Bootcamp MLOps: Entegable 2

Por: Víctor Alejandro Regueira Romero

Análisis y Documentación del Data Set de Consumo de Energía en Tetuán, Marruecos

1. Análisis Exploratorio de Datos (EDA):

El conjunto de datos proporcionado contiene información sobre el consumo de energía en la ciudad de Tetuán, Marruecos. Antes de proceder con el análisis exploratorio de datos, realizaremos las siguientes tareas:

- Importar el conjunto de datos en un entorno de trabajo como Python o R.

- Revisar las primeras filas del conjunto de datos para comprender su estructura y contenido.

- Identificar el tipo de cada variable (categórica, numérica, temporal, etc.).

- Evaluar la presencia de valores faltantes en el conjunto de datos y decidir cómo tratarlos (eliminar, imputar, etc.).

- Realizar estadísticas descriptivas básicas para comprender la distribución y variabilidad de las variables numéricas.

- Visualizar la relación entre las variables mediante gráficos como diagramas de dispersión, histogramas, matrices de correlación, etc.

2. Pregunta de Interés para el Modelo de Aprendizaje Automático:

La pregunta que deseamos abordar con un modelo de aprendizaje automático podría ser:

- ¿Cómo se verá el consumo de energía en las diferentes zonas de Tetuán en función de factores como la temperatura, la humedad, la velocidad del viento y los flujos difusos generales?

3. Necesidad de una Estrategia de MLOps:

Es necesario implementar una estrategia de MLOps para este conjunto de datos debido a las siguientes razones:

- Para garantizar la reproducibilidad y escalabilidad de los modelos de aprendizaje automático.

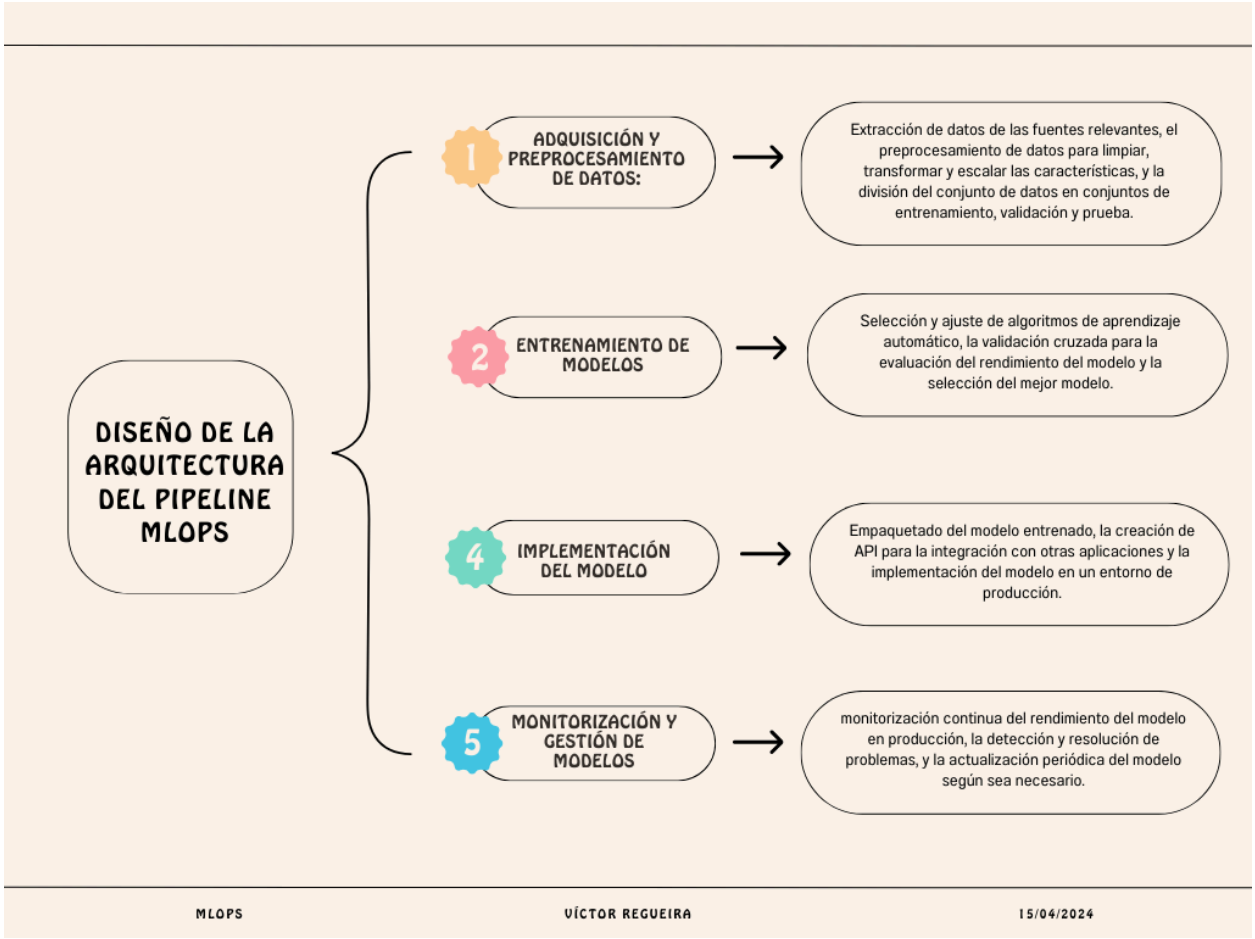
- Para automatizar y gestionar de manera eficiente todo el ciclo de vida del modelo, desde el desarrollo hasta la implementación y monitorización en producción.

- Para garantizar la colaboración efectiva entre los equipos de ciencia de datos, ingeniería de datos, desarrollo y operaciones.

- Para mejorar la transparencia, el control de versiones y la trazabilidad de los modelos.

4. Diseño de la Arquitectura del Pipeline MLOps:

El diseño del pipeline MLOps para esta nueva iniciativa de aprendizaje automático podría incluir los siguientes componentes:



5. Creación de un Modelo Base:

Para abordar tareas de predicción relacionadas con la pregunta de interés, crearemos un modelo base utilizando algoritmos de regresión, como regresión lineal o árboles de decisión. Este modelo servirá como punto de partida para futuras iteraciones y refinamientos.

6. Configurar estructura del modelo:

Para ello debemos de definir una base de modelo, en este caso usaremos XGBoost como propuesta, con base a este realizaremos ajustes de hiperparámetros y CrossVadation para evitar el entrenamiento de las mismas partes del modelo. Esto nos permitirá tener un modelo con baja probabilidad de tener un underfitting u overfitting.