A COMPARATIVE STUDY BETWEEN NAIVE BAYES AND RANDOM FOREST APPLIED TO PREDICTING HEART DISEASE

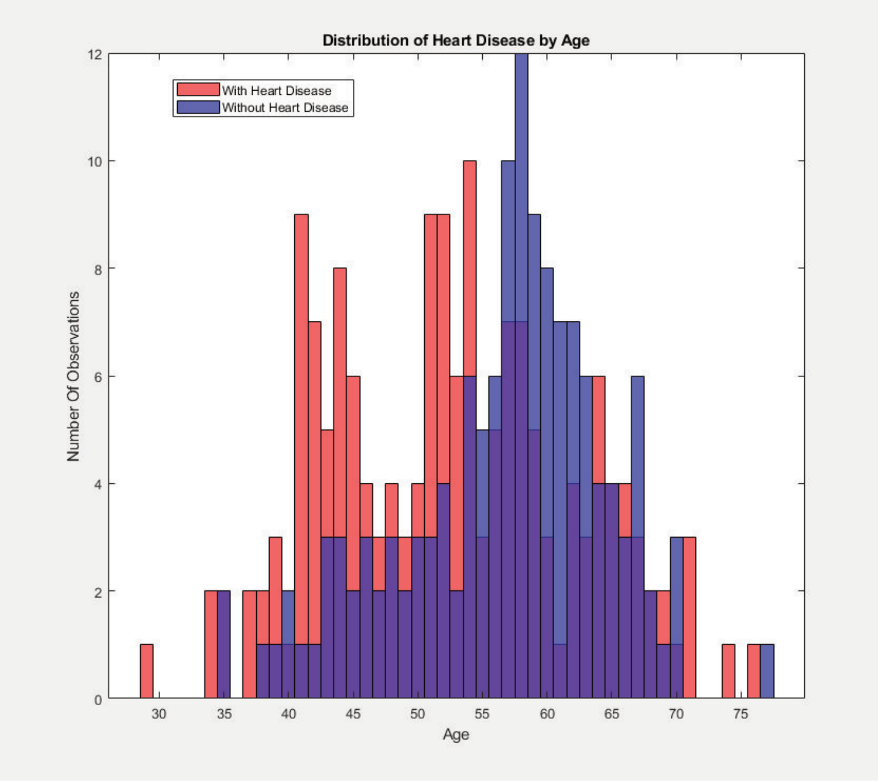
Dmitry Borisovskiy and Ozan Karatepe

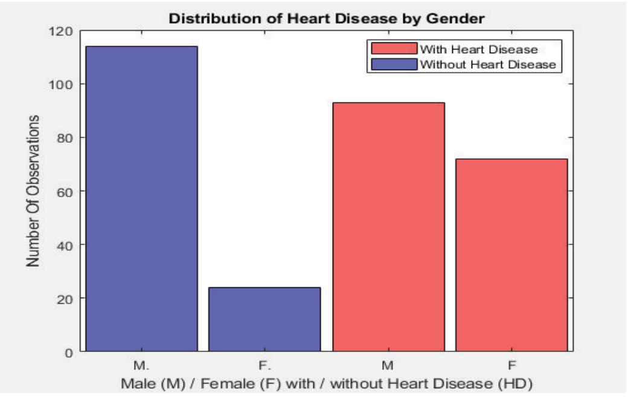
Brief description and motivation of the problem

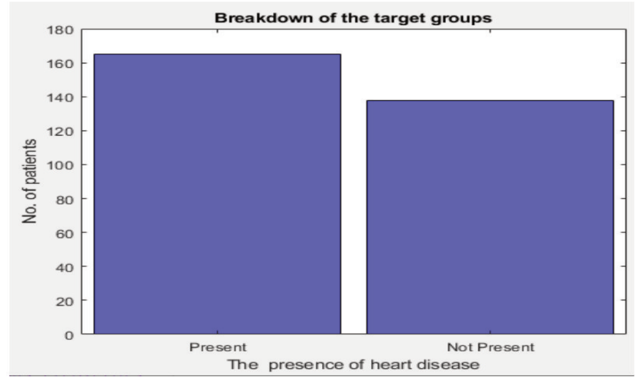
- Compare and evaluate the performance of Naïve Bayes and Random Forest in application to binary classification problem, predicting if the participants have heart disease or not.
- Heart Disease is considered to be one of the major causes of death both in the UK and across the World. Being able to detect and treat Heart Disease reduces the chance a patient having heart issues.
- Our results will be compared with the recent study done for the same dataset by Mohan, Thirumalai and Srivastava (2019).

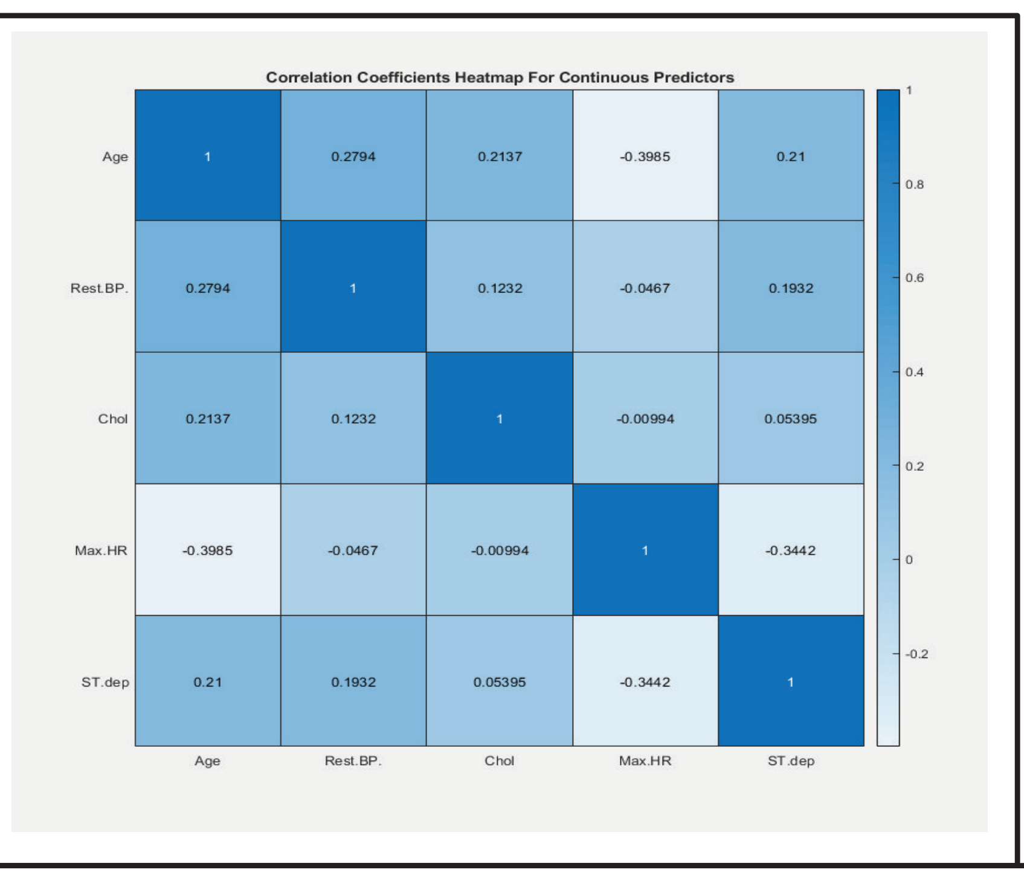
Initial analysis of the data set including basic statistics

- Dataset: Cleveland Heart Disease Dataset from the UCI Repository consisting of 303 observations, 13 predictors (5 continuous and 8 categorical) and 1 attribute called 'target' which identifies the presence of heart disease.
- 165 participants have heart disease and 138 participants don't have heart disease so the classes are well balanced as shown on the graph.
- This version of the dataset was preprocessed for the specific purpose of creating models to generalise heart disease.
- Among the 13 predictors in the dataset, 11 attributes are lifestyle predictors based on the clinical records and considered as very important in determining heart disease (Amin, Chiam and Varathan, 2018).
- The other 2 predictors – 'age' and 'sex' are often considered as less important as they refer to patient personal information. However a lot of studies have shown that both age and gender are extremely important in predicting heart disease and therefore these two predictors will not be removed from our experiment.
- The correlation coefficients heat map has been used to find correlation between continuous predictors, no high correlation has been observed so no need to remove any of the predictors.









	Age	Rest.BP.	Chol	Max.HR	ST.dep
Age	1	0.2794	0.2137	-0.3985	0.21
Rest.BP.	0.2794	1	0.1232	-0.0467	0.1932
Chol	0.2137	0.1232	1	-0.00994	0.05395
Max.HR	-0.3985	-0.0467	-0.00994	1	-0.3442
ST.dep	0.21	0.1932	0.05395	-0.3442	1

Naive Bayes

Naive Bayes (NB) is based on Bayes Theorem which looks to find the probability of heart disease occurring given the predictors have occurred. Naïve Bayes is known as one of the most widely used machine learning algorithms and a good example of generative model. It is a probabilistic model that makes the assumption that all predictors are independent from each other, this is called class conditional independence. According to Dong, Li and Xie (2014) Naive Bayes aims to assign the data object to a set of categories, it calculates the probability of the data point belonging to each category and assigns it to the category with the highest probability.

Advantages:

- Easy to compute and implement.
- Great for small datasets but can deal with big data as well.
- Time efficient.
- Proved to be effective in dealing with real-world cases.

Disadvantages:

- Based on a rarely realistic assumption that predictors are independent and therefore it's accuracy drops when predictors are not independent.
- It struggles with nonparametric continuous attributes. (Arroyo and Sucar, 2006)
- Generally is outperformed by other models including Random Forest.
- Can not learn relationships between predictors due to assumed conditional independence.

Random Forest

Random Forest (RF) is a popular machine learning tool applicable to both regression and classification tasks. Random Forest is a tree-based classifier. Depending on the task RF increases the number of decision trees which are trained on a different subset of same training data. It helps to improve efficiency and at the same time reduces the chance of over-fitting.

Advantages:

- Proved to be effective in dealing with both regression and classification tasks, its a versatile model.
- Works well in application to imbalanced data.
- Time efficient with large datasets
- Lower variance compared to Decision Trees, still with low bias.
- Works well with dimensionality.

Disadvantages:

- Not always easy to find the optimal number of decision trees.
- Low interpretability, in the medical domain being able to explain a decision is important.
- Tendency to overfit, hyperparameter tuning is necessary.
- Can take a lot of computational power on large datasets.
- Might be biased in feature selection for individual trees which can lead to a decrease in accuracy.(Paul et al., 2018)

Hypothesis statement

- We believe that Random Forest will perform better than Naive Bayes as the variables are not independent from each other. Rather the majority of predictors are related to a patient's lifestyle which is common in the medical domain.
- This hypothesis is backed up by the study by Mohan, Thirumalai and Srivastava (2019) where RF has performed better than NB in terms of accuracy and other performance measures.
- From our study of the literature RF works marginally better than NB with this dataset.
- We also expect NB to work well as there are relatively small number of observations in the dataset. The classifier generally works well with small datasets since only independent variables are assumed and therefore only variance of the variables for each class needs to be determined (Pattekari and Parveen, 2012).

Description of choice of training and evaluation methodology

- Dataset was shuffled before partition as the initial observation has shown that the dataset was vaguely split between patients with heart disease and without heart disease.
- Data was shuffled to reduce variance and generalise well without the risk of overfitting.
- Dataset was split using 'holdout' method - 70% for training and 30% for testing.
- Considering the number of observations in the dataset, a subset for validation was omitted.
- Different subsets were created to run with different Naive Bayes models due to the algorithm's inability to handle categorical and continuous variables under the same distribution.
- For Naive Bayes models, parameters were set prior to running the experiment.
- 10 fold CV cross validation was used, using training subset.
- The model chosen for testing was the model with lowest cross validation classification error over 10 runs.
- Random Forest hyperparameters were chosen using a Grid Search.
- OoB Prediction was used as the model's cross validation.
- Hyperparameters were chosen based on the model with the lowest classification error.

Choice of parameters and experimental results

Naive Bayes

- As the dataset consisted of categorical and continuous variables 3 experiments were run to analyse the impact of using variations of the dataset.
- Method 1 – All predictors were kept, mixed distribution of Gaussian and Multivariate Multinomial. Gaussian was used on normalised continuous predictors. Multivariate Multinomial distribution was used on categorical predictors.
- This method was further experimented with Kernel distribution and Kernel distribution with triangle smoothing.
- Method 2 – Categorical and continuous predictors were split. The dataset is split into subsets for categorical and continuous predictors. Multivariate Multinomial was assumed for categorical predictors.
- Continuous predictors were normalised using Z score and two models were created one assuming a Gaussian distribution and one assuming a Kernel distribution.
- Method 3 – Continuous variables were discretised through binning.
- The predictors were binned into 10 bins which transformed the dataset into categorical predictors. A Multivariate Multinomial was assumed for the dataset.
- 7 models were created with different parameters
- All models run simultaneously, to analyse what combination performed best.
- 10 K Fold cross validation was used to improve the models over 10 runs.
- Models were analysed based on k fold error and training error.

Results

- From the plot it can be seen that model 1 performed better, achieving a lower training and validation compared to other models.
- Model 1 had a mixed distribution of Gaussian and Multivariate Multinomial, with all predictors. This highlights that NB performs better for this dataset with all predictors under assumption of mixed distribution compared to models with specified predictors.
- Class prior was added to further improve model 1. Slight improvement was achieved, prior was kept in the model for testing.
- Model 1 achieved 86.6% testing accuracy.
- It took model 1 0.04 seconds to run the test dataset.
- Model 7 which consisted of just categorical predictors performed remarkably well considering the majority of predictors were excluded.
- Application of the different subsets of the dataset had a greater effect on NB performance than changes to hyperparameters.
- NB performance could be further improved with more experiments with the hyperparameters.

Random Forest

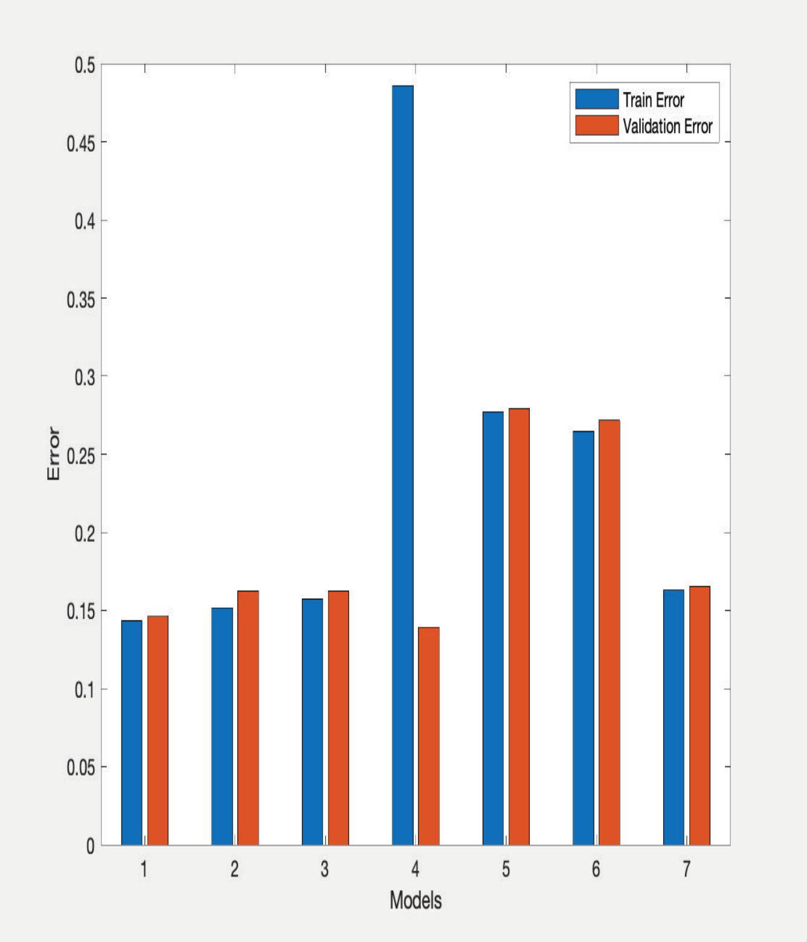
- Dataset consisted of categorical predictors and normalised continuous predictors.

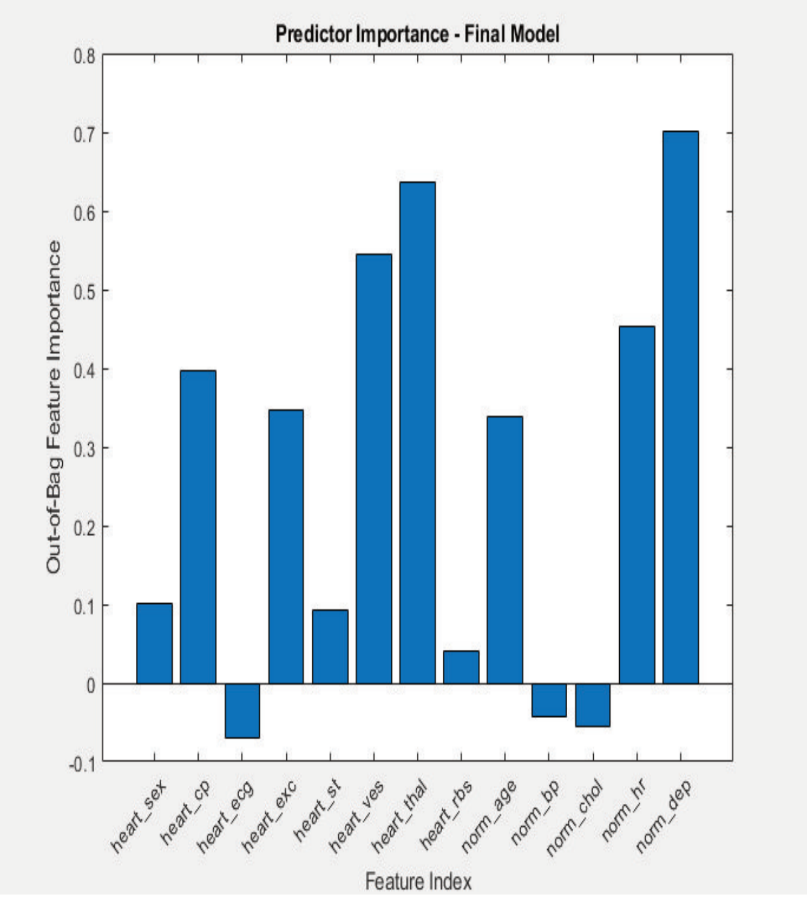
1. Baseline model was created with 500 trees and no changes to hyperparameters.
1. Out of bag error was plotted against number of trees to gain a baseline error and understand the number of trees suitable for the dataset.
1. Oob error decreased as the number of trees increased.
Plot levelled out around 50 trees with little error decrease afterwards.

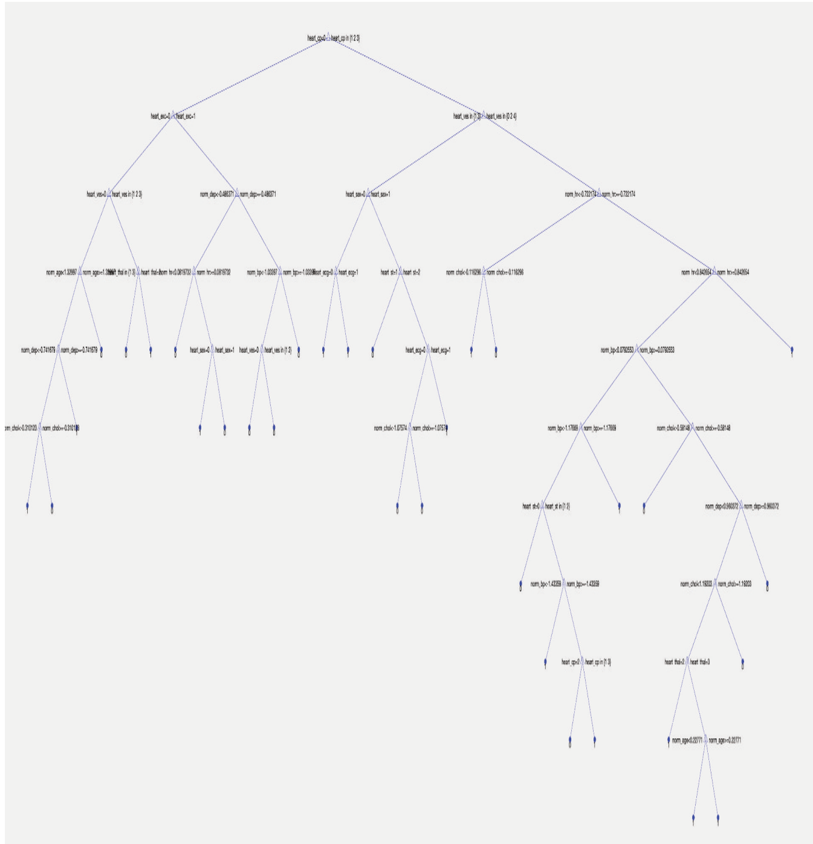
- Predictor importance was plotted using a curvature test, predictor 'Chest Pain' was very important was at the top the decision tree model.
- Grid search was used to optimise the model.
- Hyperparameters that were considered in the optimisation process:

- Number of trees in the ensemble,
- Minimum number of observations per tree leaf
- Number of predictors to sample at each decision split

- The best model consisted of 100 trees, 6 features and a minimum leaf size of 3.
- It took 96 seconds to optimise the model
- 'ST depression' was considered to be the most important predictor in the final model.



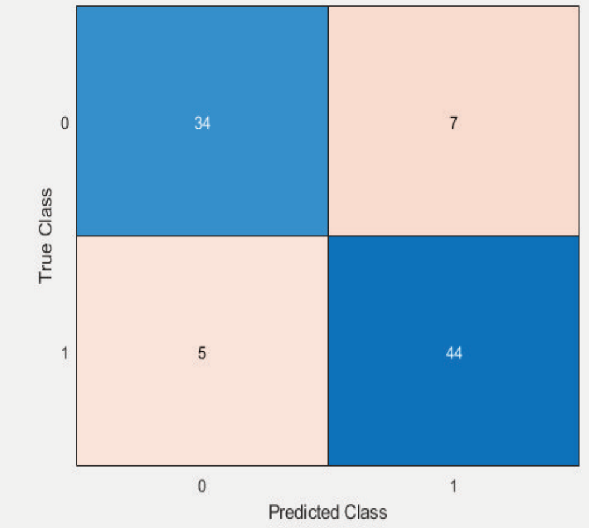




Analysis and critical evaluation of results

- Overall our experiment has shown a high accuracy results for both NB and RF as expected. This is partly because of the well pre-processed and balanced dataset which is used by many machine learning studies. The used dataset did not contain any missing values or extreme outliers that could affect the accuracy of the models.
- Predicting heart disease is an extremely important challenge for machine learning techniques. The fact that both models have shown a very high performance results proves that machine learning has a great potential in predicting heart disease and further studies and developments are needed to be able to generalise it to other medical illnesses.
- Our experiment has shown that RF has performed better than NB which is inline with our hypothesis.
- RF had a higher accuracy than NB, considering the class balance of the dataset, this is an important indicator.
- RF had better success at classifying observations correctly than NB.
- RF outperformed NB in precision, recall and F1 score. This shows that RF was better at classifying both patients with heart disease and patients with no heart disease.
- Precision is highly important in the medical field where it is vital to correctly classify true positives.
- Recall is just as important where it is vital not to miss a patient who has heart disease
- RF's ability to classify classes correctly is summarised by its high F1 score.

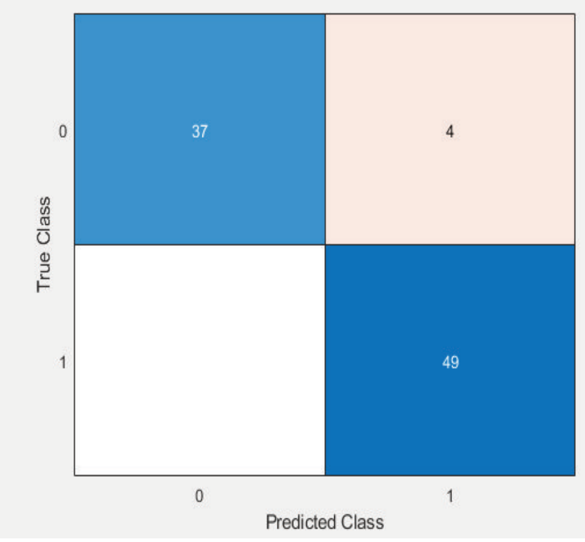
NAÏVE BAYES



Test Set Model Performance Indicators

Naïve Bayes	Models	Random Forest
86.6%	Accuracy	95.5%
82.9%	Precision	90.2%
87.1%	Recall	100%
0.8500	F1 Score	0.9487
0.9059	AUC	0.8671
0.04	Time Taken (seconds)	10.06

RANDOM FOREST

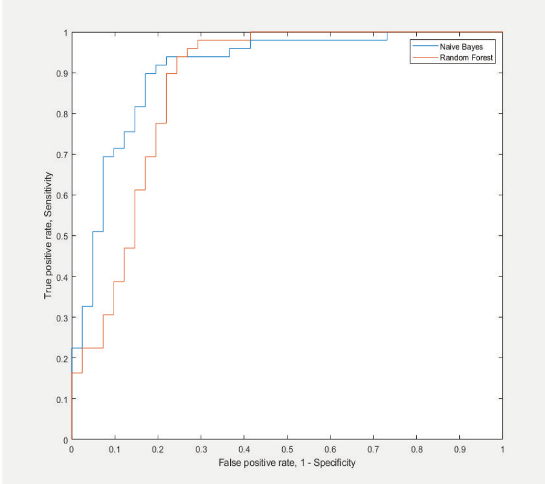


- NB outperforms RF in AUC, proving that it is able to distinguish between classes well. Both models have high discrimination capacity. NB shows thats its better at separating classes. RF shows that it better at classifying patients with heart disease or without heart disease.
- Naive Bayes impresses in its performance considering the presence of both continuous and categorical predictors and the fact that the predictors are dependant on each other. This supports the phenomenon of NB.
- NB took less time to run compared to RF, which is an important factor when considering to run the models on larger datasets. It is important to consider whether the extra time and computational power is worth the increased performance.
- Training and testing was performed on a relatively small dataset (303 observations). It would be beneficial to see how the models perform on a larger dataset.
- Shuffling the dataset helped to prevent overfitting which especially supported the RF model, which has a tendency to overfit when optimising the model.
- NB could have performed better if it was exposed to the same grid search as RF rather than having the hyperparameters preset.

Accuracy Errors For Each Model

Naïve Bayes	Models	Random Forest
17.37%	Training Error	15.40%
18.31%	Validation Error	15.49%
18.89%	Test Error	13.36%

ROC Curve



Lessons learned and future work

- It is important to investigate the results of combining different predictors together on the accuracy of the models. For example, according to (Amin, Chiam and Varathan, 2018) the highest accuracy for NB was achieved by predictor combination: sex, cp, thalach, exang, oldpeak, ca.
- Domain knowledge is extremely important, with greater knowledge we can look at feature generation from the dataset or from the other Heart Disease datasets.
- It is interesting to explore the effect of conducting both out of bag and cross validation on the accuracy of RF.
- To experiment further with different methods of binning the continuous variables.
- To apply PCA and analyse whether reducing dimensionality improves the models.
- Better knowledge of the medical domain will aid in the binning process of continuous predictors.
- To expose models to larger datasets, to how long RF will take to run on a larger
- dataset. To analyse NB's generalisation on a larger dataset.
- The test data consisted of 90 observations which is relatively small, which brings doubt to the validity of the results.

Amin, M.S., Chiam, Y.K. and Varathan, K.D. (2018) 'Identification of significant features and data mining techniques in predicting heart disease, Telematics and Informatics, pp. 82-93.
Arroyo, M.M. and Sucar, L.E. (2006) ' Learning an Optimal Naive Bayes Classifier', Conference Paper.
Bashir, S., Qamar,U., Khan, F.H and Javed, M.Y. (2014) 'MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble', Arabian Journal for Science and Engineering.
Cherian, V. and Bindu M.S. (2017) 'Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique', International Journal of Computer Science Trends and Technology (IJCST) – Volume 5 Issue 2, Mar – Apr 2017, pp.68 - 73.
Dong, L., Li, X. and Xie, G. (2014) ' Nonlinear Methodologies for Identifying Seismic Event and Nuclear Explosion Using Random Forest, Support Vector Machine, and Naive Bayes Classification', Hindawi Publishing Corporation, Abstract and Applied Analysis, Volume 2014.
El-Bialy, R., Salamay, M., Karam, O.H. and Khalifa, M.E. (2015) 'Feature Analysis of Coronary Artery Heart Disease Data Sets', Science Direct, pp. 459 - 468.
Hedeshi, G.N. and Abadeh, S.M. (2014) 'Coronary Artery Disease Detection Using a Fuzzy-Boosting PSO Approach', Comput Intell Neuroscience.
Huang, J., Lu, J. and Ling, C.X. (2003) 'Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy', The Third IEEE International Conference on Data Mining, 2003.
Kinge, J., and Gaikwad, S.K. (2018) 'Survey on data mining techniques for disease prediction', IRJET Journal, pp.630 - 636.
Mahboob, T., Irfan, S. and Karamat, A. (2016) 'A machine learning approach for Student Assessment in E-Learning Using Quinlan's C4.5, Naïve Bayes and Random Forest Algorithms', Fatima Jinnah Women University Rawalpindi, Pakistan.
Mohan, S., , Thirumalai, C. and Srivastava, G. (2019) 'Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques', Volume 7, 2019.
Mujawar, S.H. and Devale, P.R. (2015) 'Prediction of Heart Disease using Modified K-means and by using Naive Bayes', International Journal of Innovative Research in Computer and Communication Engineering, pp. 10265 - 10273.