Universidad Tecnológica de Panamá
Facultad de Ingeniería de Sistemas Computacionales
Proyecto Final de Curso
Curso: Introducción a la Ciencia de los Datos

Profesor: Juan Montenegro, M.Sc. **Semestre:** II, del 2024

Reglas del juego:

- Incluye el código fuente escrito para la tarea como un apéndice en tu informe en PDF (como texto, no como capturas de pantalla). Además, incluye un archivo zip por separado que contenga el código ejecutable y cualquier archivo de datos necesario. Los programas deben ser código funcional escrito en Python y deben cargar los datos, etc., al ejecutarse, de modo que podamos descomprimir tu entrega y ejecutarla directamente para comprobar que funciona. Mantén el código breve y limpio, con nombres de variables significativos, etc.
- Importante: Para cada problema, tu objetivo principal es demostrar que entiendes lo que estás haciendo, no solo ejecutar un programa y citar los números que genera. Generalmente, la mayor parte del crédito se otorga por la explicación/análisis en lugar del código/respuesta numérica.
- Si utilizas modelos de aprendizaje automático que no se cubrieron en el curso, debes asegurarte de demostrar que los entiendes y no simplemente ejecutas el código de manera "caja negra" (por lo tanto, explica cómo se generan las predicciones a partir de una entrada, cuál es la función de costo, cuáles son los parámetros y los hiperparámetros del modelo y cómo afectan las predicciones, etc.).
- Los informes normalmente deben tener alrededor de 5 páginas, con un límite máximo de 10 páginas (excluyendo el apéndice con el código).
- Se penalizara el uso de herramientas de IA, la idea es que entiendas lo que estas haciendo sin la ayuda de estas herramientas.

- Link del dataset: https://github.com/OmarConcepcion/Covid19-Panama/blob/master/Dataset%20Covid19%20Panama%2010-05-2020.csv

1. Comprensión del problema y del dataset

- **Objetivo**: Predecir el número de hospitalizados, fallecidos, recuperados, etc., en los corregimientos según las características geográficas y los casos registrados.
- **Tarea**: Explorar los datos de las diferentes provincias, distritos y corregimientos de Bocas del Toro y Chiriquí, y generar preguntas o hipótesis iniciales.

2. Preparación de los datos

- **Limpieza de datos**: Identificar y manejar los valores faltantes en las columnas como "hospitalizado", "aislamiento domiciliario", "fallecido", etc. Implementar estrategias de imputación o eliminación según sea necesario.
- **Tarea**: Realizar un análisis de valores nulos y presentar un plan de limpieza, detallando cómo abordarán los datos faltantes en cada columna.

3. Feature Engineering

- **Generación de nuevas características**: Crear nuevas variables como:
  - Ratios entre los diferentes casos (hospitalizados vs fallecidos, recuperados vs aislados).
  - Relación entre las coordenadas geográficas (LONG, LAT) y la cantidad de casos.
- **Tarea**: Generar al menos tres nuevas características que ayuden a mejorar el modelo predictivo.

4. División del dataset y Validación Cruzada

- **Tarea**: Dividir el dataset en un conjunto de entrenamiento y otro de prueba. Implementar una validación cruzada (k-fold) para asegurar que los modelos no están sobreajustados.

5. Modelado

- **Modelos sugeridos**:
  - **Regresión lineal**: Para predecir la cantidad de hospitalizados, fallecidos, etc.
  - **Árboles de decisión o Random Forest**: Para explorar relaciones no lineales.
  - **Regresión logística**: Si el objetivo es predecir un resultado binario (por ejemplo, si habrá o no hospitalización).
- **Tarea**: Implementar al menos dos algoritmos de machine learning (uno lineal y uno no lineal), ajustarlos y comparar los resultados.

6. Evaluación del modelo

- **Métricas**: Utilizar métricas como el *MAE, MSE, RMSE* para problemas de regresión, y *precisión, recall, AUC* para problemas de clasificación.
- **Tarea**: Evaluar el rendimiento de cada modelo y discutir cuál ofrece un mejor balance entre precisión y generalización.

7. Conclusiones

- **Tarea**: Presentar un informe donde describan el proceso completo, desde la limpieza de los datos hasta la evaluación de los modelos, incluyendo conclusiones sobre el modelo más efectivo.