Lab 1

Our first dataset: https://www.kaggle.com/datasets/adityakadiwal/water-potability.

This dataset contains water quality metrics for 3276 different water bodies. It contains the pH value, sulfate content, hardness, turbidity etc. of different water bodies and the result is if the water is potable or not.

Pros: - Large number of records - All numerical values - Output is binary only

Cons: - Multiple null values

Our second dataset: https://www.kaggle.com/datasets/michaelacorley/unemployment-and-mental-illness-survey

This dataset contains records that investigate the relation between mental illness and unemployment levels. Values include any mental illnesses, education level, annual income etc.

Pros: - Records with binary values exists - Large number of attributes
Cons: - Few attributes have values within ranges

Our third dataset: https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri

This dataset contains images to classify brain tumor into 4 categories. MRI images of different types of tumors are given. These images have similar features which cannot be distinguished by humans, so the use of ML is required to classify these images according to the tumor type.

Pros: - Its an image dataset, so CNN is to be used. - High accuracy of result

Cons: -Large number of images to process

We have chosen the water potability dataset, as it contains more records than the other two datasets. The output is also binary, so it is easier to get an accurate result. The problem with the chosen dataset is that there are almost 1000+ null values, which has to be filtered.