Predicting Heart disease from a set of attributes of patient's health

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease. Link for the dataset: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

1. Data:

Column Descriptions:

id (Unique id for each patient)
age (Age of the patient in years)
origin (place of study)
sex (Male/Female)
cp chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])
trestbps resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))
chol (serum cholesterol in mg/dl)
fbs (if fasting blood sugar > 120 mg/dl)
restecg (resting electrocardiographic results)
-- Values: [normal, stt abnormality, lv hypertrophy]
thalach: maximum heart rate achieved
exang: exercise-induced angina (True/ False)
oldpeak: ST depression induced by exercise relative to rest
slope: the slope of the peak exercise ST segment
ca: number of major vessels (0-3) colored by fluoroscopy
thal: [normal; fixed defect; reversible defect]
num: the predicted attribute

This problem is a classification problem i.e to predict the presence of heart disease or not using some other attributes about patient's health.

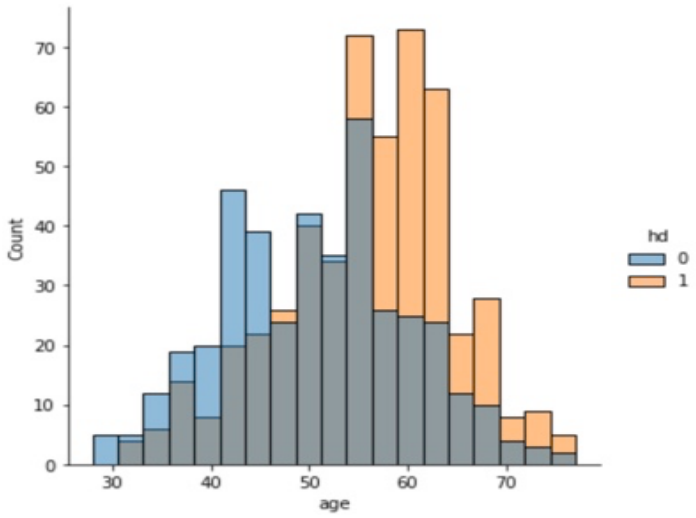
2. **Models used for prediction are:**

    1. DecisionTreeClassifier,
    2. RandomForest
    3. GradientBoostingClassifier
    4. Logistic Regression
    5. and SVM

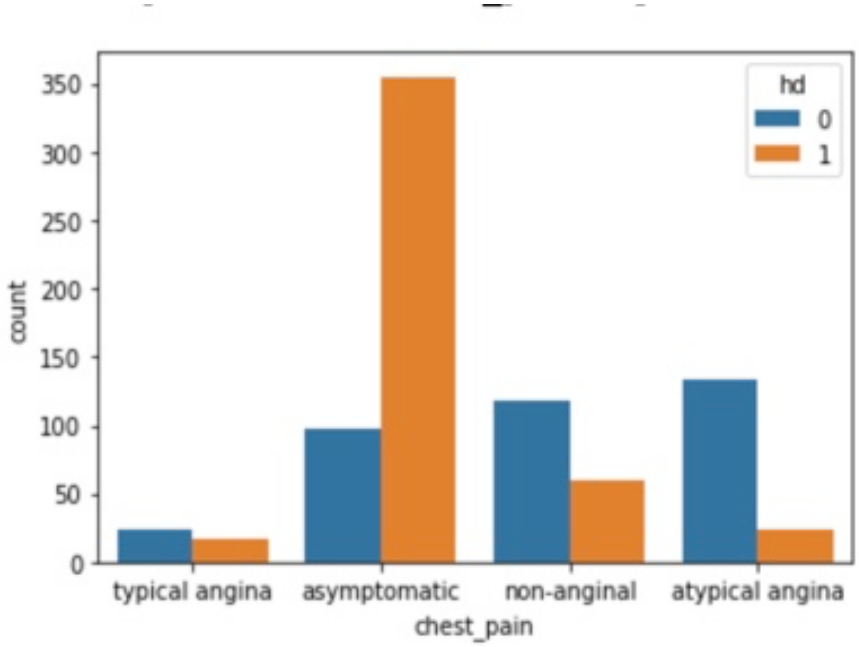
3. **Data Cleaning:**

    1. columns which are not relevant for prediction are removed like ‘id’
    2. column ‘hd’ represent different level of heart disease , here we are only predicting the heart disease and not the level so all the values >=1 are replaced as 1 else 0.

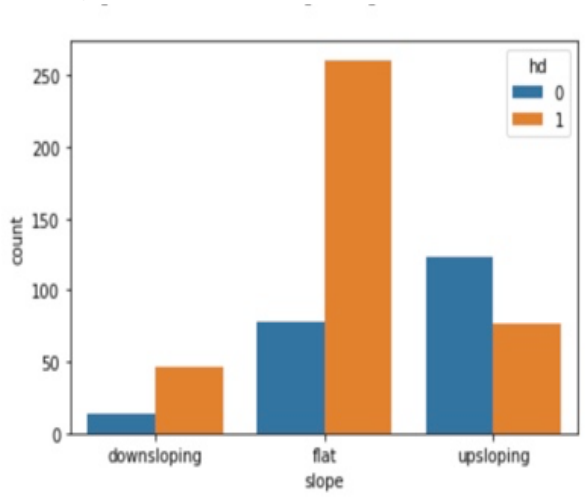
4. **EDA**



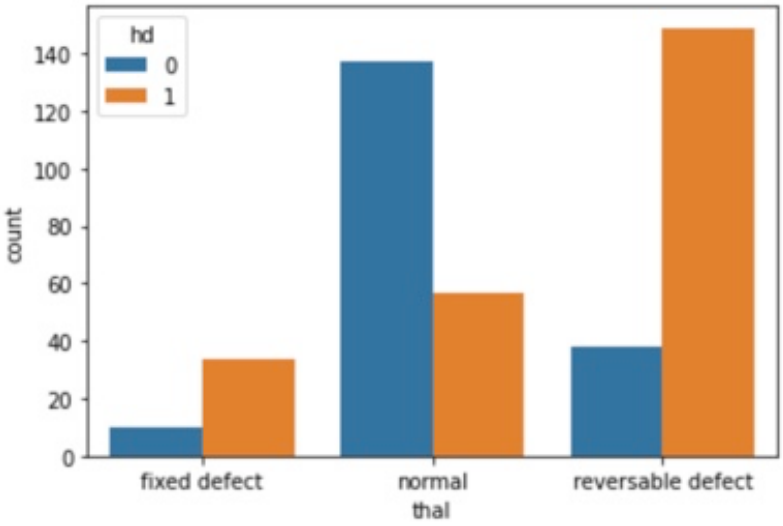
Heart disease age group wise, according to the graph heart disease is present in age group mostly between 55-70



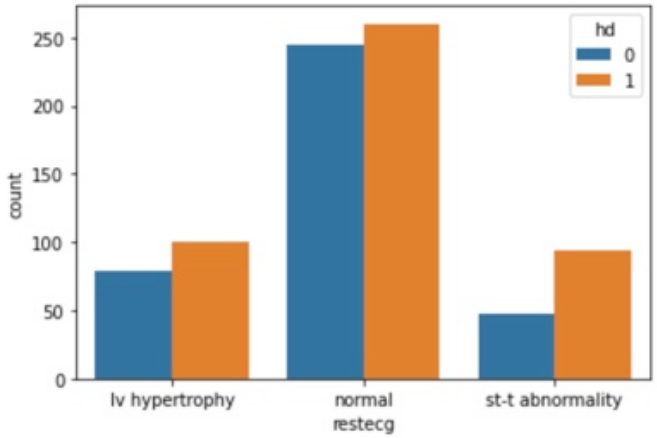
Correlation between chest pain and heart disease:
It shows heart disease in mostly those patients who have asymptomatic chest pain



Looks like ‘flat’ slope is highly correlated to heart disease.

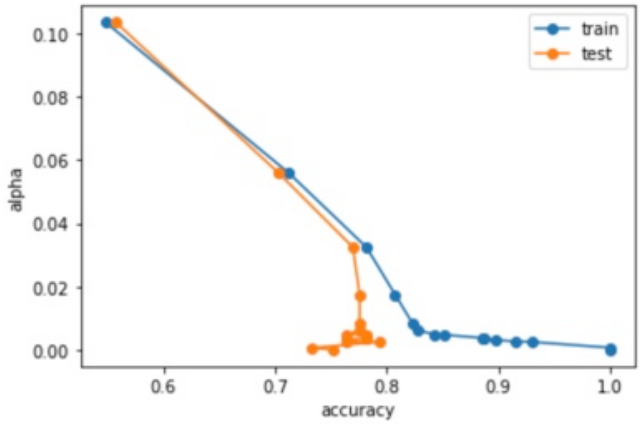


The ‘reversable defect’ has high correlation with hd.

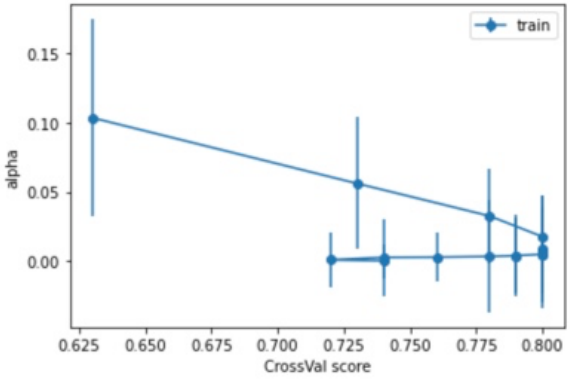


All the type of restecg are equally correlated with hd

5. **Feature Engineering:**
   Here the categorical columns like 'sex','chest_pain','fbs','restecg','exang','slope','thal' are converted into numerical columns using get_dummies method
6. After feature engineering data is divided into train and test set so that models can be trained on train set and tested on test set
7. Hyperparameter tuning for models:
   Choosing alpha parameter for decision tree



Here base decision tree model is trained on training and prediction is made on both training and test set with different alpha values and accuracy is plotted for both the set of data for each alpha value

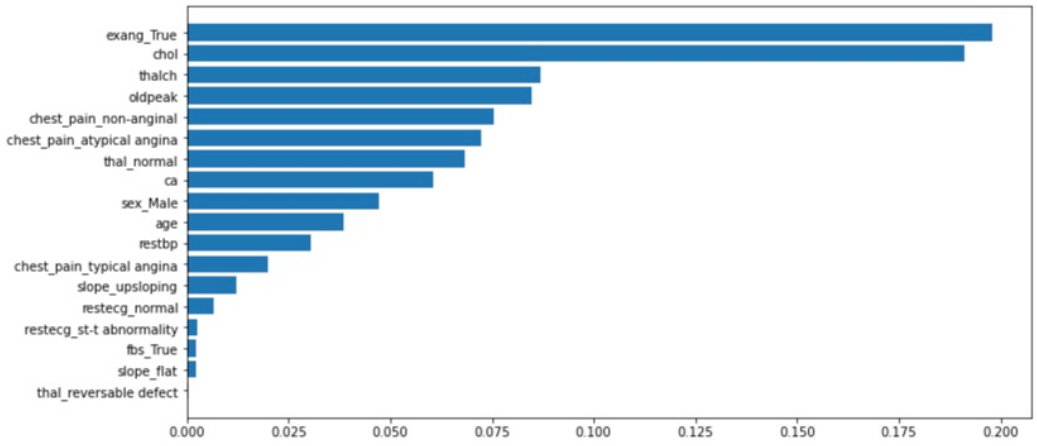


Here again base DT model is trained on training data and then cross validation score is computed with cv score = 5 on set of different alpha values. Here the standard deviation is shown as error for each value and corresponding cross validation score

From both the above graph alpha value of 0.02 is a good value as it is showing accuracy of around 80% on both training and test set

8. After Decision tree hyperparameter tunning is done for all the models like Random Forest, Gradient boosting, Logistic Regression and SVM using Grid Search CV method.

9. Feature importance



According to gradient boosting model feature importance of each feature in dataset is shown as above.
It shows that 'exang_true' and 'chol' are the top two highly correlated features

10. Comparison of models:

hyperparameter tunning is done to choose parameters of all the models like alpha , criterion: (gini, entropy), depth of the tree, max features and max leaf nodes, penalty, kernel using **grid search CV**

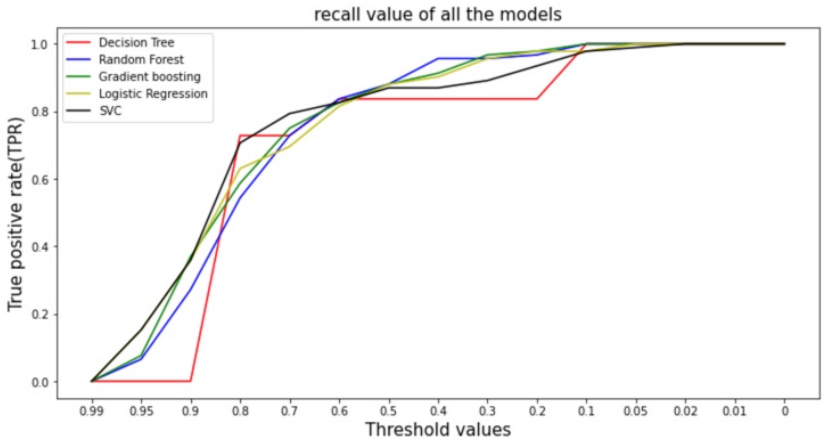
accuracy and comparison of all the models using accuracy_score

Model	Accuracy
Decision_Tree	0.775758
Random_Forest	0.818182
Gradient_Boost	0.836364
Logistic_Regression	0.824242
SVC	0.830303

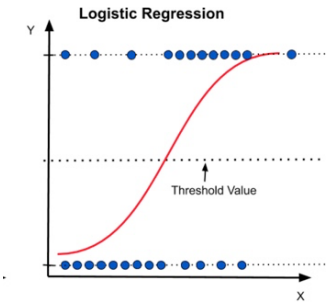
Model	NMSE Train Score	NMSE Test Score	Variance
Decision_Tree	-0.204545	-0.284848	0.080303
Random_Forest	-0.174242	-0.224242	0.050000
Gradient_Boost	-0.162121	-0.236364	0.074242
Logistic_Regression	-0.175758	-0.236364	0.060606
SVC	-0.166667	-0.218182	0.051515

Comparison of all the models on NMSE on train and test score using cross validation

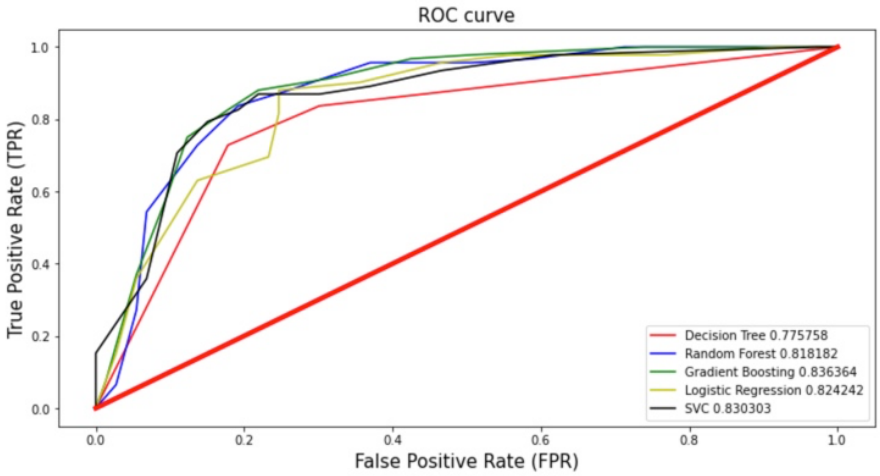
11. For predicting heart disease, we want a model with high recall value i.e. which makes less type 2 errors:



Here In this graph true positive rate i.e TPR is shown corresponding to different threshold values which helps in deciding the true or false value of the target column.
For ex : if we want all the predicted values above and equal to 0.8 are to be considered as true otherwise false . in the below picture threshold value is shown as 0.5, so if target value is above 0.5 , it is considered to be true(1) otherwise false(0)



12. ROC Curve



13. Conclusion :

After comparing accuracy of all the models it turns out that Gradient Boosting and SVC models are performing better than others with almost 83% accuracy.

14. Future Improvements:

1. We can use more complicated model like XGboost for more accurate results.
2. it might possible that different values of categorical columns like ‘trestbps,’cp’,’restecg’, ‘slope’ and ‘thal’ have different impact on possibility of heart disease, so to get better and accurate results these columns can be explored further to check if ordinal encoding can be applied on them.