CIENCIA DE DATOS

VISUALIZACION DE DATOS REALIZADA

EDISON ANDRES FORERO RIAÑO

DAMIAN DANILO NARAJO PERILLA

ESCUELA TECNOLÓGICA INSTITUTO TÉCNICO CENTRAL

FACULTAD DE INGENIERÍA

TECNOLOGÍA EN DESARROLLO DE SOFTWARE

BOGOTÁ D.C

2024

Informe sobre la Visualización de Datos Utilizando el Dataset de Iris y Titanic

La visualización de datos es una herramienta fundamental en el campo de la Ciencia de Datos, ya que permite a los analistas e investigadores interpretar y comunicar información compleja de manera efectiva y accesible. Este informe aborda la visualización de datos utilizando dos conjuntos de datos significativos: el dataset de Iris y el dataset de Titanic. Se emplearon diversas técnicas y métodos de visualización, tales como gráficos de dispersión, gráficos de pares, mapas de calor y visualizaciones específicas para analizar la supervivencia en el Titanic. A continuación, se presenta un análisis detallado de cada dataset, los métodos utilizados y los resultados obtenidos.

Visualización del Dataset de Iris

El dataset de Iris es ampliamente reconocido en el análisis de datos y clasificación de especies. Contiene información sobre las dimensiones de los sépalos y pétalos de tres especies de Iris: Iris-setosa, Iris-versicolor e Iris-virginica. Para visualizar estos datos, se utilizaron herramientas como Matplotlib y Seaborn en Google Colab, facilitando así la creación de gráficos de dispersión y gráficos de pares.

Los gráficos de dispersión fueron cruciales para observar la relación entre las dimensiones de los sépalos y pétalos. Se identificó un patrón claro en el que Iris-setosa se separa notablemente de las otras dos especies. Sus dimensiones de sépalo y pétalo son significativamente más pequeñas, lo que facilita su identificación. Esta diferenciación resalta la efectividad de las visualizaciones al permitir que los analistas reconozcan rápidamente las características distintivas de cada especie.

Además, al analizar los gráficos de pares, se evidenció una fuerte correlación positiva entre la longitud y el ancho del pétalo. Este hallazgo sugiere que a medida que aumenta la longitud del pétalo, también lo hace su ancho, especialmente en las especies Iris-versicolor e Iris-virginica. Este tipo de análisis permite a los científicos de datos no solo entender las relaciones entre las variables, sino también formular hipótesis sobre cómo estas dimensiones podrían influir en la clasificación de las especies.

Para complementar el análisis, se utilizó un mapa de calor que representa la matriz de correlación entre las variables numéricas del dataset. Este mapa reveló

correlaciones significativas, especialmente entre la longitud y el ancho del pétalo, así como entre la longitud y el ancho del sépalo. Estas correlaciones indican que las dimensiones del pétalo son más discriminativas al clasificar las especies, lo que puede ser útil para futuros análisis y modelos de clasificación. La visualización en este caso no solo destaca la importancia de las dimensiones del pétalo, sino que también puede guiar futuras investigaciones en el área de la botánica y la ecología.

Visualización del Dataset de Titanic

El segundo conjunto de datos analizado fue el del Titanic, que incluye información sobre los pasajeros y su tasa de supervivencia. Este dataset es fundamental para comprender cómo diferentes factores, como la clase, género y edad, influyeron en la supervivencia durante el hundimiento del barco. Al igual que con el dataset de Iris, se utilizaron Matplotlib y Seaborn para generar gráficos de barras, histogramas y gráficos de dispersión.

Los gráficos de barras permitieron observar las tasas de supervivencia por clase de pasajero. Se encontró que los pasajeros de primera clase tenían una tasa de supervivencia significativamente mayor en comparación con aquellos de segunda y tercera clase. Este hallazgo resalta la influencia del estatus socioeconómico en la probabilidad de supervivencia durante el desastre. Los gráficos de barras también facilitaron la comparación visual de la cantidad de sobrevivientes y no sobrevivientes en cada clase, lo que es esencial para el análisis estadístico y la toma de decisiones informadas.

Además, los histogramas de edad revelaron que los niños tenían una tasa de supervivencia más alta, lo que indica que se tomaron medidas para rescatar a los más jóvenes. Este descubrimiento subraya la importancia de la edad como variable crítica en la supervivencia. Las visualizaciones en este contexto no solo ayudan a identificar tendencias, sino que también ofrecen una perspectiva histórica sobre cómo se valoraban las vidas en situaciones de crisis.

Resultados y Conclusiones

Los resultados obtenidos a partir de las visualizaciones del dataset de Iris y Titanic proporcionan información valiosa para la toma de decisiones y la formulación de hipótesis. En el caso del Iris, la separación clara entre las especies y las correlaciones significativas entre las dimensiones del pétalo y sépalo facilitan el desarrollo de modelos de clasificación más precisos. Este tipo de análisis puede ser

crucial en aplicaciones prácticas, como en la conservación de especies y la investigación en genética.

Por otro lado, el análisis de los datos del Titanic permite comprender cómo diversos factores influyeron en la supervivencia de los pasajeros, lo que puede ser útil para estudios históricos y análisis sociales. La visualización de datos se convierte así en una herramienta esencial para el análisis de tendencias y la toma de decisiones basadas en evidencias.

En conclusión, la visualización de datos no solo facilita la interpretación de información compleja, sino que también permite identificar patrones y relaciones significativas que pueden influir en la toma de decisiones. Las técnicas y métodos utilizados en este informe, como gráficos de dispersión, gráficos de pares y mapas de calor, demostraron ser herramientas efectivas para explorar y comunicar los hallazgos obtenidos en ambos datasets. Estos análisis no solo enriquecen nuestra comprensión de los datos, sino que también brindan un fundamento sólido para futuros estudios en el ámbito de la Ciencia de Datos. Además, resaltan la importancia de la visualización como un componente crítico en el análisis y la comunicación de información relevante en diferentes contextos.