EE 257 PROJECT (SPRING 2023)
FINAL PROJECT REPORT/CODE DUE: MAY 15th

In this project, you will study a classification problem using a dataset in UCI Machine Learning Repository. The project may be done in groups of at most 2 students.

Follow the steps below to complete the project.

1. Pick one of the following datasets from the UCI repository:

- **Accelerometer Dataset** (2021)
  https://archive.ics.uci.edu/ml/datasets/Accelerometer
- **Room Occupancy Estimation Dataset** (2021)
  https://archive.ics.uci.edu/ml/datasets/Room+Occupancy+Estimation
- **AI4I 2020 Predictive Maintenance Dataset (2020)**
  https://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset
- **Early stage diabetes risk prediction dataset (2020)**
  https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.
- **Deepfakes: Medical Image Tamper Detection Data Set** (2020)
  https://archive.ics.uci.edu/ml/datasets/Deepfakes%3A+Medical+Image+Tamper+Detection

2. **Sign-up to a project group on canvas ->People->Project Groups (each project is limited to 6 students) Each student MUST sign-up. It is NOT enough only one partner signs-up.** (**DUE April 6th**)

PROJECT PARTS

a) Data Set Description

Explain in detail the dataset: attributes, input variables, output variable (only one), data types, missing data, etc.

(**DUE: APRIL 13** - Textbox entry on Canvas)

b) Data Set Visualization

Visualize your data with appropriate plot. See Hands-on Machine Learning with ScikitLearn, Keras & TensorFlow (Chapter 2: Pages 42-62)

(**DUE: APRIL 20** – Upload the code with comments)

c) Data Set Cleaning

Clean your data. See Hands-on Machine Learning with ScikitLearn, Keras & TensorFlow (Chapter 2: Pages 62-72)
Describe how you have handled the outliers if there exists any.

d) Related Work

Describe at least one paper that has used this dataset. UCI website lists related papers for each dataset. What problem did the paper solve? How did they use the data set? What results they obtained? Etc

e) Feature Extraction

Extract (eliminate) features using different methods we learned in class

f) Model development

- Classification: Select and train at least FOUR different classification models (models should be part of course syllabus, see the list below for possible models).

g) Fine-tune your models & Feature Set:

Evaluate your models using test set. Fine-tune your model, and optimize feature set. Consider also applying regularization.

e) Performance:

Measure the performance of your classifiers using different metrics

TO BE SUBMITTED:

Code [15% of the Grade]: Well documented code with a ReadMe file. I should be able to run the code, and obtain the results provided in the report.

Report [85% of the Grade]: Your report should have the following sections:

a) Data Set Description (5 points)
b) Data Set Visualization (5 points)
c) Data Set Cleaning (5 points)
d) Related Work (5 points)
e) Feature Extraction (5 points)
f) Model development (10 points)
g) Fine-tune your models & Feature Set (10 points)
h) Performance (15 points)
i) Overall discussion of results & Conclusions (15 points)

Your report will also be graded on Originality (10 points).

NOTE: EACH STUDENT MUST WRITE THEIR OWN REPORT even though they might share the same code and plots with their partner.

ALLOWED CLASSIFICATION MODELS:

1. Logistic Regression
2. LDA/QDA
3. KNN
4. Support Vector Machines
5. Decision Trees
6. Random Forest

THINGS YOU SHOULD NOT DO IN YOUR REPORT:

- Do not have exact the same sentences with your partner or anybody else (turn-it-in will check and give similarity results). Do not share code or figures with anybody other than your partner.

- Do not take figures from online or others sources without referencing

- **Do not use online codes without referencing properly ==> You will lose your 15 points (code) + 10 points (originality) automatically. Explain any code you used from other sources in your ReadMe file and also Discussions & Conclusions.**

- **Do not put CODE IN THE REPORT**

- Do not copy and paste CODE OUTPUTS into the report. Convert them to appropriate plots or tables. Make figures whenever possible, do not send me tables with more than 30 lines in it. Tables are for relatively small size.

- Provide all the details of the simulations (not the code, the parameters you use). I should be able to repeat your simulations with the given info in the report.

- Do not submit late

- Make sure you have your partner's name on the report