


Evaluación del módulo #6

Aplicación Ciencia de datos 

Proyecto: Aplicación Ciencia de datos

Evaluación del módulo Machine Learning para Ingeniero de Datos

Situación inicial

Unidad solicitante: Equipo de desarrollo de soluciones analíticas en una fintech.

La fintech Alke Wallet ha crecido exponencialmente en los últimos meses y necesita automatizar decisiones clave relacionadas con la evaluación crediticia de nuevos usuarios. Actualmente, este proceso se realiza manualmente, lo que genera demoras y errores de evaluación. El equipo de analítica ha solicitado el diseño de una solución automatizada que utilice técnicas de aprendizaje automático para predecir si un nuevo usuario será considerado apto o no para acceder a servicios financieros, utilizando datos históricos disponibles.

El equipo de ingeniería de datos ha sido convocado para diseñar un pipeline de machine learning que contemple todo el proceso: desde la preparación y análisis de datos hasta la creación de un modelo robusto y su despliegue en una API funcional que permita integrar las predicciones a las aplicaciones internas.

Nuestro objetivo

Entrenar modelos de clasificación y regresión utilizando datasets reales o simulados, aplicando buenas prácticas como la validación cruzada y el preprocesamiento de datos. El producto final será una API funcional que permita acceder a las predicciones del modelo entrenado. Esto permitirá automatizar decisiones dentro del proceso de evaluación crediticia, mejorando la eficiencia operativa.

Requerimientos

- Análisis exploratorio del dataset proporcionado o simulado.
- Implementación de técnicas de preprocesamiento (imputación, encoding, escalamiento).

- Selección y entrenamiento de un modelo de clasificación y/o regresión.
- Aplicación de técnicas de validación cruzada y análisis de ajuste.
- Evaluación del modelo con métricas apropiadas (MAE, MSE, R^2 , precisión, exactitud, recall, AUC-ROC).
- Despliegue del modelo en una API funcional utilizando Python (por ejemplo, con Flask o FastAPI).
- Documentación del pipeline completo y endpoints de la API.


Paso a paso

Este proyecto refiere exclusivamente al **módulo 6: Machine Learning para ingenieros de datos**, y se compone de **7 etapas (lecciones)**, las cuales podrás avanzar de forma progresiva y escalonada con la ayuda de los manuales teóricos y los contenidos desarrollados en las clases en vivo.

Ten en cuenta de invertir **tiempo asincrónicos** para el desarrollo de cada etapa a modo de poder finalizar el módulo y realizar la entrega formal de tu propuesta. Cualquier consulta que surja compártela en los espacios sincrónicos para resolver las dudas en equipo.

A continuación encontrarás las consignas y tareas a desarrollar:


Lección 1: Fundamentos del aprendizaje de máquina


 **Objetivo:** Comprender los conceptos de aprendizaje supervisado, diferenciando tareas de clasificación y regresión.

 **Tareas a desarrollar:**

- Leer el manual 1.
- Definir el tipo de problema a resolver (clasificación o regresión).
- Justificar la elección del tipo de modelo.


Lección 2: Validación cruzada y ajuste del modelo


 **Objetivo:** Identificar el nivel de ajuste de un modelo e implementar validación cruzada.

 **Tareas a desarrollar:**

- Leer el manual 2.
- Implementar k-fold cross-validation.
- Analizar sobreajuste y subajuste usando gráficos y métricas.

Lección 3: Preprocesamiento y escalamiento de datos


 **Objetivo:** Preparar los datos aplicando encoding, normalización y tratamiento de valores faltantes.

 **Tareas a desarrollar:**

- Leer el manual 3.
- Aplicar Label Encoding o One-Hot Encoding.
- Escalar los datos con MinMaxScaler o StandardScaler.


Lección 4: Modelado de regresión


 **Objetivo:** Entrenar un modelo de regresión y evaluar su desempeño.

 **Tareas a desarrollar:**

- Leer el manual 4.
- Entrenar un modelo de regresión (lineal, polinómica, etc.).
- Evaluar con MAE, MSE, RMSE y R^2 .

Lección 5: Modelado de clasificación

 **Objetivo:** Entrenar un modelo de clasificación y evaluarlo con métricas específicas.


 **Tareas a desarrollar:**

- Leer el manual 5.

- Entrenar un modelo K-NN o similar.
- Evaluar con matriz de confusión, precisión, recall, F1 y AUC-ROC.


Lección 6: Despliegue del modelo como API


 **Objetivo:** Convertir el modelo entrenado en un servicio accesible vía API.

 **Tareas a desarrollar:**

- Leer el manual 7.
- Guardar el modelo entrenado con `joblib`.
- Crear una API con Flask o FastAPI que reciba inputs y devuelva predicciones.
- Documentar los endpoints y probarlos.

Lección 7: Evaluación, monitoreo y cierre del proyecto

 **Objetivo:** Validar el desempeño del modelo desplegado, documentar el proyecto y preparar su presentación final.

 **Tareas a desarrollar:**

- Aplicar métricas de evaluación final utilizando datos nuevos o de validación (MAE, MSE, RMSE, R^2 , Accuracy, Precision, Recall, F1, AUC-ROC).
- Verificar el funcionamiento correcto de la API realizando pruebas integradas con herramientas como Postman o Curl.
- Documentar el pipeline completo: decisiones tomadas, problemas encontrados, métricas obtenidas, y arquitectura de la solución.
- Incorporar la solución al portafolio personal (GitHub o presentación).
- Preparar una demo o pitch del proyecto (opcional): mostrar el flujo desde los datos hasta la predicción vía API.

- Publicar README técnico con instrucciones de uso del código/API y referencias utilizadas.

¿Qué vamos a validar?

- Aplicación correcta del preprocesamiento y análisis exploratorio.
- Uso adecuado de técnicas de validación cruzada.
- Elección coherente del modelo (clasificación o regresión) en función del problema.
- Cálculo y análisis de métricas con interpretación.
- Despliegue funcional de la API y pruebas exitosas.
- Calidad del código, documentación técnica y claridad de presentación.

Referencias

<https://scikit-learn.org/stable/>

<https://numpy.org/>

<https://flask.palletsprojects.com/en/stable/>

<https://pandas.pydata.org/>

https://scikit-learn.org/stable/modules/cross_validation.html#k-fold

https://www.youtube.com/results?search_query=deploy+machine+learning+mode+flask+fastapi

Recursos

- Te invitamos a investigar el siguiente artículo: **Artículos técnicos en Towards Data Science (Medium)**
- **UCI Machine Learning Repository – Credit Approval Dataset:** Dua, D., & Graff, C. (2019). Credit Approval Data Set. UCI Machine Learning Repository: [UCI Machine Learning Repository – Credit Approval Dataset](https://archive.ics.uci.edu/ml/datasets/Credit+Approval)
- Este archivo lo pueden utilizar como ejemplo, para poder trabajar en el proyecto. Kaggle. (s. f.). Home Credit Default Risk. <https://www.kaggle.com/competitions/home-credit-default-risk>

Entregables

- Código fuente del pipeline completo.
- Notebook con análisis exploratorio, preprocesamiento y evaluación.
- Script de despliegue del modelo (API).
- Documentación técnica del proyecto.
- Video demostrativo (opcional pero recomendado).

Portafolio

Este proyecto puede ser integrado a tu portafolio profesional como ejemplo de una solución completa de Machine Learning orientada a producción. Asegúrate de incluir una descripción clara del problema, screenshots del flujo de trabajo, explicación del modelo y resultados obtenidos. Puedes subirlo a tu GitHub y enlazarlo en tu CV o LinkedIn como caso de éxito.

¡Éxitos!

Nos vemos más adelante

