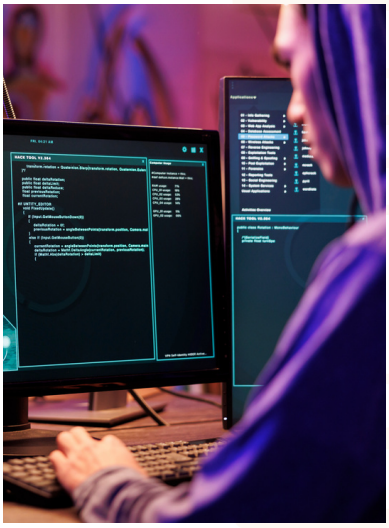
Creación de la Base de Datos a partir de datos obtenidos

Creación de la Base de Datos a partir de datos obtenidos



La creación de una base de datos a partir de los datos obtenidos es un proceso crucial en el desarrollo de proyectos de inteligencia artificial (IA). En esta etapa, los datos recolectados se preparan, procesan y transforman para formar una estructura organizada que permita su fácil acceso, consulta y análisis. Esta base de datos servirá como fundamento para entrenar modelos de IA, realizar análisis de datos y extraer información útil para la toma de decisiones.

1 Limpieza y Preprocesamiento de Datos

a) Identificación de Datos Erróneos o Faltantes:

Realizar un análisis inicial de los datos para identificar valores erróneos, nulos o faltantes.

b) Manejo de Valores Faltantes:

Decidir si eliminar filas con datos faltantes, imputar valores faltantes utilizando técnicas como la media o la mediana, o utilizar modelos predictivos para estimar los valores faltantes.

c) Detección y Tratamiento de Valores Atípicos:

Utilizar métodos estadísticos o basados en modelos para detectar y tratar valores atípicos que puedan distorsionar los resultados.

2 Transformación de Datos

a) Codificación de Variables Categóricas:

Convertir variables categóricas en variables numéricas utilizando técnicas como one-hot encoding o label encoding.

b) Escalado de Características:

Normalizar o estandarizar las características para asegurar que todas estén en la misma escala y evitar problemas de convergencia en algoritmos de aprendizaje automático.

c) Reducción de Dimensionalidad:

Aplicar técnicas de reducción de dimensionalidad como PCA (Análisis de Componentes Principales) o LDA (Análisis Discriminante Lineal) para reducir la complejidad del conjunto de datos.

3 Visualización Inicial

a) Histogramas y Gráficos de Distribución:

Visualizar la distribución de las características utilizando histogramas y gráficos de densidad para identificar posibles patrones o sesgos en los datos.

b) Diagramas de Dispersión:

Graficar diagramas de dispersión para explorar relaciones entre pares de características y detectar posibles correlaciones.

c) Visualización de Variables Categóricas:

Utilizar gráficos de barras o gráficos circulares para visualizar la distribución de variables categóricas.

Consideraciones importantes

- La limpieza y preprocesamiento de datos son pasos críticos para garantizar la calidad y confiabilidad de los modelos de inteligencia artificial.
- La transformación de datos es necesaria para preparar los datos en un formato adecuado para su análisis y modelado.
- La visualización inicial proporciona una comprensión intuitiva de la estructura y distribución de los datos, lo que puede guiar decisiones futuras en el proceso de modelado.

Desarrollo detallado

I. Identificación y Manejo de Valores Faltantes

a) Análisis de Valores Faltantes:

- Utilizaremos herramientas como **pandas** en Python para identificar la presencia de valores faltantes en nuestros datos.
- Se explorarán métodos como **isnull()** y **info()** para evaluar la cantidad y ubicación de los valores faltantes.

b) Manejo de Valores Faltantes:

- Discutiremos diferentes enfoques para manejar los valores faltantes, incluyendo la eliminación de filas o columnas, imputación de valores utilizando técnicas como la media, la mediana o la moda, y el uso de modelos predictivos para estimar valores faltantes.

II. Identificación y Manejo de Valores Duplicados

a) Detección de Valores Duplicados:

- Utilizaremos métodos como **duplicated()** en **pandas** para identificar filas duplicadas en nuestros datos.
- Exploraremos cómo verificar duplicados en columnas específicas o en todo el conjunto de datos.

b) Eliminación de Valores Duplicados:

- Discutiremos la importancia de eliminar valores duplicados para evitar sesgos en el análisis y modelado.
- Implementaremos métodos como **drop_duplicates()** para eliminar duplicados basados en criterios específicos, como todas las columnas o solo algunas columnas seleccionadas.

III. Técnicas Adicionales de Limpieza y Preprocesamiento

a) Normalización y Estandarización:

- Exploraremos la importancia de la normalización y estandarización de características para garantizar que los datos estén en la misma escala y facilitar el entrenamiento de modelos.

b) Manejo de Outliers:

- Discutiremos cómo identificar y manejar valores atípicos que pueden afectar negativamente el rendimiento de nuestros modelos.

c) Codificación de Variables Categóricas:

- Abordaremos técnicas para convertir variables categóricas en variables numéricas, como one-hot encoding o label encoding, para que puedan ser utilizadas en algoritmos de aprendizaje automático.

IV. Transformación de Datos Categóricos

a) Codificación One-Hot

- Exploraremos la técnica de codificación one-hot, que convierte variables categóricas en vectores binarios, asignando un valor de 1 a la presencia de una categoría y 0 en otros casos.
- Demostraremos cómo implementar esto utilizando funciones como **OneHotEncoder** de **scikit-learn**.

b) Codificación de Etiquetas (Label Encoding):

- Discutiremos el Label Encoding, una técnica que asigna un valor numérico a cada categoría única, convirtiendo las variables categóricas en valores enteros.
- Ejemplificaremos su uso utilizando la función **LabelEncoder** de **scikit-learn**.

V. Transformación de Datos Numéricos

a) Escalado de Características:

- Abordaremos la importancia del escalado de características para garantizar que todas las características tengan un peso igual en el modelo.
- Discutiremos métodos como la normalización y la estandarización, y cómo implementarlos con **MinMaxScaler** y **StandardScaler** de **scikit-learn**, respectivamente.

b) Transformaciones Logarítmicas:

- Introduciremos la transformación logarítmica, que puede ayudar a manejar distribuciones sesgadas o no gaussianas en datos numéricos.
- Ejemplificaremos cómo aplicar transformaciones logarítmicas utilizando la función **np.log()** de NumPy.

VI. Integración de Transformaciones

a) Pipeline de Transformación:

- Presentaremos la noción de pipelines de transformación, que permiten combinar múltiples pasos de preprocesamiento en una sola secuencia.
- Mostraremos cómo construir y utilizar pipelines de transformación con la clase **Pipeline** de **scikit-learn**.

b) Codificación de Etiquetas (Label Encoding):

- Discutiremos el Label Encoding, una técnica que asigna un valor numérico a cada categoría única, convirtiendo las variables categóricas en valores enteros.
- Ejemplificaremos su uso utilizando la función **LabelEncoder** de **scikit-learn**.

VII. Importancia de la Visualización Inicial

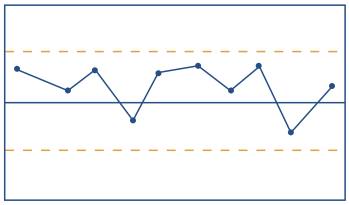
La visualización inicial de datos desempeña un papel fundamental en el análisis exploratorio de datos en proyectos de inteligencia artificial. Proporciona una comprensión intuitiva de la estructura y distribución de los datos, lo que puede revelar insights importantes y guiar decisiones en etapas posteriores del proyecto.

VIII. Herramientas de Visualización

a) Histogramas y Gráficos de Distribución:

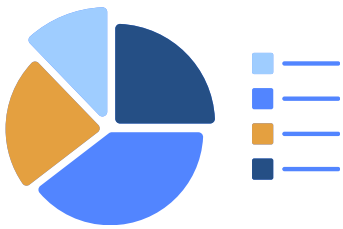
- Utilizaremos histogramas y gráficos de densidad para explorar la distribución de características numéricas en nuestros datos.
- Estas visualizaciones nos ayudarán a identificar la forma de la distribución, la presencia de outliers y la simetría de los datos.

b) Diagramas de Dispersión:



- Utilizaremos diagramas de dispersión para explorar relaciones entre pares de características numéricas.
- Estas visualizaciones nos permitirán detectar patrones de correlación o dependencia entre las variables.

c) Gráficos de Barras y Gráficos Circulares:



- Utilizaremos gráficos de barras y gráficos circulares para visualizar la distribución de variables categóricas.
- Estas visualizaciones nos ayudarán a comprender la frecuencia de cada categoría y a identificar categorías dominantes.

IX. Interpretación de Resultados

a) Identificación de Patrones y Tendencias:

- Analizaremos las visualizaciones para identificar patrones, tendencias o anomalías en los datos.
- Buscaremos distribuciones distintivas, relaciones entre variables y comportamientos inesperados que puedan requerir una investigación adicional.

b) Guía para el Preprocesamiento y Modelado:

- Utilizaremos los insights obtenidos de las visualizaciones para guiar el preprocesamiento de datos y la selección de características.
- Identificaremos posibles transformaciones o técnicas de preprocesamiento que podrían ser beneficiosas para mejorar la calidad de los datos y optimizar el rendimiento del modelo.

EJERCICIOS

¡Hora de practicar!

Ejercicio 1: Limpieza y Preprocesamiento de Datos

a) Carga el conjunto de datos "iris" utilizando la biblioteca scikit-learn.

b) Realiza un análisis inicial de los datos para identificar si hay valores nulos o faltantes.

c) Maneja los valores faltantes utilizando la media para completar los datos faltantes en las características numéricas.

d) Utiliza la moda para completar los datos faltantes en las características categóricas.

e) Elimina las filas que contengan valores nulos restantes en los datos.

Ejercicio 2: Transformación de Datos

a) Usa el método OneHotEncoder de scikit-learn para convertir las características categóricas en variables dummy.

b) Utiliza el MinMaxScaler de scikit-learn para escalar las características numéricas en un rango de 0 a 1.

c) Aplica una transformación logarítmica a una de las características numéricas y observa cómo cambia su distribución.

d) Crea un pipeline de transformación que incluya la codificación one-hot y el escalado de características.

Ejercicio 3: Visualización Inicial de Datos

a) Utiliza histogramas para visualizar la distribución de una característica numérica en los datos.

b) Crea un diagrama de dispersión para explorar la relación entre dos características numéricas.

c) Utiliza gráficos de barras para visualizar la frecuencia de una característica categórica.

d) Combina múltiples visualizaciones en una única figura utilizando matplotlib o seaborn para obtener una visión más completa de los datos.

Ejercicio 4: Integración de Limpieza, Transformación y Visualización

a) Carga un conjunto de datos que contenga tanto características numéricas como categóricas.

b) Realiza la limpieza y preprocesamiento de los datos, manejando los valores faltantes y aplicando transformaciones según sea necesario.

c) Utiliza las técnicas de transformación aprendidas para preparar los datos para el modelado.

d) Visualiza los datos preprocesados utilizando las técnicas de visualización aprendidas y analiza las distribuciones y relaciones entre las características.