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:

- ANOVA (Análisis de Varianza): El análisis de varianza se utiliza para determinar si hay diferencias significativas entre las medias de tres o más grupos. Puede ser útil en el análisis de datos para comparar los resultados de diferentes tratamientos o condiciones en un experimento o estudio. Por ejemplo, si se está investigando el efecto de diferentes dosis de un medicamento en un grupo de pacientes, el ANOVA podría utilizarse para determinar si hay diferencias significativas en los resultados de los pacientes que recibieron diferentes dosis.
- Regresión lineal: La regresión lineal se utiliza para modelar la relación entre una variable dependiente y una o más variables independientes. Puede ser útil en el análisis de datos para identificar las variables que tienen una mayor influencia en un resultado o resultado deseado. Por ejemplo, si se está estudiando la relación entre el salario y la educación, la regresión lineal podría utilizarse para determinar la influencia de la educación en el salario de una persona.
- Regresión logística: La regresión logística se utiliza para modelar la relación entre una variable dependiente binaria y una o más variables independientes. Puede ser útil en el análisis de datos para predecir la probabilidad de que un evento ocurra o para identificar los factores que influyen en la ocurrencia de un evento. Por ejemplo, si se está investigando el efecto de la edad, el género y el historial médico en el riesgo de desarrollar una enfermedad, la regresión logística podría utilizarse para identificar los factores que tienen una mayor influencia en el riesgo.

- Análisis de series de tiempo: El análisis de series de tiempo se utiliza para modelar la evolución de una variable a lo largo del tiempo. Puede ser útil en el análisis de datos para identificar patrones y tendencias a lo largo del tiempo y para predecir futuros valores de la variable. Por ejemplo, si se está analizando la evolución de las ventas de un producto a lo largo del tiempo, el análisis de series de tiempo podría utilizarse para identificar patrones estacionales en las ventas y para predecir las ventas futuras.
- Análisis de cluster: El análisis de cluster se utiliza para agrupar conjuntos de observaciones similares en grupos o clusters. Puede ser útil en el análisis de datos para identificar patrones y grupos en los datos y para identificar las características que definen cada grupo. Por ejemplo, si se está analizando la satisfacción del cliente en una empresa, el análisis de cluster podría utilizarse para identificar grupos de clientes con características similares y para identificar los factores que influyen en la satisfacción del cliente en cada grupo.
- Análisis de componentes principales (PCA): El análisis de componentes principales se utiliza para reducir la dimensionalidad de un conjunto de datos con múltiples variables. Puede ser útil en el análisis de datos para identificar las variables más importantes en un conjunto de datos y para visualizar la estructura de los datos en un espacio de menor dimensión. Por ejemplo, si se tiene un conjunto de datos con muchas variables relacionadas con la salud de un paciente, el PCA podría utilizarse para identificar las variables más importantes que influyen en la salud del paciente y para visualizar la relación entre estas variables en un espacio de dos o tres dimensiones.
- Análisis discriminante: El análisis discriminante se utiliza para clasificar observaciones en dos o más grupos en función de las características observadas. Puede ser útil en el análisis de datos para identificar las características que distinguen a los diferentes grupos y para predecir la pertenencia de nuevas observaciones a un grupo determinado. Por ejemplo, si se tiene un conjunto de datos con información sobre los clientes de una

empresa, el análisis discriminante podría utilizarse para identificar las características que distinguen a los clientes fieles de los que no lo son y para predecir la lealtad de nuevos clientes.

- Análisis de correlación: El análisis de correlación se utiliza para medir la relación entre dos o más variables. Puede ser útil en el análisis de datos para identificar las variables que están fuertemente relacionadas entre sí y para entender la naturaleza de esta relación. Por ejemplo, si se está estudiando la relación entre el nivel de educación y el ingreso, el análisis de correlación podría utilizarse para determinar si hay una relación positiva o negativa entre estas dos variables y para medir la fuerza de esta relación.
- Análisis de supervivencia: El análisis de supervivencia se utiliza para analizar el tiempo hasta que ocurre un evento. Puede ser útil en el análisis de datos para entender la duración de eventos importantes, como la supervivencia de pacientes con una enfermedad determinada o el tiempo que tarda una empresa en alcanzar un cierto nivel de rentabilidad. El análisis de supervivencia puede ayudar a identificar los factores que influyen en la duración del evento y a predecir cuánto tiempo le tomará a un nuevo evento alcanzar cierto nivel de duración.
- Análisis de regresión no lineal: El análisis de regresión no lineal se utiliza para modelar la relación entre una variable dependiente y una o más variables independientes cuando la relación no es lineal. Puede ser útil en el análisis de datos para entender cómo cambia la variable dependiente en función de las variables independientes y para identificar patrones no lineales en los datos. Por ejemplo, si se está estudiando la relación entre la temperatura y la tasa de crecimiento de una planta, el análisis de regresión no lineal podría utilizarse para identificar patrones en la tasa de crecimiento que no son lineales en función de la temperatura.

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.