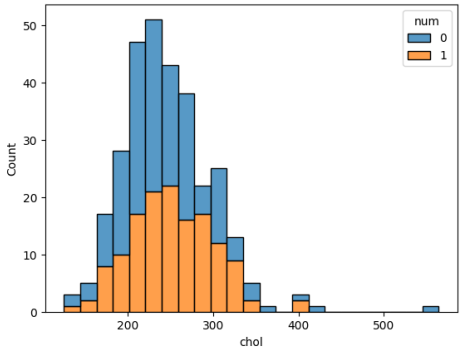
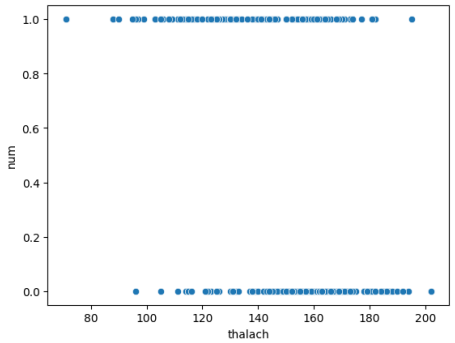
CSE 343 : Machine Learning

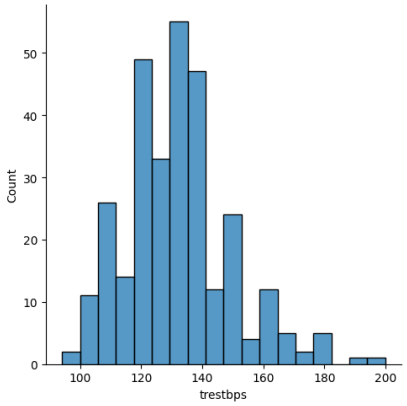
By Deepanshu Dabas (2021249)

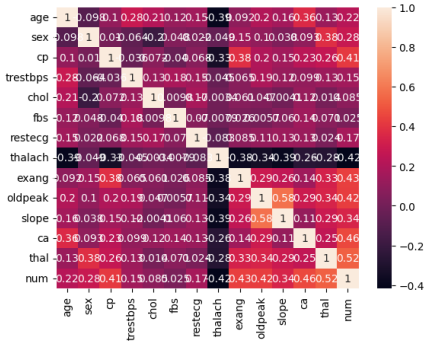
Section B

- Filled missing values in rows using mean and deleted duplicate rows also.
- EDA with scatter plots, heatmap correlation, bar graph, etc









- Grid search with various value of cv , max_features, min_samples_split, with best performing being cv=6
- Accuracy = 0.90 on test data and 0.86 training data for random forest

  0.79 for Decision Tree after grid search and parameters find.

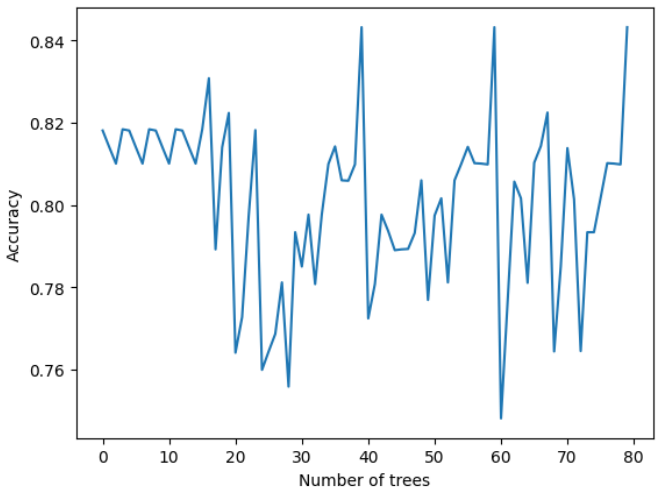
- Used sckit learn package for this problem.

	precision	recall	f1-score	support
0	0.84	0.92	0.88	135
1	0.88	0.79	0.83	107
accuracy			0.86	242
macro avg	0.86	0.85	0.86	242
weighted avg	0.86	0.86	0.86	242

Accuracy on training data: 0.859504132231405

	precision	recall	f1-score	support
0	0.87	0.93	0.90	29
1	0.93	0.88	0.90	32
accuracy			0.90	61
macro avg	0.90	0.90	0.90	61
weighted avg	0.90	0.90	0.90	61

Accuracy on test data: 0.9016393442622951



Section C:

- Used recursion for splitting nodes and used cost function to calculate best feature and selecting next feature.
- Both Information gain and gini impurity have the same score.
- Model has 0.99 accuracy and 0.025 MSE
- Used only numpy for model implementation
- Done pre-processing of given data also to test model better by encoding categorical features into numerical values.
- The code is well commented and every step logic is written clearly
- All the necessary functions were implemented
- Max depth can be passed during the initialization of function as well as later on by calling the function. You need to supply cost function criterion during initialization. By default ,max depth will be set to none and cost function will be set to gini impurity.
- Accuracy using Mean: 0.9909255898366606

  Error using MSE: 0.025408348457350273