CS350-Machine Learning

Project

Due Date: 24 Dec, 2020.

Projects are to be done individually. No late project will be accepted. Submit your project by 11:59 pm on the day that it is due.

Submissions that do not comply with the specifications given in this document will not be marked and a zero grade will be assigned.

1. Introduction:

Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Machine learning is a commonly used technique for processing enormous data in the healthcare domain. Researchers apply several machine learning techniques to analyses huge complex medical data, helping healthcare professionals to predict heart disease. In this project you are provided dataset related to heart disease, and you have to train the model on basis of supervised learning algorithms as Naïve Bayes, decision tree, K-nearest neighbor, SVM, and random forest algorithm etc.

2. Project Task

You have to perform a classification on provided dataset of Heart Attack prediction. For that prediction, first of all you have to preprocess dataset (handle missing data, irrelevant data removal, noise removal etc.) and convert it into appropriate numeric form. Then you have to perform any Machine Learning model for the prediction of Label.

Dataset files Description:

X_Train.csv: It contains all Training Dataset on which you are going to build your ML Model
X_text.csv: It contains total Test Dataset on which you are going to test your ML Model
Y_train.csv: It contains all Training Dataset label.

Here are some outcomes of the project that I expect from that I expect from you.

1. Data Preprocessing and Over the course of the semester we've discussed situations that are problematic for our algorithms, such as the "curse of dimensionality" or the difficulty of learning from unbalanced training data and missing values problem. You might want to explore techniques for handling these problems in the given dataset.
2. Applying more than 5 Machine Learning algorithms for the prediction and apply majority voting concept for the final output. Majority voting means to take Predicted labels from all applied classifiers and assign final label according to majority vote.
3. Exploratory Data Analysis: Lot of graphs to understand the Data insights. Minimum 10 different Graphs required

4. Accuracy, F Score more than 65 percent. The more the F-score, more marks will be awarded.
5. Join the kaggle competition with the below link
   https://www.kaggle.com/c/ml-heast-disease-prediction-project-fa20/leaderboard
   And upload your model predicted myprediction.csv file on it. After that you can check the performance of your model on leaderboard.
6. Project Report in which you mention the detail of following questions

    i. What features did you use and why?
    ii. What classification techniques did you try?
    iii. Which of the methods (and for what hyper parameters) showed best performance and why? Explore an "Issue" in Machine Learning
    iv. What test accuracy are you expecting?

You have to submit the following files:

1. Training.ipynb: The file containing the code to train and dump the final model using the chosen hyper parameters.
2. Testing.ipynb: The file in which you load the saved model and generate and save predictions for test data.
3. myPredictions.csv: The csv file containing your predicted labels for the **test set.**
4. Project Report

Note:

Please name your submission files in the following manner:

<roll no>_<name>_<section><MLProject>

For example: 150915_Ahmad_Ali_ML_A_MLProject

Note: You can use Numpy, Pandas, Sklearn, Matplotlib library for that.

Honor Policy

This assignment is a learning opportunity that will be evaluated based on your ability to work through a problem in a logical manner and write a research report on your own. You may however discuss verbally or via email the assignment with your classmates or the course instructor, but you are to write the actual code for this assignment without copying or plagiarizing the work of others. You may use the Internet to do your research, but the written work should be your own. **Plagiarized code will get a zero.**