Machine Learning Challenge -1: Description

Task Details:

You are given a dataset as a .txt file “Parkinsons disease raw data.txt”

Your task is to convert this to a more **structured data** and use it to train **Machine Learning models** to predict whether a person is affected with Parkinson’s disease or not. You also need to determine which model has the highest accuracy in this case and why.

Dataset Details:

This dataset is composed of a range of biomedical voice measurements. There are totally 195 observations and 24 features.

If we number the features from 0, then the 17th feature is your target column.

1 --> Parkinson’s; 0 --> Healthy

Dataset Link:
https://drive.google.com/file/d/1XdcyDqylj3rTnO4pZd5YHW84HFULeCEm/view?usp=sharing

Expected Work Flow:

(0. Do a basic research on what is meant by Parkinson’s Disease. What are all the symptoms & other details)

1. As the data is in an unstructured format in a txt file, convert it to a more structured DataFrame in Python. You are requested not to use any other tools like Excel. Please use Python alone.

2. Once you structure the data, do some basic data pre-processing & data analysis. Check whether the dataset contains any missing values. If it contains missing values, see what you can do about it.

3. Standardize the dataset and split it to training data and testing data.

4. Important: You are expected to try 4 classification models:

1. Logistic Regression

2. Support Vector Machine Classifier : kernel = Linear

3. K – Nearest Neighbors

4. Random Forest Classifier

5. Create a function that can use these 4 models separately and give the accuracy score of each of these models separately. Make sure that you are training your model with the Training data & you are evaluating it with thee Test data.

Your function’s output should look something like this:

```
Accuracy score for the  LogisticRegression()  =
Accuracy score for the  SVC(kernel='linear')  =
Accuracy score for the  KNeighborsClassifier()  =
Accuracy score for the  RandomForestClassifier()  =
```

6. Once you complete this, check which model has the highest accuracy and write an inference about why that particular model can perform better.

NOTE: You don’t need to do any optimization or tuning to increase the accuracy score. That’s a topic for a different day.

“You can also try to deploy your trained model if you want.”

OUTCOME: This challenge is not a competition. There is no winning or losing. This challenge is completely for your learning purpose. Once you complete this challenge, you will have an idea on how you can deal with a dataset that needs a lot of processing before training Machine Learning models. If you can deploy the trained model, you will have an end-to-end Machine Learning use-case.

Reference Videos:

1. Parkinson’s Disease Prediction: https://youtu.be/HbyN_ey-JVc
2. Model Deployment: https://youtu.be/WLwjvWq0GWA

Challenge Video: December 3, 2021. 5:30 pm; Link:

(You can refer the Solution video after you have tried the challenge for 2 or 3 days)

Solution Video: December 6, 2021. 5:30 pm

Google Form Link: https://forms.gle/2yRR1woHWiWFg8W46