Competencia en Plataforma de Ciencia de Datos: Kaggle

(Caso Integrador 2024)

Trabajo práctico en forma Individual

1_ Primero estudiar y realizar la ejercitación de los videos de las Clases 1, 2 y 3

2_ Resolver el mini caso y mandar la solución a la plataforma kaggle

Clasificador de Riesgo de Enfermedad Coronaria

Riesgo de Enfermedad Coronaria

Link de la competencia:
https://www.kaggle.com/t/4746c883bac144fdba1ed235a7832f87

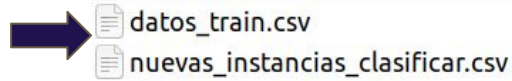
(*ENTREGAR EL DÍA: 30/05/2024 hasta las 11:00(am) hay tiempo)

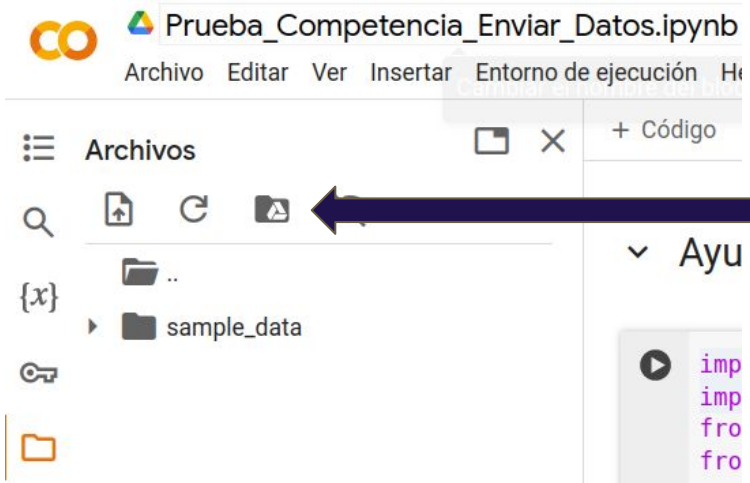
*****Recordar hacer las prácticas en forma INDIVIDUAL*****

*****La evaluación es individual*****

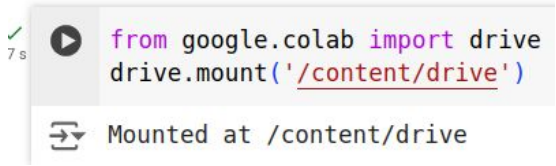
Para Leer los archivos: train, test

***1_ Pasar los archivos de la competencia a Drive**





***3_ Conectar con Drive**



***4_ Ejecutar**

***5_ Copiar ruta de drive y colocarla en pd.read_csv()**

```
pandas as pd
datos1 =
pd.read_csv('/content/drive/MyDrive/Gestion_2024/ruta_en_drive/datos_train.csv')
datos1.head(2)
```

***2_ En colab seleccionar**

Reglas y Puntaje

1_ Se debe trabajar en forma individual

2_ La evaluación es individual

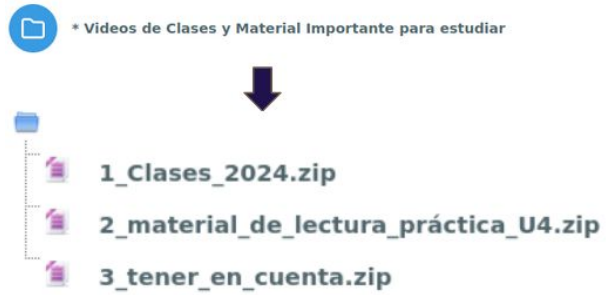
3_ Hay tiempo hasta el 30/05/2024 11:00 a.m. para mandar la solución del caso a kaggle.

4_ Se permiten 4 intentos de envío por día en la plataforma de ciencia de datos.

5_ Cumplir con el flujo de trabajo que se muestra en las filminas siguientes.

Flujo de Trabajo para Caso Integrador Kaggle

1_ Aplicar lo visto en las clases: 1,2 y 3.
2_ Seguir el método



Método

- Pipelines
- Utilizar Distintos Algoritmos de Minería de Datos. Definir línea base.
- Optimización de Hiper Parámetros
- Calibrar: Poner a competir a todos los métodos y posibles parámetros para determinar cual da el menor error
- Aplicar métodos de consenso y de potenciación

1. 1_ Importar las librerías
2. 2_ Obtener los datos y Analizarlos
3. 3_ Realizar el preprocesamiento de los Datos
4. 4_ Dividir el Conjunto de Datos
5. 5_ Crear el Modelo
6. 6_ Entrenar, Calibrar y Validar el Modelo
7. 7_ Probar el Modelo
8. 8_ Realizar Predicciones

Flujo de Trabajo para Caso Integrador Kaggle

La Notebook utilizada para enviar la solución a la competencia de Kaggle debe tener en tener en cuenta:

Análisis y exploración de datos.
El preprocesamiento:
- **Variables numéricas y categóricas**
- **Atributos importantes**
- **Valores nulos, faltantes o atípicos**
- **Desbalanceo de clases**
- **(Estudiar la clase 3)**

División de los datos. Si es posible utilizar validación cruzada.
Métodos y algoritmos utilizados para entrenar y obtener el modelo. Comparar varios modelos utilizando métricas.
Optimización de hiperparámetros. Calibración del modelo.
Validación y prueba del modelo
Utilizar las siguientes métricas: Matriz de confusión, classification report(recall, F1), kappa, curva roc.
Integrar la teoría con la práctica
Evitar el Overfitting y el Underfitting