Heart Disease Prevention

By Alejandro Harrison

Business Overview

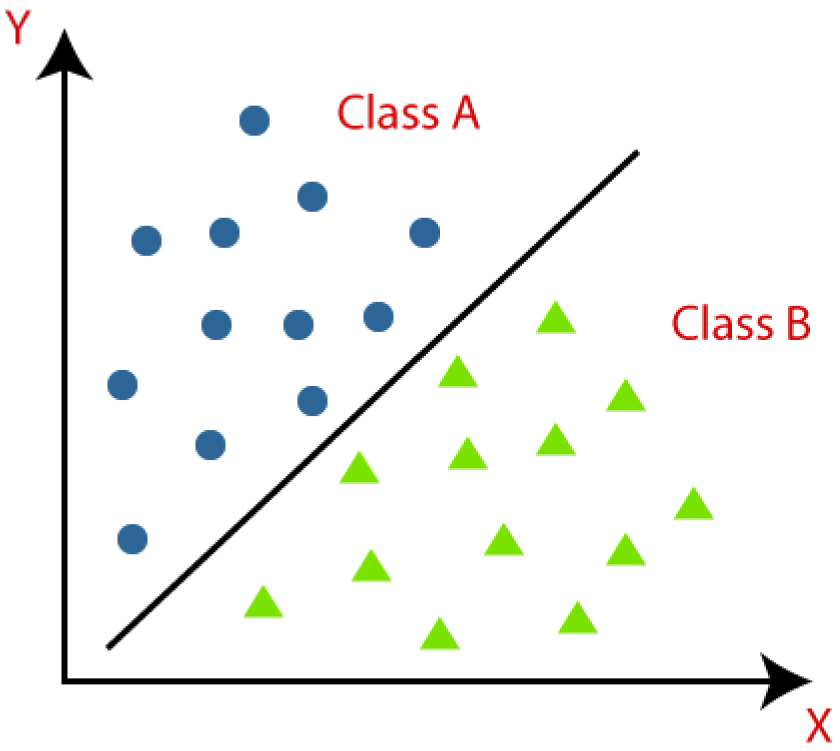
- Heart disease is leading cause of death in US (CDC)
- Coronary heart disease (CHD) most common type
- Hospital looks at variety of factors
- Identify risk of CHD

Method

- Use of classification models
- Pick best one
- Able to identify subjects at risk for CHD
- What is classification?

Classification

- Process of dividing data into classes or categories
- Can be used to predict future values
- Yes/no for risk of CHD



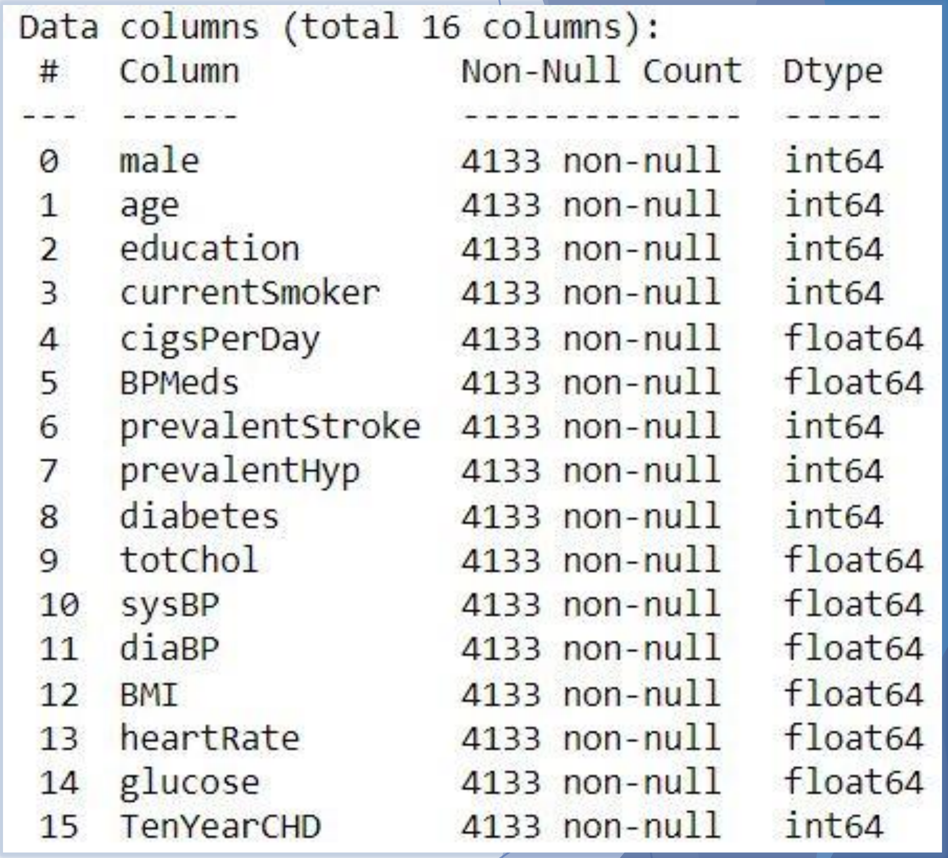
The Data

- Framingham Heart Study
- Started in 1948
- Long term, ongoing study
- Used by more than 1200 scientific journals
- Identifies factors that lead to Cardiovascular disease

Buffalo
Framingham, MA
Boston
Cleveland
New York
Pittsburgh
Philadelphia
Baltimore
Virginia Beach
45°N
40°N
80°W
75°W
70°W
65°W

Data Continued

- Dataset had 4133 entries
- Demographic, behavioral and medical
- Age, sex, cholesterol, etc
- 1 yes/no factor for CHD risk



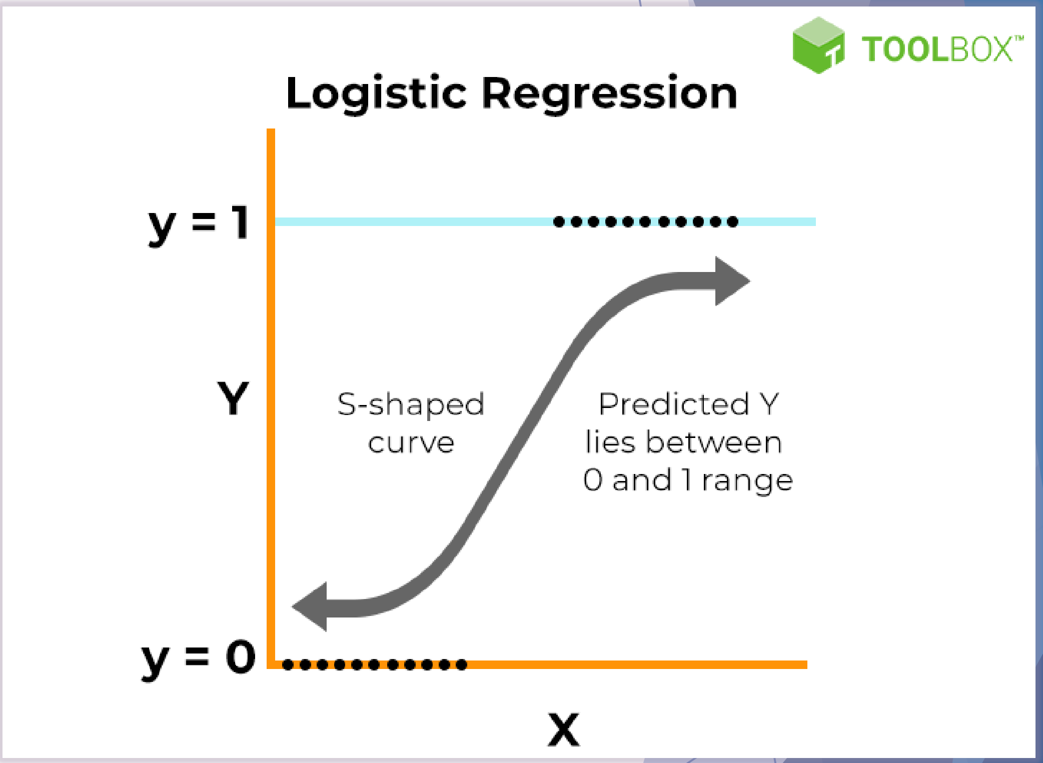
```
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   male             4133 non-null   int64  
 1   age              4133 non-null   int64  
 2   education        4133 non-null   int64  
 3   currentSmoker    4133 non-null   int64  
 4   cigsPerDay       4133 non-null   float64
 5   BPMeds           4133 non-null   float64
 6   prevalentStroke  4133 non-null   int64  
 7   prevalentHyp     4133 non-null   int64  
 8   diabetes         4133 non-null   int64  
 9   totChol          4133 non-null   float64
 10  sysBP            4133 non-null   float64
 11  diaBP            4133 non-null   float64
 12  BMI              4133 non-null   float64
 13  heartRate        4133 non-null   float64
 14  glucose          4133 non-null   float64
 15  TenYearCHD       4133 non-null   int64  
```

Models

- Build initial models then improve
- Logistic Regression
- Decision Tree
- Random Forest

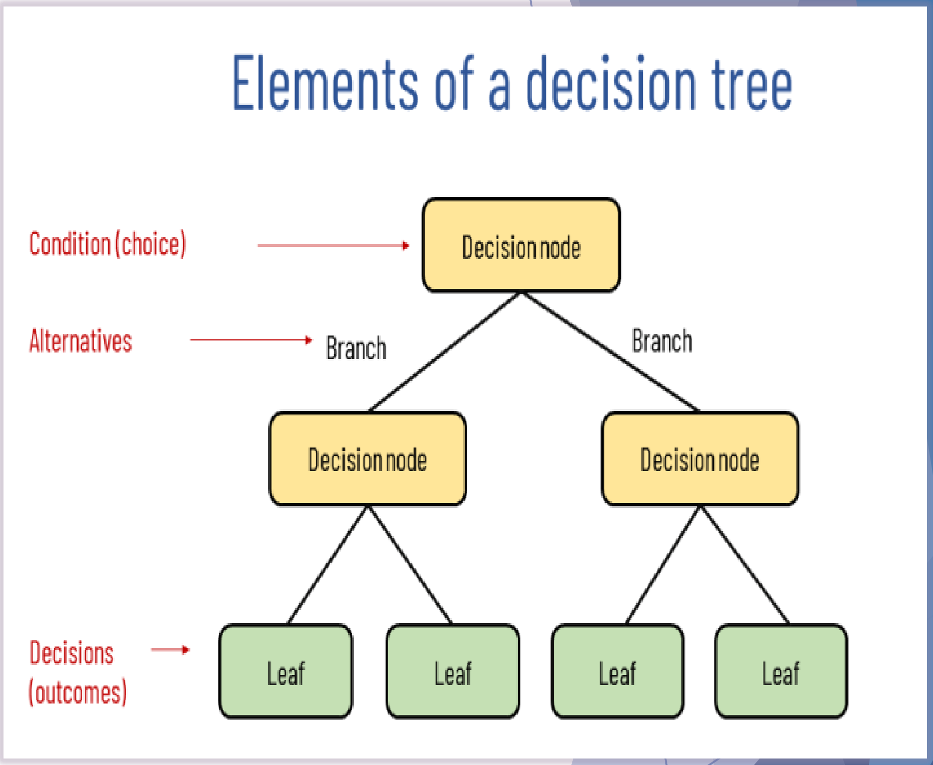
Logistic Regression

- Want to predict a binary outcome
- Involves fitting data to an S shaped curve
- Values fall between 0-1 range
- 0 and 1 represent binary choices/categories



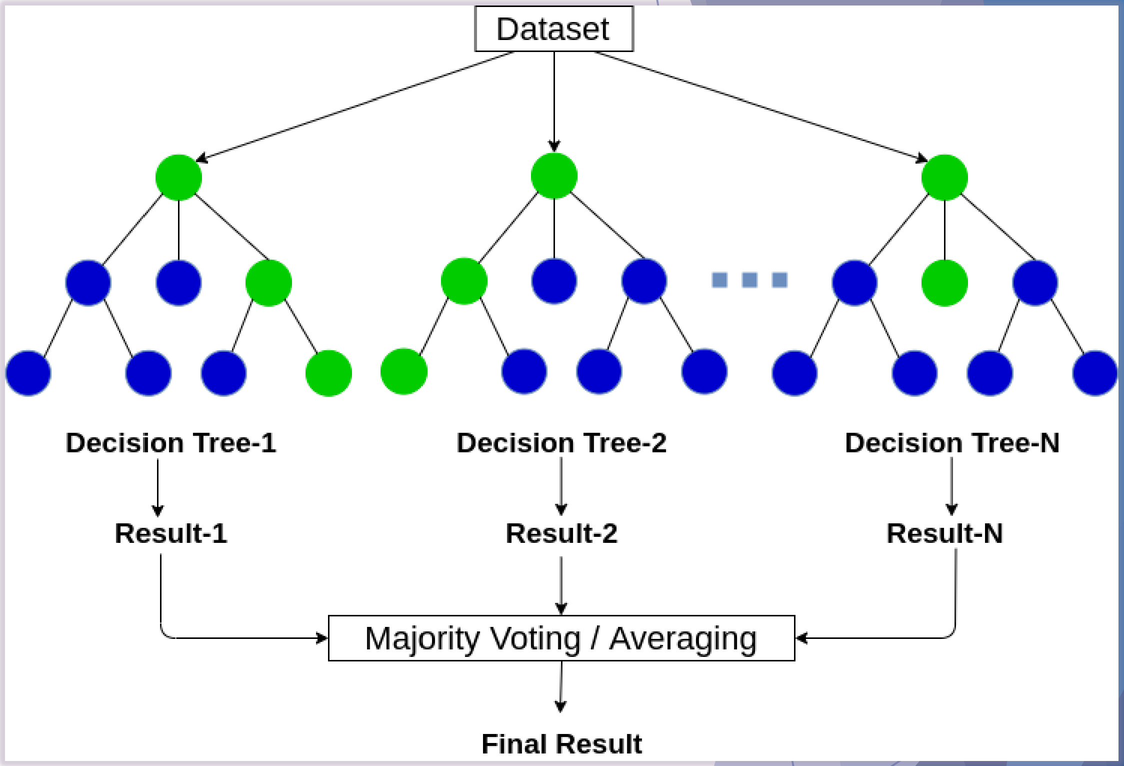
Decision Tree

- All possible outcomes and decisions
- Multiple branches and multiple outcomes
- E.g. smoking
- Decision reached and data categorized



Random Forest

- Creates random samples from data
- Individual decision trees made for each sample
- Compares all tree predictions together
- Majority vote for best prediction

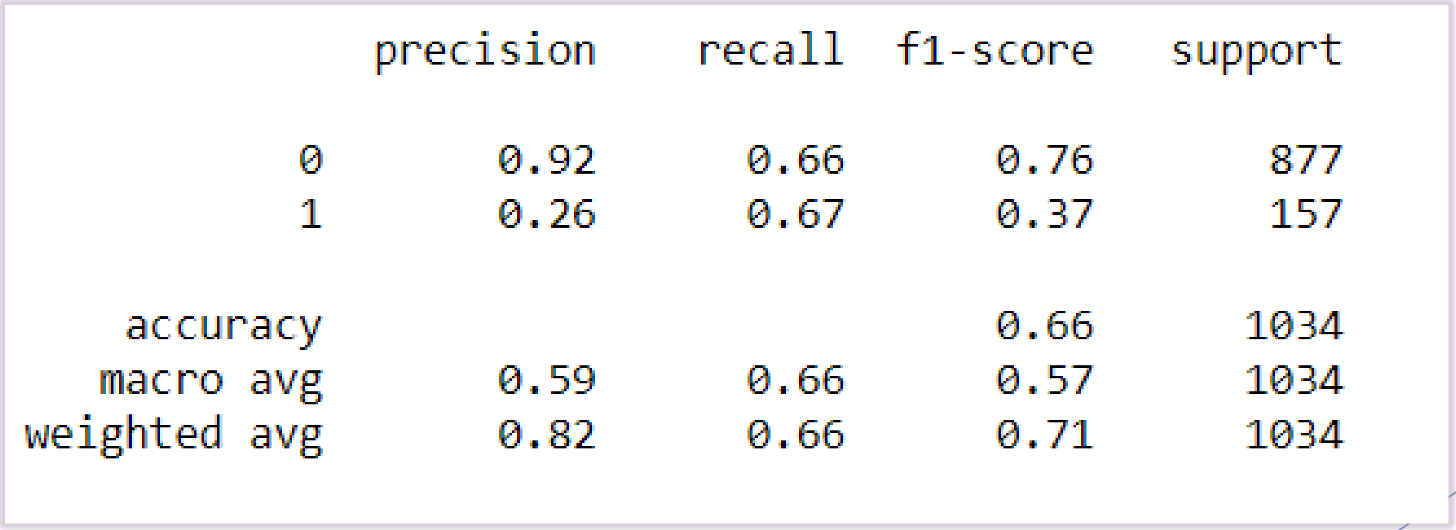


Best Model

- Found that our best algorithm was Random Forest
- Addresses lack of randomness in decision trees
- Room for interpretability
- Results split into classification report, confusion matrix and feature importance sections

Classification Report

- Measures quality of predictions
- Precision, recall, f1 score
- Out of total predicted at risk CHD, how many were actually at risk(precision)
- Out of actual total at risk CHD, how many at risk CHD predictions were correct (recall)
- How accurate is our model (F1 score)



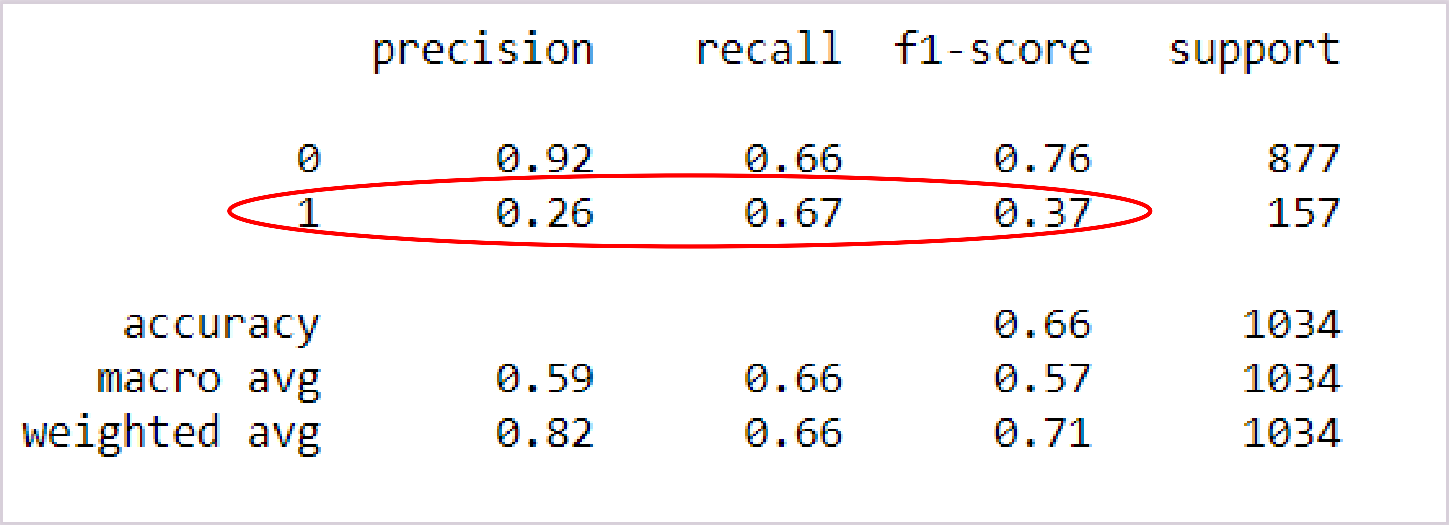
```
              precision    recall  f1-score   support

           0       0.92      0.66      0.76       877
           1       0.26      0.67      0.37       157

    accuracy                           0.66      1034
   macro avg       0.59      0.66      0.57      1034
weighted avg       0.82      0.66      0.71      1034
```

Classification Report Continued

- Precision was at 26%, which is similar to our initial model.
- Recall was at 67%, which is a 40% improvement from the initial model
- Our model had an f1 score of 37%, which is a 7% improvement from the initial model.



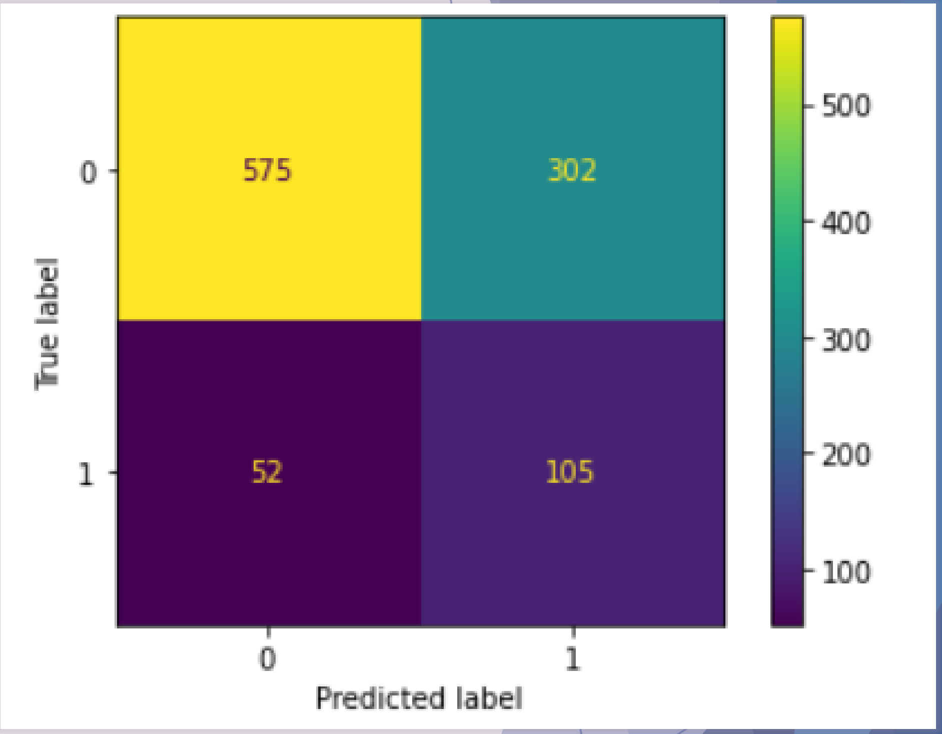
```
              precision    recall  f1-score   support

           0       0.92      0.66      0.76       877
           1       0.26      0.67      0.37       157

    accuracy                           0.66      1034
   macro avg       0.59      0.66      0.57      1034
weighted avg       0.82      0.66      0.71      1034
```

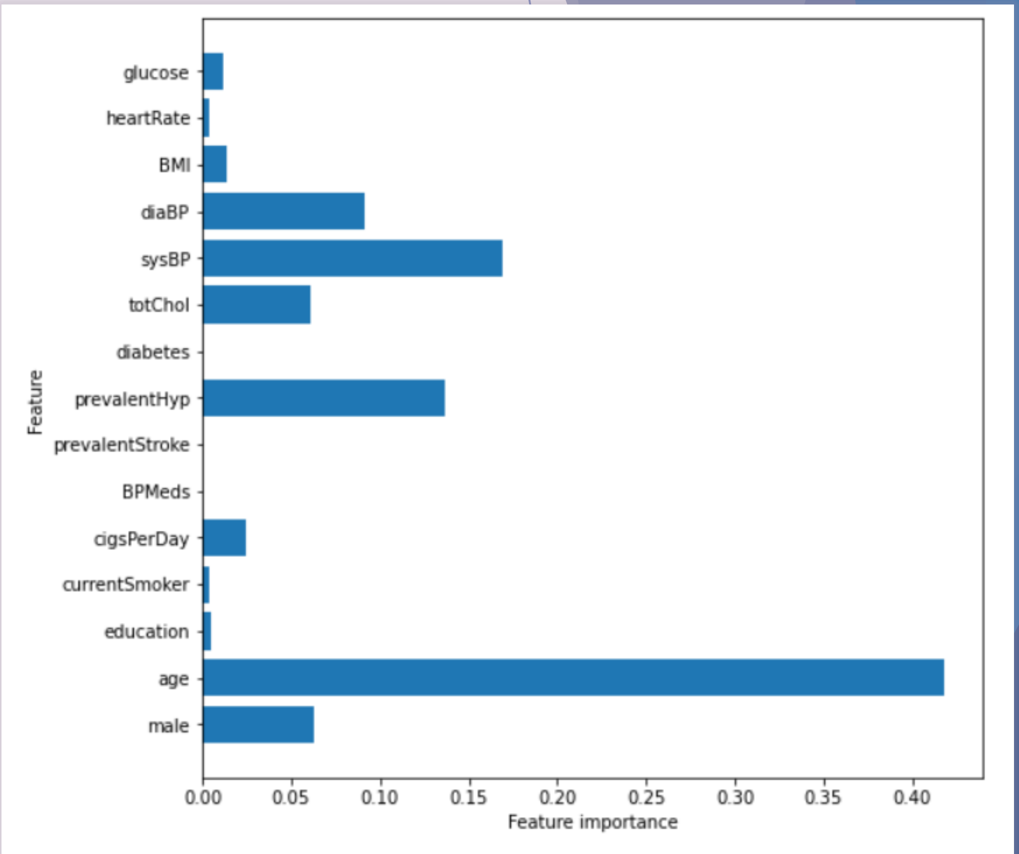
Confusion Matrix

- 575 patients were correctly predicted as not being at risk for CHD (TN)
- 52 patients were wrongly predicted as not being at risk for CHD (FN)
- 302 patients were wrongly predicted as being at risk for CHD (FP)
- 105 patients were correctly predicted as being at risk for CHD (TP)
- 62 more instances of true positives
- 62 less instances of false negatives



Feature Importance

- How useful is each variable in predicting target variable
- Most indicative feature of whether someone is at risk for CHD was age
- Other important features included :
  - Systolic and diastolic blood pressure (sysBP and diaBP)
  - History of high blood pressure (prevalentHyp)
  - Total cholesterol level (totChol)



Results

- Best model was Random Forest
- Showed Increases in true positives
- Showed decreases in false negatives
- Age was most indicative factor of CHD risk

Recommendations

- Run entirely new model with only the top important features
- Screen earlier in adulthood for high blood pressure and high cholesterol
- Screening earlier allows for treatment of risk factors preventing heart disease