Proyecto Machine Learning

Curso 98D

Docente: Susana Medina G.

Objetivos Generales

- **Comprender los fundamentos del Machine Learning:** Los estudiantes deben demostrar una comprensión sólida de los conceptos básicos del Machine Learning, incluyendo los diferentes tipos de aprendizaje (supervisado, no supervisado, por refuerzo), los algoritmos clave y las métricas de evaluación.
- **Desarrollar habilidades prácticas en Machine Learning:** Los estudiantes deben ser capaces de aplicar sus conocimientos teóricos para construir, entrenar y evaluar modelos de Machine Learning utilizando herramientas y bibliotecas populares (por ejemplo, Python, Scikit-learn).
- **Resolver problemas del mundo real con Machine Learning:** Los estudiantes deben aprender a identificar problemas que pueden ser abordados con Machine Learning, seleccionar los algoritmos apropiados y adaptar sus modelos a escenarios del mundo real.
- **Despliegue de modelos de Machine Learning en una aplicación web sencilla**. Los estudiantes deben aprender a utilizar herramientas de software (MLOps) para despliegue de proyectos de Machine Learning en entornos controlados como docker.

Los grupos deben estar conformados máximo por 3 estudiantes.

Conjunto de datos (dataset)

En este proyecto cada grupo podrá escoger su conjunto de datos (dataset), bajo las siguientes normas.

Requisitos del dataset

- Mínimo 2000 registros o filas. Recomendado: 5000 registros o más.
- Mínimo 10 variables o columnas.
- Formatos de archivos recomendados para datos de tipo texto: **CSV**, **TXT** , **JSON**.
- La carga correcta del dataset es responsabilidad del estudiante/ grupo de trabajo. Si hay inconvenientes, será responsabilidad de los estudiantes solucionarlos.
- Tenga en cuenta que: usar un dataset distribuido en varios archivos o bases de datos requiere más tiempo de procesamiento y limpieza, por lo cual si escogen varios archivos será responsabilidad de los estudiantes demostrar que hay coherencia en el procesamiento de todas las fuentes de datos.

- Datos como imágenes requieren un manejo especial para aplicar técnicas especiales de limpieza y visión computacional (como filtros y segmentación) y luego aplicación de Deep Learning como redes neuronales (Convolucionales, Recurrentes, Autoencoder, etc.). Es importante que los estudiantes con este tipo de proyectos vayan adelantados y avanzando en estos temas debido a que son los últimos del contenido del curso.
- Tamaño recomendado del archivo de dataset: 500MB máximo 1 GB. Archivos más pesados (en especial de más de 1 GB) requieren máquinas o entornos con mayor capacidad de procesamiento y memoria.
- No se permiten datasets de datos privados o sensibles que no cumplan las leyes de protección de datos (Ley 1581 de 2012). Si tiene un Trabajo de grado especial con este tipo de datos debe garantizar que cuenta con los permisos necesarios (Comité de ética*) para almacenar esos datos y trabajar con ellos. También debe tener claro hasta dónde le será permitida la divulgación de los resultados de su trabajo (ej: carta con permisos de divulgación abierta, limitada o totalmente restringida*).
- Fuentes de datos de **repositorios públicos** sugeridas: Kaggle, Github, GitLab, UCI Machine Learning Repository, Datos abiertos del gobierno, entre otros.
- **Cada grupo/estudiante debe tener un dataset diferente**. **Registrar y validar el dataset** seleccionado por cada grupo de estudiantes con la docente.
- Hasta la etapa de Preprocesamiento de Datos le será permitido a cada grupo cambiar de dataset para el proyecto, presentando argumentos contundentes.

Cronograma

Semana 1-2: Introducción y Preparación

- Selección y definición del proyecto.
- Recopilación del *dataset* y exploración inicial.
- Instalación de *software* y configuración del entorno de desarrollo:
    - python versión 3.13 o anteriores
    - librerías: pandas, numpy, scikit-learn, matplotlib, seaborn.
- Revisión de conceptos básicos de Machine Learning y estadística.
- **Entrega 1**: "selección de conjunto de datos" **febrero 24 de 2025**.

Semana 3-4: Preprocesamiento de Datos

- Limpieza y transformación de datos.
- Manejo de valores faltantes y datos atípicos.
- Análisis Exploratorio de Datos (EDA).
- Ingeniería de características (Feature Engineering).

Semana 5-7: Modelado (Parte 1)

- Selección de algoritmos de Machine Learning relevantes para el problema.
- Entrenamiento y evaluación de modelos básicos.

- Ajuste de hiperparámetros inicial.
- Validación cruzada.
- **Entrega 2** “modelos básicos de ML”: **marzo 24 de 2025**.

Semana 8-10: Modelado (Parte 2)

- Experimentación con diferentes algoritmos y combinaciones.
- Optimización de modelos con técnicas avanzadas.
- Análisis de errores y ajuste fino de modelos (*tunning*).
- Interpretación de resultados y métricas de evaluación.
- **Entrega 3** “modelos mejorados de ML y comparaciones”: **abril 3 de 2025**.

Semana 11-12: Implementación y Pruebas

- Desarrollo de un prototipo o aplicación web básica*.
- Pruebas exhaustivas del modelo con datos nuevos.
- Ajustes finales al modelo basados en los resultados de las pruebas.
- Documentación del código y el proceso.
- **Entrega 4** “prototipo de app web para ML”: **abril 29 de 2025**.

Semana 13-14: Presentación y Entrega

- Preparación de la presentación del proyecto.
- Elaboración del informe final.
- Entrega del proyecto y el código fuente.
- Asesorías por proyecto*

Semana 15: Evaluación y Reflexión

- **Presentación final del proyecto**.
- **Entrega Final “modelo ML en app web”: mayo 22 de 2025**.
- Evaluación del proyecto por parte del docente y los compañeros.
- Reflexión sobre el proceso de aprendizaje y los resultados obtenidos.

Reportes de avances y final

Para los Notebooks puede hacerse en el mismo archivo, de manera ordenada.

- ➔ Utilizar bloques de texto para resumir los hallazgos.
- ➔ Resumen de sus hallazgos en un informe estructurado.
- ➔ Secciones a incluir en el reporte final:
    1. Introducción al dataset y objetivos del análisis.

2. Resumen descriptivo de las variables.
3. Visualizaciones clave que respalden los resultados.
4. Conclusiones basadas en los patrones identificados.
5. Posibles preguntas para análisis futuros.