Guía: Proyecto Final Individual

Machine Learning Operations Bootcamp

VALIDACIÓN TÉCNICA

En la validación técnica, abordarás tu propio desafío. Resolverás un problema común de aprendizaje automático utilizando el **data set** proporcionado y aplicarás lo que has aprendido en el entrenamiento.

Objetivos:

- Evaluar tu capacidad para aplicar conocimientos teóricos en un contexto práctico.
- Demostrar tu competencia en los temas cubiertos en el programa.
- Evaluar tus habilidades para resolver problemas en escenarios del mundo real utilizando un data set.

Descripción del problema:

En este desafío, te sumergirás en la exploración de un data set único. La tarea inicial implica analizar y comprender completamente esta información a través de un detallado ***Análisis Exploratorio de Datos***. La singularidad de este desafío radica en en identificar las razones fundamentales que lo justifican.

Tareas:

1. ***Analizar y comprender*** el data set proporcionado a través de un Análisis Exploratorio de Datos.

2. ***Determinar*** la pregunta que deseas abordar con un modelo de aprendizaje automático.

3. ***Identificar*** por qué se necesita una estrategia de MLOps para este data set. (Módulo 1)

4. ***Diseñar*** la arquitectura del pipeline para esta nueva iniciativa de aprendizaje automático.

5. ***Crear*** un modelo base para abordar tareas de predicción (clasificación, regresión, etc.) relacionadas con la pregunta. Este modelo no necesita una alta precisión, recall o puntuación F1; el objetivo es crear un modelo rápido para iteración. (Módulo 4)

6. ***Configurar*** la estructura correcta del modelo con la idea de dejarlo listo para el despliegue. (Módulo 6)

7. ***Crear*** nuevas versiones del modelo, que incluyan cambios en las características o ajustes de hiper parámetros para la reproducibilidad y el seguimiento experimental. Agregar estrategias reproducibles y de prueba que se aplicarán a tu pipeline de aprendizaje automático. (Módulo 7)

8. ***Implementar*** un modelo en su entorno local utilizando contenedores y planificar el reentrenamiento, el drift, el redeploy, el escalado y el monitoreo. Orquestar y registrar el pipeline de aprendizaje automático, y se realizar el despliegue inicial. (Módulo 8)

9. ***Demostrar*** estrategias de prueba y versionamiento. Evaluar posibles derivas que puedan ocurrir e incluir tareas de monitoreo y pruebas adicionales para abordarlas. (Módulos 7 y 8).

Requisitos:

- **Completar** con éxito el desafío asignado utilizando el data set proporcionado.
- **Aplicar** los conceptos aprendidos en el programa para el desarrollo integral de tareas cotidianas de aprendizaje automático.
- **Documentar** de manera clara y concisa el proceso de resolución de problemas en una documentación/framework.
- **Presentar** las soluciones y adherencia a las pautas e instrucciones dadas.
- **Incluir una explicación** detallada de lo siguiente:
    - ¿Por qué eligiste un modelo en particular?
    - ¿Qué elementos matemáticos se consideraron en esta decisión?
    - ¿Cómo se integrarán los nuevos datos?
    - ¿Cómo se medirá el drifting?
    - ¿Se considera la prueba en el desarrollo del pipeline?

Entregables:

Accede al siguiente forrmulario para para realizar un seguimiento de tu progreso e ir registrando tu avance.

- Soluciones a desafíos individuales que demuestren conocimientos aplicados.
- Implementaciones de código que demuestren el uso de técnicas aprendidas.
- Ideas y análisis derivados de los conjuntos de datos proporcionados.
- Diagramas o elementos gráficos para ilustrar y facilitar la comprensión de la solución general.
- La URL del repositorio donde has estado trabajando en tu desafío.

Fechas de entrega:

- Entregable 1: **Viernes 15 de marzo, 2024**
- Entregable 2: **Viernes 5 de abril, 2024**

- Entregable 3: **Viernes 26 de abril, 2024**
- Entrega final: **Viernes 10 de mayo, 2024**

LISTADO DE DATA SETS

Dataset Name	Type of problem	Num of records	Num of Features	Participant
Real-time Election Results: Portugal 2019	Regression	21,640	29	Alejandra Moreno Morales
Dry Bean Dataset	Classification	13,610	16	Angel Enrique Medina Pérez
Wine Quality	Classification	4,900	12	Carlos Joya Hernández
Forest Fires	Regression	517	13	Christian Ramirez
CDC Diabetes Health Indicators	Classification	253,680	21	Cristian Alejandro Hernandez
Product Classification and Clustering	Classification	35,310	7	Daniel Isai Yañez Torres
Air quality	Regression	9,360	15	Daniel Martínez Escobosa
Predict students' dropout and academic success	Classification	4,420	36	Edgar Alberto González Ambriz
Solar Flare	Regression	1,390	10	Eric Evaristo Nieves
Steel Industry Energy Consumption	Regression	35,040	11	Erick Alexei Cambray Servin
Accelerometer	Regression	153,000	5	Fabio Solorzano Flores
Amphibians	Classification	189	23	Fernando Alfredo Rojas Estrella
Secondary Mushroom Dataset	Classification	61,070	20	Fernando Maytorena
Room Occupancy Estimation	Classification	10,130	16	Fernando Sebastián Sánchez Cardona
Algerian Forest Fires Dataset	Regression	244	12	Francisco Felipe Andrade Sánchez
Early stage diabetes risk prediction dataset.	Classification	520	17	Francisco Gonzalez
Regensburg Pediatric Appendicitis	Classification	782	59	Francisco José Arellano Montes
Estimation of obesity levels based on eating habits and physical	Classification	2,110	17	Francisco Ramirez

condition				
Divorce Predictors data set	Classification	170	54	Francisco Roman Peña de la Rosa
Online News Popularity	Classification	39,800	61	Gamaliel Torres Vargas
Risk Factor prediction of Chronic Kidney Disease	Regression	202	28	Guadalupe Vidal Cruz
Maternal Health Risk	Classification	1,010	6	Irlanda Ordoñez Kelly
Higher Education Students Performance Evaluation	Classification	145	33	Jesus Joel Buenrostro Jiménez
Residential Building Data Set	Regression	372	105	Jesús Ramseths Echeverría Rivera
Iranian Churn Dataset	Classification	3,150	13	Juan Carlos Becerril Cabrera
Bike Sharing Dataset	Regression	17,390	16	Laura Lizbeth Salazar Ortiz
Seoul Bike Sharing Demand	Regression	8,760	14	Leonardo Sánchez Bojórquez
Student Performance on an entrance examination	Classification	666	11	Marco Medina
AI4I 2020 Predictive Maintenance Dataset	Regression	10,000	14	Mauricio Daniel Tellez Nava
Glioma Grading Clinical and Mutation Features	Classification	839	23	Porfirio Basaldúa González
Heart failure clinical records	Classification	299	12	Rocio del Pilar Najera Lopez
Forty soybean cultivars from subsequent harvests	Classification	320	11	Salvador Guillermo Cruz González
Gender Gap in Spanish WP	Classification	4,750	21	Samuel Yoshua Márquez Galindo
Parkinsons Telemonitoring	Regression	5,880	19	Sebastián Arturo Uribe Martínez
Insurance Company Benchmark (COIL 2000)	Regression	9,000	86	Sergio Ramírez Peña
Power consumption of Tetouan city	Regression	52,420	9	Victor Alejandro Regueira Romero

