Trabajo práctico de introducción a la IA.

En este proyecto vamos a estar realizando el análisis y la preparación de datos tabulares (dataframe) para con ellos entrenar modelos de clasificación que sirvan para determinar la potabilidad del agua.

Herramientas y librerías

- Python 3.0
- Google Colabs (Jupyter Notebook).
- Pandas
- Seaborn
- Matplotlib
- SKlearn
- Imbalanced Learn

Proceso de trabajo:

1) Búsqueda del Dataframe:

Para este trabajo se utilizó un dataframe de Kaggle de clasificacion de agua en base a su potabilidad, este dataframe cuenta con las siguientes variables:

- PH `float64`
- Dureza `float64`
- Cloraminas `float64`
- Sulfatos `float64`
- Conductividad `float64`
- Carbón Orgánico `float64`
- Trialometanos `float64`
- Turbiedad `float64`
- Potabilidad `int64` *

> * Téngase en cuenta que la potabilidad se encuentra en formato int64, teniendo la no potabilidad el valor de 0 y la potabilidad en valor de 1

> Puede obtener el dataframe con el siguiente comando:

```
wget https://raw.githubusercontent.com/laaledesiempre/machine_learning_PW/main/archive.zip
```

2) Análisis de datos

El dataframe contaba con múltiples datos nulos, los cuales, después de analizar si llenarlos con otra informacion, se decidio directamente eliminar, reduciendo considerablemente la muestra.

Un análisis superficial de la relación de las variables por medio de un heatmap, se indicó que ninguna era potencial de ser eliminada en ese momento. Posterior a ello se escalaron los valores para evitar inclinar la balanza con las variables de valor más alto.

También se pudo ver que los datos eran, principalmente, de cuerpos de agua no potables, lo que se tomó como posible dificultad para los modelos, por los que, en la primera etapa se realizaron dos entrenamientos diferentes para cada modelo, uno con los datos sin balancear y otros balanceados.

3.1) Entrenamiento de modelos.

> No se detallaran los resultados de todos los modelos en esta seccion, para una referencia mayor, dirijirse a el dataframe original

Modelos probados:

- Regresión Lineal
- Árbol de decisión
- K hermanos vecinos
- Red neural
- Support Vector Machine
- Random Forest

De todos los modelos pudimos sacar las siguientes conclusiones:

- El conjunto de datos balanceado no beneficio, sino perjudicó a los modelos a la hora de clasificar
- Los modelos mas prometedores para la siguiente etapa eran: Árbol de Decisión, Red neural y SVM
- Los modelos antes mencionados no parecían diferenciar bien los potables pero si diferenciar muy bien los no potables

3.2) Entrenamiento de modelos con informacion reducida

En base a lo indicado por el modelo de Árbol de decisión se extrajo las 5 caracteristicas mas notables y se volvió a entrenar los modelos, destacando principalmente el modelo de Arbol de desicion, que nos dio una muy alta precisión para detectar los no potables.

4) Conclusiones y objetivo final del modelo.

> Esta seccion esta directamente extraida del cuaderno de entrenamiento.

De todos los modelos analizados, el arbol de decision donde se extrajo la mayoria de columnas y se dejo solo las reelevantes nos dio lo siguiente:

La mayoria de los positivos son, efectivamente, positivos en potabilidad En cambio, si bien casi todos los negativos estan efectivamente marcados como negativos, algunos positivos tambien lo estan habria que contrastar esto con otros dataset pero, en base a la informacion presente, se podria utilizar este modelo para tener un primer aproximamiento al descubrimiento de cuerpos de agua potable.

al pasar un grupo de cuerpos de agua por este modelo, si el resultado de "Potable". entonces es muy probable que sea potable, por lo que el hacer un analisis de potabilidad puede ser prioritario en caso de estar buscando cuerpos de agua potables, y a su vez, puede quitarsele prioridad en caso de buscar identificar no potables.

suponiendo la siguiente situacion:

Un grupo de cientificos toma muestras de diferentes cuerpos de agua en una zona donde el acceso al agua potable es limitada y principalmente es obtenidad de cuerpos de agua natural. pasan los resultados de sus muestras por este modelo, y ponen como prioridad el analizar los cuerpos de agua resultantes como positivos, asi se puede asegurar que los primeros en ser analizados van a resultar, en su mayoria, potables, y asi asegurar el que la poblacion de este sector extraiga su agua de estas fuentes con seguridad. y deje de tomarla de otras potencialmente no potables, y luego de analizar todas las designadas como "Potables" por este modelo, ya se puede proceder a analizar las "no potables" asi se puede marcar definitivamente cuales verdaderamente no son potables y deben ser suspedido su uso para consumo humano y cuales son falsos negativos.