EEP 596: AI and Health Care || Programming 1

Dr. Karthik Mohan

Univ. of Washington, Seattle

Mar 28, 2022

Submission guidelines

- You have to submit a Jupyter/IPython notebook file, report and Kaggle predictions as part of your submission.
- You can use the template notebook given and add your solutions to it.
- The report should be in pdf format and have all plots, correlation matrix's and tables added in it. Feel free to use either latex or word for creating it. It should answer all the theoretical questions in the write up below, and show your learnings and insights.
- There is a Kaggle competition as well for this assignment, submit your predictions on the "held out" test data set for a fun peer learning experience!
- You may discuss/brainstorm ideas to solve the assignment with peers - However, your submission should be your own and show your code implementation.

Problem overview and dataset description

The main objective is to do binary classification on the given dataset. All the patients have diabetes, and a 1 in the target refers to a high-risk of chronic diabetes, while a 0 refers to a low-risk of chronic diabetes. We have 10 numeric features - age, sex, body mass index, average blood pressure and six blood serum measurements for each of diabetes patient. You can work on data.csv, it has both the feature and target values. The kaggle.csv only has feature values, you need to find out the target for it and upload it on kaggle to enter the leaderboard.

Diabetes Classification

1. Plot the correlation matrix for the data set, and discuss how the features affect the target, and show a few scatter plots to substantiate your findings. In the scatter plot, do you observe that there is a correlated change in target values with the change of most correlated features? (15 marks)
2. For the binary classification problem, use logistic regression and decision tree models to get F1score, precision, recall and accuracy scores. Also mention hyperparameter tuning done to get the best results. You can use sklearn library for this. (15 marks)
3. Feature selection - Find out the top 5 features for both the models. For decision tree, be sure to discuss about the information gain measure. (20 marks)
4. Use decision tree to explain a few positive and a few negative examples and why the model chose to predict that way. (20 marks)

Diabetes Classification

1. Plot the F1score for different threshold values for both the models on the same graph. Analyze and discuss the AUC, and which model performs better for different thresholds? (20 marks)
2. Kaggle submission (10 marks)

Google forms for bonus!

Please fill up the following 3 google forms if you haven't already for 5 bonus marks!

Pre-course Survey

Class and OH timings

New healthcare topics?