Framingham Heart Study

Anna Puk, Jaya Deonandan | Prof. MacDonald | SCS 3253 – Machine Learning | University of Toronto

Predicting a Ten-Year Risk of Developing Coronary Heart Disease:

Problem: What risks and causes contribute to developing coronary heart disease, and are they good predictors in determining whether someone will develop this health condition?

Project Overview

- The goal of this project is to examine epidemiological data, specifically potential causes and risk factors that contribute to developing coronary heart disease throughout a ten-year span
- Within this project, the following factors are examined:
  - Sex
  - Age
  - Education
  - Whether someone currently smokes
  - Number of cigarettes smoked per day
  - Whether or not they are taking blood pressure medication
  - Presence of a stroke
  - Presence of diabetes
  - Glucose levels
  - Total Cholesterol levels (mg/dL)
  - Systolic Blood Pressure (mmHg)
  - Diastolic Blood Pressure (mmHg)
  - BMI (Body Mass Index)
  - Heart Rate (beats/min)

Steps of Analysis

Data Exploration

- Data is explored to examine patterns, trends, and points of interest between categorical and numerical variables

Data Preparation

- Data is standardized and encoded as preparation for models

Model Selection and Tuning

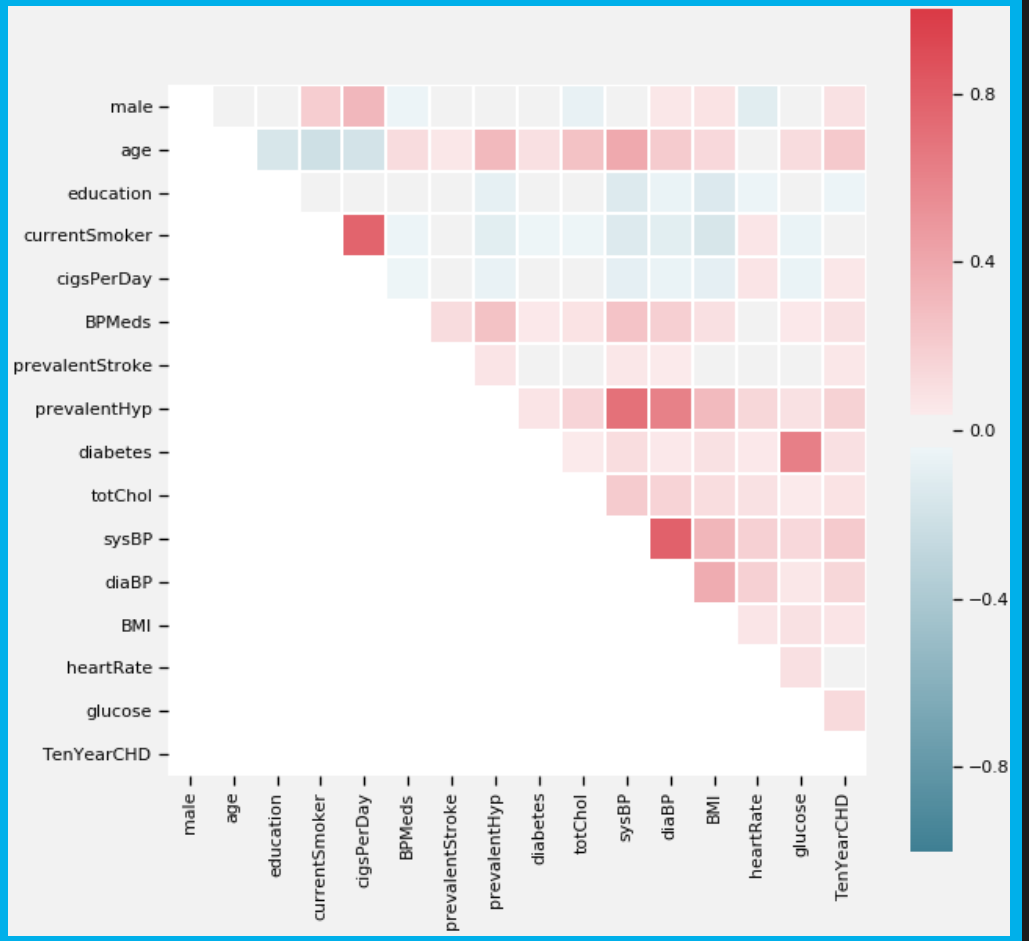
- The following models are examined:
  - K-Nearest Neighbours
  - SVM Classification
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Ensemble Methods

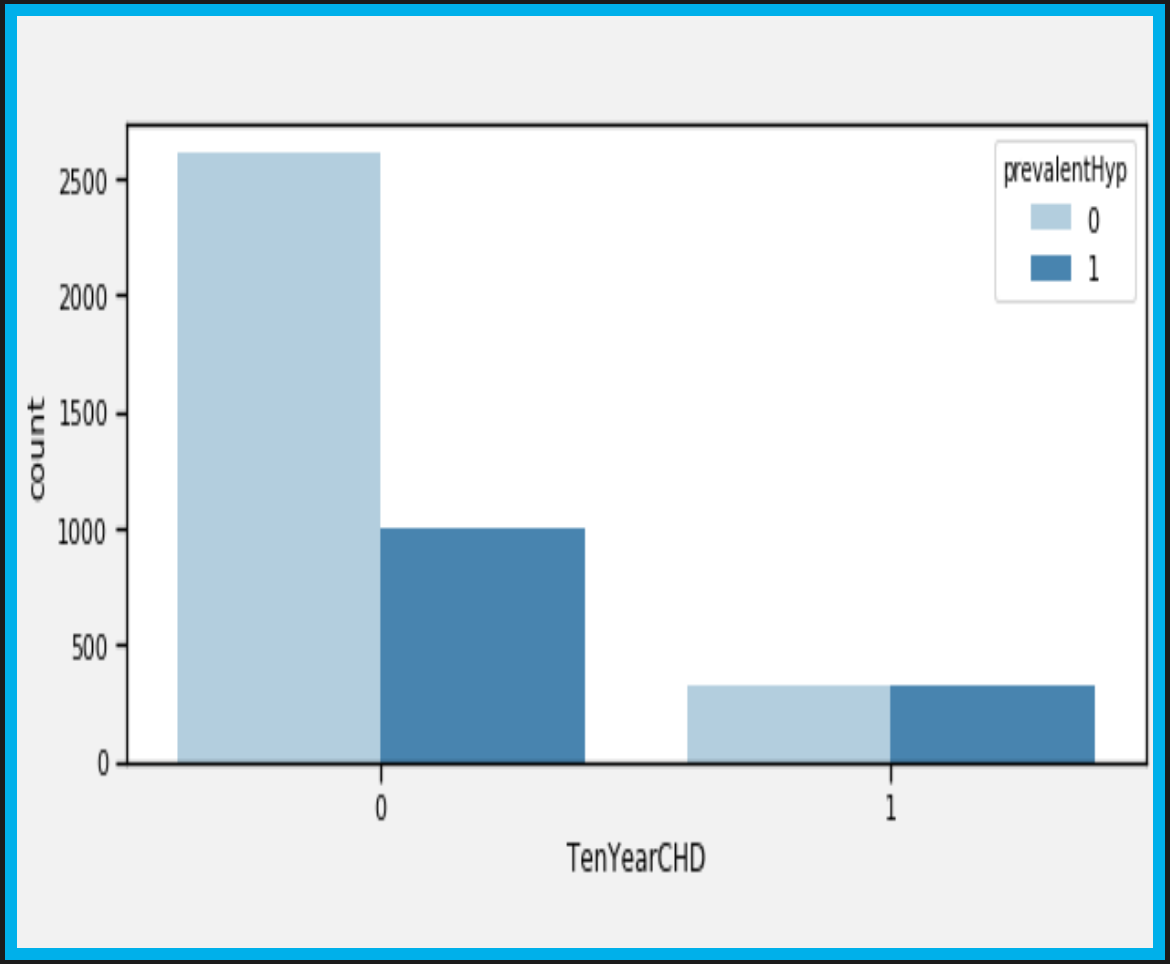
Comparison of Models and Results

- Performance of models is evaluated and compared

Data Exploration

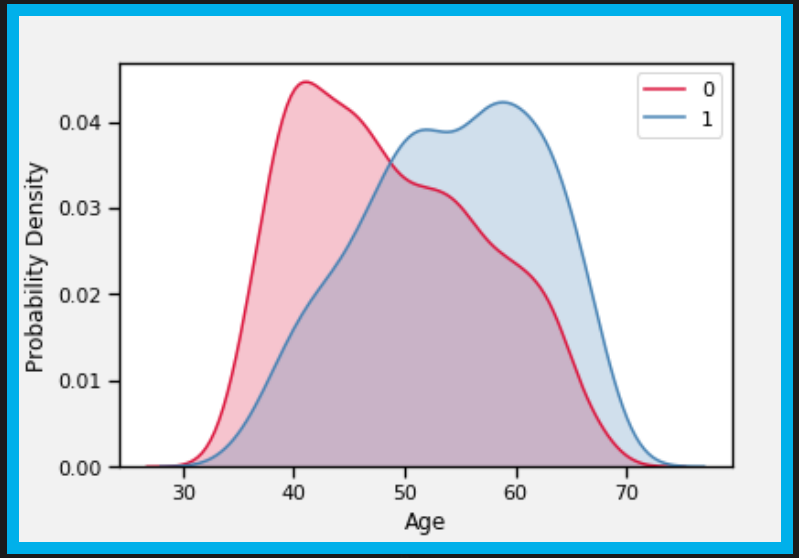
Correlations and Categorical Data Trends:

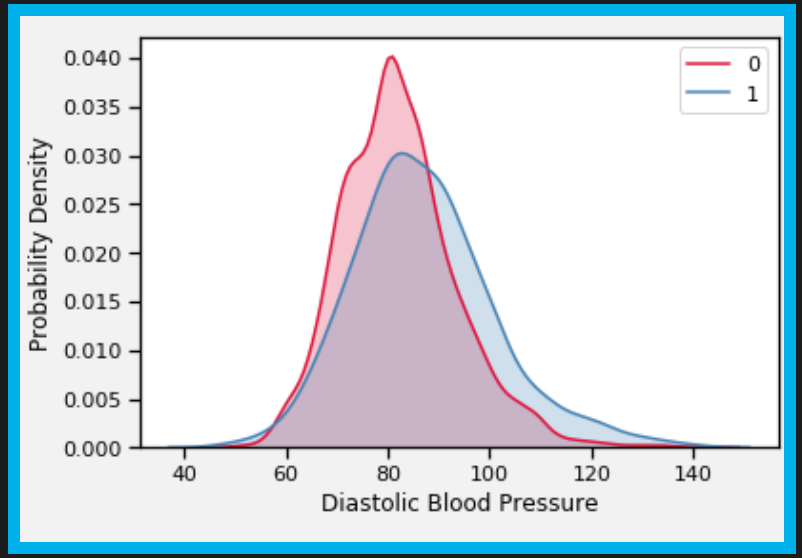


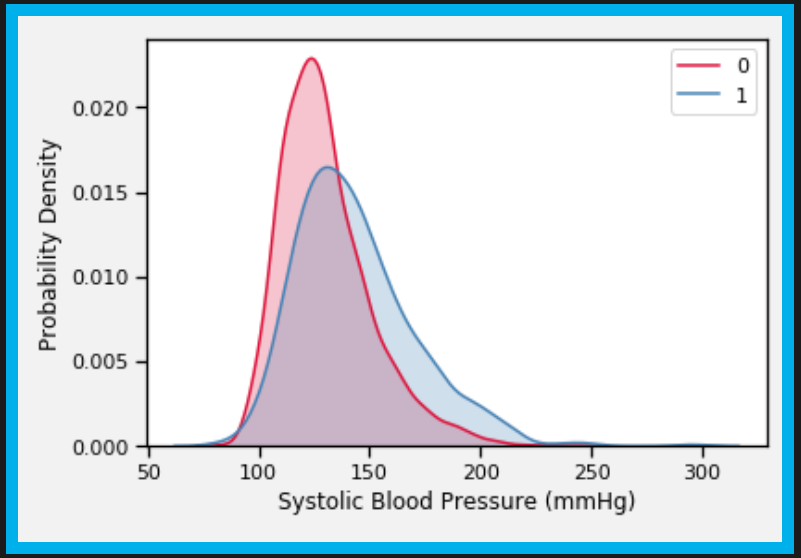


Data Exploration

Numerical Data Trends:







Data Preparation

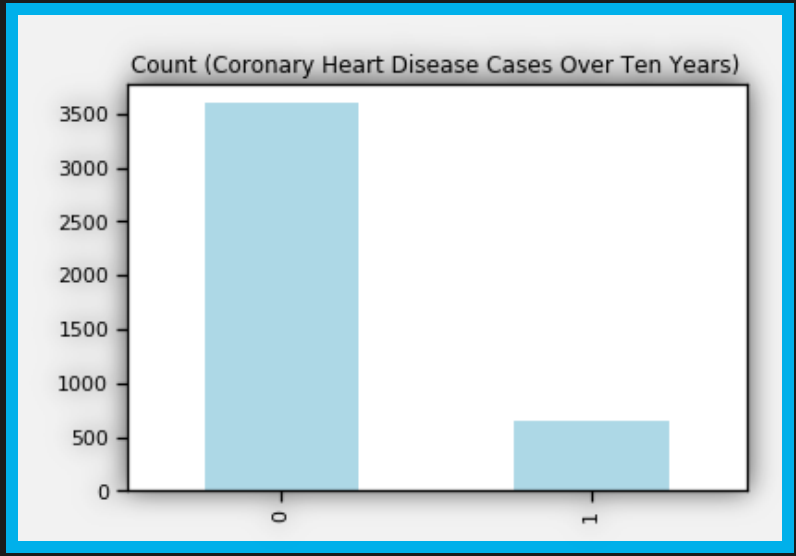
The following steps were taken to prepare our data for our models:

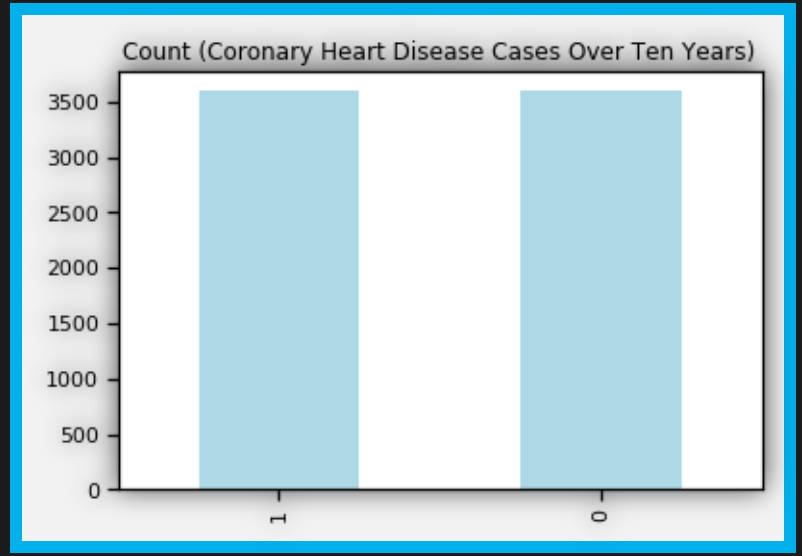
1. For numerical data, replace missing values with the median, and standardize the data.

2. For categorical data, replace missing values with the most frequent value and use one-hot encoding to encode the data.

3. Created a second data set to balance the underrepresented class (i.e. having CHD)

4. Split the data into training and testing data for both a balanced and unbalanced version of our dataset.



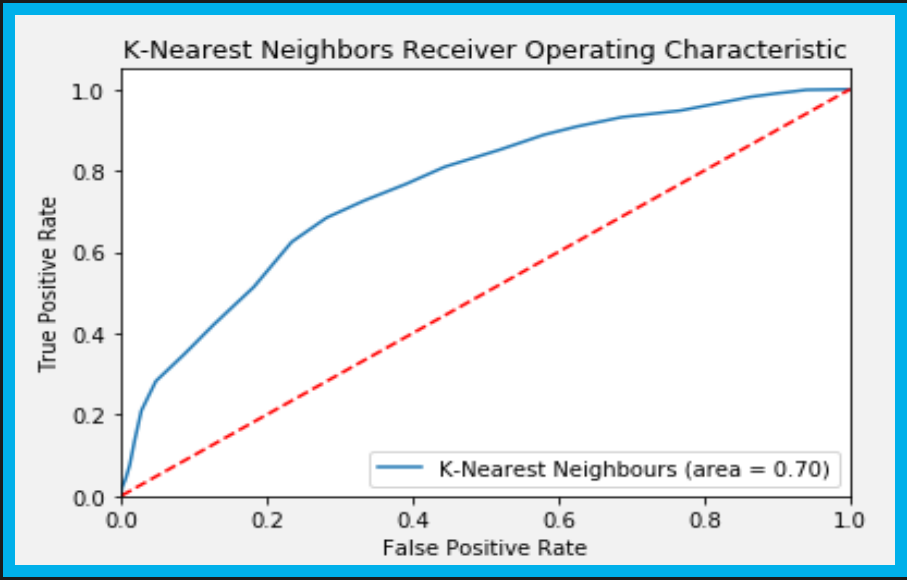


Model Selection

The following models were explored and tested for accuracy in predicting CHD:

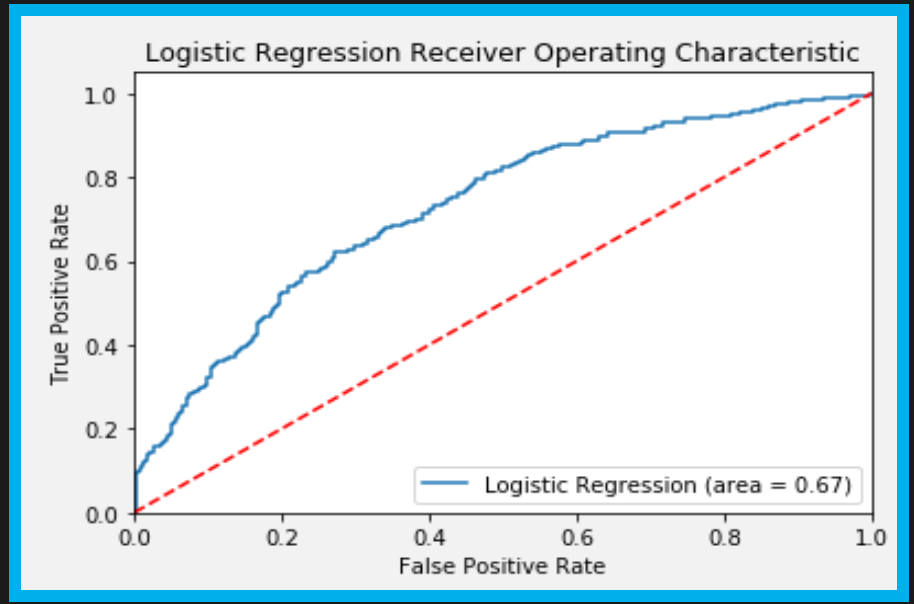
1. K-Nearest Neighbours

	Precision	Recall	F1-score
0	0.73	0.67	0.70
1	0.67	0.73	0.69



2. Logistic Regression

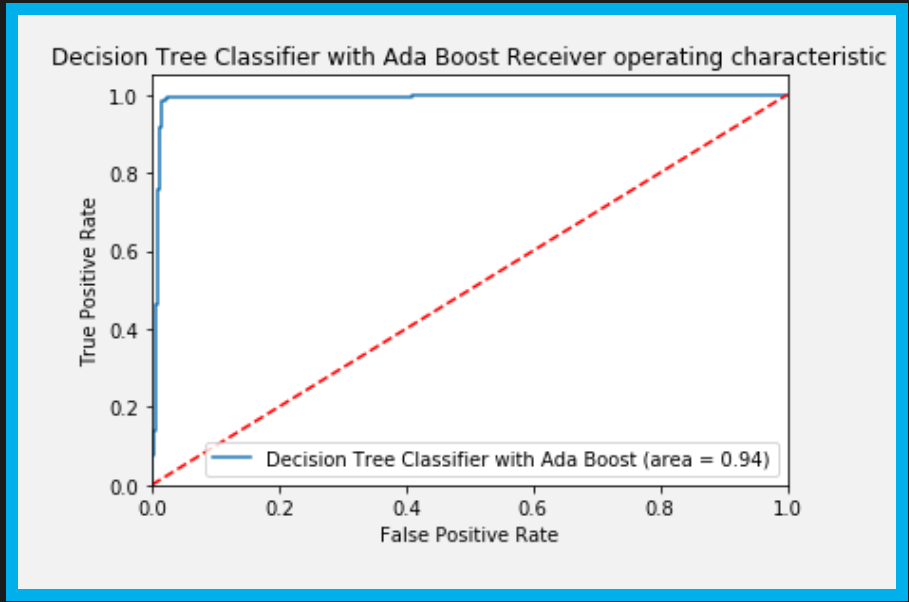
	Precision	Recall	F1-score
0	0.69	0.66	0.68
1	0.65	0.67	0.66



Model Selection

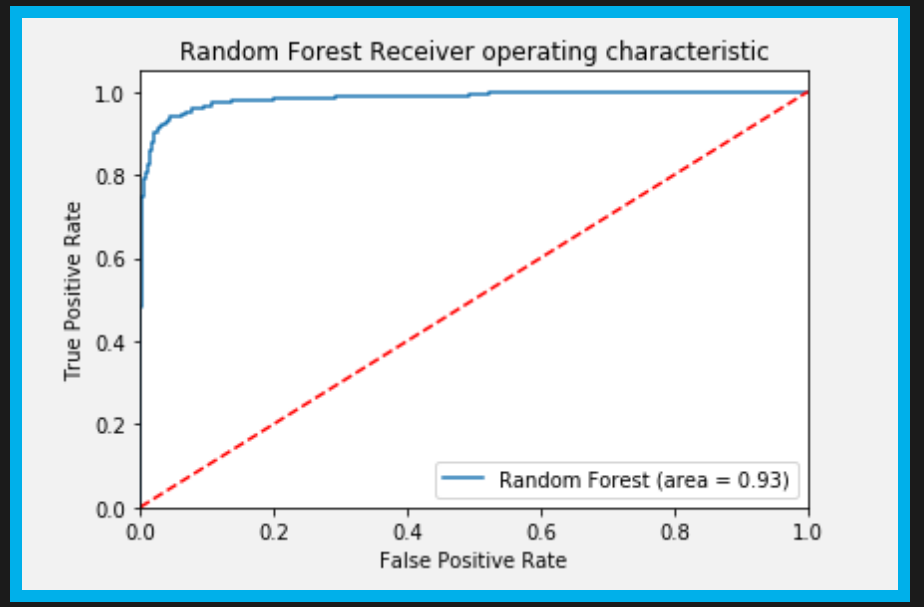
3. Decision Tree Classification using AdaBoost

	Precision	Recall	F1-score
0	0.99	0.88	0.93
1	0.88	0.99	0.93



4. Random Forest

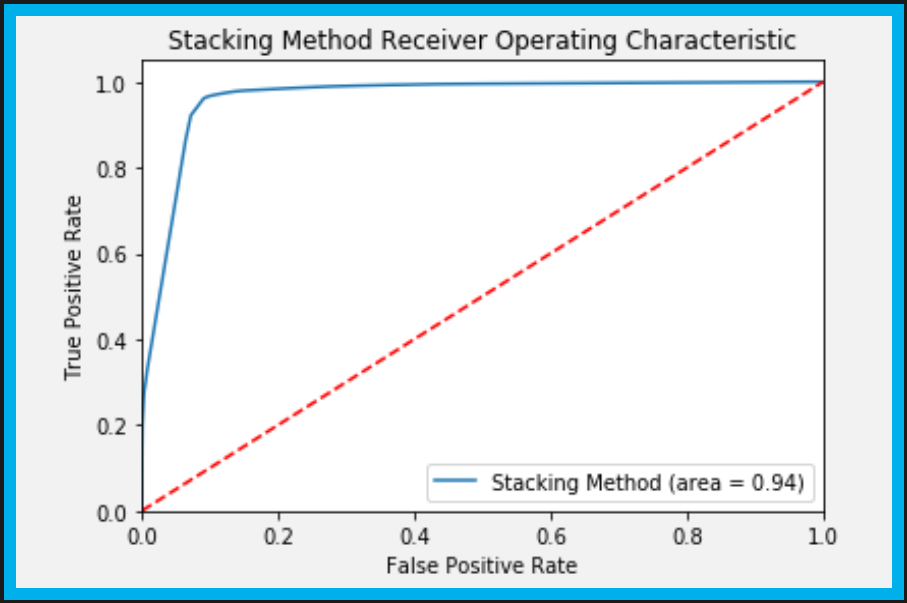
	Precision	Recall	F1-score
0	0.97	0.90	0.93
1	0.90	0.97	0.93



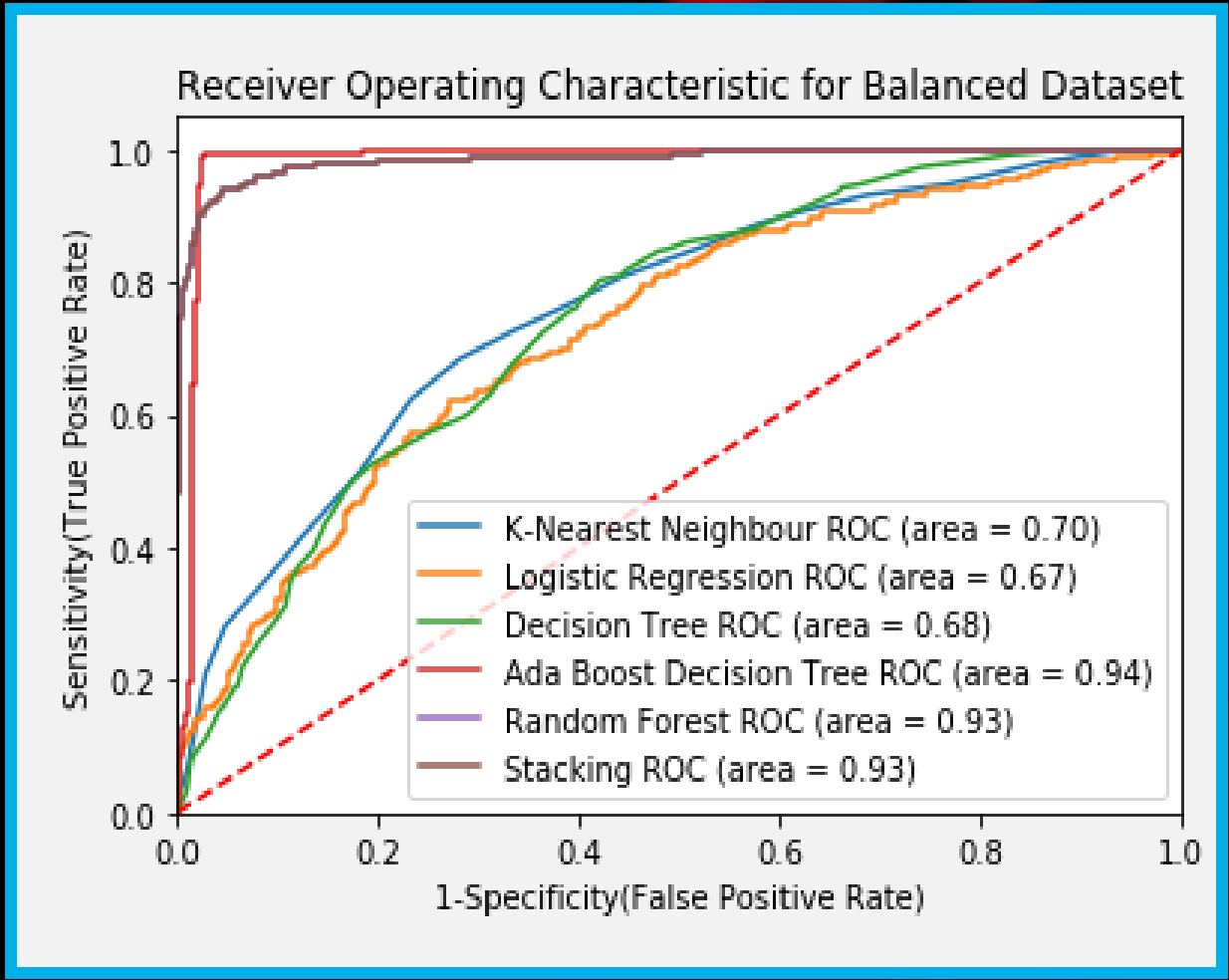
Model Selection

6. Ensembles (Hard Voting and Stacking)

	Precision	Recall	F1-score
0	0.98	0.95	0.97
1	0.95	0.98	0.96



Comparison of Models



The top performing models with the highest ROC scores are:

- **Ada Boosted Decision Tree (area=0.94)**
- **Random Forest (area= 0.93)**
- **Stacking (area= 0.93)**

Comparison of Models

		F1-score	Accuracy
K-Nearest Neighbour	**No CHD**	0.70	69%
	CHD	0.69	
Logistic Regression	**No CHD**	0.68	67%
	CHD	0.66	
Decision Tree Classification using AdaBoost	**No CHD**	**0.93**	**93%**
	CHD	**0.93**	
Random Forest	**No CHD**	**0.93**	**93%**
	CHD	**0.93**	
Stacking	**No CHD**	**0.97**	**97%**
	CHD	**0.96**	

Conclusion

- The following models produced the most accurate results:
  - AdaBoosted Decision Tree Classification with **93%** accuracy and **0.93** f1 score
  - Random Forest with **93%** accuracy and **0.93** f1 score
  - Stacking with **97%** accuracy and **0.96** f1 score