TIC

TALENTO TECH

Misión 1

INTELIGENCIA ARTIFICIAL

Explorador

Tema 6: Análisis Exploratorios de Datos

iTraining | UNIVERSIDAD DE ANTIOQUIA | Universidad de Caldas | ubicua Technology

Campista, llegó el momento de retar tus conocimientos y que los pongas aprueba a través de los diferentes recursos que encontraras en este espacio como son: conceptos, ejemplos, herramientas, actividades prácticas y retos, los cuales te ayudaran alcanzar los objetivos trazados en el nivel explorador.

Calidad de datos en el Ciclo de Vida de una aplicación de Machine Learning

La visualización de datos es una herramienta esencial para entender, analizar y comunicar información compleja de manera efectiva. A través de gráficos y representaciones visuales, se pueden identificar patrones, tendencias y anomalías en los datos que podrían no ser evidentes de otra manera. A continuación, se describen algunos tipos comunes de gráficos utilizados en la visualización de datos, junto con sus aplicaciones y ejemplos.

https://youtu.be/_-KW4gT_oGU

DIMENSIONES DE LA CALIDAD DE DATOS

- **Complentitud**

- *Descripción:* Mide si todos los datos necesarios están presentes en el conjunto de datos.

  Ejemplo:

  En un proyecto de análisis de consumo energético, un conjunto de datos completo debería incluir información sobre el consumo energético diario de cada hogar, así como datos sobre el tipo de vivienda, el número de ocupantes, y el clima. Si faltan datos sobre algún aspecto importante, el análisis podría estar incompleto.

- **Consistencia**

- *Descripción:* Se refiere a la coherencia de los datos en todo el conjunto de datos.

  Ejemplo:

En un proyecto de evaluación de la eficiencia de paneles solares, los datos de generación de energía deben ser consistentes en términos de unidades de medida (por ejemplo, siempre en

kWh). Si algunos datos están en MWh, habrá inconsistencias que podrían llevar a errores en el análisis.

1. Exactitud

- *Descripción:* Mide si los datos representan correctamente la realidad.

Ejemplo:

En un estudio sobre la producción de hidrógeno verde, los datos sobre la cantidad de hidrógeno producido deben ser exactos y reflejar la producción real. Si los datos están errados, por ejemplo, debido a errores en la calibración de los sensores, los resultados del estudio serán incorrectos.

DETECCIÓN Y TRATAMIENTO DE DATOS AUSENTES

Reemplazo de valores ausentes con valores estimados. Las técnicas incluyen:

- **Media o Mediana**

En un conjunto de datos sobre el rendimiento de diferentes tipos de turbinas eólicas, si algunos registros de eficiencia están ausentes, se puede reemplazar esos valores con la media o mediana de los registros disponibles para mantener el conjunto de datos utilizable.

- **Regresión**

Si faltan datos sobre el consumo energético en ciertos periodos, se puede usar un modelo de regresión que tenga en cuenta variables como la temperatura y el día de la semana para predecir los valores ausentes basados en los registros completos.

- **Hot Deck**

Ejemplo: En un conjunto de datos de encuestas sobre la aceptación de tecnologías limpias, si algunos encuestados no proporcionaron información sobre su nivel de satisfacción, se puede imputar esos valores utilizando la información de encuestados con características similares (por ejemplo, en la misma región).

- Videos complementarios
    - **¿Cómo manejar datos faltantes?** https://youtu.be/ARwHkq4t2q0
    - **Tutorial: MANEJO DE DATOS CATEGÓRICOS FALTANTES con Python, Pandas y Scikit-Learn** https://youtu.be/G3tNCSQUoXw

NORMALIZACIÓN DE DATOS

- **MIN - MAX**

Escala los datos a un rango específico, generalmente [0, 1].

- *Ejemplo:* En un proyecto de comparación de la eficiencia de diferentes tipos de celdas solares, los datos de eficiencia pueden ser escalados a un rango de [0, 1] para facilitar la comparación entre celdas con diferentes rangos de eficiencia.

- **Robust**

Utiliza los valores de los cuartiles para escalar los datos, reduciendo la influencia de los valores atípicos.

- *Ejemplo:* En un proyecto de comparación de la eficiencia de diferentes tipos de celdas solares, los datos de eficiencia pueden ser escalados a un rango de [0, 1] para facilitar la comparación entre celdas con diferentes rangos de eficiencia.

- Video complementario
    - **Escalamiento, Normalización y Estandarización de Datos con Python para Ciencia de Datos** https://youtu.be/-VuR14Qyl7E

ANÁLISIS UNIVARIADO

El análisis univariable se centra en el estudio de una sola variable en un conjunto de datos. Se emplea para comprender la distribución, tendencia y dispersión de la variable en cuestión. Aquí están los aspectos clave:

- **Validación de Distribución normal**

Este paso evalúa si los datos de una variable siguen una distribución normal (gaussiana). La normalidad es una suposición común en muchos métodos estadísticos y modelos. La validación se puede realizar mediante:

  - *Histogramas:* Graficar los datos en un histograma ayuda a visualizar si los datos se distribuyen en forma de campana.
  - *Gráfico Q-Q (Quantile-Quantile):* Compara los cuantiles de los datos con los cuantiles de una distribución normal teórica. Los puntos deben alinearse a lo largo de la línea diagonal si los datos siguen una distribución normal.
  - *Pruebas de Normalidad:* Métodos estadísticos como la prueba de Shapiro-Wilk o la prueba de Kolmogorov-Smirnov pueden determinar si los datos se desvían significativamente de la normalidad.

- **Estadísticas descriptivas**

Las estadísticas descriptivas proporcionan un resumen numérico de las características principales de una variable. Incluyen:

  - *Media:* La medida de tendencia central que indica el valor promedio.
  - *Mediana:* El valor central que divide los datos en dos mitades iguales; útil cuando los datos no están simétricamente distribuidos.
  - *Moda:* El valor que ocurre con mayor frecuencia en el conjunto de datos.
  - *Desviación Estándar:* Mide la dispersión de los datos respecto a la media. Una desviación estándar alta indica mayor dispersión.
  - *Varianza:* El cuadrado de la desviación estándar, también mide la dispersión.

- *Rango Intercuartílico (IQR):* La diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1); mide la dispersión de la mitad central de los datos.
- *Coeficiente de Asimetría:* Indica si los datos están sesgados hacia la derecha o la izquierda.
- *Curtosis:* Mide la "altitud" de la distribución de los datos. Una curtosis alta indica colas pesadas.

- **Video complementario**
  - **The normal distribution, Clearly Explained**

    https://youtu.be/rzFX5NWojp0

  - **Mediana, Media y asimetría en curvas de densidad**

    https://youtu.be/88lpvX0YEso

El análisis univariable es fundamental para entender las características de los datos antes de proceder con análisis más complejos o modelos predictivos.

ANÁLISIS BIVARIADO: CORRELACIÓN

El análisis bivariado se enfoca en estudiar la relación entre dos variables para entender cómo una variable puede afectar o estar relacionada con otra. Uno de los métodos más comunes para llevar a cabo este análisis es la correlación. A continuación, se describe el concepto de correlación, sus tipos y se proporcionan ejemplos relevantes.

- Correlación

La correlación mide la fuerza y la dirección de la relación entre dos variables. Un coeficiente de correlación cuantifica esta relación, con valores que van desde -1 hasta 1.

- *Valor de 1:* Correlación positiva perfecta; cuando una variable aumenta, la otra también aumenta en la misma proporción.

- *Valor de -1:* Correlación negativa perfecta; cuando una variable aumenta, la otra disminuye en la misma proporción.
- *Valor de 0:* No hay correlación; no hay relación lineal entre las variables.

- **Importancia**
  - Permite identificar y cuantificar relaciones lineales entre variables.
  - Ayuda a entender cómo una variable puede influir en otra, lo cual es crucial para la modelización y la toma de decisiones.

- **Video complementario**
  - **Correlación de variables en Machine Learning**

    https://youtu.be/XDE6fWJUo-0

- **Tipos de Correlación**

Correlación Positiva

Ambas variables aumentan o disminuyen juntas.

- *Ejemplo:* En un estudio sobre la generación de energía solar, es probable que haya una correlación positiva entre la cantidad de horas de sol y la producción de energía. A medida que aumentan las horas de sol, la producción de energía solar también aumenta

Correlacion Negativa

Una variable aumenta mientras que la otra disminuye.

- *Ejemplo:* En un análisis de consumo energético en diferentes regiones, puede haber una correlación negativa entre el uso de energía no renovable y la adopción de tecnologías limpias. A medida que aumenta el uso de energías limpias, el consumo de energía no renovable disminuye.

Sin correlación

No hay una relación lineal significativa entre las variables.

- *Ejemplo:* En un estudio sobre el impacto de la tecnología en la calidad del aire, puede no haber una correlación clara entre la cantidad de

dispositivos tecnológicos y la calidad del aire si otros factores como el tráfico y la industria son más influyentes.

- **Medidas de correlación**

Coeficiente de Correlación

Mide la relación lineal entre dos variables continuas.

- *Ejemplo:* Se puede usar el coeficiente de correlación de Pearson para analizar la relación entre la temperatura media y la producción de energía en una planta de energía solar. Un coeficiente cercano a 1 indicaría una fuerte relación positiva.

Coeficiente de Correlación de Spearman

Mide la relación monotónica entre dos variables, no necesariamente lineal.

- *Ejemplo:* En el análisis de la relación entre la clasificación de eficiencia de diferentes tecnologías de energía y su adopción en el mercado, el coeficiente de Spearman puede ayudar a evaluar si hay una tendencia general a medida que una variable aumenta.

Coeficiente de Correlación de Kendall

Mide la asociación entre dos variables ordinales.

- *Ejemplo:* Puede ser utilizado para analizar la relación entre el ranking de satisfacción del usuario y el nivel de inversión en energías renovables en diferentes comunidades.

ANÁLISIS MULTIVARIADO: ANÁLISIS DE COMPONENTES PRINCIPALES

- **Análisis Multivariado: Análisis de Componentes Principales (PCA)**

El análisis multivariado se utiliza para entender relaciones entre múltiples variables y reducir la complejidad de los datos. Uno de los métodos más importantes en el **análisis multivariado es el Análisis de Componentes Principales (PCA)**. A continuación, se describe el PCA, su propósito, y se proporcionan ejemplos relevantes.

- **Análisis de Componentes Principales (PCA)**

El PCA es una técnica de reducción de dimensionalidad que transforma un conjunto de variables correlacionadas en un conjunto de variables no correlacionadas, conocidas como componentes principales. Estos componentes principales son combinaciones lineales de las variables originales y están ordenados por la cantidad de varianza que explican en los datos.

  - *Componente principal 1 (PC1)*: Captura la mayor parte de la varianza en los datos.
  - *Componente principal 2 (PC2)*: Captura la siguiente mayor parte de la varianza, y es ortogonal a PC1.
  - *Etc.:* Cada componente adicional captura una menor proporción de la varianza restante.

- **Importancia**
  - Reduce la complejidad de los datos al disminuir el número de variables manteniendo la mayor parte de la información.
  - Facilita la visualización y el análisis de datos de alta dimensión.
  - Mejora la eficiencia de los algoritmos de aprendizaje de máquina al reducir el número de características.

- **Video complementario**
  - **StatQuest: Principal Component Analysis (PCA), Step-by-Step**

PROCESO DE PCA

- **Estandarización de Datos**

Antes de aplicar PCA, los datos deben ser estandarizados para que cada variable tenga una media de 0 y una desviación estándar de 1.

- *Ejemplo:* En un estudio sobre la eficiencia de diferentes tipos de paneles solares, las métricas de rendimiento, como la producción de energía y el costo, deben estandarizarse antes de aplicar PCA para asegurar que cada métrica contribuya de manera equitativa al análisis.

- **Cálculo de la Matriz de Covarianza**

Se calcula la matriz de covarianza de los datos estandarizados para entender cómo varían las variables entre sí.

- *Ejemplo:* En el análisis de datos sobre el consumo energético en distintas regiones, la matriz de covarianza ayudará a identificar la relación entre variables como la demanda de energía y la temperatura.

- **Cálculo de los Componentes Principales**

Se calculan los vectores propios (eigenvectors) y los valores propios (eigenvalues) de la matriz de covarianza para obtener los componentes principales.

- *Ejemplo:* En un análisis de factores que afectan la eficiencia de cultivos agrícolas, PCA puede identificar los principales factores (componentes) que explican la variabilidad en los rendimientos.

- **Proyección de Datos en el Espacio de Componentes Principales**

Los datos originales se proyectan en el espacio de los componentes principales para obtener una representación reducida.

- *Ejemplo:* En un estudio de impacto ambiental de diferentes tecnologías de energía, los datos proyectados en el primer y segundo componente principal pueden revelar patrones sobre cómo cada tecnología afecta diferentes indicadores ambientales.

- **Visualización de datos**

PCA ayuda a visualizar datos de alta dimensión en un espacio de menor dimensión, típicamente 2D o 3D.

  - *Ejemplo:* En la evaluación de la transición hacia energías limpias, PCA puede visualizar las relaciones entre múltiples indicadores de sostenibilidad en un gráfico 2D.

- **Reducción de dimensionalidad para Modelos de aprendizaje de Máquina**

PCA puede reducir el número de características en un modelo de aprendizaje de máquina, lo que puede mejorar el rendimiento y reducir el tiempo de computación.

  - *Ejemplo:* En un modelo predictivo de demanda energética, PCA puede reducir el número de variables utilizadas sin perder información clave, mejorando la eficiencia del modelo.

- **Identificación de Patrones y Tendencias**

PCA puede reducir el número de características en un modelo de aprendizaje de máquina, lo que puede mejorar el rendimiento y reducir el tiempo de computación.

  - *Ejemplo:* En un modelo predictivo de demanda energética, PCA puede reducir el número de variables utilizadas sin perder información clave, mejorando la eficiencia del modelo.

- **Identificación de Patrones y Tendencias**

PCA puede revelar patrones subyacentes en los datos que no son evidentes en el espacio de características original.

- *Ejemplo:* En un análisis de datos de calidad del aire, PCA puede identificar patrones de contaminación que son comunes a varias regiones y ayudarlos a enfocar las estrategias de mitigación.

Aprendizajes Prácticos

Estos ejercicios permitirán a los estudiantes aplicar los conceptos aprendidos en situaciones prácticas y específicas.

- **Práctica: Evaluación y Mejora de la Calidad de Datos**

Objetivo: El estudiante aprenderá a evaluar la calidad de un conjunto de datos mediante el análisis de su completitud, consistencia y exactitud. Además, se abordará la detección y tratamiento de datos ausentes utilizando diversas técnicas, y se realizará la normalización de los datos para prepararlos adecuadamente para un modelo de machine learning.

TAREAS

PRÁCTICA: EVALUACIÓN Y MEJORA DE LA CALIDAD DE DATOS

1. DIMENSIONES DE CALIDAD DE DATOS

- *Objetivo:* Evaluar la calidad del conjunto de datos considerando su completitud, consistencia y exactitud.
- *Tareas:*
    - **Completitud: Verificar que no falten datos en ninguna de las variables críticas del conjunto.**
        - *Ejemplo:* En un conjunto de datos sobre consumo de energía, se analiza la variable "Consumo energético mensual" para asegurarse de que no haya valores faltantes, ya que la completitud es esencial para un análisis preciso.

- **Consistencia: Asegurarse de que los datos sean internamente coherentes.**
  - *Ejemplo:* Se revisan los datos de "Horas de uso diario de energía solar" y "Consumo energético mensual". La inconsistencia se detecta si un hogar reporta un uso diario alto de energía solar pero muestra un bajo consumo energético total, lo que indicaría un error en los datos.
- **Exactitud: Confirmar que los datos reflejen fielmente la realidad.**
  - *Ejemplo:* Se comparan los valores reportados de "Costo de instalación de energía solar" con datos oficiales de fuentes gubernamentales para asegurar que los valores sean precisos y no estén subestimados o sobreestimados.

2. DETECCIÓN DE DATOS AUSENTES

- *Objetivo:* Identificar y tratar los datos ausentes en el conjunto, utilizando diferentes técnicas para mejorar la calidad del dataset.
- *Tareas:*
  - **Detección de Datos Ausentes: Identificar qué variables tienen datos faltantes y cuantificar el porcentaje de valores ausentes.**
    - *Ejemplo:* Se detecta que la variable "Ingreso anual del hogar" tiene un 10% de datos ausentes en un conjunto de datos sobre adopción de energías limpias.
  - **Descarte de Datos Ausentes: Considerar eliminar observaciones o variables con muchos valores ausentes.**
    - *Ejemplo:* Se decide descartar la variable "Ocupación del jefe de hogar" porque tiene un 40% de datos ausentes y no es crucial para el análisis de adopción de energías renovables.
  - **Imputación de Datos Ausentes:**
    - *Media o Mediana:* Rellenar los datos faltantes con la media o mediana de la variable..

- *Ejemplo:* Se imputan los valores faltantes de "Ingreso anual del hogar" con la mediana, ya que la variable presenta una distribución sesgada.

- **Regresión: Utilizar un modelo de regresión para predecir los valores faltantes.**
  - *Ejemplo:* Se usa regresión lineal para predecir los valores faltantes de "Consumo energético mensual" basado en variables como "Tamaño del hogar" y "Horas de uso diario de energía".
- **Hot Deck: Imputar valores faltantes utilizando observaciones similares.**
  - *Ejemplo:* Se usa hot deck para imputar los valores faltantes de "Horas de uso diario de energía solar" basándose en hogares similares en la misma región.

3. NORMALIZACIÓN DE DATOS

- *Objetivo:* Escalar los datos para asegurar que todas las variables tengan un rango comparable, mejorando así el rendimiento del modelo de machine learning. (Seleccionar una de las dos técnicas)
- *Tareas:*
  - **Normalización Minmax: Escalar los datos para que todas las variables tengan valores en un rango de 0 a 1.**
    - *Ejemplo:* Se normaliza la variable "Costo de instalación de energía solar" utilizando Minmax, de modo que los costos más altos no dominen el análisis en un modelo de machine learning.
  - **Normalización Robust: Escalar los datos para que las variables sean robustas frente a outliers.**
    - *Ejemplo:* Se aplica normalización Robust a la variable "Ingreso anual del hogar", que presenta varios outliers que podrían sesgar los resultados si no se manejan adecuadamente.

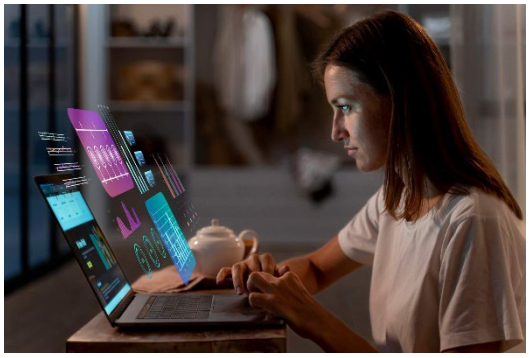
4. DOCUMENTACIÓN

- *Objetivo:* Presentar de manera clara y estructurada el proceso de evaluación, tratamiento de datos ausentes y normalización, destacando su importancia para el modelo de machine learning.
- *Tareas:*
    - **Redactar un informe que incluya:**
    - La evaluación de la calidad de los datos.
    - El proceso seguido para la detección y tratamiento de datos ausentes.
    - La técnica de normalización utilizada y su justificación.

Asegurarse de que el informe sea claro, conciso y bien estructurado, con ejemplos y gráficos que respalden los puntos principales.

Práctica: Análisis Univariable del Conjunto de Datos

Objetivo: El estudiante realizará un análisis univariable del conjunto de datos seleccionado, enfocándose en la validación de la distribución normal y el cálculo de estadísticas descriptivas. El objetivo es evaluar cómo estas características de los datos influyen en el desarrollo y la precisión del modelo de machine learning.



TAREAS

PRÁCTICA: EVALUACIÓN Y MEJORA DE LA CALIDAD DE DATOS

1. VALIDACIÓN DE LA DISTRIBUACIÓN NORMAL

- *Objetivo:* Determinar si las variables en el conjunto de datos siguen una distribución normal.

- *Tareas:*
    - **Gráficos de Distribución: Utilizar histogramas y gráficos de probabilidad normal (Q-Q plots) para visualizar la distribución de las variables..**
        - *Ejemplo:* Para un conjunto de datos sobre la eficiencia de paneles solares, se crea un histograma para la variable "Eficiencia energética" y un Q-Q plot para comparar la distribución con una normal teórica. Si el histograma muestra una forma de campana y los puntos del Q-Q plot siguen aproximadamente una línea recta, esto sugiere que la variable puede seguir una distribución normal.
    - **Pruebas Estadísticas: Aplicar pruebas como la prueba de Shapiro-Wilk o Kolmogorov-Smirnov para verificar la normalidad..**
        - *Ejemplo:* Se aplica la prueba de Shapiro-Wilk a la variable "Horas de sol anual" para determinar si los datos se ajustan a una distribución normal. Un valor p alto (mayor a 0.05) indicaría que no se rechaza la hipótesis nula de que los datos siguen una distribución normal.

2. CÁLCULO DE ESTADÍSTICAS DESCRIPTIVAS

- *Objetivo:* Calcular y analizar las estadísticas descriptivas para cada variable.
- *Tareas:*
    - **Cálculo de Estadísticas Descriptivas: Calcular la media, mediana, desviación estándar, y otros parámetros descriptivos para cada variable.**
        - *Ejemplo:* En un conjunto de datos sobre el consumo energético de diferentes hogares, se calcula que la media del "Consumo energético mensual" es de 300 kWh, la mediana es de 290 kWh, y la desviación estándar es de 50 kWh. Estos valores proporcionan una visión general de la tendencia central y la dispersión de los datos.

- **Interpretación de Resultados: Evaluar cómo estas estadísticas ayudan a entender la distribución y variabilidad de los datos.**
    - *Ejemplo:* Una alta desviación estándar en "Costo de instalación de energía solar" sugiere una gran variabilidad en los costos, lo que puede indicar diferencias significativas en los tipos de instalaciones o en los proveedores. Esto es importante para modelar correctamente el impacto del costo en la adopción de tecnologías solares.

3. INTERPRETACIÓN Y RELEVANCIA PARA EL MODELO DE MACHINE LEARNING.

- *Objetivo:* Explicar cómo la normalidad de la distribución y las estadísticas descriptivas afectan el desarrollo y precisión del modelo de machine learning.
- *Tareas:*
    - **Influencia de la Normalidad: Discutir cómo la normalidad (o la falta de ella) en la distribución de variables puede influir en la selección y rendimiento del modelo.**
        - *Ejemplo:* Si la variable "Eficiencia energética" no sigue una distribución normal, podría ser necesario aplicar transformaciones (como logaritmos) para normalizar los datos antes de usarlos en un modelo de machine learning, ya que muchos modelos asumen normalidad en los datos.
    - **Importancia de las Estadísticas Descriptivas: Evaluar cómo las estadísticas descriptivas influyen en la comprensión del dataset y en la preparación de los datos para el modelado.**
        - *Ejemplo:* La alta desviación estándar en "Consumo energético mensual" puede sugerir la necesidad de técnicas de escalado o normalización para mejorar el rendimiento del modelo de machine learning, ya que variables con alta variabilidad pueden afectar el entrenamiento del modelo.

4. DOCUMENTACIÓN

- *Objetivo:* Presentar de manera clara y estructurada los hallazgos del análisis univariable, destacando su importancia para el modelo de machine learning..
- *Tareas:*
    - **Redactar un informe que incluya:**
    - La validación de la distribución normal con gráficos y pruebas estadísticas.
    - Las estadísticas descriptivas calculadas para cada variable.
    - La interpretación de los resultados y cómo estos influyen en el desarrollo y precisión del modelo de machine learning.

Asegurarse de que el informe sea claro, conciso y bien estructurado, utilizando gráficos y explicaciones que respalden los puntos principales.

Material Complementario

Campista, en este espacio encontraras ayudas en diferentes formatos que pueden potenciar tu proceso de aprendizaje.

- ***The Main Ideas behind Probability Distributions***

- ***The Normal Distribution, Clearly Explained!!!***

- ***Calculating the Mean, Variance and Standard Deviation, Clearly Explained!!!***

- ***Escalamiento, Normalización y Estandarización de Datos con Python para Ciencia de Datos***