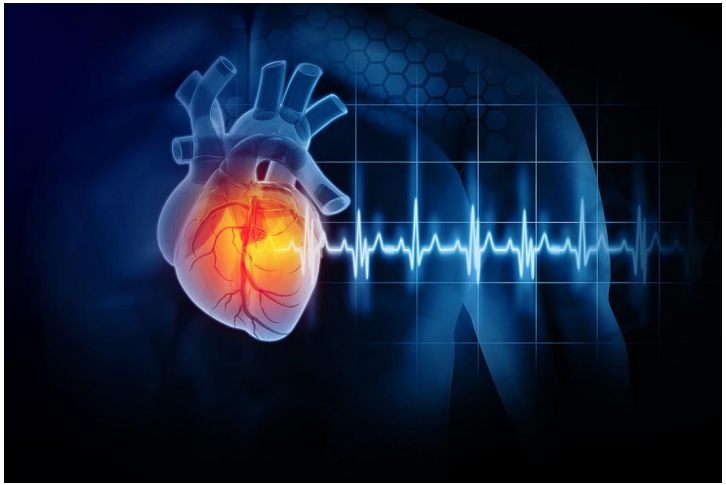
Machine Learning for Heart Disease Prediction



STAT 451 Group Project - Group 5

[Anke Li, Ce Guo, Samuel Merten, Yongseong Kim]

Index

Problem Recognition

Why is it Worth Solving?

Our Goals

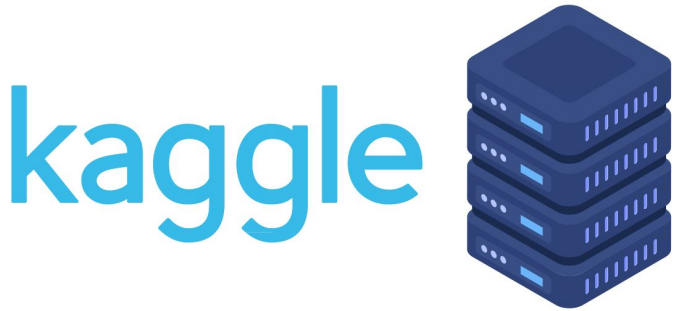
1. To find whether an individual has had an incidence of heart disease.
2. To identify which machine learning algorithm is best for identifying heart disease.
3. To determine which dataset features are the most relevant re: predictive power

Data Description

Data Source: Kaggle

Key Variables: BMI, Age, HighBP, Stroke, GenHlth, etc.

Response variable: HeartDiseaseorAttack(Indicates whether the individual has had a heart disease or heart attack)

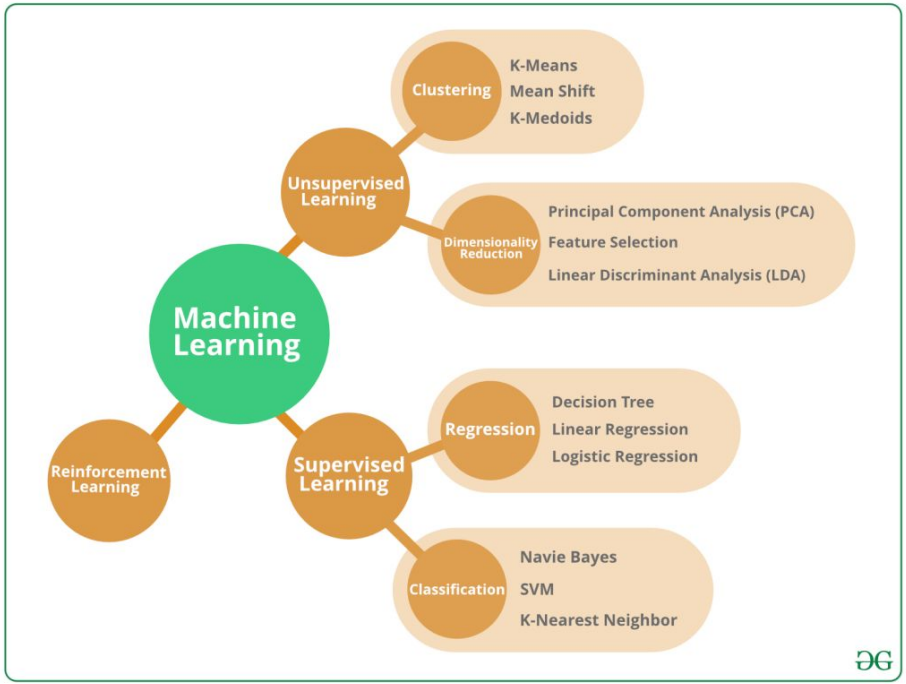


Kaggle Dataset: https://www.kaggle.com/datasets/bhaveshmisra/heart-disease-indicators/data

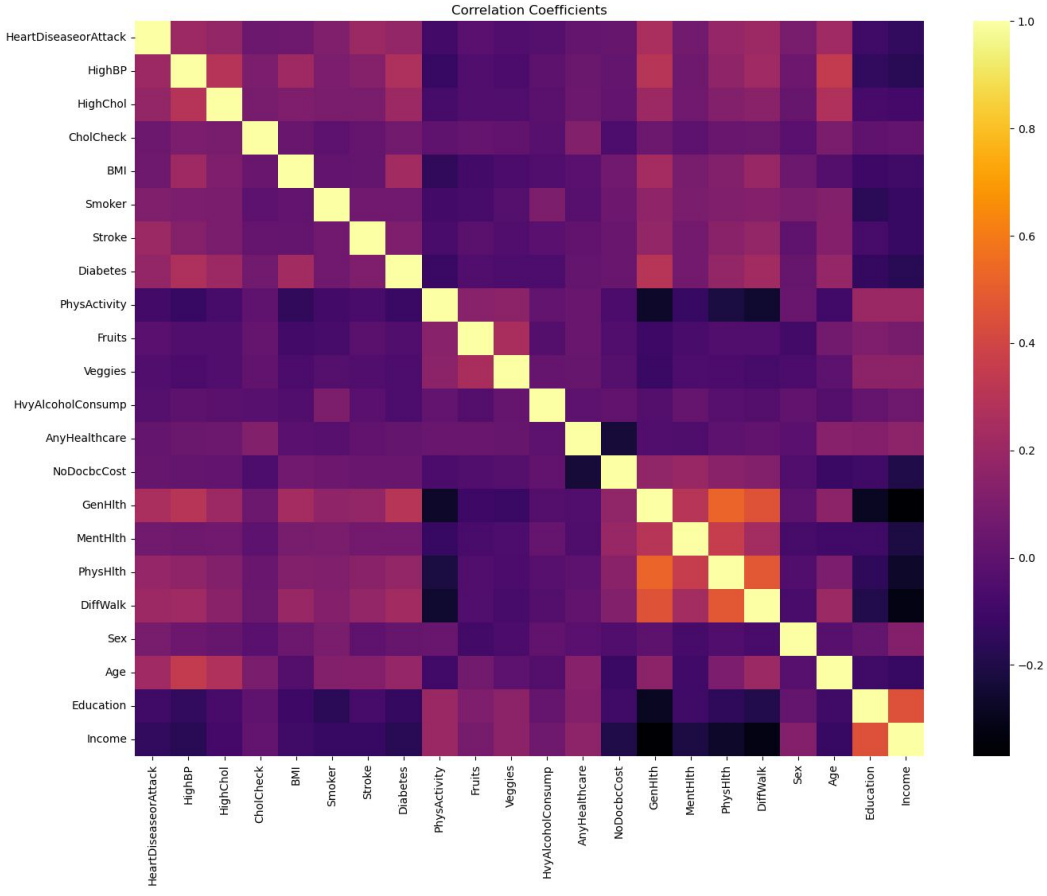
Predictive Modeling

Which Model is best for the Predict?

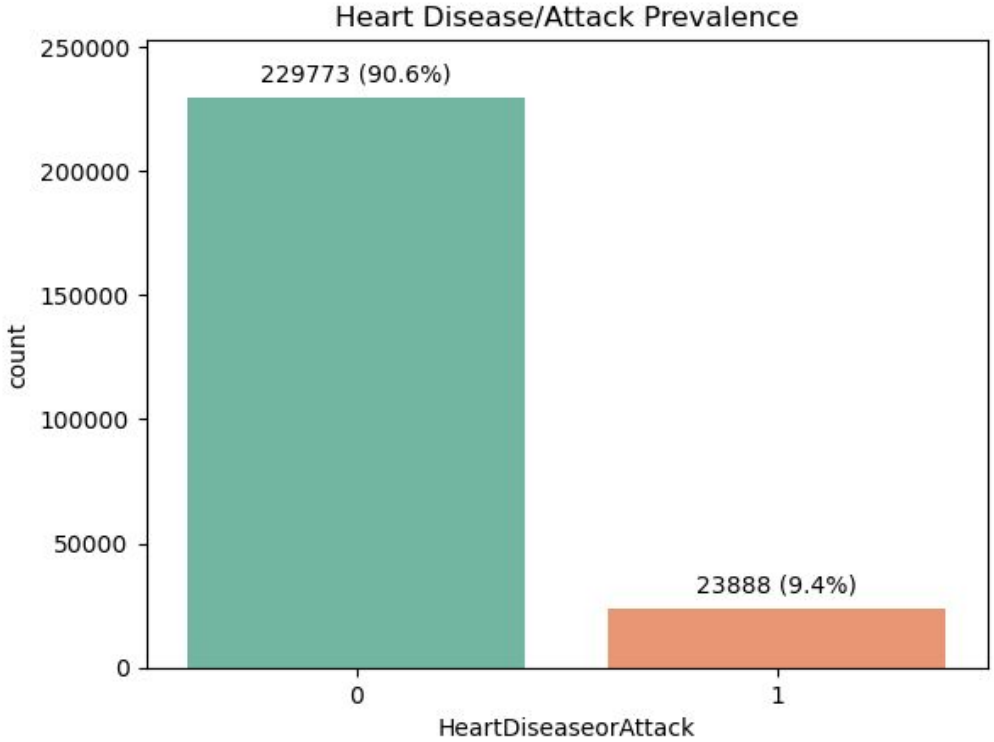
1. K-Nearest Neighbors (KNN)
2. Logistic Regression
3. Support Vector Machine (SVM)
4. Decision Tree
5. Random Forest



Exploratory data analysis



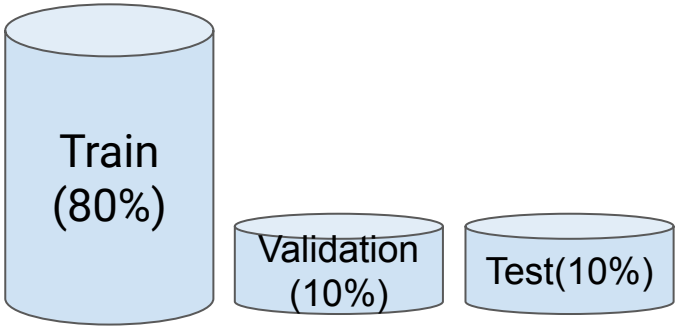
Problem of Data Set



Heavy Class Imbalance!

Solution; Pre-Modeling Data Treatment

Train-Validation-Test Split



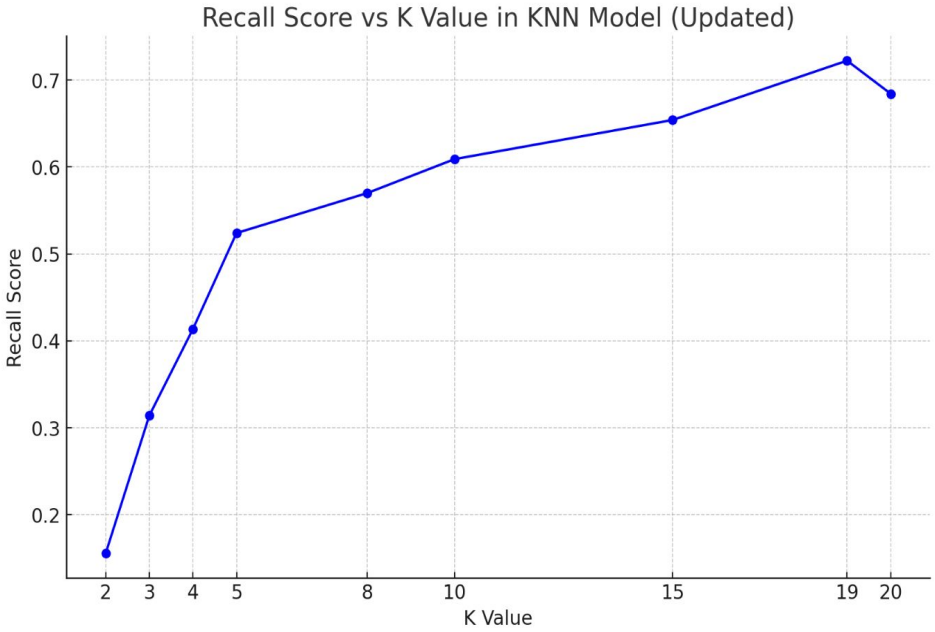
Due to heavy class imbalance (low incidence of heart disease), We had to perform class balancing via oversampling

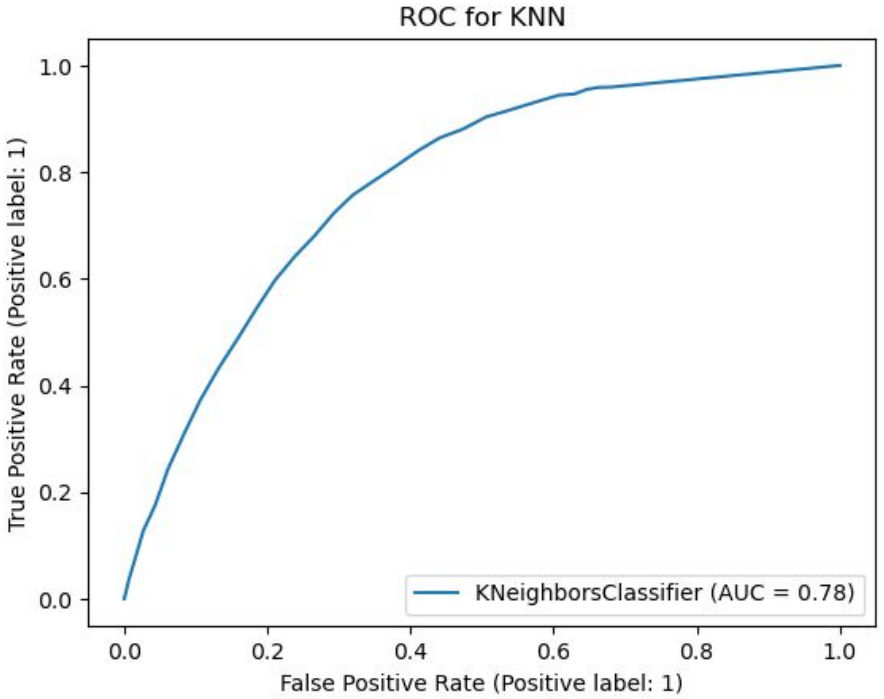
Original ratio of heart disease
[9.4%]

Ratio after oversampling minority
[50%]

K-Nearest Neighbors (KNN)

Hyperparameter: K = 19 (Find the best Recall score)

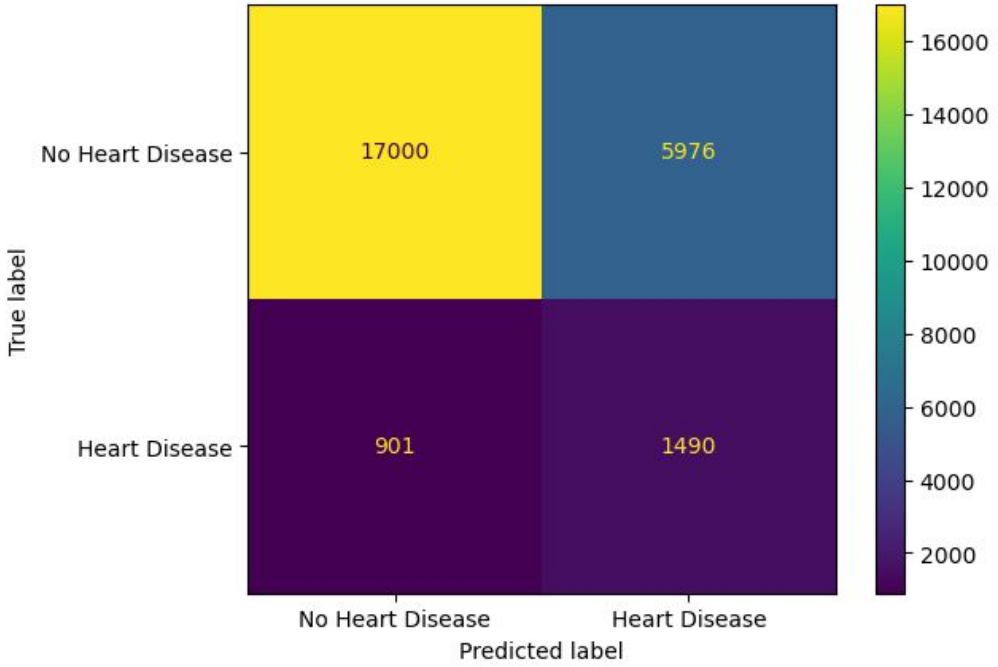




K-Nearest Neighbors (KNN)

- *Hyperparameter Grid Search based on 'recall'*
  - Optimal k = **19**

- ***Results**:*
  - *AUC:* 0.77
  - *Accuracy:* 0.70
  - *Precision:* 0.20
  - *Recall:* 0.72

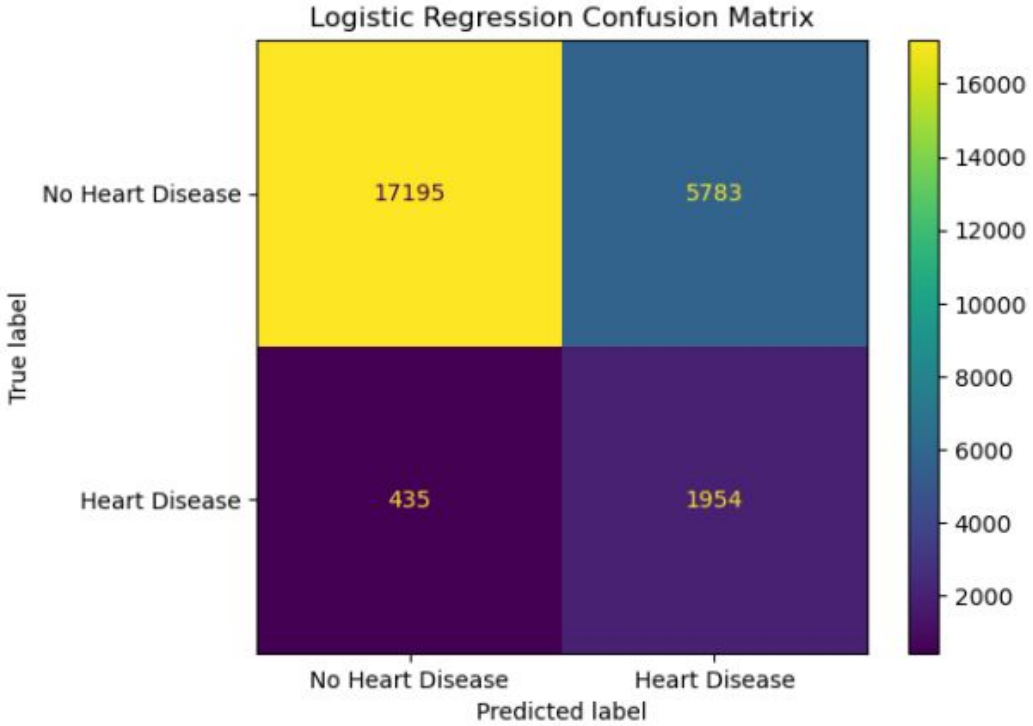


Logistic Regression

- *Hyperparameter Grid Search based on 'recall'*
  - Optimal C = **280**

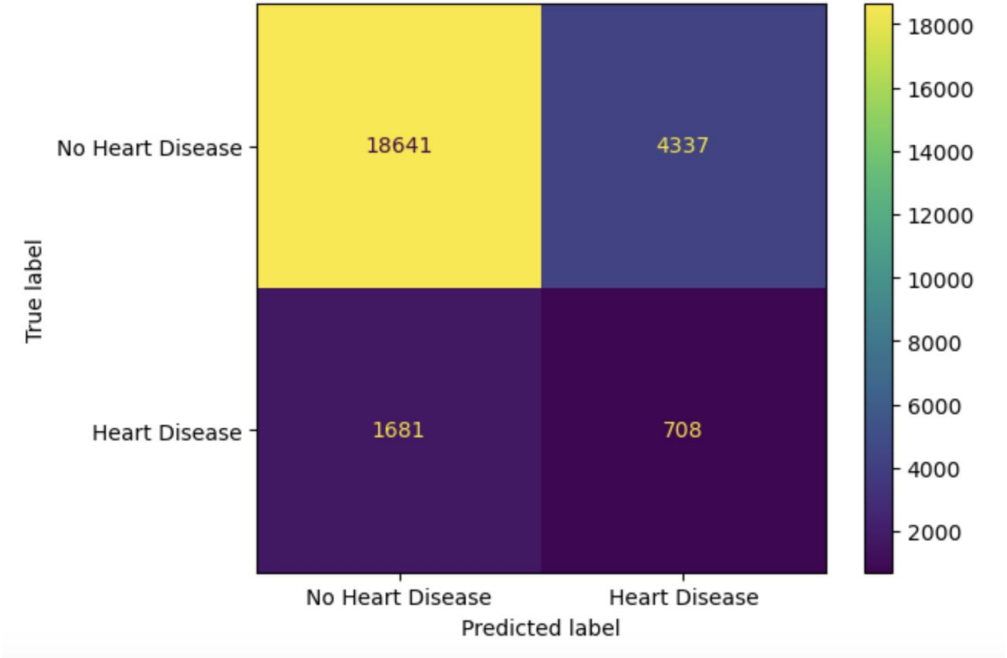
- ***Results**:*
  - *AUC:* 0.86
  - *Accuracy:* 0.76
  - *Precision:* 0.25
  - *Recall:* 0.82

	precision	recall
0	0.98	0.75
1	0.25	0.82



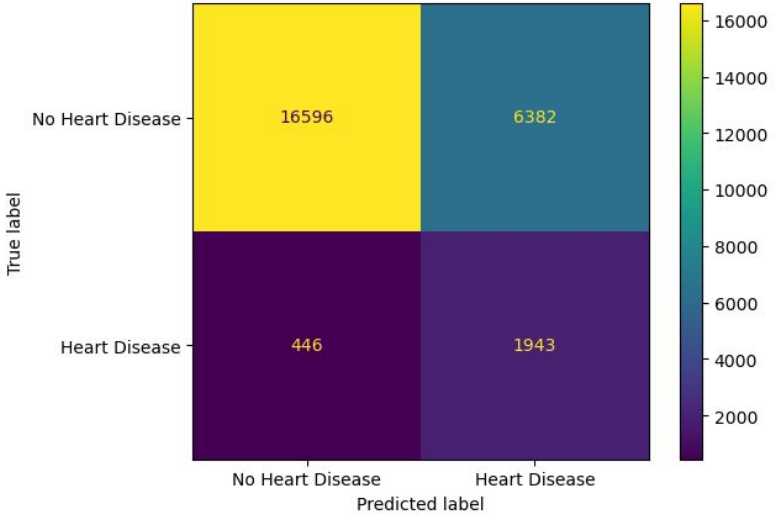
Support Vector Machine (SVM)

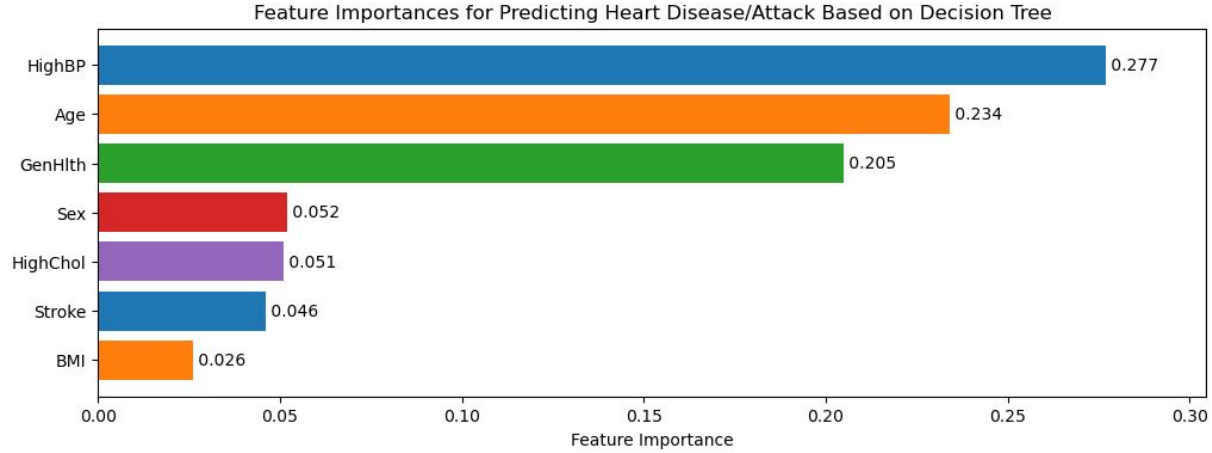
- *Hyperparameter Grid Search based on 'linear kernel'*
  - Optimal C = **20**
- ***Results**:*
  - *AUC:* 0.27
  - *Accuracy:* 0.76
  - *Precision:* 0.14
  - *Recall:* 0.29

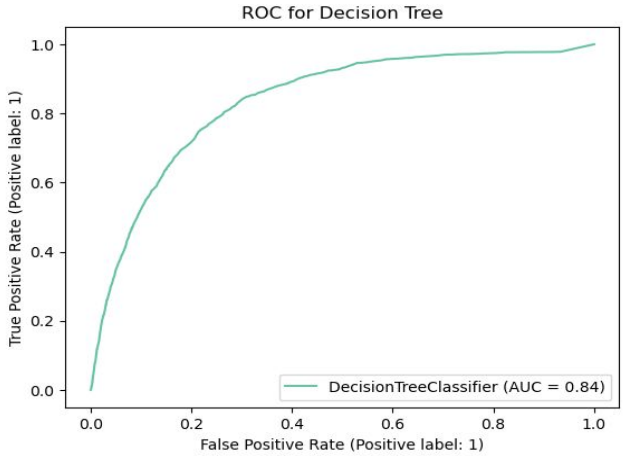


Decision Tree

- *Hyperparameter Grid Search based on “recall ”*
  - *Criterion = “entropy”*
  - *Max_depth = 10*
- *Result*
  - *AUC: 0.84*
  - *Accuracy: 0.74*
  - *Precision: 0.24*
  - *Recall: 0.80*

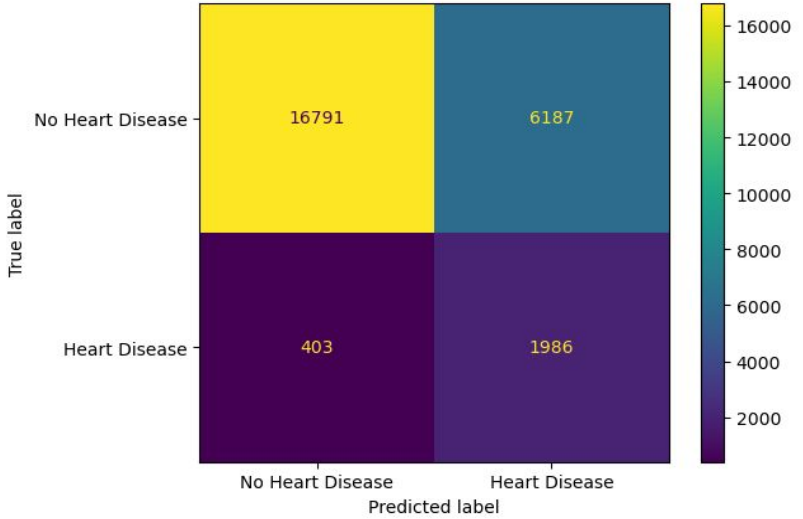


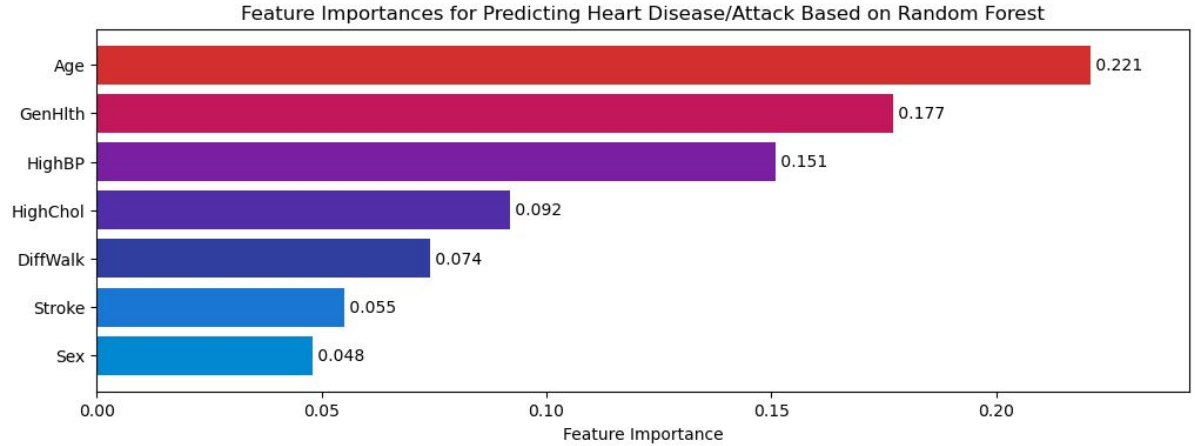


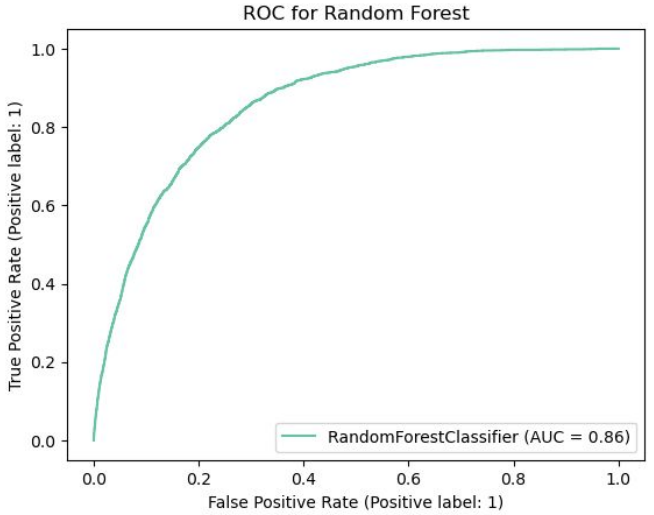


Random Forest

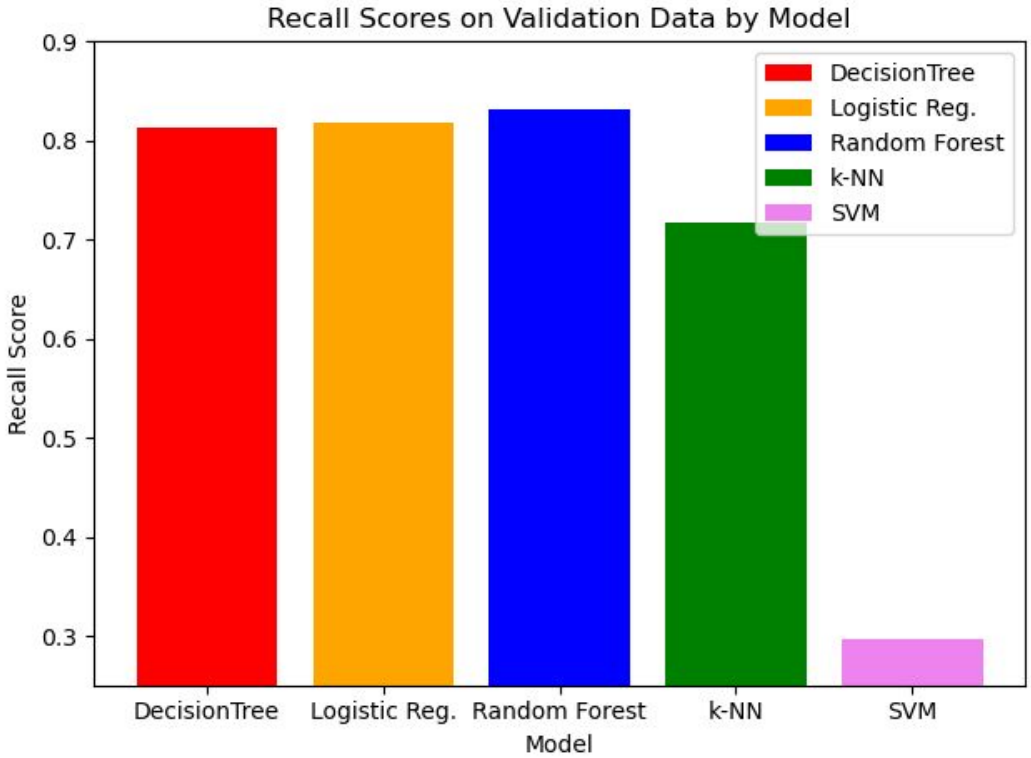
- *Hyperparameter Grid Search based on “recall ”*
  - *criterion = “entropy”*
  - *n_estimators=200*
  - *max_depth = 10*
- *Result*
  - *AUC: 0.86*
  - *Accuracy: 0.74*
  - *Precision: 0.24*
  - *Recall: 0.83*





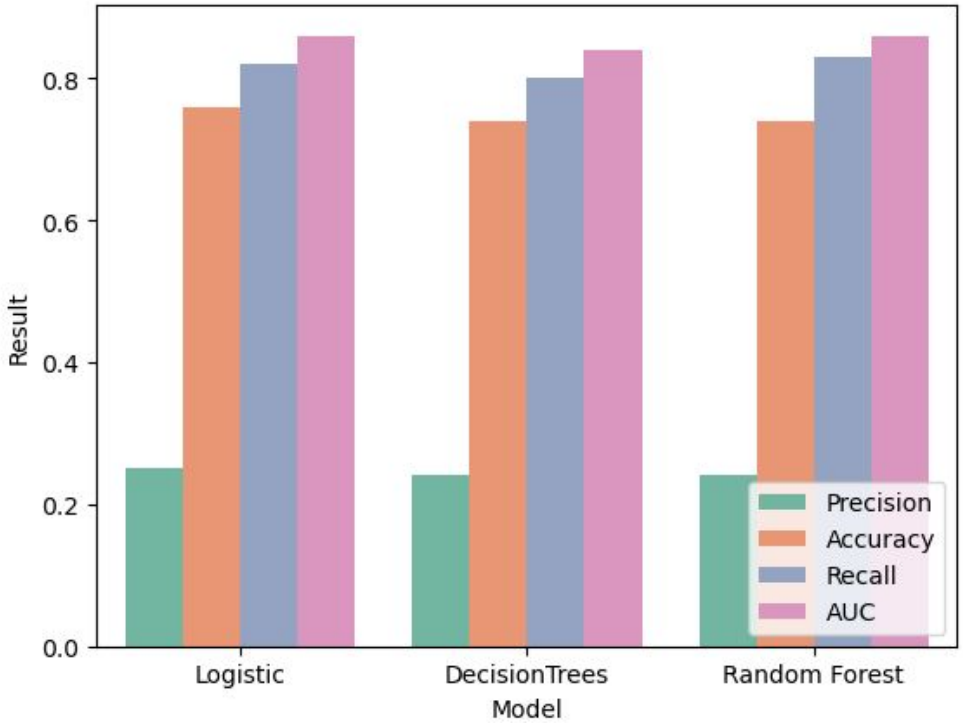


Model Results



Which is the best choice among the top 3 recall scores?

Conclusion



Our analysis shows that logistic regression, decision tree, and random forest models demonstrate comparable performance in our dataset.

Discussion

- Seeking the best Recall Score often leads to more false positives. Further research is needed to balance accuracy, recall, and precision, considering doctors' needs.
- In the further research, implementing feature selection could be beneficial. With numerous variables, feature selection can help reduce code running time.