PASOS PROYECTO INTEGRADOR ML

Paso 1: Entender el Problema y los Datos

Antes de comenzar cualquier proyecto de ciencia de datos, es crucial comprender completamente el problema y los datos disponibles.

1.1. Lee y comprende la descripción del proyecto y la problemática planteada.

1.2. Examina el archivo de datos "ML\_cars.csv" para familiarizarte con las columnas y los valores que contiene.

1.3. Asegúrate de comprender la diferencia entre los vehículos de gama alta y baja en función de la mediana de los precios.

Paso 2: Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos te ayudará a comprender mejor tus datos y a tomar decisiones informadas sobre cómo reprocesarlos y diseñar tus modelos.

2.1. Carga los datos en un marco de datos (por ejemplo, en Python usando pandas) y comienza a explorarlos. Puedes utilizar funciones como head(), info(), y describe() para obtener una vista previa de los datos.

2.2. Visualiza los datos utilizando gráficos (histogramas, diagramas de dispersión, etc.) para identificar tendencias, distribuciones y relaciones entre las variables.

2.3. Limpia los datos eliminando valores nulos o duplicados si es necesario.

2.4. Realiza un análisis estadístico básico para comprender la distribución de las características y cómo se relacionan con la variable objetivo (precio).

Paso 3: Preparación de Datos

Antes de construir los modelos, debes preparar los datos de acuerdo con las necesidades de cada modelo.

3.1. Codifica las variables categóricas, como "fueltype" y "carbody", en variables numéricas utilizando técnicas como one-hot encoding o label encoding.

3.2. Divide los datos en conjuntos de entrenamiento y prueba para que puedas evaluar tus modelos de manera adecuada.

3.3. Escala las características numéricas para que tengan una misma escala, lo que puede mejorar el rendimiento del modelo.

Paso 4: Modelamiento y Evaluación

Ahora estás listo para construir y evaluar tus modelos de Machine Learning.

4.1. Implementa un modelo de clasificación para distinguir vehículos de gama alta y baja utilizando la mediana de los precios como punto de corte. Puedes utilizar algoritmos como Regresión Logística, Árboles de Decisión o Random Forest.

4.2. Evalúa el modelo de clasificación utilizando métricas como precisión, recall y F1-score. Ajusta el modelo si es necesario.

4.3. Implementa un modelo de regresión para predecir el precio de los vehículos. Puedes utilizar algoritmos como Regresión Lineal, Regresión de Bosque Aleatorio o Regresión de Soporte Vectorial.

4.4. Evalúa el modelo de regresión utilizando métricas como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2). Ajusta el modelo si es necesario.

Paso 5: Entrega de Resultados

5.1. Genera archivos de texto plano con las predicciones de tus modelos de clasificación y regresión.

5.2. Prepara un informe final que incluya una descripción detallada de tus hallazgos, los modelos utilizados, las métricas de rendimiento y recomendaciones para el cliente.

5.3. Presenta tus resultados al cliente de manera clara y comprensible, destacando las conclusiones clave.

Paso 1: Entender el Problema y los Datos

1.1. Lee y Comprende la Descripción del Proyecto y la Problemática Planteada:

Lee detenidamente la descripción del proyecto que proporcionaste al principio. Asegúrate de entender la tarea que se te ha asignado, que es analizar características de vehículos para una automotriz china que desea ingresar al mercado de automóviles. El objetivo es comprender las diferencias entre los vehículos de gama alta y baja en función de la mediana de los precios.

Enfatiza en los puntos clave, como la clasificación de vehículos en baratos y caros, y la predicción de precios. Comprende por qué estas tareas son importantes para el cliente y cómo pueden ayudar en la toma de decisiones.

1.2. Examina el Archivo de Datos "ML\_cars.csv":

Abre el archivo de datos "ML\_cars.csv" utilizando una herramienta de análisis de datos como pandas en Python o el software de análisis de datos que prefieras.

Examina las primeras filas del conjunto de datos utilizando la función head() para ver ejemplos de las primeras observaciones y las columnas disponibles.

Utiliza info() para obtener información sobre el tipo de datos de cada columna y verificar si hay valores nulos.

Emplea describe() para obtener estadísticas resumidas de las columnas numéricas y tener una idea de la distribución de los datos.

1.3. Comprende la Diferencia Entre los Vehículos de Gama Alta y Baja:

Presta atención a la columna "price" (precio) en el conjunto de datos. Esta columna será fundamental para distinguir entre vehículos de gama alta y baja utilizando la mediana de los precios como punto de corte.

Calcula la mediana de los precios para entender cuál será ese punto de corte específico que se utilizará en el proyecto. La mediana dividirá los vehículos en dos grupos: aquellos por debajo de la mediana se considerarán de gama baja, y los que estén por encima se considerarán de gama alta.

Comprende por qué se ha elegido este enfoque y cómo ayudará al cliente a tomar decisiones estratégicas en la cotización de vehículos.