HENRY

LET'S GO!/>

Despliegue de proyecto modelo de riesgo crediticio

PIM5 - Data Science

Proyecto Integrador

Contexto del proyecto y rol del estudiante:

Has iniciado tu labor en el **equipo de Datos y Analítica de una empresa financiera, desempeñándote como Científico de Datos Junior Advanced.** Tu primera asignación consiste en desarrollar un modelo predictivo mediante técnicas de aprendizaje automático, utilizando información histórica de créditos, con el objetivo de anticipar el comportamiento de nuevos usuarios.

La **empresa** opera bajo un esquema estructurado de proyectos, en el cual cada iniciativa debe seguir una arquitectura de carpetas estrictamente definida. Esta estructura no puede ser modificada, ya que los procesos de despliegue a producción están automatizados a través de pipelines de validación en Jenkins. Cualquier alteración en la organización de carpetas podría generar retrasos significativos en el paso a producción.

Como **primer paso** se te solicita **crear un repositorio público en GitHub que contenga el desarrollo del proyecto.** El enlace a este repositorio será el entregable final que compartirás al concluir el proyecto.

Objetivos del PI

- **Documentar y versionar correctamente un proyecto en GitHub**, incluyendo archivos como README.md, .gitignore, y requirements.txt.
- **Implementar scripts para el despliegue del modelo en producción** respetando la estructura del proyecto.
- **Automatizar tareas con scripts** para facilitar la ejecución del pipeline y el monitoreo.
- **Entrenar modelos de aprendizaje automático supervisado** para predecir el comportamiento de nuevos usuarios.
- **Evaluar el rendimiento de los modelos utilizando métricas apropiadas** (precisión, recall, F1, ROC-AUC, etc.).
- **Integrar principios de MLOps para asegurar reproducibilidad, trazabilidad y escalabilidad del modelo.**

<Proyecto Integrador>

Entregable Final

Proyecto Integrador

Repositorio en GitHub, mediante link

El repositorio debe tener esta estructura de carpetas y tres ramas: developer, certification, master. Se te dará mayor información en los demás avances.

Asegúrate que se pueda acceder sin problemas al enlace

```
├── mlops_pipeline/
|   └── src/
|       |   └── Cargar_datos.ipynb
|       |   └── comprension_eda.ipynb
|       |   └── ft_ engineering.py
|       |   └──model_training_evaluation.py
|       |   └──model_deploy.py
|       |   └──model_monitoring.py
├── Base_de_datos.csv
├── requirements.txt
├── .gitignore
├── readme.md
```

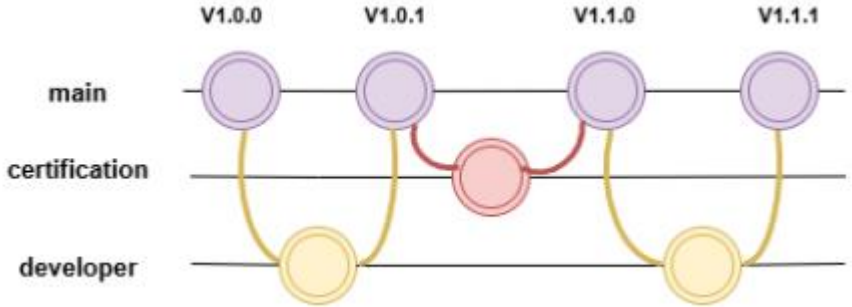
<Proyecto Integrador>

Detalle de Avances

Proyecto Integrador

Detalle de Avance #1:

- ➔ **Generar la estructura de carpetas tal y como se indica para la entrega.** Esta estructura debe estar en el repositorio creado anteriormente y no debe cambiarse.
- ➔ **Clona tu repositorio**
- ➔ **Genera un archivo de requirements.txt**
- ➔ **Configura un entorno virtual.**
- ➔ **Se deben hacer modificaciones a los archivos siguiendo la estructura de versiones y ramas siguientes:**



Para la V 1.0.1 se te pedirá avanzar en los notebooks **cargar_datos.ipynb** y **comprension_eda.ipynb**

Para la V1.0.0 crea la estructura de carpetas idéntica para cada rama. Este será el punto de partida.

Proyecto Integrador

Detalle de Avance #1:

- **Desarrollo de los siguientes pasos desde la rama developer:**

1. **Cargar_datos.ipynb**
2. **Comprensión_eda.ipynb:** teniendo en cuenta:

- Exploración inicial de datos
- Exploración de datos y descripción (EDA)
  * Análisis univariable.
  *Análisis bivariable.
  *Análisis multivariable.

- **Conocimientos necesarios:** Versionamiento en GitHub, conocimiento de ramas y pull request, trabajo colaborativo, análisis exploratorio, limpieza de datos, manejo de librerías Python.
- **Tech Stack necesario:** GitHub, Python, Seaborn, pandas, matplotlib.

Proyecto Integrador

Detalle de Avance #2:

- **Realizar el proceso de ingeniería de características.**
  - Empezar a entrenar los primeros modelos.
  - Realizar la evaluación de los modelos supervisados, seleccionando el de mejor performance.
- **En este avance se generará la versión 1.1.0** y **V1.0.1 como se muestra en el gráfico inicial**, en la que se abordará el proceso de ingeniería de características y modelamiento, respectivamente. A partir de allí puedes crear las versiones que desees y seguir un flujo que te sea lógico y cómodo.
- **Conocimientos necesarios:** Ingeniería de características, imputación, tipos de variables, pipelines, modelamiento supervisado.
- **Tech Stack necesario:** pandas, numpy, sk-learn, feature-engine, seaborn.

Proyecto Integrador

Detalle de Avance #3:

- **Realizar procesos de monitoreo y detección de data drift.**
- **Desarrollar una aplicación en streamlit.**
- **Generar el archivo REAMDE.md para tu ejercicio**, documentando el caso de negocio y los principales hallazgos y proceso
- **Conocimientos necesarios:** Data Drift, CI/CD

Proyecto Integrador

Detalle de Avance #4:

- **Disponibilizar mediante una API el modelo creado.**
- **Crear una imagen que contenga las librerías y el código para una app.**

  model_deploy.py

  Imagen Docker
- **Conocimientos necesarios:** Fundamentos de las APIs y FastAPI, Funciones básicas, métodos y parámetros en FastAPI, Creación de API para un modelo de ML con FastAPI, Visualización de modelos con Streamlit, Desarrollo de una aplicación web de Streamlit para ML, Docker.
- **Tech Stack necesario:** Fast API, Docker, Streamlit.

Proyecto Integrador

EXTRACREDIT

- **Configura Sonar Cloud https://sonarcloud.io en el repositorioy ejecuta pruebas para validar:**

1. **Calidad del código:** Evalúa la mantenibilidad del código fuente.
2. **Seguridad:** Detecta vulnerabilidades y puntos débiles que podrían ser explotados por atacantes.
3. **Cobertura de Pruebas:** Mide qué porcentaje del código está cubierto por pruebas unitarias o de integración.
4. **Integridad y Estilo:** Verifica que el código siga convenciones de estilo y buenas prácticas. Evalúa: