step 1

your task is to use machine learning to make some predictions about something of your choosing. The steps of this assignment are:

1. Select a dataset. You can select one of the datasets from Kaggle's dataset collection (https://www.kaggle.com/datasets), or from any other source you wish.

2. Select a subset of features from the dataset to use for your learning and select an outcome (label) to estimate.

3. Build at least three classifiers in Scikit Learn that estimates what you set out to estimate.

Note

• You should make sure to go through each step of "End-to-End Machine Learning Project". (except fine-tune model).

• You should make sure to apply the concepts for training and building the classifiers. (Ex: confusion matrix, precision, recall, F1 score, ROC curve, AUC, error analysis, … etc.)

Step 2

Based on the classifiers applied on step1 , you should perform two methods form the following (choose two methods from 1-4):

1- Develop an ensemble learning algorithm of the chosen classifiers based on the majority voting rule.
2- 2- Develop an ensemble learning algorithm of the chosen classifiers based on the soft voting rule.
3- 3- Develop an ensemble learning algorithm on one of the chosen classifiers based on bagging method.
4- 4- Develop an ensemble learning algorithm on one of the chosen classifiers based on pasting method.

Then:

5- 5- Do comparative analysis of the created ensemble learning algorithms with the single classifiers in terms of the accuracy, precision, recall, … etc.

Report:

The assignment should be documented as a technical report.

The report should include the following:

• Abstract: It should consist of 1 paragraph consisting of the motivation for your work, the explanation of the methodology you used, and the results obtained.

• Introduction: Explain the problem and why it is important. Clearly state what the input and output is. For example: “The input of the used algorithm is an {image, patient age, etc.}. We then use a {SVM, decision tree, linear regression, etc.}. The output of the used algorithm is a predicted {age, stock price, cancer type, music genre, etc.}.” This is a very important point that needs to be clarified since different teams have different inputs/outputs about different application domains.

• Dataset: Describe your dataset: how many training/validation/test examples do you have? Did you do any preprocessing? Did you do any normalization? What is the resolution of your images? Add a citation on where you obtained your dataset from, present some examples from your dataset, write about the features you used, and why you choose them

• Methods: Describe your learning algorithms/proposed algorithm(s). For each algorithm, give a short description of how it works.

• Experiments/Results/Discussion: If you use (hyper)parameters, you should give details about what (hyper)parameters you chose. Did you do cross validation, if so, how many folds? List and explain what your primary metrics are: accuracy, precision, recall, confusion matrix. etc. Make sure to discuss the figures/tables in your report. Any plot should include legends, axis labels, and have font sizes that are clear enough.

• Conclusion/Future Work: Summarize your report. Which algorithms were the highest performing? Why do you think that some algorithms worked better than others? For future work, what would you do?

• References and Contributions! The final report should be at most 5 pages long. Please include a section that describes what each team member worked on and contributed to the project.