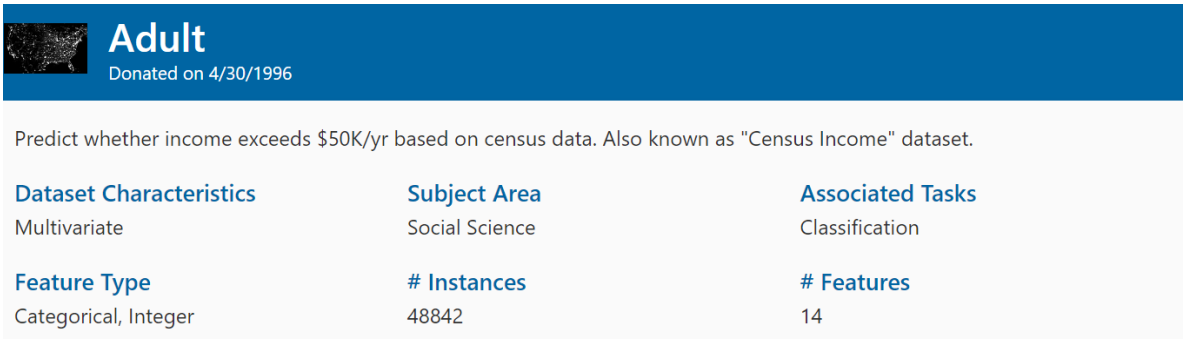
Taller de modelado matemático I

Proyecto: Modelos de aprendizaje estadístico

Este proyecto consiste en aplicar la metodología sobre modelos de aprendizaje estadístico para construir **un modelo de clasificación** empleando un conjunto de datos en particular. Los conjuntos de datos a seleccionar serán los siguientes:

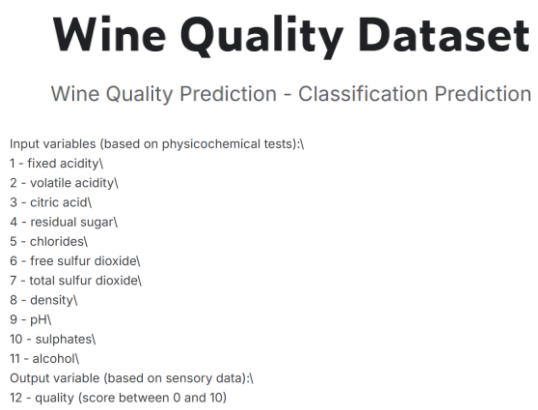
Equipo 1

https://archive.ics.uci.edu/dataset/2/adult



Equipo 2

https://www.kaggle.com/datasets/yasserh/wine-quality-dataset



Observación: Discretizar la variable quality para tener mala calidad (< 6) y buena calidad (>=6), y volverlo un problema de clasificación binario.

Equipo 3

https://archive.ics.uci.edu/dataset/105/congressional+voting+records

Congressional Voting Records

Donated on 4/26/1987

1984 United Stated Congressional Voting Records; Classify as Republican or Democrat

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Social Science	Classification
Feature Type	**# Instances**	**# Features**
Categorical	435	16

Equipo 4

https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Heart Failure Prediction Dataset

11 clinical features for predicting heart disease events.

Attribute Information

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

Equipo 5

https://www.kaggle.com/datasets/rakeshkapilavai/extrovert-vs-introvert-behavior-data

Extrovert vs. Introvert Behavior Data

Explore and Predict Social Behaviors and Personality Types

Size: The dataset contains 2,900 rows and 8 columns.

Features:

```
- Time_spent_Alone: Hours spent alone daily (0-11).
- Stage_fear: Presence of stage fright (Yes/No).
- Social_event_attendance: Frequency of social events (0-10).
- Going_outside: Frequency of going outside (0-7).
- Drained_after_socializing: Feeling drained after socializing (Yes/No).
- Friends_circle_size: Number of close friends (0-15).
- Post_frequency: Social media post frequency (0-10).
- Personality: Target variable (Extrovert/Introvert).*
```

Observación: Variable de respuesta binaria es Personality.

Equipo 6

https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset

📚 Students Performance Dataset 📚

Academic Success Factors in High School Students

About Dataset

This dataset contains comprehensive information on 2,392 high school students, detailing their demographics, study habits, parental involvement, extracurricular activities, and academic performance. The target variable, `GradeClass`, classifies students' grades into distinct categories, providing a robust dataset for educational research, predictive modeling, and statistical analysis.

Observación: Reagrupar la variable GradeClass para tener una variable de respuesta binaria: A,B,C,D en nivel aprobado y F en reprobado.

Observación: *LEAN bien los diccionarios de datos y la información que vienen de las variables en las ligas que les pasé. Es necesario entender bien que significa cada columna de marco de datos.*

Comentarios sobre el proyecto:

Para cada conjunto de datos se tiene que crear el modelo con el mejor poder de predicción para la clasificación correspondiente. Tienen que aplicar las técnicas de preprocesamiento así como los pasos correspondientes para ajustar cada modelo. Entre los modelos a implementar se tienen que considerar los que vimos en clase: modelos con regularización (Ridge y LASSO), árboles de decisión, bosques aleatorios, y modelos de gradiente potenciado.

SE tiene que formular un trabajo escrito, que debe estar bien estructurado (no importa que no sea muy extenso, de hecho, entre más conciso mejor) pero si debe contener su introducción, objetivo, análisis, resultados y conclusiones. Todo lo referente al código debe estar en un anexo. Si bien el objetivo es muy claro, siempre habrá algo que decir de todo el proceso de modelación, así como de

los potenciales modelos finales y del modelo final seleccionado, y de las variables más influyentes (desde el punto predictivo).

El trabajo escrito se envía más tardar el martes 1 de julio 10 pm.

Presentaciones

El lunes 30 de junio cada equipo va a presentar el proyecto, con un tiempo máximo de 20 min y 5 min de preguntas. Tienen que ser concisos y apegarse al tiempo, para que todos los equipos puedan pasar a exponer, por lo que se usará temporizador para monitorear la presentación.