Report:

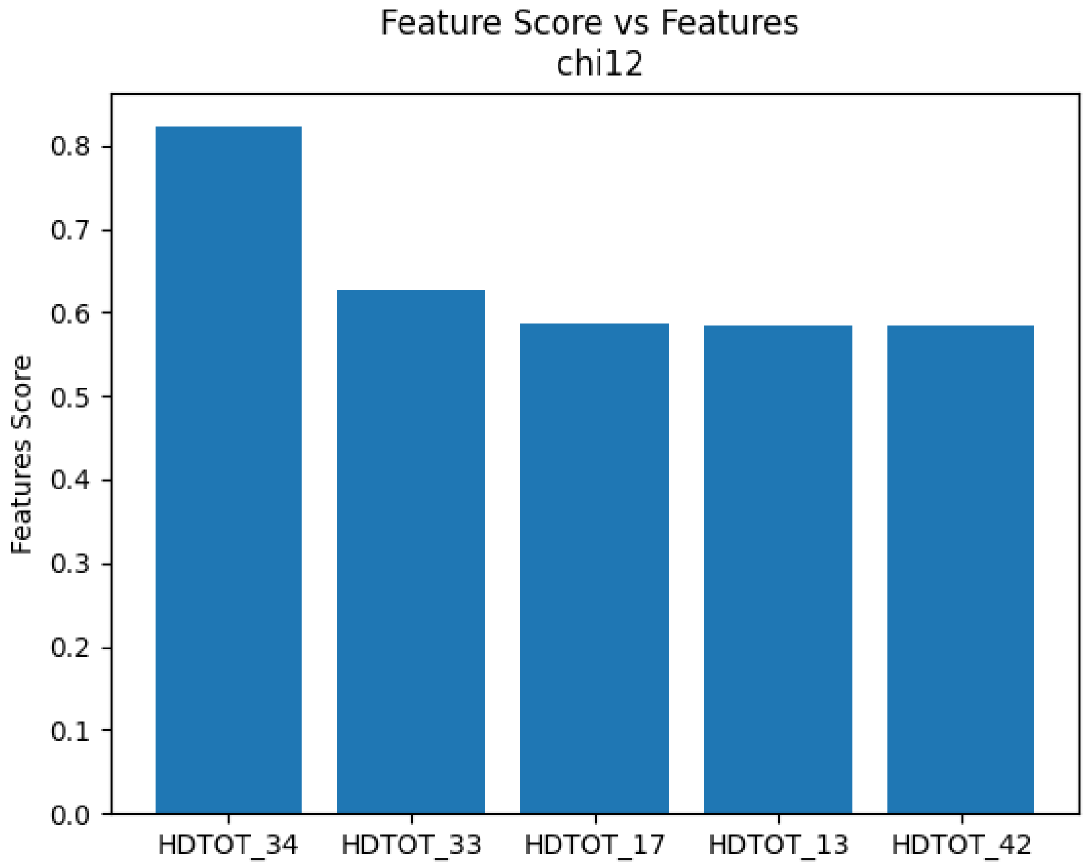
Assumptions (which are not necessarily true):

- Medicine features are independent - there is no effect of one medicine on the other or a combined effect of both.
- There is no correlation between medicine features and demographics. No impact of location on the medicine given.

Data preprocessing:

Delete all unmatched IDs. Delete the row with -2 value in features matrix, only 1 row with such value.
Return the row with -2 in features matrix to delete afterwards from y_teacher.
Binomial labeling based on a given threshold per subject.

Training results:
Most significant features found:



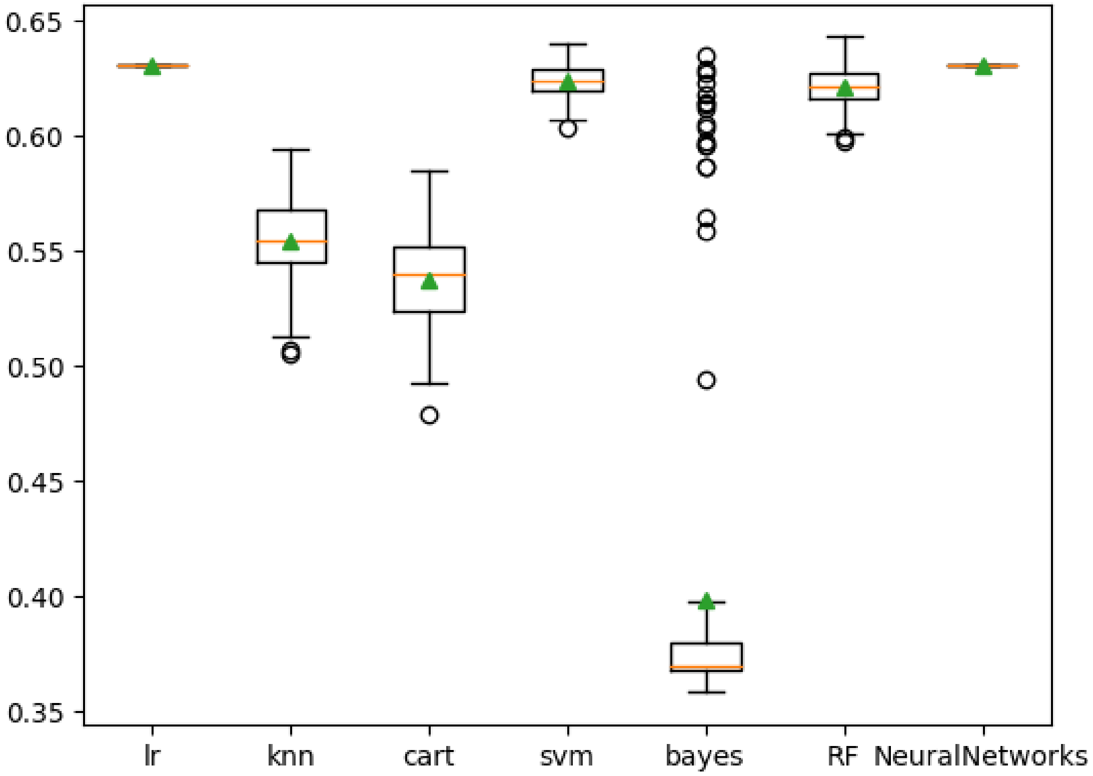
They were found using the chi^2 algorithm for probability distribution, a significant difference between features based on the binomial classifier.

Modeling:

In order to cross reference the demographic data, I took the data of the patients who got significant improvement and checked whether they have a significant difference in demographic information compared to the whole group.

Unfortunately significant of <0.05 was not found for the demographic value.

In order to find the best model, a cross model ensemble was done. The best models achieved 63.1% accuracy.



Low accuracy can be explained by lack of data and insufficient variation of data.

Explanation:
We can see that both logistic regression and neural networks achieved the same results. We can infer that NN neurons have learned the same model as LR. That means that the weights of the NN composes a LR curve to classify the data.
From that we can assume that bigger databases with deeper layers of NN would be able to reach better accuracy results by learning more sub-features, differences, correlations and patterns.

Features to add:
Implement week based non-binomial classifier based on slope of week vs depression improvement. Set 2 different models - one for patients who did not pass the threshold. And one for patients who had and it is possible to track their week vs depression slope. (code is commented)

Add more data to improve SVM model accuracy

** There is vast potential in using more complex supervised models with DL to get better model accuracy. For that more data is required than given in the exercise.