Differential Privacy Project on a Healthcare Dataset

In this project, the main aim is to allow the students to further explore the different classes and methods available and provided in the python library “diffprivlib”. You are expected to implement the following:

- Find a medical dataset of your choice in Kaggle (https://www.kaggle.com/).
- Consider a classification supervised task and an unsupervised task which is clustering.
- In the classification task, create multiple machine learning models with epsilons ranging from [0,01, 0.5], then create a voting classifier model based on the multiple models and find the best voting classifier that yields the best performance.
- As a performance metric consider macro F1-score.
- For the clustering task, use the DP version of KMeans provided by diffprivlib (https://diffprivlib.readthedocs.io/en/latest/modules/models.html#diffprivlib.models.KMeans) try multiple epsilons and create the elbow curve (https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/) for each epsilon of your choice and find the best number of clusters.