Diplomado en Ciencia de Datos y Análisis Avanzado

Unidad 5: Modelado Predictivo I-Regresión y Clasificación

Proyecto Práctico

Competencia de Predicción de Abandono de Clientes

Objetivos

- **Aplicar** los conceptos fundamentales de los modelos de clasificación (Regresión Logística, kNN, Naive Bayes) en un problema de negocio real**.**
- **Incorporar** un flujo de trabajo completo de Machine Learning, incluyendo la exploración de datos, el preprocesamiento, el entrenamiento y la evaluación de modelos.
- **Analizar** y comparar el rendimiento de diferentes algoritmos de clasificación utilizando métricas apropiadas como ROC AUC.
- **Aprender** a interpretar los resultados de los modelos y a preparar una entrega para una plataforma de competencia como Kaggle.
- **Desarrollar** habilidades de trabajo en equipo, colaboración y comunicación técnica a través de un proyecto práctico.

Consigna

Se les solicita a los equipos desarrollar un proyecto de clasificación para predecir el abandono de clientes de una empresa de telecomunicaciones. Para ello, deberán seguir los siguientes pasos:

1. **Equipos:** Trabajan en equipo de acuerdo a los grupos que ya están preestablecidos.

2. **Descargar los Materiales:** Acceder al directorio **Tareas Prácticas Asincronas** para descargar los archivos de datos (train.csv, test.csv) y la plantilla de trabajo (Plantilla_de_Jupyter_Notebook_para_el_Proyecto.ipynb).
3. **Análisis y Modelado:** Utilizando la plantilla proporcionada como referencia, realizar el análisis exploratorio de datos, el preprocesamiento y el entrenamiento/evaluación de los modelos de Regresión Logística, kNN y Naive Bayes.
4. **Participar en la Competencia**:
   - Registrar al equipo en la competencia de Kaggle: https://www.kaggle.com/t/57b70c381e4d451b8ae38e164b91a2aa
   - Seleccionar el mejor modelo basado en la validación local.
   - Generar un archivo de sumisión (submission.csv) con las **probabilidades** de abandono para el conjunto test.csv.
   - Subir el archivo a Kaggle para obtener una puntuación y competir en el leaderboard. Se permiten hasta 2 sumisiones por día por equipo.
5. **Elaborar el Informe Final:** Completar el cuaderno Jupyter Notebook (.ipynb) documentando todo el proceso, análisis, código, decisiones tomadas y conclusiones del equipo.
6. **Realizar la Entrega Final:** Comprimir el cuaderno final y el archivo de sumisión en un único archivo .zip y subirlo a esta tarea en Moodle antes de la fecha límite.

Fecha límite de entrega:

La fecha de entrega será hasta el 29 de Julio de 2025

Criterios de evaluación

La calificación final del proyecto se calculará en base a la siguiente ponderación:

- **70% - Informe de Resultados (Jupyter Notebook):**
- **Análisis Exploratorio de Datos (15%):** Profundidad del análisis, calidad y pertinencia de las visualizaciones.
- **Preprocesamiento (15%):** Correcta implementación de la limpieza, codificación y escalado de datos. Justificación de las decisiones tomadas.
- **Modelado y Evaluación (25%):** Implementación correcta de los tres modelos solicitados. Rigurosidad en la evaluación local y comparación de los modelos.
- **Calidad General del Notebook (15%):** Claridad del código, calidad de los comentarios, estructura del informe, reproducibilidad y conclusiones.
- **30% - Rendimiento en la Competencia de Kaggle:**

Bibliografía utilizada y sugerida

- **Documentación de Scikit-learn:**
  - Logistic Regression
  - KNeighbors Classifier
  - Gaussian Naive Bayes
- **Libros y Recursos Adicionales:**
  - Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow. 2nd ed. O'Reilly Media.
  - Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.
  - Foro de la competencia en Kaggle para discusión de ideas generales.