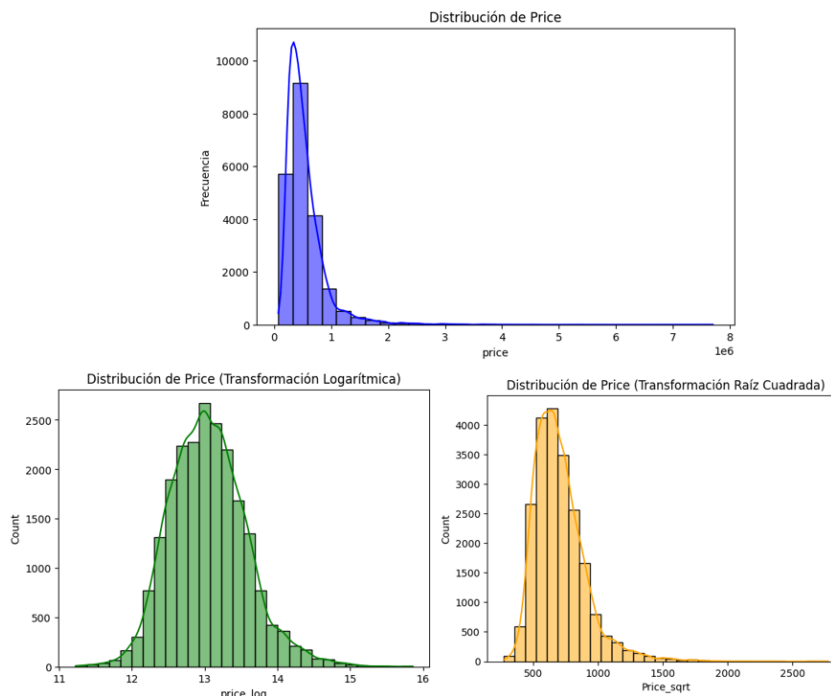


Figura 3.1: Distribución de la variable `price` antes y después de las transformaciones.

El siguiente paso consistió en realizar un análisis exhaustivo de las demás características del conjunto de datos para verificar que los valores estuvieran de acuerdo con la descripción proporcionada en la documentación del conjunto de datos en Kaggle. Para ello, se utilizó el método `.describe()` de `pandas` en combinación con `.get_numeric_data()`, obteniendo un resumen estadístico inicial de las variables numéricas. En este análisis se identificaron las siguientes inconsistencias y valores atípicos, como se observa en la Tabla 3.1.

Tabla 3.1: Resumen descriptivo de las características numéricas del conjunto de datos.

Variable	Count	Mean	Std	Min	25 %	50 %	75 %	Max
bedrooms	21613.0	3.37	0.93	0.00	3.00	3.00	4.00	33.00
bathrooms	21613.0	2.11	0.77	0.00	1.75	2.25	2.50	8.00
sqft_living	21613.0	2079.9	918.4	290.0	1427.0	1910.0	2550.0	13540.0
sqft_lot	21613.0	15106.9	41420.5	520.0	5040.0	7618.0	10688.0	1651359.0
floors	21613.0	1.49	0.54	1.00	1.00	1.50	2.00	3.50
waterfront	21613.0	0.007	0.087	0.00	0.00	0.00	0.00	1.00
view	21613.0	0.23	0.77	0.00	0.00	0.00	0.00	4.00
condition	21613.0	3.41	0.65	1.00	3.00	3.00	4.00	5.00
grade	21613.0	7.66	1.18	1.00	7.00	7.00	8.00	13.00
sqft_above	21613.0	1788.4	828.1	290.0	1190.0	1560.0	2210.0	9410.0
sqft_basement	21613.0	291.5	442.6	0.00	0.00	0.00	560.0	4820.0
yr_built	21613.0	1971.0	29.4	1900.0	1951.0	1975.0	1997.0	2015.0
yr_renovated	21613.0	84.4	401.7	0.00	0.00	0.00	0.00	2015.0
zipcode	21613.0	98077.9	53.5	98001.0	98033.0	98065.0	98118.0	98199.0
lat	21613.0	47.56	0.14	47.16	47.47	47.57	47.68	47.78
long	21613.0	-122.21	0.14	-122.52	-122.33	-122.23	-122.13	-121.32
sqft_living15	21613.0	1986.6	685.4	399.0	1490.0	1840.0	2360.0	6210.0
sqft_lot15	21613.0	12768.5	27304.2	651.0	5100.0	7620.0	10083.0	871200.0
price_log	21613.0	13.05	0.53	11.23	12.68	13.02	13.38	15.86

Entre las observaciones más relevantes, se encontraron los siguientes puntos:

- Las variables `bedrooms` y `bathrooms` presentan valores mínimos de 0, lo cual carece de sentido en un contexto de viviendas, ya que al menos debería haber una habitación y un baño en cualquier propiedad.
- La variable `bedrooms` también muestra un valor máximo de 33 habitaciones, lo

3. APLICACIÓN DE LAS TÉCNICAS SELECCIONADAS

cual es extremadamente alto y poco común para viviendas estándar.

- En `sqft_living`, el valor máximo de 13,540 pies cuadrados parece corresponder a una mansión o una vivienda de lujo, fuera de lo común respecto a los valores promedio.
- La variable `sqft_lot` muestra un rango extremadamente amplio, con un máximo de 1,651,359 pies cuadrados, lo cual indica un terreno desproporcionadamente grande en comparación con el resto de los registros.

Para abordar estas inconsistencias, se utilizó el rango intercuartil IQR, como criterio para la detección y eliminación de valores atípicos. Se aplicó este criterio a las variables donde se detectaron anomalías, tales como `bedrooms`, `bathrooms`, `sqft_living` y `sqft_lot`. Este procedimiento resultó en la eliminación de 1,180 registros, reduciendo el conjunto de datos a 20,433 viviendas. Al filtrar estas observaciones extremas, se logró mejorar la calidad y la representatividad de los datos, garantizando así una base más adecuada para el entrenamiento del modelo.

Además de tratar los valores atípicos, se realizaron transformaciones en ciertas características para mejorar la calidad y relevancia del conjunto de datos. La variable `id`, que no aportaba información significativa al modelo, fue eliminada. Por su parte, la característica `date` fue descompuesta en dos nuevas variables: `year` (año de venta) y `month` (mes de venta), dado que el contexto temporal es crucial para entender las fluctuaciones en el mercado inmobiliario. Adicionalmente, la característica `zipcode`, que proporciona información sobre la ubicación geográfica de las viviendas, fue transformada para facilitar su integración en el modelo, ya que la ubicación es un factor altamente influyente en el precio de venta de una propiedad. Estas modificaciones contribuyeron a capturar mejor las relaciones subyacentes en los datos y mejorar el desempeño del modelo predictivo.

Además de tratar los valores atípicos, se realizaron transformaciones en ciertas características para mejorar la calidad y relevancia del conjunto de datos. La variable `id`, que no aportaba información significativa al modelo, fue eliminada. Por su parte, la característica `date` fue descompuesta en dos nuevas variables: `year` (año de venta) y `month` (mes de venta), dado que el contexto temporal es crucial para entender las fluctuaciones en el mercado inmobiliario. Adicionalmente, la característica `zipcode`, que proporciona información sobre la ubicación geográfica de las viviendas, fue transformada en variables dummies mediante el método `pd.get_dummies`, eliminando la primera categoría para evitar multicolinealidad. Esto permitió representar los códigos postales como variables categóricas relevantes.

Posteriormente, se procedió a la normalización de las variables continuas del conjunto de datos con el objetivo de estandarizar sus valores. Para ello, se seleccionaron las columnas `bedrooms`, `bathrooms`, `sqft_living`, `sqft_lot`, `floors`, `sqft_above`, `sqft_basement`, `sqft_living15` y `sqft_lot15`. Estas variables fueron transformadas utilizando la técnica de estandarización mediante el `StandardScaler` de la librería `sklearn.preprocessing`, la cual ajusta cada característica para que tenga una media de 0 y una desviación estándar de 1. Esta transformación es crucial para garantizar que

las variables tengan la misma escala y que los algoritmos de aprendizaje automático basados en distancia o gradiente no se vean sesgados por diferencias en las magnitudes de las características. La verificación posterior de los estadísticos descriptivos mostró que las transformaciones fueron aplicadas correctamente, con todas las variables continuas presentando una media cercana a 0 y una desviación estándar unitaria.

A continuación, se realizó un análisis de correlación entre las variables del conjunto de datos con el objetivo de identificar relaciones lineales significativas entre ellas. Para esto, se calculó la matriz de correlación utilizando el método `corr()`, que evalúa la relación entre cada par de variables numéricas. Posteriormente, se filtraron las correlaciones con valores absolutos mayores a 0.8, indicando relaciones lineales muy fuertes. Entre los resultados obtenidos, se identificó una correlación de 0.824 entre las variables `sqft_living` y `sqft_above`, lo que sugiere una alta redundancia de información entre ambas características. Para evitar problemas de multicolinealidad y simplificar el modelo, se tomó la decisión de eliminar la variable `sqft_above`, manteniendo únicamente `sqft_living` como representante de esta relación.

3.1.2. Generación del Modelo

En esta etapa, se procedió a la construcción de un modelo de regresión lineal con el fin de predecir el precio de las viviendas a partir de las características previamente seleccionadas y transformadas del conjunto de datos. Para ello, se utilizó un enfoque de mínimos cuadrados, que es un método estadístico estándar en la regresión lineal que busca minimizar la suma de los errores al ajustar el modelo, es decir, reducir la diferencia entre los valores reales y los valores predichos por el modelo.

A continuación, se describen los pasos seguidos en el proceso de generación del modelo:

```
1 | from sklearn.model_selection import train_test_split
2 | from sklearn.linear.model import LinearRegression
3 |
4 | y = df['variable objetivo']
5 | X = df.drop(columns=['variable objetivo'])
6 |
7 | X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
8 |                                                    =0.2, random_state=42)
9 |
10 | model = LinearRegression()
11 | model.fit(X_train, y_train)
12 |
13 | y_pred = model.predict(X_test)
```

Código Fuente 3.1: Generación del modelo mediante la técnica de regresión lineal

A continuación, se describen los pasos seguidos en el proceso de generación del modelo:

1. **Importación de las bibliotecas necesarias:** Se importaron las librerías esenciales de `scikit-learn` para la construcción y evaluación del modelo. En particular, se usaron `train_test_split` para dividir el conjunto de datos en subcon-

3. APLICACIÓN DE LAS TÉCNICAS SELECCIONADAS

juntos de entrenamiento y prueba, y `LinearRegression` para crear el modelo de regresión lineal.

2. **Preparación de los datos:** Se definieron las variables independientes (X) y la variable dependiente (y). La variable dependiente, y , correspondía al precio de las viviendas, mientras que X contenía todas las características de las viviendas seleccionadas y transformadas.
3. **División de los datos:** Se dividió el conjunto de datos en un subconjunto de entrenamiento y uno de prueba utilizando `train_test_split`. Este proceso permitió asignar el 80 % de los datos al conjunto de entrenamiento y el 20 % restante al conjunto de prueba, lo que ayuda a evaluar la capacidad de generalización del modelo. El parámetro `random_state` se estableció en 42 para garantizar la reproducibilidad de los resultados.
4. **Ajuste del modelo:** Se instancia un objeto de la clase `LinearRegression` y se ajustó al conjunto de datos de entrenamiento utilizando el método `fit()`. Este método calculó los coeficientes de la regresión lineal que minimizan el error cuadrático entre los valores predichos y los reales en el conjunto de entrenamiento.
5. **Predicción:** Una vez ajustado el modelo, se utilizaron los datos de prueba para hacer predicciones sobre el precio de las viviendas mediante el método `predict()`. Este método devuelve un array con los valores predichos para las observaciones del conjunto de prueba.

Este proceso proporcionó un modelo de regresión lineal que puede utilizarse para prever el precio de las viviendas en función de las características seleccionadas. La evaluación del rendimiento del modelo se puede realizar utilizando métricas de error como el error cuadrático medio (RMSE), el coeficiente de determinación (R^2), entre otras, para asegurar su efectividad y precisión en la tarea de predicción.

3.1.3. Evaluación del modelo

Una vez entrenado el modelo, se evaluó su desempeño utilizando el conjunto de prueba, calculando métricas de error y de ajuste. Inicialmente, se midió el error cuadrático medio (MSE) en la escala logarítmica en la que había sido transformada la variable objetivo. Este análisis arrojó un valor de 0,0275, indicando que las predicciones del modelo presentan un nivel de error bajo respecto a los valores reales transformados. Asimismo, el coeficiente de determinación (R^2) alcanzó un valor de 0,8661, lo cual sugiere que el modelo explica aproximadamente el 86.61 % de la variabilidad de los datos de prueba.

Para interpretar las predicciones en términos de la escala original, se retransformaron tanto los valores predichos como los valores reales mediante la función exponencial. Esto permitió calcular métricas en la escala original, que son más intuitivas para interpretar el rendimiento del modelo en el contexto de precios de viviendas. En esta escala, el error cuadrático medio (MSE) fue de 8,349,476,812,59, y la raíz del error

cuadrático medio (RMSE) alcanzó un valor de 91,375,47. Este último valor refleja el error promedio absoluto del modelo al predecir el precio de las viviendas, medido en dólares.

Adicionalmente, se calculó el error porcentual absoluto medio (MAPE), el cual resultó ser del 12,62 %. Este resultado indica que, en promedio, las predicciones del modelo presentan un desvío del 12.62 % respecto a los valores reales, lo que es aceptable en problemas de esta naturaleza.

El modelo de regresión lineal ajustado presentó un desempeño satisfactorio, logrando capturar de manera adecuada las relaciones entre las características seleccionadas y el precio de las viviendas. Esto demuestra la utilidad de este enfoque para problemas de predicción en este contexto.

3.2. Problemas de clasificación: Técnica de regresión logística

Como parte del ejercicio práctico, se seleccionó un conjunto de datos público, disponible en la plataforma Kaggle, con el nombre: **Cardiovascular Disease dataset** accesible a través del siguiente enlace: <http://surl.li/fetuqn>.

Este conjunto de datos se compone de diversas características que describen el estado de salud cardiovascular de individuos. Las variables están organizadas en tres categorías principales:

- **Características Objetivas:** Se incluyen variables como la *edad* (en días), *altura* (en centímetros), *peso* (en kilogramos) y *género* (código categórico), las cuales aportan información básica y cuantificable del individuo.
- **Resultados de Exámenes Médicos:** Variables como la *presión arterial sistólica* (ap_hi), *presión arterial diastólica* (ap_lo), *nivel de colesterol* (clasificado en tres niveles) y *nivel de glucosa* (clasificado en tres niveles) proporcionan una medición directa del estado cardiovascular de los individuos.
- **Información Subjetiva Proporcionada por el Paciente:** Incluye datos sobre el *hábito de fumar*, *consumo de alcohol* y *actividad física*, representados en formato binario. Estos factores, aunque subjetivos, tienen un impacto significativo en la salud cardiovascular.
- **Variable Objetivo:** La variable *cardio* indica la presencia (valor 1) o ausencia (valor 0) de enfermedad cardiovascular en el individuo, y es el principal foco del análisis y modelado.

El conjunto de datos fue compartido por **Svetlana Ulianova**, estudiante de Ciencia de Datos en la Universidad de Ryerson, quien resalta la importancia de que los valores fueron recopilados durante una evaluación médica directa. Esto permite un

3. APLICACIÓN DE LAS TÉCNICAS SELECCIONADAS

análisis de los factores relacionados con las enfermedades cardiovasculares, basado en una combinación de datos objetivos y subjetivos, obtenidos durante las evaluaciones de salud.

Cada registro en el conjunto de datos representa a un individuo y su estado de salud cardiovascular en un momento específico de la evaluación. La variable objetivo, denominada **cardio**, permite identificar la presencia o ausencia de enfermedad cardiovascular en función de las características mencionadas.

Este conjunto de datos es crucial para la creación y evaluación de modelos de aprendizaje automático, dado que ofrece una combinación de características cuantitativas y cualitativas que permiten aplicar técnicas de clasificación.

3.2.1. Limpieza y Preprocesamiento del Conjunto de Datos

Para asegurar la efectividad de la técnica de clasificación mencionada anteriormente, la *Regresión Logística*, es fundamental realizar una preparación cuidadosa del conjunto de datos. Este proceso tiene como objetivo principal eliminar sesgos potenciales (sin introducir nuevos) y garantizar que los resultados obtenidos sean lo más precisos posible. En esta sección, nos enfocaremos en el proceso de limpieza del conjunto de datos para asegurar su adecuación antes de aplicar el modelo de *Regresión Logística*.

Para iniciar la limpieza y análisis del conjunto de datos, utilizamos la librería **pandas** para convertir el conjunto de datos en un **DataFrame**. Al aplicar el método `.info()`, mediante este método se pudo determinar que el **DataFrame** contienen un total de 70,000 valores no nulos, lo que indica que no existen datos faltantes. Además, los tipos de datos de cada columna coinciden con los parámetros que representan. Sin embargo, la ausencia de valores nulos no elimina la necesidad de un procesamiento adicional de los datos. Es esencial realizar un análisis detallado para detectar posibles valores atípicos o inconsistencias que puedan afectar el rendimiento del modelo de *Regresión Logística*.

Para detectar posibles valores atípicos o inconsistencias en el conjunto de datos, se utilizó el siguiente código: `.get_numeric_data().describe().T`, el cual permite calcular estadísticas descriptivas de las variables numéricas presentes en el **DataFrame** cargado desde un archivo CSV.

Este código combina tres métodos: el método `.get_numeric_data()`, que selecciona únicamente las columnas numéricas del conjunto de datos, devolviendo un nuevo **DataFrame** con las variables relevantes para el análisis; el método `.describe()`, que calcula estadísticas descriptivas como la media, desviación estándar, valor mínimo y máximo para cada variable; y, finalmente, el método `.T`, que transpone el **DataFrame** para una presentación más clara de los resultados. La información obtenida se muestra en la Tabla 3.2, donde se puede observar la media, desviación estándar, valor mínimo y máximo de cada variable, lo que permite identificar valores atípicos o inconsistentes.

Analizando la variable objetivo en la Tabla 3.2, observamos que presenta una media de 0,4997 y una desviación estándar de 0,5000, lo que sugiere una distribución balanceada entre las dos clases. Este comportamiento se debe a que, en una variable binaria con clases igualmente representadas, la media tiende a ser cercana a 0,5 y la

	Count	Mean	Std	Min	Max
Age	70000.0	19468.9	2467.3	10798.0	23713.0
Gender	70000.0	1.3496	0.4768	1.0	2.0
Height	70000.0	164.3592	8.2101	55.0	250.0
Weight	70000.0	74.2057	14.3958	10.0	200.0
Ap_hi	70000.0	128.8173	154.0114	-150.0	16020.0
Ap_lo	70000.0	96.6304	188.4725	-70.0	11000.0
Cholesterol	70000.0	1.3669	0.6803	1.0	3.0
Gluc	70000.0	1.2265	0.5723	1.0	3.0
Smoke	70000.0	0.0881	0.2835	0.0	1.0
Alco	70000.0	0.0538	0.2256	0.0	1.0
Active	70000.0	0.8037	0.3972	0.0	1.0
Cardio	70000.0	0.4997	0.5000	0.0	1.0

Tabla 3.2: Descripción estadística de los datos.

desviación estándar alcanza su valor máximo en 0,5, correspondiente a la mayor variabilidad posible en una distribución uniforme de clases. Para confirmar este equilibrio, realizamos un análisis adicional utilizando el método `value_counts`, que corrobora que las frecuencias relativas de las clases son aproximadamente iguales. Por lo tanto, no es necesario aplicar técnicas de manejo de desequilibrio de clases en este caso.

En la Tabla 3.2, se identifican valores atípicos o inconsistencias en los parámetros de *Altura* (Height), donde se observan un valor mínimo de 55,0 cm y un valor máximo de 250,0 cm, lo cual excede los rangos normales de altura humana. En el caso del *Peso* (Weight), se presentan valores mínimos y máximos de 10 kg y 200 kg, respectivamente, los cuales también son inusuales en términos de distribución del peso corporal. Además, en los parámetros de *Presión arterial alta* (Ap_hi) y *Presión arterial baja* (Ap_lo), se encuentran valores que son completamente incompatibles con la vida, como un valor mínimo de -150 para la presión alta y un máximo de 16020 para la presión baja, lo que indica la presencia de errores en la medición o en el registro de los datos. Estos valores atípicos requieren una revisión detallada para determinar su origen y decidir si deben ser corregidos o eliminados del conjunto de datos.

Al dividir el conjunto de datos por géneros, se observa que las alturas promedio son de aproximadamente 170 cm para hombres y 161 cm para mujeres. Estas cifras son consistentes con las estimaciones previas basadas en la edad, lo cual contribuye a

3. APLICACIÓN DE LAS TÉCNICAS SELECCIONADAS

validar la coherencia de los datos de altura en el conjunto, considerando la edad de los individuos como un factor relevante.

	Valor Mínimo	Valor Máximo	Datos Atípicos
Mujeres - IQR	144	179	497
Hombres - IQR	151	189	225
Mujeres - Z-Score	134	189	96
Hombres - Z-Score	142	198	57

Tabla 3.3: Resultados del análisis de valores atípicos para la altura.

Para identificar posibles valores atípicos en la variable de altura, se emplearon las técnicas del *Rango Intercuartílico* (IQR) y el *Z-Score*. Ambas metodologías definen intervalos de valores considerados normales y permiten cuantificar la cantidad de valores atípicos presentes en el conjunto de datos. Tras analizar los resultados expuestos en la Tabla 3.3, se optó por el intervalo de valores determinado mediante el Z-Score, ya que los resultados obtenidos muestran una distribución más coherente con la población en general.

Posteriormente, los valores que se encuentran fuera del intervalo establecido mediante el Z-Score, es decir, aquellos identificados como atípicos, fueron reemplazados por la media para preservar la integridad y validez del conjunto de datos.

Para la característica de peso (*weight*), se siguió un proceso similar al utilizado para la altura, con el objetivo de identificar y tratar posibles valores atípicos en el conjunto de datos. En este caso, el análisis se realizó sobre el conjunto completo de datos, sin distinción por género, dado que el peso presenta una mayor variabilidad y no se dispone de información de referencia específica para la población.

Para la detección de valores atípicos, se emplearon nuevamente las técnicas de *Rango Intercuartílico* (IQR) y *Z-Score*, definiendo así intervalos de valores considerados normales dentro de la distribución observada.

	Valor Mínimo	Valor Máximo	Datos Atípicos
IQR	39.5	107.5	1819
Z-Score	31.02	117.39	701

Tabla 3.4: Resultados del análisis de valores atípicos para el peso.

Como se observa en la Tabla 3.4, la técnica del Z-Score resultó en un intervalo de valores más amplio (31.02 a 117.39), identificando 701 valores atípicos. Por otro lado,

el método IQR generó un intervalo más restringido (39.5 a 107.5), lo cual resultó en un mayor número de datos atípicos, alcanzando un total de 1819.

Dado que el Z-Score ofrece un intervalo que refleja mejor la distribución general del peso en la población, se optó por utilizar este método para el tratamiento de los datos. Los valores de peso identificados como atípicos mediante el Z-Score fueron reemplazados por la media de la distribución, con el fin de preservar la integridad y representatividad del conjunto de datos.

Para continuar con el proceso de limpieza del conjunto de datos, se analizó las características de presión arterial sistólica (*ap_hi*) y presión arterial diastólica (*ap_lo*). Ambas variables muestran una correlación de 0.72, lo cual sugiere una relación lineal moderada y permite implementar imputación por regresión lineal en aquellos valores considerados atípicos en función de los rangos clínicamente aceptables.

Para identificar estos valores atípicos, se establecieron límites críticos basados en valores de presión arterial incompatibles con la vida. Según estos límites, se consideraron fuera de rango los valores que se encuentran en los siguientes intervalos:

- **Presión arterial sistólica (PAS):** menos de 70 mmHg o superior a 250 mmHg.
- **Presión arterial diastólica (PAD):** menos de 30 mmHg o superior a 150 mmHg.

Con base en estos criterios, se identificaron un total de 1,028 valores atípicos para la presión diastólica (*ap_lo*) y 229 para la presión sistólica (*ap_hi*), correspondientes a aquellos valores que se encuentran fuera de los rangos críticos descritos. Para estos valores fuera de rango, se aplicó imputación mediante regresión lineal, aprovechando la relación entre *ap_hi* y *ap_lo* para corregir las inconsistencias.

Por otra parte, los valores que se encuentran dentro de los rangos clínicos, aunque pudieran mostrar variabilidad, se mantuvieron tal cual, sin ningún tipo de modificación. Este proceso de identificación y tratamiento permite mejorar la calidad y la consistencia de los datos, asegurando que los valores de presión arterial utilizados en el análisis final sean lo más representativo y realistas posible.

	Count	Mean	Std	Min	Max
Height	70000.0	164.4204	7.7152	140.0	191.0
Weight	70000.0	73.6606	13.1928	31.0	117.0
Ap_hi	70000.0	127.0025	17.0847	70.0	240.0
Ap_lo	70000.0	81.3538	9.5026	30.0	150.0

Tabla 3.5: Descripción estadística de las características ajustadas.

El proceso de limpieza de valores atípicos aplicado a las variables de altura, peso y presión arterial mejoró notablemente la coherencia y plausibilidad de los datos.

3. APLICACIÓN DE LAS TÉCNICAS SELECCIONADAS

Los ajustes realizados mantuvieron las medias generales dentro de rangos similares a los originales, asegurando que el perfil promedio de la población no se alterara significativamente, como se observa en la tabla 3.5. Sin embargo, se lograron reducciones importantes en los valores mínimos y máximos, eliminando datos extremos que no correspondían con valores fisiológicamente plausibles. Este refinamiento garantiza que el conjunto de datos final refleje de manera más fiel a la población en estudio, aumentando así la confiabilidad de los análisis posteriores.

Para la aplicación de la técnica de *regresión logística*, se decidió proceder con la *estandarización* de las variables continuas del conjunto de datos.

La *regresión logística* se ve favorecida por la estandarización debido a que el algoritmo puede ser sensible a las diferencias de escala entre las variables. En particular, cuando se utiliza *regularización L2*, la estandarización asegura que todas las variables tengan una contribución equivalente al modelo. De no estandarizar, las características con un rango mayor podrían dominar la contribución al modelo, lo que dificultaría la interpretación de los coeficientes y podría sesgar los resultados.

En este proceso, se empleó la función `StandardScaler` de la librería `scikit-learn` para aplicar la estandarización a las variables continuas. Esto transforma cada característica para que tenga una media de 0 y una desviación estándar de 1, asegurando que cada variable contribuya de manera equilibrada al proceso de modelado.

Además, dado que las variables `ap_hi` y `ap_lo` están correlacionadas entre sí con un valor de 0.720662, se optó por eliminar la variable `ap_lo` debido a su mayor cantidad de valores atípicos en comparación con `ap_hi`. Este enfoque se realizó utilizando el método `drop()` de `pandas`, eliminando la columna correspondiente de manera eficiente.

Por último, como parte del procesamiento final de los datos, el conjunto se dividió en dos particiones: el conjunto de datos de **entrenamiento** que comprende el 90 % de los datos y el conjunto de **prueba** que contiene el 10 % restante. Para asegurar que ambas particiones se generaran de manera aleatoria, se utilizó el método `train_test_split()` de `scikit-learn`, que permite obtener una distribución arbitraria de los datos, garantizando que tanto el conjunto de entrenamiento como el de prueba sean representativos del conjunto de datos original.

3.2.2. Generación del Modelo

Una vez completado el procesamiento y la partición del conjunto de datos en entrenamiento y prueba, el siguiente paso consiste en la generación del modelo de regresión logística. Este proceso comienza con la selección de los hiperparámetros a ajustar, los cuales son fundamentales para optimizar el rendimiento del modelo. El ajuste de estos hiperparámetros se realiza mediante técnicas específicas como **Grid Search** o **Random Search**, que permiten explorar diversas combinaciones de parámetros para encontrar la configuración que maximice la precisión y minimice el error. A lo largo de este proceso, se busca no solo mejorar la capacidad predictiva del modelo, sino también garantizar su robustez y generalización al aplicarlo a nuevos datos.

Para optimizar el modelo de regresión logística, se llevó a cabo un ajuste de hiper-

parámetros utilizando el método **Grid Search** con validación cruzada. Este enfoque sistemático permite explorar combinaciones de valores para los hiperparámetros del modelo y seleccionar aquella configuración que maximiza el desempeño predictivo según la métrica seleccionada. En este caso, se utilizó la métrica F_1 -score, ideal para problemas de clasificación binaria debido a su capacidad para equilibrar precisión y exhaustividad (recall).

Entre los hiperparámetros evaluados se incluyeron:

- C : regula la intensidad de la regularización, con valores pequeños favoreciendo una mayor penalización.
- Penalización (L_1 , L_2 , ElasticNet o sin penalización): controla el tipo de regularización aplicada.
- Solver: algoritmos como `liblinear`, `saga`, `lbfgs` y `newton-cg` para optimizar los parámetros del modelo.
- `max_iter`: el número máximo de iteraciones para la convergencia.
- `tol`: tolerancia que define el criterio de parada para la optimización.

La elección de F_1 -score como métrica de evaluación responde a su capacidad para proporcionar una medida equilibrada del desempeño del modelo, combinando precisión y recall en una única métrica armónica. Esto resulta crítico en escenarios donde los costos asociados a falsos positivos y falsos negativos tienen un impacto significativo. Aunque en este caso particular las clases del conjunto de datos están equilibradas, el F_1 -score sigue siendo útil para evaluar modelos en situaciones donde ambas métricas son igualmente importantes.

Para la penalización ElasticNet, que combina L_1 y L_2 , se evaluó adicionalmente el hiperparámetro `l1_ratio`, el cual controla la proporción entre las dos regularizaciones. La validación cruzada con 5 particiones (`cv=5`) garantizó una evaluación robusta de las configuraciones probadas, minimizando el riesgo de sobreajuste.

El proceso de ajuste arrojó la combinación óptima de hiperparámetros que maximiza el desempeño del modelo en el conjunto de entrenamiento, asegurando un balance adecuado entre capacidad predictiva y generalización.

Dado que la explicación detallada de cada uno de estos hiperparámetros y su aplicación práctica se encuentra fuera del alcance de este trabajo de grado, se recomienda al lector consultar la documentación oficial de Scikit-learn ¹. Esta fuente proporciona información detallada, ejemplos prácticos y explicaciones que permiten una comprensión más profunda y personalizada de los parámetros utilizados, facilitando su correcta aplicación y adaptación a problemas específicos.

Con los hiperparámetros óptimos seleccionados mediante el ajuste de hiperparámetros con Grid Search, se procedió a la generación y evaluación del modelo de regresión

¹https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.

[LogisticRegression.html](#)

3. APLICACIÓN DE LAS TÉCNICAS SELECCIONADAS

logística. Este proceso involucró el entrenamiento del modelo con los datos particionados previamente en conjunto de entrenamiento y prueba, asegurando así la generalización y robustez del modelo. Los hiperparámetros seleccionados que maximizan el desempeño del modelo son los siguientes:

- **C (Regularización):** 0.001
- **$l1_ratio$:** 0.1 (Para ElasticNet, mezcla entre L_1 y L_2)
- **max_iter :** 100 (Número máximo de iteraciones para la convergencia del algoritmo)
- **Penalización:** L_2 (Penalización para controlar el sobreajuste)
- **Solver:** lbfgs (Algoritmo de optimización utilizado)
- **Tolerancia (tol):** 0.01 (Criterio de convergencia del algoritmo)

Con los hiperparamtros obtenidos, se procedió a la construcción de un modelo de regresión logística con el objetivo de predecir la variable de interés a partir de las características seleccionadas y transformadas previamente del conjunto de datos. Para maximizar el rendimiento del modelo, se llevó a cabo un proceso de selección de los mejores hiperparámetros, asegurando que la configuración utilizada fuera la más adecuada para los datos y la tarea de predicción.

```
2 from sklearn.model_selection import train_test_split
3 from sklearn.linear_model import LogisticRegression
4 y = df['variable objetivo']
5 X = df.drop(columns=['variable objetivo'])
6 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
7     =0.2, random_state=42)
8 model = LogisticRegression()
9 model.fit(X_train, y_train)
10 y_pred = model.predict(X_test)
```

Código Fuente 3.2: Generación del modelo mediante la tecnica de regresión logistica

A continuación, se describen los pasos seguidos en el proceso de generación y ajuste del modelo:

1. **Importación de las bibliotecas necesarias:** Se importaron las librerías esenciales de `scikit-learn` para la construcción y evaluación del modelo de regresión logística. En particular, se utilizó `train_test_split` para dividir el conjunto de datos en subconjuntos de entrenamiento y prueba, y la clase `LogisticRegression` para crear el modelo de regresión logística.
2. **Preparación de los datos:** Se definieron las variables independientes (X) y la variable dependiente (y). La variable dependiente, y , representaba la clase o categoría a predecir, mientras que X contenía las características seleccionadas y transformadas.

3. **División de los datos:** Se dividió el conjunto de datos en subconjuntos de entrenamiento y prueba utilizando `train_test_split`. Se asignó el 80 % de los datos al conjunto de entrenamiento y el 20 % al conjunto de prueba para evaluar la capacidad de generalización del modelo. El parámetro `random_state` se estableció en 42 para garantizar la reproducibilidad de los resultados.
4. **Selección de hiperparámetros:** Se realizó un proceso de búsqueda de los mejores hiperparámetros utilizando técnicas como la búsqueda en cuadrícula (`GridSearchCV`) o búsqueda aleatoria (`RandomizedSearchCV`). Este proceso consistió en probar diferentes combinaciones de hiperparámetros y seleccionar aquellas que optimizaban las métricas de evaluación, como la precisión, el F1-score o el área bajo la curva ROC (AUC).
5. **Ajuste del modelo:** Una vez seleccionados los mejores hiperparámetros, se ajustó el modelo de regresión logística al conjunto de entrenamiento utilizando el método `fit()`. Este proceso permitió calcular los coeficientes internos que mejor se ajustaban a los datos.
6. **Predicción y evaluación:** Tras el ajuste del modelo, se realizaron predicciones sobre el conjunto de prueba y se evaluó su rendimiento con métricas como la precisión (accuracy), la matriz de confusión y el área bajo la curva ROC (AUC). Estas métricas proporcionaron una visión cuantitativa de la efectividad del modelo en la tarea de clasificación.

Este proceso resultó en un modelo de regresión logística optimizado que puede emplearse para la predicción de la variable objetivo. La evaluación del rendimiento permitió confirmar la capacidad de generalización y la precisión del modelo, garantizando que la configuración de hiperparámetros elegida fuera la más adecuada para el problema en cuestión.

3.2.3. Evaluación del modelo

Con los hiperparámetros óptimos, el modelo fue evaluado utilizando la matriz de confusión y varias métricas de desempeño. A continuación, se presentan los resultados:

Matriz de Confusión

$$\begin{pmatrix} 2744 & 774 \\ 1125 & 2357 \end{pmatrix}$$

- Verdaderos Negativos (TN): 2744
- Falsos Positivos (FP): 774
- Falsos Negativos (FN): 1125
- Verdaderos Positivos (TP): 2357

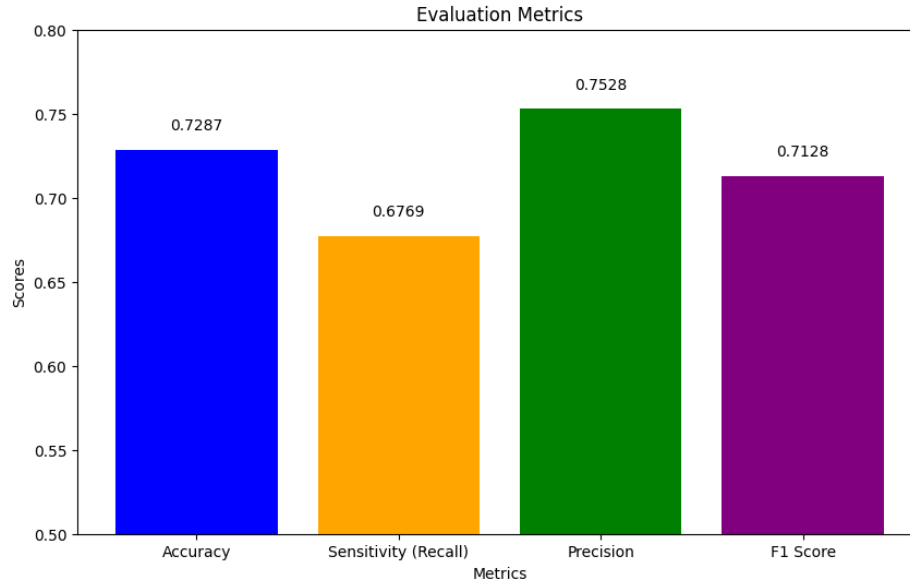


Figura 3.2: Métricas: regresión logística

Los resultados obtenidos muestran que el modelo de regresión logística, ajustado mediante Grid Search y validación cruzada, tiene un desempeño moderado en la clasificación 3.2. La exactitud del 72.87 % indica que el modelo es relativamente bueno, aunque no perfecto. La combinación de una buena precisión (75.28 %) y una sensibilidad moderada (67.69 %) sugiere que el modelo es adecuado para evitar falsos positivos, pero aún tiene espacio para mejorar en la identificación de las instancias positivas.

El F_1 -score de 0.7128 refleja un equilibrio entre las métricas de precisión y recall, lo que lo convierte en una métrica útil para problemas en los que tanto los falsos positivos como los falsos negativos deben ser gestionados adecuadamente.

Los resultados son prometedores, pero el modelo podría beneficiarse de una mayor optimización, como el ajuste de otros hiperparámetros, el uso de técnicas de balanceo de clases o la incorporación de más características relevantes en el conjunto de datos, así como la generación de modelos mediante otras técnicas de aprendizaje automático.

3.3. Problemas de agrupamiento (clustering): Técnica de K-Nearest Neighbors (KNN)

En esta sección, se explora la aplicación de la técnica de *clustering K-Nearest Neighbors (KNN)* sobre el conjunto de datos **Cardiovascular Disease dataset**, el cual ya fue empleado en el análisis de *regresión logística*. El objetivo de este análisis es determinar si la técnica de KNN es capaz de formar dos grupos que se asemejen a los

3.3 Problemas de agrupamiento (clustering): Técnica de K-Nearest Neighbors (KNN)

patrones observados en el conjunto de datos utilizado anteriormente. Para llevar a cabo esta tarea, se elimina la variable objetivo, ya que el enfoque de clustering no requiere etiquetas de clase. Este proceso permite evaluar cómo KNN agrupa los datos en función de sus características y si los grupos resultantes son coherentes con las categorías de riesgo y patrones identificados en el análisis de clasificación previo.

3.3.1. Limpieza y Preprocesamiento del Conjunto de Datos

En esta etapa, se utilizará el mismo conjunto de datos empleado en la generación del modelo de clasificación, aplicando los mismos procesos de preprocesamiento realizados anteriormente. La única diferencia es la eliminación de la variable objetivo, `cardio`, para crear un nuevo dataframe. Este paso es esencial para permitir la comparación entre la variable `cardio` y los grupos formados por el modelo de clustering. De esta manera, se podrá evaluar si los grupos generados por el modelo son coherentes con los patrones de riesgo y categorías previamente identificados.

3.3.2. Generación del Modelo

Para la generación del modelo de agrupamiento, se utilizó la técnica K-Means con el objetivo de dividir los datos en dos grupos. Esta técnica es un algoritmo de aprendizaje no supervisado que busca agrupar datos en un número k de grupos. :

```
2 | from sklearn.model_selection import train_test_split
   | from sklearn.neighbors import KNeighborsClassifier
4 | y = df['variable objetivo']
   | X = df.drop(columns=['variable objetivo'])
6 |
   | X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
   | =0.2, random_state=42)
8 |
   | model = KNeighborsClassifier(n_neighbors=k)
10 | model.fit(X_train, y_train)
12 | y_pred = model.predict(X_test)
```

Código Fuente 3.3: Generación del modelo mediante la técnica de KNN

A continuación, se describen los pasos realizados para la creación y ajuste del modelo de K-Means

1. **Importación de las bibliotecas necesarias:** Se importaron las bibliotecas esenciales de `scikit-learn` para la construcción y evaluación del modelo de clustering. En particular, se usó la clase `KMeans` para aplicar la técnica de K-Means.
2. **Preparación de los datos:** Se definieron las variables independientes (X) que representan las características seleccionadas para realizar el agrupamiento. En este caso, no se incluyó una variable objetivo, ya que la técnica de clustering no requiere etiquetas de clase, aunque se genera un `DataFrame` con conjunto de

3. APLICACIÓN DE LAS TÉCNICAS SELECCIONADAS

clases (variable objetivo), para su posterior comparación con las clases(grupos) generadas mediante el modelo.

3. **Configuración del número de clústeres:** Se definió el hiperparámetro `n_clusters` con un valor de 2, indicando que los datos se dividirían en dos grupos distintos. Este valor se seleccionó con base en un análisis previo de los datos y el objetivo del estudio.
4. **Ajuste del modelo:** Se ajustó el modelo de K-Means al conjunto de datos utilizando el método `fit()`. Este proceso implicó la asignación de los datos a los clústeres más cercanos y la actualización iterativa de los centroides hasta que se alcanzó la convergencia.
5. **Predicción y evaluación:** Tras el ajuste del modelo, se realizaron predicciones sobre el conjunto de datos, asignando a cada observación el clúster al que pertenecía. Para evaluar el desempeño del modelo, se analizaron las características de cada grupo y se utilizó la visualización de los resultados para observar cómo se distribuyen los clústeres en el espacio de características.

3.3.3. Evaluación del Modelo

Ya obtenidos un agrupamiento por parte del modelo, se procede a comparar con las clases reales del conjunto de datos para analizar la coherencia de los patrones identificados. A continuación, se presentan los resultados obtenidos:

Clase	Precisión	Recall	F1-Score
0.0	0.66	0.83	0.74
1.0	0.78	0.58	0.66

Tabla 3.6: Métricas de evaluación del modelo K-Means.

La matriz de confusión muestra los siguientes valores:

- Clase 0.0: 29,176 predicciones correctas y 5,845 incorrectas.
- Clase 1.0: 20,161 predicciones correctas y 14,818 incorrectas.

La exactitud global alcanzada por el modelo fue del 70.48 %. Esto indica que el modelo tiene un nivel aceptable de coherencia con las etiquetas reales, aunque existen diferencias en el desempeño entre las dos clases. La precisión para la clase 1.0 (78 %) es mayor que para la clase 0.0 (66 %), lo que sugiere que el modelo identifica con mayor certeza las instancias pertenecientes a esta clase. Sin embargo, el *recall* para la clase 1.0 (58 %) es considerablemente menor, lo que implica que el modelo no detecta correctamente un número significativo de ejemplos de esta categoría.

3.3 Problemas de agrupamiento (clustering): Técnica de K-Nearest Neighbors (KNN)

Los resultados obtenidos reflejan que el modelo de clustering es capaz de identificar patrones relacionados con las clases reales, pero su desempeño está limitado por la estructura intrínseca de los datos. La alta precisión en ambas clases sugiere que los clústeres generados son consistentes, pero el *recall* más bajo en la clase 1.0 podría deberse a un solapamiento entre las clases o una mayor dispersión de los datos de esta categoría.

En términos de las métricas adicionales obtenidas, el *Adjusted Rand Index (ARI)* es 0.1678, lo que indica un acuerdo relativamente bajo entre las etiquetas reales y los clústeres generados, después de ajustar por coincidencias aleatorias. Este valor sugiere que los clústeres capturan cierta información sobre la estructura de los datos, aunque su capacidad para reflejar con precisión las clases reales es limitada.

La *homogeneidad* (0.1351) y la *completitud* (0.1419) también reflejan la capacidad del modelo para asignar correctamente los datos a clústeres únicos y garantizar que todas las instancias de una misma clase estén contenidas dentro de un único clúster, respectivamente. Estos valores relativamente bajos indican que, aunque hay cierta agrupación de los datos basada en características comunes, el modelo no logra separar completamente las clases reales.

Finalmente, la métrica *V-measure* (0.1384), que combina la homogeneidad y la completitud en una media armónica, confirma la existencia de una relación moderadamente débil entre los clústeres generados y las etiquetas reales. Estos resultados sugieren que, aunque el modelo de clustering aporta información útil sobre la estructura de los datos, no logra capturar completamente las relaciones complejas presentes en el conjunto de datos.

El modelo de clustering aporta una perspectiva interesante sobre la segmentación de los datos, aunque los resultados deben ser interpretados con cautela, ya que las técnicas no supervisadas no garantizan una correspondencia directa con las etiquetas reales. La comparación entre las métricas de este modelo y las obtenidas con la técnica de regresión logística proporciona una visión más completa sobre las relaciones y patrones presentes en el conjunto de datos.

Capítulo 4

Conclusiones

Este trabajo de grado ha culminado con éxito en la exploración y aplicación de la ciencia de datos, desde sus fundamentos teóricos hasta la implementación práctica de técnicas de aprendizaje automático (machine learning). A través de un análisis comparativo, se han evaluado diferentes enfoques para abordar problemas específicos, como la clasificación de enfermedades cardiovasculares y la predicción de precios de viviendas, demostrando la utilidad de los conceptos teóricos descritos en el marco teórico y su relevancia en contextos prácticos. Este enfoque integral proporciona una base sólida para futuros estudios en ciencia de datos y refuerza la importancia de seleccionar técnicas adecuadas para alcanzar resultados óptimos.

Entre los logros y hallazgos más destacados se encuentran:

- La exploración de los fundamentos teóricos de la ciencia de datos en el contexto de las matemáticas aplicadas permitió identificar que disciplinas como la estadística, el álgebra lineal y la optimización son pilares esenciales en el desarrollo de técnicas de modelado de datos. Estas áreas proporcionan las bases necesarias para formular modelos matemáticos, analizar datos y evaluar resultados, estableciendo un vínculo directo entre la teoría matemática y las aplicaciones prácticas en la ciencia de datos. Este análisis subraya la importancia de un enfoque interdisciplinario que combine conceptos teóricos sólidos con herramientas computacionales para abordar problemas complejos de manera eficiente.
- La selección de técnicas representativas para las tres principales aplicaciones del aprendizaje automático, clasificación, regresión y agrupamiento permitió abordar problemas concretos utilizando herramientas adecuadas. La regresión logística fue elegida para clasificación, destacando su capacidad para predecir la presencia de enfermedades cardiovasculares; la regresión lineal se empleó para predecir precios de viviendas, aprovechando su simplicidad y efectividad en problemas de regresión; y el método de agrupamiento K-Means permitió analizar la segmentación de datos en grupos coherentes. Los conjuntos de datos seleccionados fueron cruciales para garantizar resultados significativos y alineados con los objetivos, demostrando la importancia de una adecuada correspondencia entre técnica y problema.

4. CONCLUSIONES

- La aplicación del proceso completo de generación y evaluación de modelos permitió consolidar una metodología estructurada para abordar problemas de aprendizaje automático. Desde la limpieza y preparación de los conjuntos de datos hasta el particionado en conjuntos de entrenamiento y prueba, se garantizaron datos de calidad para construir modelos precisos y confiables. Cada modelo, desarrollado a partir de técnicas seleccionadas, fue evaluado utilizando métricas como precisión, sensibilidad y error cuadrático medio, lo que permitió validar su desempeño y comparar su efectividad. Este enfoque integral resaltó la importancia de cada etapa del proceso, asegurando que los resultados obtenidos sean relevantes y escalables a problemas reales.

Este trabajo ha aportado una visión integral sobre la aplicabilidad de la ciencia de datos, ofreciendo herramientas teóricas y prácticas que fortalecen las competencias en este campo y abren oportunidades para su uso en diversos dominios. La comparación entre técnicas y su evaluación rigurosa han demostrado ser estrategias clave para avanzar en el desarrollo de soluciones basadas en datos.

Bibliografía

- Abdi, H. (2007). Singular value decomposition (svd) and generalized singular value decomposition. *Encyclopedia of measurement and statistics*, 907(912):44. [41](#)
- Ares, L. G., Brisaboa, N. R., Bustos, B., Pereira, A. O., and Fernández, Ó. P. (2010). Optimización de las búsquedas knn en espacios métricos. In *JISBD*, pages 153–162. [30](#)
- Bisong, E. and Bisong, E. (2019). Introduction to scikit-learn. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, pages 215–229. [12](#)
- Bobadilla, J. (2021). *Machine Learning y Deep Learning: Usando Python, Scikit y Keras*. Ediciones de la U. [20](#), [21](#), [22](#), [23](#), [34](#), [42](#)
- Bunge, M. (1960). La ciencia. *Su método y su filosofía*, 22:6–23. [5](#)
- B.V., I. (2019). Data science competence framework (cf-ds) edsf cf-ds - release 2. Technical report, EDISON DATA SCIENCE FRAMEWORK EUROPEAN COMMISSION. [14](#)
- Challenger-Pérez, I., Díaz-Ricardo, Y., and Becerra-García, R. A. (2014). El lenguaje de programación python. *Ciencias Holguín*, 20(2):1–13. [10](#)
- Cueli, J. M. R., Sendra, A. L., Marín, D. F. L., and Concejero, J. B. (2019). Machine learning para la optimización inteligente de la recogida de residuos. In *Avances en la investigación en ciencia e ingeniería*, pages 233–241. 3ciencias. [17](#)
- Díaz-Ramírez, J. (2021). Aprendizaje automático y aprendizaje profundo. *Ingeniare. Revista chilena de ingeniería*, 29(2):180–181. [20](#)
- Dombrowski, Q., Gniady, T., and Kloster, D. (2023). Introdução ao jupyter notebook. *The Programming Historian em Português*. [10](#)
- Epp, S. S. (2012). *Matemáticas discretas con aplicaciones*. Cengage learning. [17](#)

BIBLIOGRAFÍA

- Glass, R. and Callahan, S. (2014). *The Big Data-driven business: How to use big data to win customers, beat competitors, and boost profits*. John Wiley & Sons. [10](#)
- Herrero, J. G., de Jesús, A. B., López, J. M. M., Ángel Patricio Guisado, M., Álvaro Luis Bustamante, and Padilla, W. R. (2018). *Ciencia de datos : técnicas analíticas y aprendizaje estadístico*. Alfaomega. [7](#)
- Igual, L., Seguí, S., Igual, L., and Seguí, S. (2017). *Introduction to data science*. Springer. [7](#)
- Jiménez, A., Murillo, A., Piza, E., Villalobos, M., and Trejos, J. (2010). Reducción de la dimensionalidad en análisis de datos. análisis de componentes principales. [38](#)
- Leal, H. C. (2022). Estadística y ciencia de datos.¿ qué hay de nuevo? *Actualidad Económica*, 32(107):55–75. [17](#)
- Lemus-Delgado, D. and Pérez Navarro, R. (2020). Ciencia de datos y estudios globales: aportaciones y desafíos metodológicos. *Colombia Internacional*, (102):41–62. [17](#)
- Linares, P., Ramos, A., Sánchez, P., Sarabia, A., and Vitoriano, B. (2001). Modelos matemáticos de optimización. *Madrid, España*. [18](#)
- Medina, F. and Galván, M. (2007). *Imputación de datos: teoría y práctica*. Cepal. [31](#), [32](#), [33](#)
- Moreno, A., Armengol, E., Béjar Alonso, J., Belanche Muñoz, L. A., Cortés García, C. U., Gavaldà Mestre, R., Gimeno, J. M., Martín Muñoz, M., and Sànchez-Marrè, M. (1994). Aprendizaje automático. [18](#)
- Müller, A. C. and Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. .O'Reilly Media, Inc.". [21](#)
- Muñoz, R. S. (2021). El origen y evolución de la ciencia de datos (data science). *Fundación iS+ D*. [10](#)
- Naur, P. (1966). The science of datalogy. *Communications of the ACM*, 9(7):485. [9](#)
- Norman, A. T. (2019). *Aprendizaje automático en acción*. Litres. [19](#), [28](#)
- Press, G. (2013). A very short history of data science. *Forbes*. [12](#)
- Provost, F., Tom Grindrod, F., Gutttag, J. P., Delen, D., Solka, J. L., and Wong, R. K. (2023). *A Guide to the Data Science Body of Knowledge*. Data Science Society, 2nd edition. [14](#)
- Ringnér, M. (2008). What is principal component analysis? *Nature biotechnology*, 26(3):303–304. [41](#)
- Roque López, J. et al. (2021). Técnicas de selección de variables en regresión lineal múltiple. Master's thesis, Universidad Internacional de Andalucía. [24](#)

- Strang, G. (2022). *Álgebra Lineal en Ciencia de Datos*. SIAM. [17](#)
- Tukey, J. W. (1962). The future of data analysis. *Source: The Annals of Mathematical Statistics*, 33:1–67. [12](#)
- VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. .O'Reilly Media, Inc.". [6](#)
- Vega Rojas, R. A. and Hernández Martinez, J. (2023). Comparación de métodos de reducción de dimensionalidad para algoritmos de clasificación supervisada aplicados a la detección de intrusiones. B.S. thesis, Universidad de las Ciencias Informáticas. Facultad 4. [38](#)