Ciencia de datos

Big data

John Matthew Espinosa Rojas

Profesor: Sebastián Perdomo

Escuela Tecnológica Instituto Técnico Central

Bogotá D.C 2024

John Matthew Espinosa. Escuela Tecnológica Instituto Técnico Central. Semestre 7.

Correspondencia: jmespinosar@itc.edu.co Tel. 3157118023

Tabla de contenidos

1. **Introducción**

La ciencia de datos ha transformado la manera en que las industrias analizan información y toman decisiones. Un ejemplo claro es el sector de ventas al por menor, donde la visualización de datos ha permitido una comprensión más precisa de las preferencias del cliente y optimización del inventario. Al utilizar gráficos de calor y análisis de tendencias, los minoristas pueden observar patrones de compra a lo largo del tiempo y detectar qué productos tienen mayor demanda en diferentes temporadas.

Por ejemplo, la cadena de tiendas Walmart usa visualización de datos para analizar millones de transacciones y optimizar la cadena de suministro. A través de herramientas de visualización avanzadas, la compañía puede monitorear en tiempo real los niveles de inventario en diferentes ubicaciones y ajustar la distribución de productos para satisfacer la demanda sin tener excedentes. Esta estrategia no solo reduce costos operativos sino que también mejora la experiencia del cliente, asegurando que los productos más demandados estén siempre disponibles. La implementación de estos sistemas de visualización ha generado una ventaja competitiva para Walmart al permitirles actuar rápidamente en función de datos precisos y en tiempo real, maximizando tanto la satisfacción del cliente como los márgenes de ganancia.

2. Desarrollo

2.1. El Impacto de la ciencia de datos en la industria financiera

La ciencia de datos ha transformado significativamente la industria financiera, impulsando mejoras en la toma de decisiones, la gestión de riesgos y la personalización de productos financieros. A medida que el volumen de datos financieros ha crecido exponencialmente, los bancos y otras instituciones financieras han adoptado herramientas avanzadas de análisis de datos para procesar y extraer valor de esta enorme cantidad de información.

Una de las principales áreas donde la ciencia de datos ha marcado una diferencia es en la gestión de riesgos. Antes, los bancos dependían de modelos tradicionales para evaluar la capacidad crediticia de los clientes, basándose en información limitada como ingresos y historial crediticio. Hoy en día, los algoritmos de machine learning permiten a las instituciones financieras analizar grandes cantidades de datos no estructurados, como el comportamiento en redes sociales o patrones de consumo, lo que lleva a decisiones de crédito más informadas y precisas.

Además, la detección de fraudes ha mejorado considerablemente con el uso de la ciencia de datos. Algoritmos sofisticados analizan transacciones en tiempo real y detectan patrones anómalos que podrían indicar actividades fraudulentas. Gracias a estas herramientas, los bancos pueden reaccionar más rápido y prevenir pérdidas significativas.

El aumento de los datos también ha permitido la personalización de servicios financieros. A través del análisis de datos de comportamiento, las instituciones pueden ofrecer productos

financieros más personalizados, como préstamos a medida o recomendaciones de inversión basadas en el perfil de riesgo y objetivos financieros de cada cliente.

2.2. **Informe del análisis descriptivo del dataset Iris**

En este ejercicio, analizamos el dataset de Iris, uno de los conjuntos de datos más comunes en visualización de datos. Usamos Matplotlib para crear un gráfico de dispersión de la longitud y anchura del sépalo. En este gráfico, cada punto representa una flor de iris, y los colores indican las distintas especies. Este gráfico permite observar las diferencias entre especies en términos de tamaño de sépalo, mostrando una clara agrupación que facilita la identificación de cada especie.

Luego, creamos un histograma para mostrar la distribución de la longitud del sépalo. Este gráfico permite observar cómo se distribuyen las medidas dentro de las tres especies, mostrando que la longitud del sépalo tiende a concentrarse en un rango específico, lo cual es útil para clasificar las especies.

A continuación, utilizamos Seaborn para generar un gráfico de pares, que muestra relaciones entre todas las combinaciones de variables. Este tipo de visualización es excelente para identificar correlaciones y patrones. Por último, creamos un mapa de calor de la correlación entre las variables. Este mapa de calor muestra cómo las variables (longitud y anchura de sépalo y pétalo) están relacionadas entre sí, siendo particularmente útil para detectar relaciones fuertes o débiles entre características.

Estas visualizaciones no solo permiten una mejor comprensión de las características del dataset de Iris, sino que también facilitan la identificación de patrones y correlaciones que serían difíciles de detectar solo con análisis numéricos.

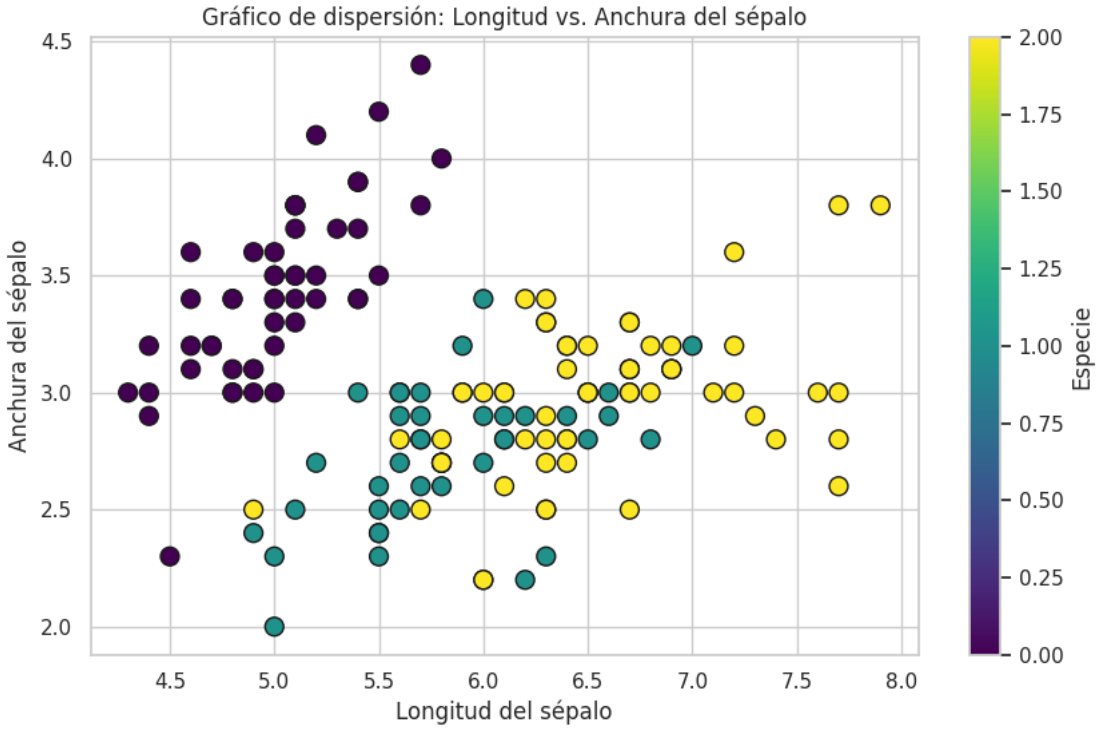
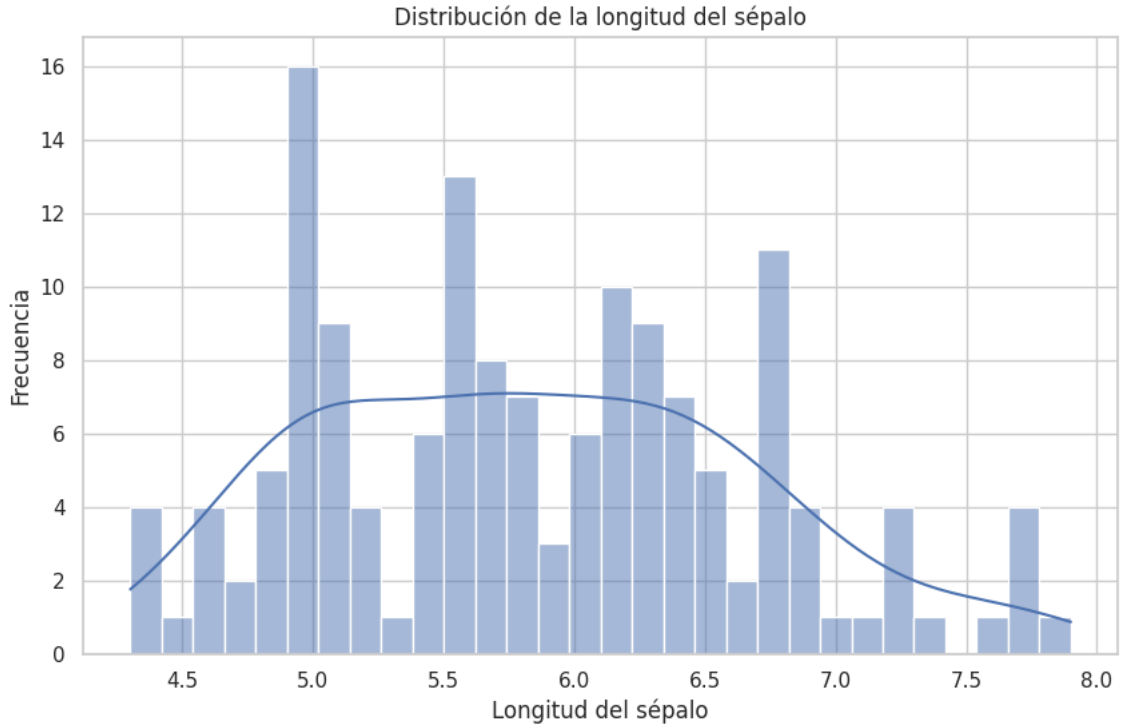


Gráfico de dispersión: Longitud vs. Anchura del sépalo
Anchura del sépalo
Longitud del sépalo
Especie
4.5
4.0
3.5
3.0
2.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00


Distribución de la longitud del sépalo
Frecuencia
Longitud del sépalo
16
14
12
10
8
6
4
2
0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

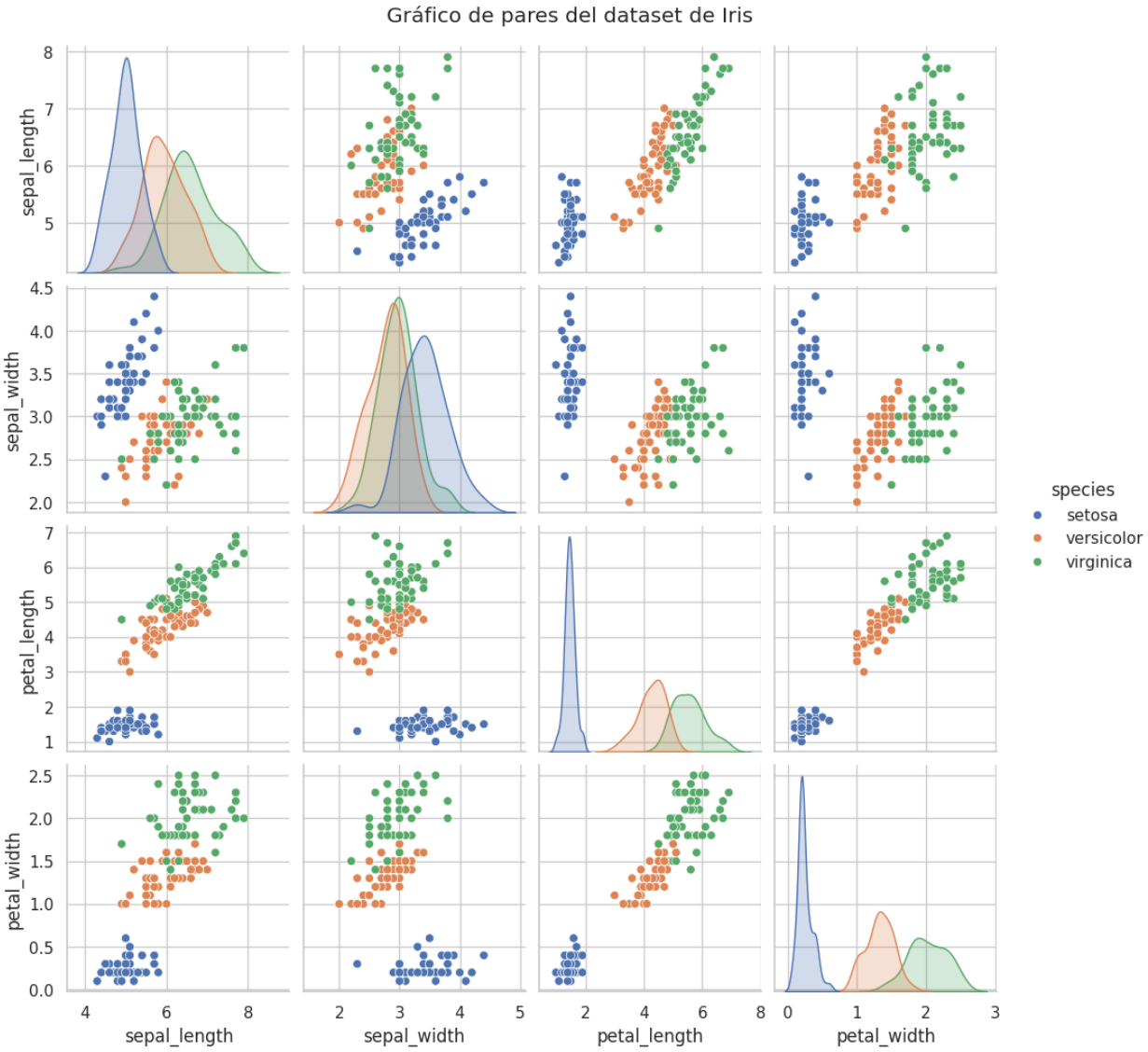
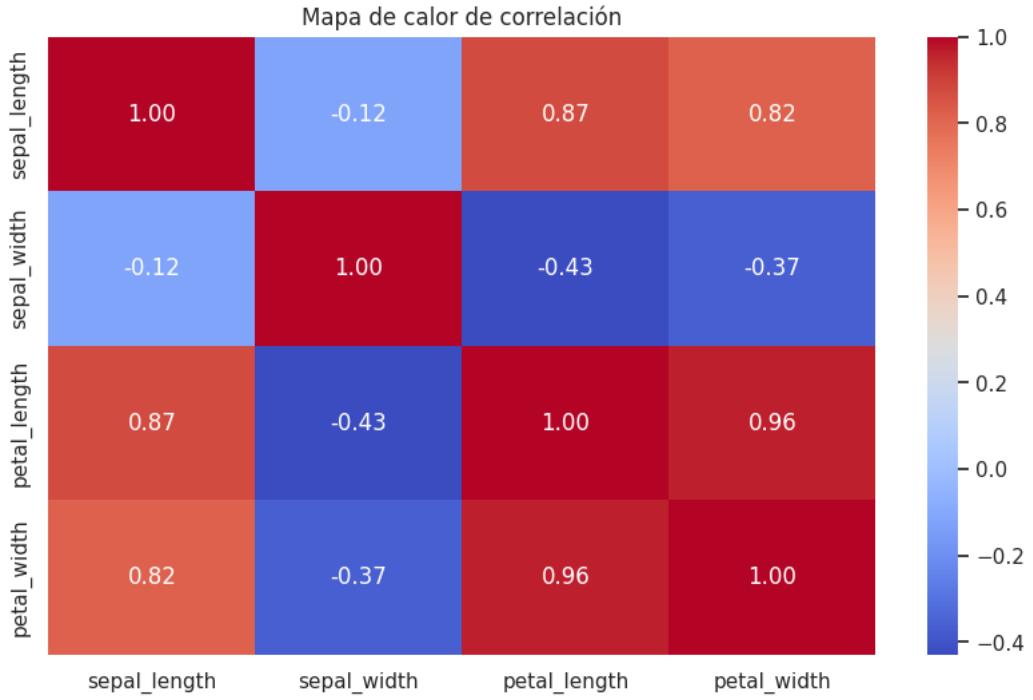


Gráfico de pares del dataset de Iris
sepal_length
sepal_width
petal_length
petal_width
species
setosa
versicolor
virginica



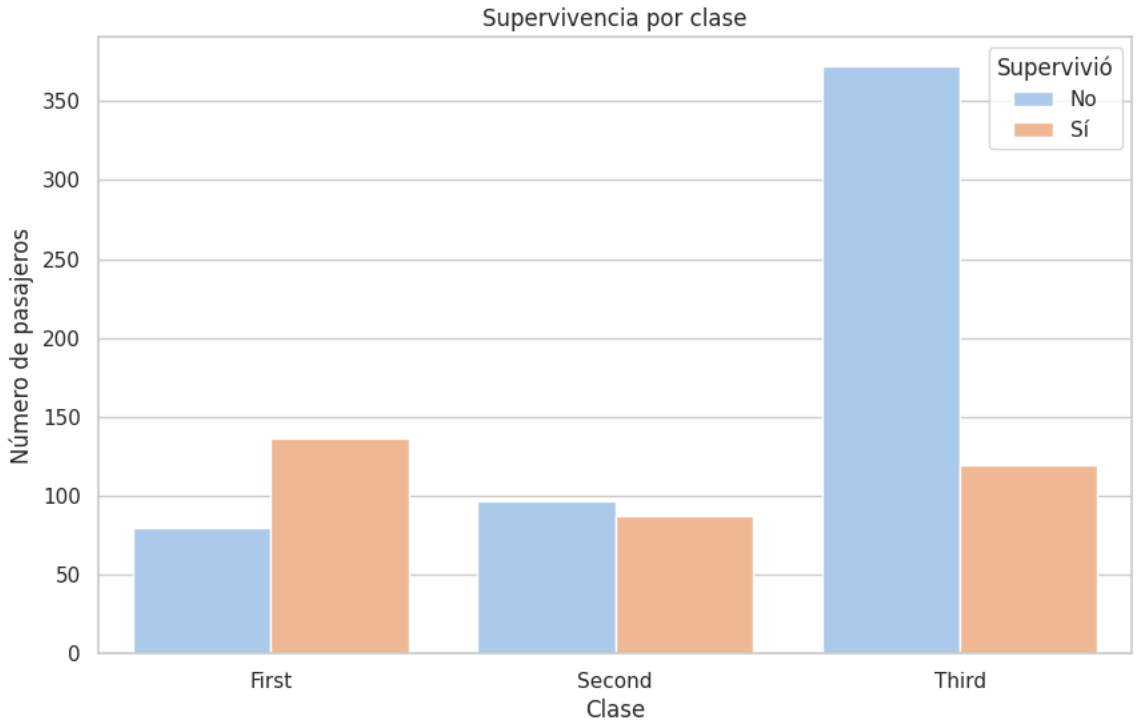
2.3. **Informe algoritmo de optimización simple del dataset Titanic**

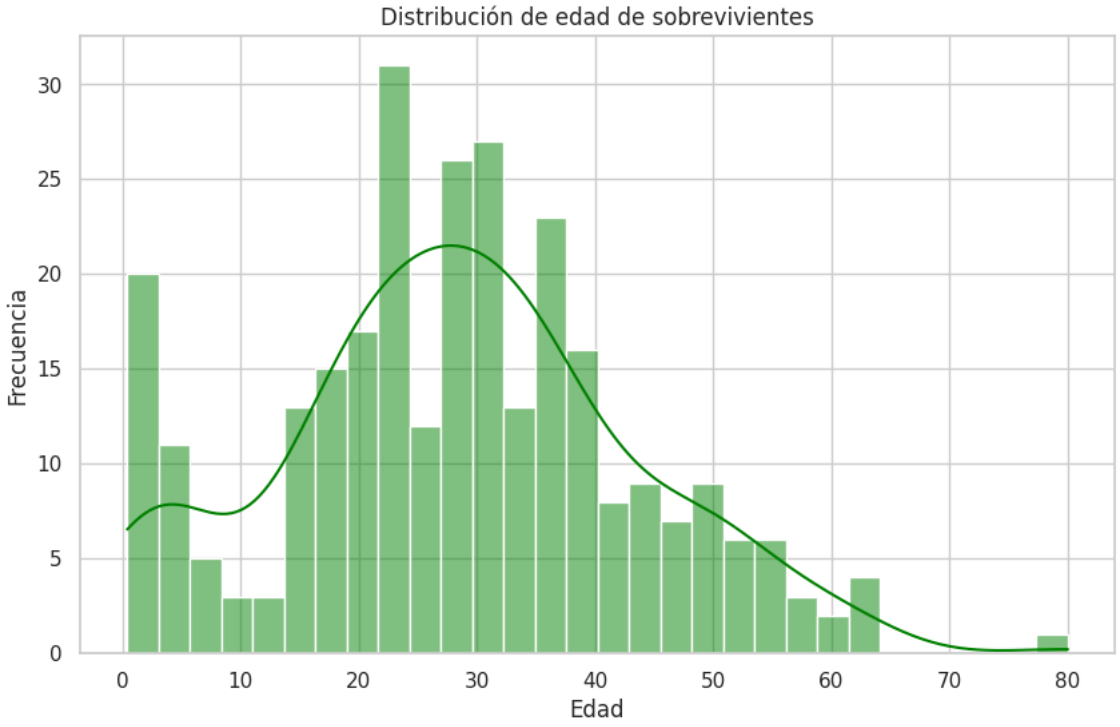
La visualización de datos es fundamental tanto en Ciencia de Datos como en Big Data, aunque sus enfoques y técnicas pueden variar. Usando el dataset de Titanic, observamos algunas de estas diferencias. En Ciencia de Datos, se tiende a trabajar con conjuntos de datos estructurados y más manejables. Para este análisis, usamos gráficos de conteo y distribuciones para visualizar aspectos como la supervivencia por clase y género. Estos gráficos, generados con Seaborn, permiten obtener información precisa y detallada sobre el comportamiento de variables específicas dentro del dataset, lo que facilita el análisis en entornos académicos o de investigación.

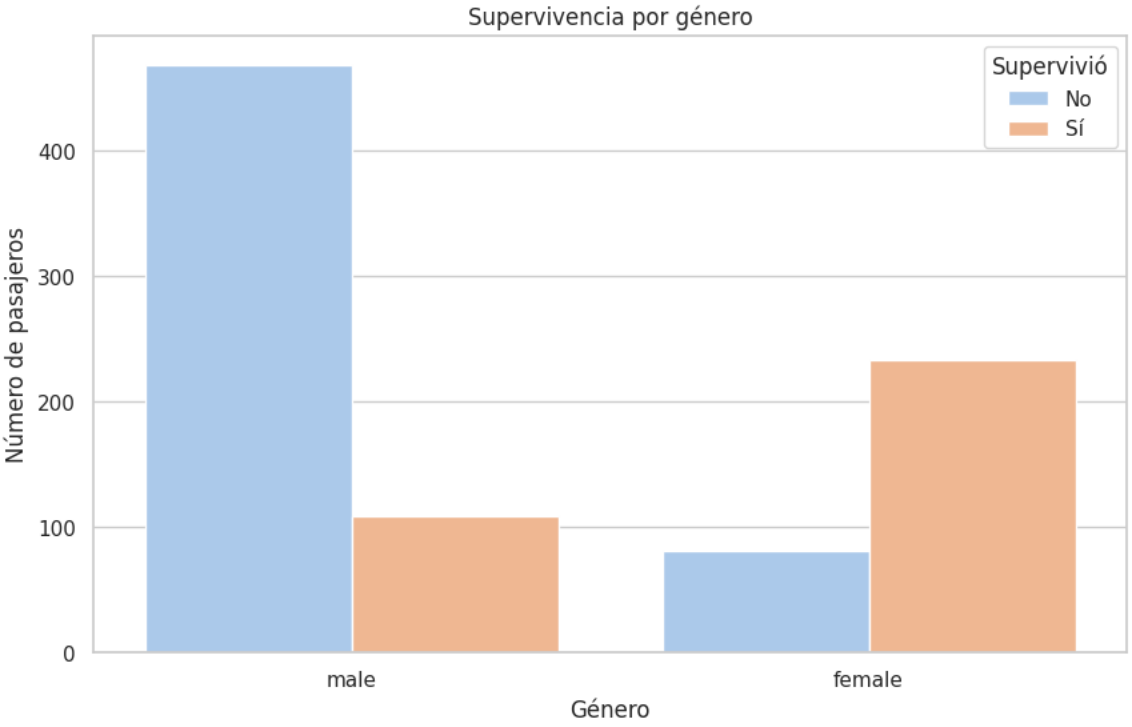
En contraste, la visualización en Big Data suele manejar volúmenes mucho mayores y menos estructurados de información. Las herramientas para visualización en Big Data a menudo se orientan a procesar y presentar datos en tiempo real y a una escala mayor, mediante sistemas distribuidos como Apache Hadoop o Spark. Estas herramientas permiten que, en vez de analizar

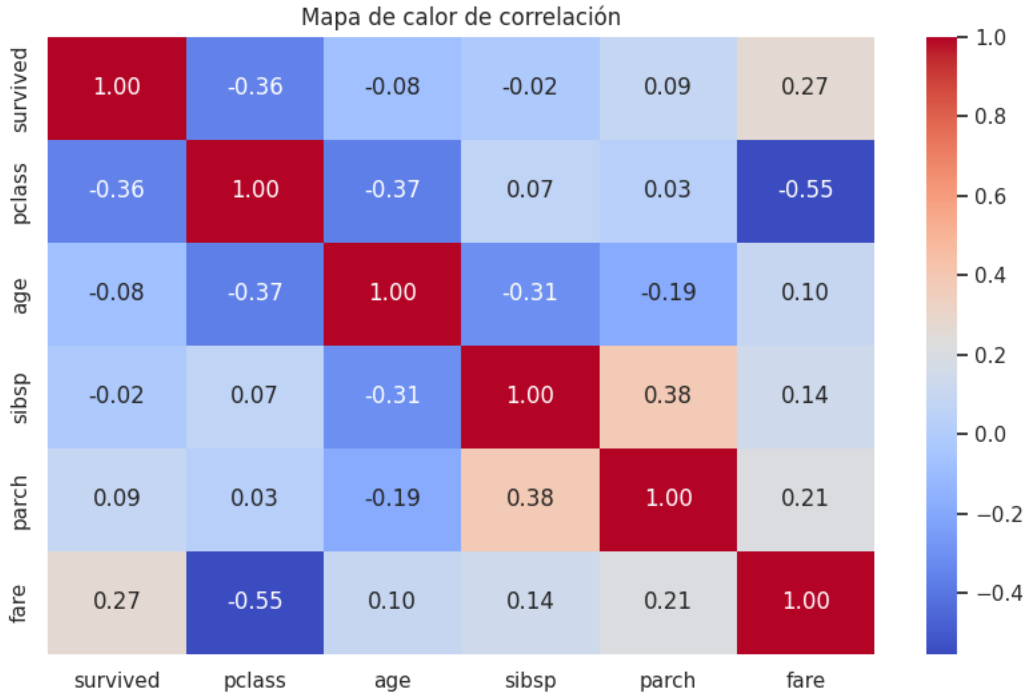
variables específicas, se pueda analizar el comportamiento masivo de datos con técnicas como dashboards en tiempo real o mapas de calor de grandes volúmenes de información.

Mientras que en Ciencia de Datos el enfoque está en el análisis profundo de un conjunto limitado de datos, en Big Data se prioriza la velocidad y el procesamiento de grandes volúmenes, lo cual es crucial para aplicaciones que requieren decisiones en tiempo real, como la detección de fraudes o el análisis de redes sociales.




Distribución de edad de sobrevivientes
Frecuencia
Edad
0
5
10
15
20
25
30
0
10
20
30
40
50
60
70
80


Supervivencia por género
Supervivió
No
Sí
Número de pasajeros
Género
0
100
200
300
400
male
female


Mapa de calor de correlación
survived
pclass
age
sibsp
parch
fare
1.00 -0.36 -0.08 -0.02 0.09 0.27
-0.36 1.00 -0.37 0.07 0.03 -0.55
-0.08 -0.37 1.00 -0.31 -0.19 0.10
-0.02 0.07 -0.31 1.00 0.38 0.14
0.09 0.03 -0.19 0.38 1.00 0.21
0.27 -0.55 0.10 0.14 0.21 1.00
survived
pclass
age
sibsp
parch
fare
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

3. **Conclusiones**

- En conclusión, la evolución de la ciencia de datos ha permitido a la industria financiera optimizar sus operaciones, mejorar la experiencia del cliente y gestionar riesgos de manera más efectiva. Con el continuo crecimiento de los datos y el avance tecnológico, es probable que la ciencia de datos siga desempeñando un papel crucial en el futuro de las finanzas, además ha demostrado ser un factor clave en la transformación y modernización de la industria financiera.
- La visualización de datos es una herramienta fundamental en la toma de decisiones estratégicas. Su implementación en el sector de ventas al por menor, por ejemplo, permite detectar patrones de consumo, optimizar inventarios, y mejorar la satisfacción del cliente. Este caso demuestra que la visualización de datos permite a las empresas actuar de manera más rápida y precisa, otorgándoles una ventaja competitiva al reducir costos y maximizar la eficiencia operativa.
- El análisis del dataset de Iris utilizando Matplotlib y Seaborn mostró que la visualización de datos facilita la comprensión de relaciones y patrones en un conjunto de datos. Las técnicas de dispersión, histogramas, gráficos de pares y mapas de calor permitieron identificar características únicas y correlaciones entre las variables de cada especie de flor. Este ejercicio confirma que la visualización es crucial en Ciencia de Datos para detectar patrones y tomar decisiones fundamentadas sobre los datos.
- Comparar la visualización de datos en Ciencia de Datos y Big Data revela que aunque ambos campos utilizan visualización para entender los datos, sus enfoques difieren significativamente. La Ciencia de Datos se enfoca en el análisis profundo de conjuntos de datos manejables, mientras que Big Data se orienta a procesar grandes volúmenes en

tiempo real. La visualización en Big Data permite tomar decisiones rápidas en escenarios de alto impacto, como detección de fraudes o monitoreo de redes sociales, demostrando que cada enfoque tiene un rol específico según la naturaleza y escala de los datos.

4. **Bibliografía**

Iuvity. (s. f.). Ciencia de datos: ¿conoces su aporte al sector financiero? *iuvity — TODO1 Services Inc. DBA iuvity*. https://www.iuvity.com/es/blog/ciencia-de-datos-conoces-su-aporte-al-sector-financiero#:~:text=La%20ciencia%20de%20datos%20tiene,comprensi%C3%B3n%20de%20los%20procesos%20econ%C3%B3micos.

Gabayet, C. (2024, 3 junio). *Ciencia de datos: ¿Cómo los datos están revolucionando el sector financiero?* DigDash. https://www.digdash.com/es/news-articles-es/business-intelligence-es/ciencia-de-datos-como-los-datos-estan-revolucionando-el-sector-financiero/

Joeportilla. (2023, 5 abril). *Analisis Exploratorio de Datos dataset Iris*. Kaggle. https://www.kaggle.com/code/joeportilla/analisis-exploratorio-de-datos-dataset-iris

Torres, A. (2023, 19 mayo). *Descenso de gradiente: ejemplo de algoritmo de aprendizaje automático*. freeCodeCamp.org. https://www.freecodecamp.org/espanol/news/descenso-de-gradiente-ejemplo-de-algoritmo-de-aprendizaje-automaticod/