Proyecto NZES

Esquema Dataset

1. Análisis Exploratorio de Datos

Lo primero que se debe realizar es explorar el conjunto de datos, esto para entender su estructura, distribución y características. Así se podrán identificar anomalías, patrones e incluso tendencias en los datos. Estos ayudaran en la toma de decisiones.

Este tipo de análisis puede darse en dos tipos, el análisis univariado y el análisis multivariado. En la primera se enfoca solo en una variable, para poder comprender su distribución y características. Se suelen usar histogramas, bloxplots, gráficos de densidad. Junto a las medidas estadísticas descriptivas, como lo son media, mediana y desviación estándar.

En el segundo caso, se enfoca en la relación de varias variables, entendiendo la relación de estas entre sí. Se suelen utilizar gráficos de dispersión, matrices de correlación, análisis de componentes principales y análisis de conglomerados. Junto a esto, se suelen responder las siguientes preguntas:

- ¿Cuál es el rango de valores de las variables?
- ¿Qué variables tienen la mayor varianza o desviación estándar?
- ¿Hay valores atípicos o outliers en los datos?
- ¿Cuáles son las relaciones entre las variables?
- ¿Hay alguna variable que esté altamente correlacionada con otra?
- ¿Cómo se distribuyen los datos en cada variable?

2. Visualizaciones

Las visualizaciones son las herramientas que ayudan en el análisis de datos, mostrando la información de maneras sencillas de comprender, pero que permita la identificación de tendencias, patrones y anomalías. Muy similar a lo mencionado anteriormente se suelen utilizar los siguientes gráficos:

- Gráficos de barras: Este tipo de gráfico se utiliza para mostrar la distribución de una variable categórica. Cada barra representa la frecuencia o proporción de cada categoría en la variable.
- Gráficos de dispersión: Este tipo de gráfico se utiliza para mostrar la relación entre dos variables numéricas. Cada punto en el gráfico representa una observación y la posición en el eje x e y representa los valores de las dos variables.
- Histogramas: Este tipo de gráfico se utiliza para mostrar la distribución de una variable numérica. Los datos se agrupan en intervalos y la altura de cada barra representa la frecuencia o proporción de observaciones en cada intervalo.
- Gráficos de líneas: Este tipo de gráfico se utiliza para mostrar la evolución de una variable numérica a lo largo del tiempo. Cada punto en el gráfico representa una observación y las líneas conectan los puntos para mostrar la tendencia de la variable a lo largo del tiempo.
- Bloxplot – Diagrama de caja y bigotes: Este tipo de gráfico se utiliza para mostrar la distribución de una variable numérica y detectar valores atípicos. El diagrama muestra la mediana, el rango intercuartílico y los valores mínimos y máximos.

3. Modelos Estadísticos

Permiten analizar datos para entender las relaciones entre variables y hacer predicciones. Se utilizan para construir un modelo que relacione una o más variables de entrada (predictores) con una variable de salida (respuesta). Permiten entender las relaciones entre las variables y predecir valores futuros. Los más usuales son:

- Regresión lineal: Es un modelo estadístico que se utiliza para modelar la relación entre una variable de respuesta y una o más variables predictoras. El objetivo de la regresión lineal es encontrar la mejor línea recta que pueda predecir la variable de respuesta a partir de las variables predictoras.
- Regresión logística: Es un modelo estadístico que se utiliza para modelar la relación entre una variable binaria de respuesta y una o más variables predictoras. El objetivo de la regresión logística es encontrar la mejor curva logística que pueda predecir la variable binaria de respuesta a partir de las variables predictoras.
- Análisis de varianza (ANOVA): Es un modelo estadístico que se utiliza para determinar si hay diferencias significativas entre las medias de dos o más grupos.
- Análisis de componentes principales (PCA): Es un modelo estadístico que se utiliza para reducir la dimensionalidad de los datos mediante la identificación de las variables principales que explican la variabilidad en los datos.
- Análisis de conglomerados (clustering): Es un modelo estadístico que se utiliza para agrupar observaciones similares en grupos o clústeres.

4. Machine Learning

Campo de la inteligencia artificial enfocado en desarrollar algoritmos y modelos, que, a partir de unos datos, permite generar predicciones o clasificaciones. Las principales técnicas que se suelen usar son:

- Árboles de decisión: Se utilizan para la clasificación y la regresión. El modelo se construye mediante la partición recursiva del conjunto de datos en subconjuntos más pequeños, utilizando una serie de reglas de decisión que se basan en las características del conjunto de datos. El resultado final es un árbol que puede ser utilizado para hacer predicciones sobre nuevos datos.
- Bosques aleatorios: Se utilizan para la clasificación y la regresión. El modelo se construye a partir de múltiples árboles de decisión, donde cada árbol se entrena con una muestra aleatoria del conjunto de datos y con un subconjunto aleatorio de las características. Luego, el resultado final se obtiene mediante la combinación de las predicciones de todos los árboles.
- Regresión logística: Se utiliza para la clasificación binaria. El modelo utiliza una función logística para predecir la probabilidad de que una observación pertenezca a una de las dos clases posibles. A partir de esta probabilidad, se puede asignar una clase a la observación.
- Redes neuronales: Se utilizan para la clasificación y la regresión. El modelo se construye a partir de múltiples capas de neuronas, donde cada neurona se activa mediante una función matemática y se conecta con las neuronas de la capa siguiente. La última capa de neuronas produce la salida del modelo.
- Análisis de componentes principales (PCA): Se utiliza para la reducción de la dimensionalidad. El objetivo es encontrar las características más importantes del conjunto de datos, reduciendo la cantidad de variables que se utilizan para representar los datos. Esto se logra mediante la transformación de las variables originales en un conjunto de nuevas variables, llamadas componentes principales, que explican la mayor parte de la varianza del conjunto de datos.