Tarea 1 INF-398

1. PHISHING DATASET

DATASET OVERVIEW:
https://archive.ics.uci.edu/dataset/327/phishing+websites

1. Cargar datos y chequear integridad
2. Pensando en un pipeline de ML, en este paso es necesario preparar los datos, ¿se debe estandarizar primero y separar la data luego, o separar la data primero y luego estandarizar por separado? ¿Por qué?.
3. ¿Es recomendable estandarizar los datos en este caso en específico? explique su respuesta, y proceda según corresponda.
4. Entrenar los modelos NB, QDA, LDA y Logistic Regression utilizando correctamente los conjuntos de entrenamiento, test y validación.
5. Elija 3 métricas de clasificación que encuentre relevante y explique cuál y por qué encuentra que un modelo es mejor que los demás.
6. Realice una matriz de confusión para visualizar cuales fueron los resultados de cada modelo, con el fin de realizar comparaciones. Comente lo que observa.
7. De una pequeña explicación a por qué cree que los dos modelos con peor desempeño tuvieron malos resultados.

2. Home Equity Line of Credit (HELOC)

DATASET OVERVIEW: https://www.openml.org/search?type=data&sort=runs&id=45026&status=active

1. Cargar datos y chequear integridad.
2. Realizar un grafico scatter de las primeras 5 dimensiones y pintando en colores las clases (RiskPerformance) y responda priori y solo viendo como se distribuyen las clases, ¿cree que los modelos obtendrán buenos o malos

resultados? ¿por qué? (HINT: para graficar, revisar la función pairplot de la biblioteca Seaborn).

3. Pensando en un pipeline de ML, en este paso es necesario preparar los datos, ¿se debe estandarizar primero y separar la data luego, o separar la data primero y luego estandarizar por separado? ¿Por qué? (PUEDE SALTAR LA RESPUESTA ESCRITA SI YA RESPONDIO EN LA PREGUNTA DEL OTRO DATASET).
4. ¿Es recomendable estandarizar los datos en este caso en específico? explique su respuesta, y proceda según corresponda.
5. Entrenar los modelos NB, QDA, LDA y Logistic Regression utilizando correctamente los conjuntos de entrenamiento, test y validación.
6. Realice una matriz de confusión para visualizar cuales fueron los resultados de cada modelo, con el fin de realizar comparaciones. Comente lo que observa.
7. Elija 3 métricas de clasificación que encuentre relevante y explique cuál y por qué encuentra que un modelo es mejor que los demás.
8. ¿Los resultados obtenidos, coinciden con la suposición realizada en el punto 2.2?, Tanto en caso positivo como negativo, explique su respuesta en base los datos y el funcionamiento de cada modelo.

3. Desafío Kaggle

DATASET OVERVIEW:

Todos los datos provienen de una medición continua de EEG con el Emotiv EEG Neuroheadset. La duración de la medición fue de 117 segundos. El estado del ojo fue detectado a través de una cámara durante la medición de EEG y luego agregado manualmente al archivo después de analizar los fotogramas del video. '1' indica el estado de ojos cerrados y '0' el estado de ojos abiertos. Todos los valores están en orden cronológico con el primer valor medido en la parte superior de los datos.

Las características corresponden a 14 mediciones de EEG del auricular, originalmente etiquetadas como AF3, F7, F3, FC5, T7, P, O1, O2, P8, T8, FC6, F4, F8, AF4, en ese orden.

Se espera que los estudiantes sean capaz de aplicar correctamente las técnicas de exploración de datos, preparación de datos y entrenamiento de modelos de clasificación vistos en clase (NB, QDA, LDA y Logistic Regression) de forma correcta. Optimice sus modelos con base en la métrica F1-Score y suba sus mejores predicciones a la competencia de Kaggle: https://www.kaggle.com/t/9a7ab69d866a4d779fc51e5fbade0fed.

Recuerde utilizar el formato que se encuentra en el archivo Submission_example.csv (no cambie el orden del conjunto de test), en donde en la primera columna llamada ID corresponde a el índice de sus respuestas, mientras que la segunda llamada Class, es la predicción realizada con su modelo.

¡Éxito!