Actividad 3.4 – Representación plot de datasets, selección de características y entrenamiento de modelos.

El objetivo de esta actividad es poner en práctica los conocimientos adquiridos para el preprocesamiento de datos, selección de características y entrenamiento de modelos.

Para ello es necesario seleccionar un Dataset que consideres oportuno, de clasificación o regresión, y distinto a los utilizados en clase.

Los puntos a desarrollar son los siguientes:

1. **(5%)** Describir el origen y breve explicación del Dataset, así como de cada una de las características.
2. **(5%)** Procesamiento de datos en el dataset: ajustes en características con datos no informados, conversión de variables categóricas, etc…
3. Utilizar las siguientes herramientas explicadas en clase para la selección de características:
   3.1. **(10%)** Matriz de gráficos de correlación.
   3.2. **(10%)** Matriz de gráficos de dispersión.
   3.3. **(10%)** SelectKBest.
4. **(5%)** Una pequeña reflexión sobre la elección de las características elegidas.
5. Con las librerías para NaiveBayes vistas en clase, entrenar el modelo que consideres más adecuado.
   5.1. **(10%)** Sin utilizar Cross Validation.
   5.2. **(15%)** Utilizando Cross Validation.
6. **(5%)** Obtener una conclusión sobre los resultados obtenidos en la predicción y evaluación al utilizar o no Cross Validation.
7. **(10%)** Además de las herramientas indicadas anteriormente, se valorará la utilización de alguna otra herramienta o técnica no vista en el curso para la selección de las características.

(10%) Formato:

- El pdf final tiene una portada.
- Se utiliza un índice, apartado de conclusiones y referencias (web).
- Se hace mención a referencias externas, no recogidas en el material suministrado.

(5%) Aportaciones personales:

- Se enriquece la actividad con aportaciones personales distintas a las solicitadas en la propia actividad: de opinión, estrategia, herramientas utilizadas en la resolución de la actividad o elementos gráficos (dibujos, fórmulas, etc) en el propio cuaderno de Google Colab

Además de las diapos de clase, quizás te puedan ayudar los siguientes recursos:

Título: seaborn.load_dataset
Url: http://seaborn.pydata.org/generated/seaborn.load_dataset.html#seaborn-load-dataset

Título: seaborn.pairplot
Url: https://seaborn.pydata.org/generated/seaborn.pairplot.html

Título: mwaskom/seaborn-data
Url: https://github.com/mwaskom/seaborn-data

Cada pregunta se evaluará entre 0 y 10 atendiendo a los siguientes criterios:

Puntos	Clasificador
De 0 a 1	Nada adecuado
De 2 a 3	Mínimamente adecuado
De 4 a 5	Algo adecuado
De 6 a 7	Moderadamente adecuado
De 8 a 9	Muy adecuado
10	Excelente

Formato de entrega

- Nombrar el archivo siguiendo el siguiente patrón:

  SNS_ACT3_4_NombreApellidos.pdf

- Entregar un fichero en formato pdf con el siguiente contenido:

  - Incluir en el pie de cada página el nombre y apellidos del autor/a, así como el número de página y el total de páginas que contiene el documento.
  - Las imágenes capturadas han de tener la resolución necesaria para una buena visualización.
  - Indicar la url del GitHub donde se encuentran el/los diferentes cuadernos que has utilizado, con el objeto de consultarlos para descargarlo y verificar su funcionamiento. **En el caso de no indicar la url de Github se restará un punto en la nota final de esta actividad.**
  - **Importante**: Respetar la estructura de las preguntas, con el objeto de permitir una corrección homogénea para todas las actividades. De lo contrario se valorará con ceros puntos.
  - Al comienzo del notebook poner el nombre y apellidos del autor/a.