Ryerson University
CKME 136 Data Analytics Capstone Course

Submitted By: Todd Bethell
July 30, 2019

Coronary Artery Disease Prediction Using Classification Algorithms

Introduction

Using Angiography for the diagnosis of Coronary Artery Disease (CAD) is an accurate, but expensive method with many side effects. A more cost-effective and less invasive method is needed in order to make accurate diagnoses of CAD accessible for doctors and patients. This research paper will assess the predictive capability of three different types of machine learning algorithms. The effectiveness of these three types of algorithms will be compared against known results established using angiography. The three types of classification algorithms that will be evaluated are the Random Forest, the Naïve Bayes, and the Logistic Regression algorithms.

The dataset used in this paper is called Z-Alizadeh Sani [1] and was collected from 303 random visitors to Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Center who were suspected of having CAD. Angiography was used to find that 87 visitors were healthy and 216 had CAD. These results are used as the basis of comparison to assess the predictive capability of the three classification algorithms.

Literature Review

In 2018, heart disease is the second leading cause of death in Canada [2] , the United States [3] , and in fact it remains the leading cause of death in industrialized nations around the world [4]. Coronary heart disease is also known as ischaemic heart disease and it refers to the build-up of plaque in the heart's arteries which can cause a heart attack, a stroke or heart failure. Heart disease is more common in men, people who smoke, people who are overweight or obese, people over the age of 55 and for those with a family history of heart disease or heart attack. As such, there are risk factors that can be controlled and others that cannot. Making lifestyle changes can reduce the chance of having heart disease. Some controllable risk factors include smoking, high low-density lipoprotein (LDL, often called 'bad' cholesterol), low high-density lipoprotein (HDL, often called 'good' cholesterol), uncontrolled high blood pressure, physical inactivity, obesity, uncontrolled diabetes, and uncontrolled stress.

Angiography is considered an accurate method used to diagnose the presence of coronary heart disease. An angiogram allows doctors to view blood flow through the heart by injecting a special dye into the coronary arteries. The dye is injected into arteries of the heart through a flexible catheter that is threaded through an artery, usually in the leg. This dye shows narrow spots and blockages on X-ray images [5]. There are many risks and potential side-effects of angiograms. There is the extremely small chance of developing cancer in the long term due to the exposure to the radiation. There are also potential side-effects with some medications, such as blood thinning and diabetic medications. In addition, there are the risks of allergic reactions, infection, blood clot, and weakness of the blood vessel wall [6].

Two other diagnostic tests that are referred to in the dataset used for this paper are the Electrocardiogram (ECG) and the Echocardiogram. An ECG records electrical signals as they travel through the heart and can show evidence of a previous heart attack or one that is in progress and can also show abnormalities indicating inadequate blood flow to the heart. An echocardiogram uses sound waves to produce images of the heart. Such images are used to determine whether all parts of the heart are contributing normally to the heart's pumping activity [7].

There have been numerous studies completed using various data mining algorithms for the purpose of heart disease prediction. The use of such algorithms in health care applications has gained in popularity over recent years given their efficiency, cost and non-invasive advantages. Heart disease prediction by using a number of different machine learning algorithms on a few different datasets has been successfully achieved in these studies. Sonam Nikhar and A.M. Karandikar (2016) [8] concluded that the Decision Tree classifier provided better results in the diagnosis of heart disease than Neural Network, Support Vector Machine (SVM) or k-Nearest Neighbour (kNN) classifier techniques. Aamanpreet Kaur (2017) [9, 10] found that the Random Forest algorithm is quite effective at correctly classifying instances of heart disease. N. Bhatla and K. Jyoti (2012) [11] were able to achieve high accuracy with the Naïve Bayes and Decision Tree classifiers while also reducing the number of attributes used from 15 to 4. Mai Shouman et al (2012) [12] was able to show that applying the k-Nearest Neighbour (kNN) algorithm can achieve higher accuracy than neural network ensemble in the diagnosis of heart disease.

However, while there has been significant success predicting heart disease with various data mining techniques, Mudasir Kirmani (2017) [13] found that there is no single classifier which produces the best results for every dataset and no single data mining technique which gives consistent results for all types of healthcare data. As such, further investigation is still needed given the availability of huge amounts of medical data and the subsequent need for data analysis tools to extract useful knowledge.

The three classification algorithms that will be used in this paper are the Random Forest, the Naïve Bayes, and the Logistic Regression algorithms. All three of these machine learning algorithms are and/or can be non-parametric methods that can be used for classification and regression [14]. A confusion matrix is an available output for all three types of algorithms which will be used to evaluate their accuracy, sensitivity, specificity, precision, F1-Score and Matthew's Correlation Coefficient. Sensitivity measures the true positive prediction rate, specificity measures the true negative prediction rate, and accuracy is calculated as the number of all correct positive and negative predictions divided by the total number of observations in the dataset [15]. The other evaluation measures are described later in this paper.

Dataset

The dataset that is used for this paper is called the Z-Alizadeh Sani dataset. It has been downloaded from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/machine-learning-databases/00412/). This dataset contains the records of 303 patients, each of which have 54 attributes. All of the attributes are considered as indicators of Coronary Artery Disease (CAD). These attributes are split into four groups: Demographic, Symptom & Examination, ECG, and Laboratory & Echo Features. The Demographic group includes 16 attributes, the Symptom & Examination group includes 14 attributes, the ECG group includes 7 attributes and the Laboratory & Echo Features group includes 17 attributes. This dataset shows the results of angiography testing in which each patient is diagnosed into one of two possible categories: CAD or Normal. A patient is diagnosed as CAD if their coronary artery diameter narrowing is greater than or equal to 50%, otherwise they are categorized as Normal. The actual diagnoses of CAD or Normal was determined using angiography. The results showed that 87 patients were healthy (that is, 'Normal') and 216 had CAD.

The following four tables present all of the attributes used in the Z-Alizadeh Sani dataset, broken into the four groups, and include their valid ranges. Table 5 defines the discretized features and ranges of some of the attributes in tables 1, 2 and 4. Table 6 provides a summary of the attributes *before* there any changes made to the data types and before any discretization is completed.

DEMOGRAPHIC ATTRIBUTES

Feature Name	Column Header	Range
Age	Age	30 – 86
Weight (kgs)	Weight	48 – 120
Sex	Sex	Male, Female
History of Diabetes Mellitus	DM	Yes, No
History of Hypertension (High Blood Pressure)	HTN	Yes, No
Current Smoker	Smoker	Yes, No
Ex-Smoker	ExSmoker	Yes, No
FH (Family History of Heart Disease in First-Degree Relatives)	FH	Yes, No
Obesity	Obesity	Yes, if BMI > 25, No otherwise
Chronic Renal Failure	CRF	Yes, No
Cerebrovascular Accident (often referred to as a Stroke)	CVA	Yes, No
Airway Disease (such as asthma, COPD and bronchiectasis)	AD	Yes, No
Thyroid Disease	TD	Yes, No
Congestive Heart Failure	CHF	Yes, No
Dyslipidemia (elevated cholesterol, triglycerides, or both)	DLP	Yes, No

Table 1

Note: 'Age' will change to a categorical attribute with 2 levels, as defined on Table 5. 'Weight' will be discretized as 'Average' or 'High' using the below or above mean results using the data from this attribute since it was not discretized in the original dataset.

SYMPTOM & EXAMINATION ATTRIBUTES

Feature Name	Column Header	Range
Blood Pressure (mmHg)	BP	90 - 190
Pulse Rate (pulse per minute)	PR	50 - 110
Edema (excess fluid trapped in the body's tissues)	Edema	Yes, No
Weak Peripheral Pulse (weak pulse in extremities)	WPP	Yes, No
Lung Rales (rattling/bubbling sound in lungs)	LR	Yes, No
Systolic Murmur	SysM	Yes, No
Diastolic Murmur	DiaM	Yes, No
Typical Chest Pain (pressure/squeezing in chest)	TCP	Yes, No
Dyspnea (shortness of breath)	Dyspnea	Yes, No
Heart Failure Functional Class [20] *	Fclass	1, 2, 3, 4
Atypical Chest Pain (not heart-related pain)	ACP	Yes, No
Nonanginal Chest Pain (typically lasting over 30 minutes)	NCP	Yes, No
Exertional Chest Pain (pain from exertion or excitement)	ECP	Yes, No
Low Threshold Angina (low threshold for angina pain)	LTAng	Yes, No

Table 2

** Heart Failure Functional Class: https://www.chf-solutions.com/heart-failure-classifications/*

Note: Blood Pressure and Pulse Rate will be changed to categorical attributes (with 3 levels). Heart Failure Functional Class will change from a categorical attribute with 4 levels, to having 2 levels. These changes and levels are defined on Table 5.

ECG ATTRIBUTES

Feature Name	Column Header	Range
Q Wave (Present = Yes, not present = No)	Qwave	Yes, No
ST Elevation (Elevation present = Yes, not present = No)	Stelev	Yes, No
ST Depression (Depression present = Yes, not present = No)	Stdep	Yes, No
T Inversion (Inverted = Yes, not inverted = No)	Tinv	Yes, No
Left Ventricular Hypertrophy	LVH	Yes, No
Poor R Wave Progression	PoorR	Yes, No
Bundle Branch Block	BBB	LBBB, RBBB, No

Table 3

LABORATORY & ECHOCARDIOGRAPHIC ATTRIBUTES

Feature Name	Column Header	Range
Fasting Blood Sugar (mg/dl)	FBS	62 - 400
Creatine (mg/dl)	Cr	0.5 - 2.2
Triglyceride (mg/dl)	TG	37 - 1050
Low Density Lipoprotein (mg/dl)	LDL	18 - 232
High Density Lipoprotein (mg/dl)	HDL	15 - 111
Blood Urea Nitrogen (mg/dl)	BUN	6 to 52
Erythrocyte Sedimentation Rate (mm/h)	ESR	1 to 90
Hemoglobin (g/dl)	Hb	8.9 - 17.6
Potassium (mEq/lit)	K	3.0 - 6.6
Sodium (mEq/lit)	Na	128 - 156
White Blood Cell (cells/ml)	WBC	3700 - 18000
Lymphocyte (%)[21]	Lymph	7 to 60
Neutrophil (%)[22]	Neut	32 - 89
Platelet (1000/ml)	PLT	25 - 742
Ejection Fraction (%)	EF	15 - 60
Region With Regional Wall Motion Abnormality	RWMA	0, 1, 2, 3, 4
Valvular Heart Disease	VHD	Normal, Mild, Moderate, Severe

Table 4

<u>Note:</u> For those attributes in Table 4 that are not already categorical, all but two of them will be changed to categorical attributes, as defined in Table 5. Lymphocyte and Neutrophil will be discretized using accepted ranges provided by the medical community per the references provided.

Discretized Features and Their Range of Values

Feature Name	Low	Medium	High
Fasting Blood Sugar	FBS < 70	70 ≤ FBS ≤ 105	FBS > 105
Creatine	Cr < 0.7	0.7 ≤ Cr ≤ 1.5	Cr > 1.5
Triglyceride		TG ≤ 200	TG > 200
Low Density Lipoprotein		LDL ≤ 130	LDL > 130
High Density Lipoprotein	HDL < 35	HDL ≥ 35	
Blood Urea Nitrogen	BUN < 7	7 ≤ BUN ≤ 20	BUN > 20
Erythrocyte Sedimentation Rate		If male & ESR ≤ age/2 or if female & ESR ≤ (age/2) + 5	If male & ESR > age/2 or if female & ESR > (age/2) +5
Hemoglobin	If male & Hb < 14 or if female & Hb < 12.5	If male & 14 ≤ Hb ≤ 17 or if female & 12.5 ≤ Hb ≤15	If male & Hb > 17 or if female & Hb > 15
Potassium	K < 3.8	3.8 ≤ K ≤ 5.6	K > 5.6
Sodium	Na < 136	136 ≤ Na ≤ 146	Na > 146
White Blood Cell	WBC < 4000	4000 ≤ WBC ≤ 11000	WBC > 11000
Lymphocyte (%)[21]	Lymph < 18	18 ≤ Lymph ≤ 45	Lymph > 45
Neutrophil (%)[22]	Neut < 44	45 ≤ Neut ≤ 75	Neut > 75
Platelet	PLT < 150	150 ≤ PLT ≤ 450	PLT > 450
Ejection Fraction (%)	EF ≤ 50	EF > 50	
Regional Wall Motion Abnormality		RWMA = 0	RWMA ≠ 0
Blood Pressure	BP < 90	90 ≤ BP ≤ 140	BP > 140
Pulse Rate	PR < 60	60 ≤ BP ≤ 100	PR > 100
Heart Failure Functional Class		1	2, 3, 4
Weight		≤ 70.52 kg if female, or ≤ 76.21 kg if male	> 70.52 kg if female, or > 76.21 kg if male
* Age		If male & age ≤ 45 or if female & age ≤ 55	If male & age > 45 or if female & age > 55

Table 5

** Since women under age 55, and men under age 45 are less affected by CAD, the range of age is partitioned at these values.*

Approach

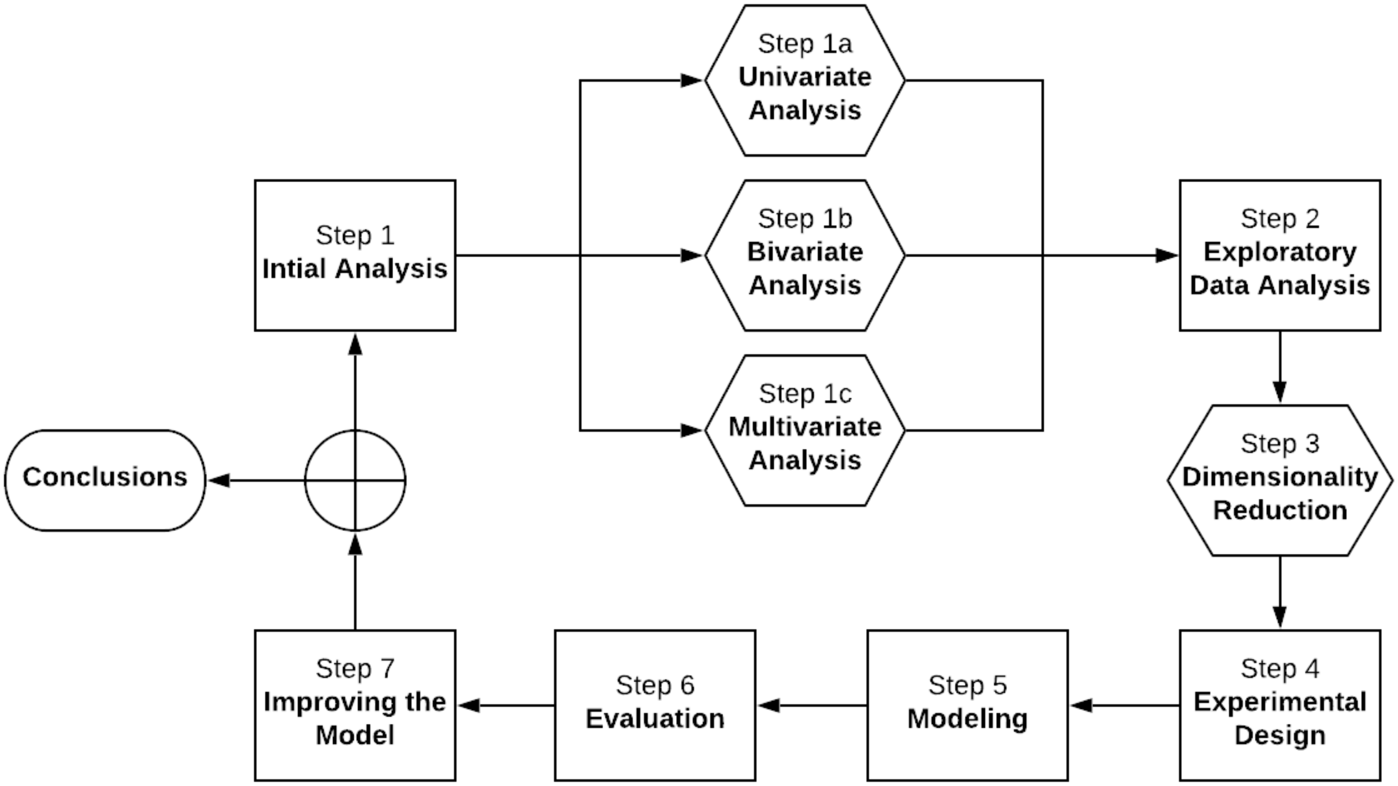


Diagram 1

Approach Details

Step 1: Initial Analysis

- Download the Z-Alizadeh Sani dataset
- Review attributes to understand, create and update the data dictionary
- Define the four groups of attributes and how each set of attributes are related
- Clarify the dependent variable that is going to be predicted (Coronary Artery Disease)
- Import dataset as a .csv file into R Studio
- Update column headers, as needed, for clarity and/or correspondence with data dictionary
- Remove 'Length' attribute from data set. This attribute was ultimately not included in the original study, presumably due to the fact that they have the Body Mass Index values, and as such, 'length' (that is, height) is redundant.
- Removed 'BMI' from the original dataset, given that this attribute is captured in the 'obesity' variable (that is, BMI scores over 25 are considered 'obese')
- Remove dependent attribute from data set in R – where not used for algorithms
- Determine whether there are any missing values
- Determine whether there are any errors, duplicates and/or incomplete or inappropriate entries
- Assign the correct data types (numeric, categorical, etc)
- Create four attribute groups as four data frames to be used throughout the project
- Need to change the attribute scale from numeric to factor, and assign them as 'low', 'medium' and 'high' as defined in the 'Discretized Features & Their Range of Values' table shown in the Dataset section (Table 5). Will discretize lymphocyte and neutrophil percentages into 'low', 'medium' and 'high' limits using accepted medically defined cut-offs. **

- Change the data types to ‘ordered factor’ for those that are factor data types, but retain some ordering (such as ‘low, ‘medium’, ‘high’)
- Will need to further discretize categorical attributes with greater than 2 levels (refer to chart made up of all attributes and their initial data types, levels, understandings).

Step 1a: Univariate Analysis

- For numeric attributes, check the min, mean, max, first and third quartiles.
- Create box plots to visualize these numeric summary values and see outliers (if any). Determine appropriate actions for any outliers (some attributes have been discretized – which addresses those outliers)
- Check the number of levels and their meanings for categorical attributes and make sure they are correct at this stage.
- Check on the distributions of the attributes. Given the high number of attributes (at this stage), the Shapiro-Wilk test of normality will be more useful than visualizing all of the individual distributions. Run the Shapiro-Wilk test on all continuous variables to determine whether they are normally distributed.
- Check for any ‘very low’ variance for individual attributes. This will help determine which attributes can be removed later (if any).
- Found that ‘Exertional Chest Pain’ had no variation. That is, all of the entries for this attribute were the same (‘No’), so this variable was removed.
- Check whether there is an imbalance in the dependent variable. Since the number of ‘Normal’ results to ‘Coronary Artery Disease (CAD)’ results are 87 to 216, respectively, any imbalance will need to be addressed with the algorithm parameters.
- Create visualizations of the attributes within each of the four attribute groups. Keep visualizations that appear to display something ‘interesting’ about the dataset.

Step 1b: Bivariate Analysis

- Investigate the pairwise relations between the input variables as well as between the input variables and output variable.
- Complete correlation analysis and check the significance of relationships between attributes. As appropriate, use Pearson Correlation Coefficient for quantitative variables; Chi-square for two categorical variables – using Cramer’s V test to show the strength of any association between all factor variables.
- Provide graphs visualizing these continuous variable correlations, and factor variables associations.

Step 1c: Multivariate Analysis

- Will consider the One-Way-ANOVA for the parametric attributes and the Kruskal-Wallis test for the non-parametric attributes.
- Given that none of the continuous variables have a normal distribution, the One-Way-ANOVA is not being used.
- Given that almost all of the factor variables are dichotomous (yes/no and/or ‘1’/’0’), The Kruskal-Wallis test is also not being used. There are some variables that are being converted to ordered factor variables, to provide order to those where order should be meaningful (such as ‘low’/’medium’/’high’) – but none of the factor variables are ‘ranked’.

Step 2: Exploratory Data Analysis

- None of the attributes are to be normalized. After discretizing the attributes, as needed, it does not appear that there will be a need to normalize any other attributes (such as age, weight, neutrophil and lymphocyte - since these will ultimately be discretized during the initial analysis).
- Review whether additional sub-setting can clarify attribute relationships further. For example, consider the subsets of the 4 different attribute groups as defined in the original study. Is there enough information in the

features of each of these datasets to make it beneficial to use 4 separate groups of features and 4 sets of algorithms to test each of these 4 subgroups? The feature selection step of this paper will help answer these questions.

- Can also subset those patients who have been diagnosed with CAD versus those that have not. Subset further between men and women. The small size of this dataset may limit the effectiveness of such subsetting. Save these dataframes early on for later investigation and/or predictions.

Step 3: Dimensionality Reduction

- This will be an important step in this project. There are a lot of attributes to consider, even within the four attribute groups. I anticipate that the correlation and variation analyses will enable to removal of some attributes.
- Feature Selection will be used to aid in the dimensionality reduction.
- FSelector and FSelectorRcpp R programming packages will be used for feature selection. I will use these packages on the discretized data set, as well as the data set before discretization – to compare the results. I will also compare those results with the feature selection done with the caret package in R (using Recursive Feature Elimination).
- Save final dataframes for each of the four attribute groups. Also save a separate dataframe with all four attribute groups combined, both discretized and before discretization.
- Will run these feature selection functions on the dataframes before and after factorization and discretization – to compare the results.

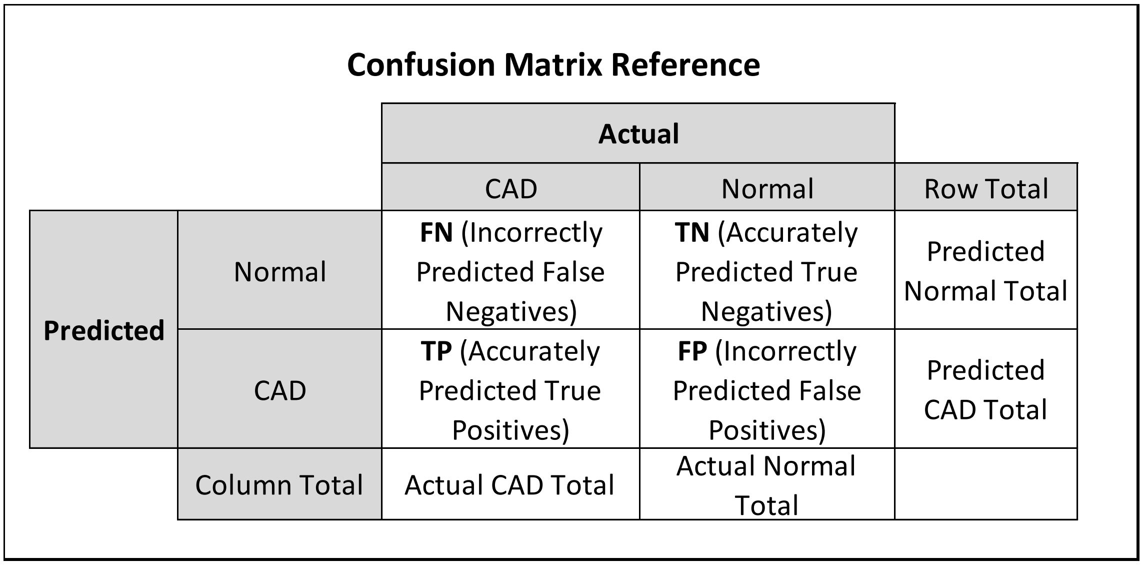
Step 4: Experimental Design

- In order to ensure that there are balanced splits between the training and testing splits of the data set, I am using the createDataPartition function in R to create the initial split for 75% testing and 25% testing. This ensures that the distribution of outcome variable classes will be similar in both sets.
- Given that this data set is not very large, the training of the algorithms will be the done by using the 'k-fold cross-validation' method. Specifically, I will use a 10-fold cross validation, repeated 3 times for each of the three algorithms.
- The 10-fold cross validation should help with addressing any 'overfitting' concerns – since there are a lot of attributes (especially given that the random forest algorithm will be evaluated)
- Address the class imbalance for the output variable (the dependent variable). That is, given that there was a total of 87 patients (28.7%) that were found to be 'Normal' and 216 (71.3%) were diagnosed with 'CAD', that means that there is some class imbalance. Will be using the Synthetic Minority Over-Sampling Technique ('SMOTE' function in R) and 'Undersampling' of the majority class to assist with this problem. SMOTE will synthesize new minority instances between the existing (actual) minority instances. Evaluation of the results will determine which method to use on which algorithm type (and whether 'oversampling' the minority class is a better approach)

Step 5: Modeling

- The research question in this paper - that is, the prediction of the presence of coronary artery disease using a number of independent variables - requires that we use a classification algorithm for the answer. In addition, given that the data set is non-linear, it is appropriate to use the Random Forest, the Naïve Bayes, and Logistic Regression algorithms. The effectiveness, efficiency and stability of these three algorithms will be compared.
- Will be running the three algorithms on the 'combined' dataframe (that is, with the four final attribute group dataframes together).

Step 6: Evaluation



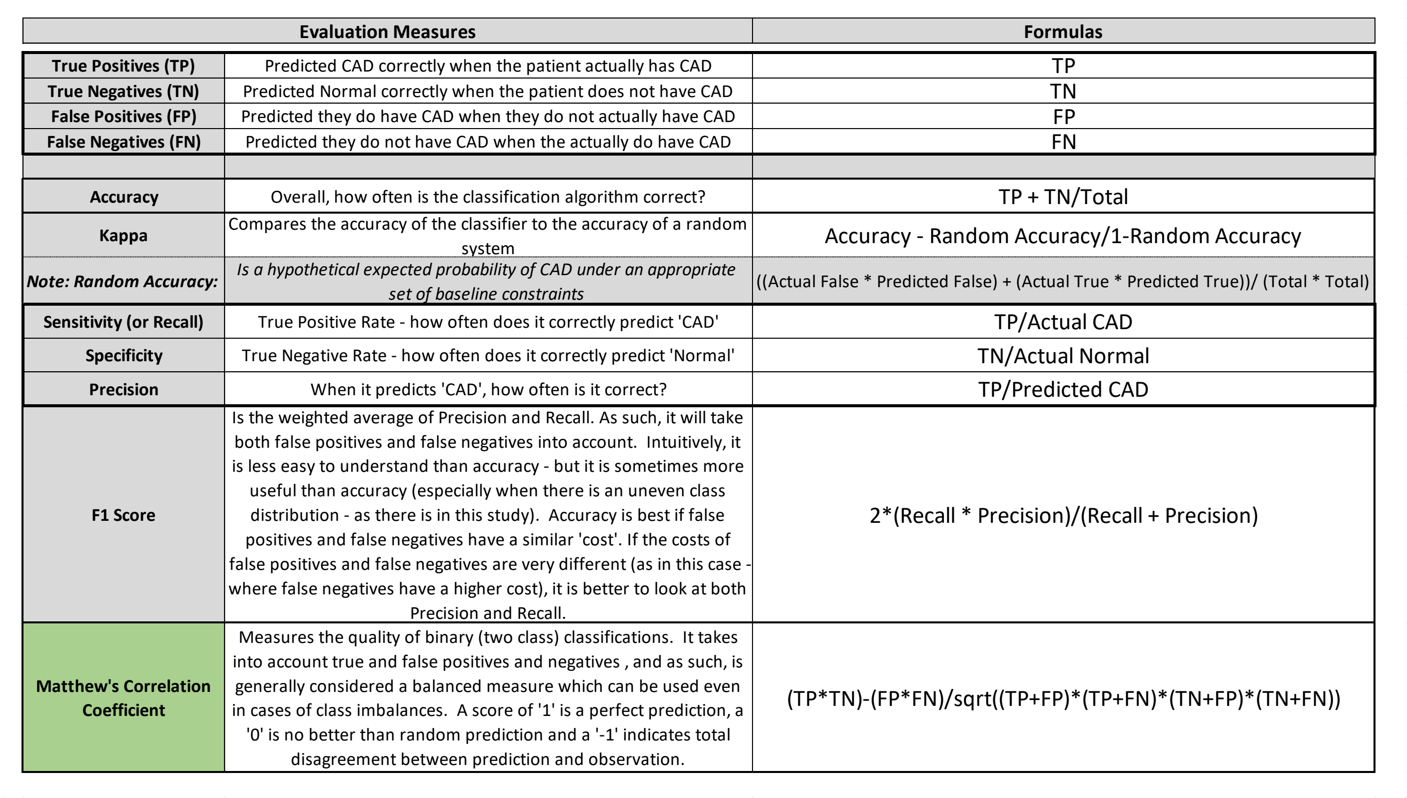
Confusion Matrix Reference

		Actual		
		CAD	Normal	Row Total
Predicted	Normal	**FN** (Incorrectly Predicted False Negatives)	**TN** (Accurately Predicted True Negatives)	Predicted Normal Total
	CAD	**TP** (Accurately Predicted True Positives)	**FP** (Incorrectly Predicted False Positives)	Predicted CAD Total
	Column Total	Actual CAD Total	Actual Normal Total	

Diagram 2

- The accuracy, sensitivity (true positive prediction rate) and specificity (true negative prediction rate) are very relevant metrics to use for evaluating algorithms used for predicting medical diagnoses. The Confusion Matrix provides these evaluation results and all three algorithms output a confusion matrix. This is another reason to run these three machine learning algorithms for this research question.
- Given the class imbalance, I will also be using the confusion matrix results to evaluate the kappa, precision, F1-score, and Matthew's Correlation Coefficient. Emphasis will be given to the Matthew's Correlation Coefficient as a single measure of performance – it is a balanced measure of the quality of binary classification algorithms. It takes into account true and false positives and negatives. A score of '1.0' is a perfect prediction, a '0' is no better than random prediction and a '-1.0' indicates total disagreement between prediction and observation.
- I will also generate the F-score (F1) to help evaluate the results. That is, the F1-Score considers <u>both</u> 'Precision' and Sensitivity. It is the 'harmonic mean' of the precision and sensitivity. In addition, F1 is a good evaluation measure to use when there is an uneven class distribution (as is the case with this data set).
- For additional evaluation and understanding of the interpretation of the impacts of the final variables selected for use on these three algorithms. I will use the logistic regression algorithm in the base R package to see whether the log odds of being diagnosed with Coronary Artery Disease increases or decreases with a one unit increase of each of the final attributes.
- The Brier Score will be calculated for the Logistic Regression algorithm. It is a function that measures the accuracy of probabilistic predictions. It is appropriate where predictions assign probabilities to a set of mutually exclusive discrete outcomes (CAD or Normal).
- I will measure the performance efficiency of the machine language algorithms using the 'tictoc' package in R.
- Lastly, I will evaluate the stability of the algorithms and confirm whether they generate consistent results. I will graph the results to help determine stability and any evidence of overfitting.

Overview of Confusion Matrix Calculations for Performance Evaluation



Evaluation Measures		Formulas
True Positives (TP)	Predicted CAD correctly when the patient actually has CAD	TP
True Negatives (TN)	Predicted Normal correctly when the patient does not have CAD	TN
False Positives (FP)	Predicted they do have CAD when they do not actually have CAD	FP
False Negatives (FN)	Predicted they do not have CAD when the actually do have CAD	FN
Accuracy	Overall, how often is the classification algorithm correct?	TP + TN/Total
Kappa	Compares the accuracy of the classifier to the accuracy of a random system	Accuracy - Random Accuracy/1-Random Accuracy
Note: Random Accuracy:	*Is a hypothetical expected probability of CAD under an appropriate set of baseline constraints*	((Actual False * Predicted False) + (Actual True * Predicted True))/ (Total * Total)
Sensitivity (or Recall)	True Positive Rate - how often does it correctly predict 'CAD'	TP/Actual CAD
Specificity	True Negative Rate - how often does it correctly predict 'Normal'	TN/Actual Normal
Precision	When it predicts 'CAD', how often is it correct?	TP/Predicted CAD
F1 Score	Is the weighted average of Precision and Recall. As such, it will take both false positives and false negatives into account. Intuitively, it is less easy to understand than accuracy - but it is sometimes more useful than accuracy (especially when there is an uneven class distribution - as there is in this study). Accuracy is best if false positives and false negatives have a similar 'cost'. If the costs of false positives and false negatives are very different (as in this case - where false negatives have a higher cost), it is better to look at both Precision and Recall.	2*(Recall * Precision)/(Recall + Precision)
Matthew's Correlation Coefficient	Measures the quality of binary (two class) classifications. It takes into account true and false positives and negatives , and as such, is generally considered a balanced measure which can be used even in cases of class imbalances. A score of '1' is a perfect prediction, a '0' is no better than random prediction and a '-1' indicates total disagreement between prediction and observation.	(TP*TN)-(FP*FN)/sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))

Table 6

Step 7: Improving the Model

- I will check whether building a machine learning ensemble improves my results.
- In the event of very high accuracy results, I will need to determine whether overfitting is a problem with the Random Forest algorithm, even though there have been preventative measures taken (such as 10-fold cross validation and dimensionality reduction).
- Will need to pay attention to the number of 'events per variable' (EPV) on the predictive accuracy performance for the logistic regression algorithm. The EPV is equal to the number of events (87, being the lesser of 216 and 87, from the 303) divided by the number of predictor variables considered in developing the prediction model. [16]. It is the number of events divided by the number of degrees of freedom required to represent all of the variables in the model. For example, a three-level categorical variable would require two degrees of freedom.
- Related to the previous point, I will want to continue to determine whether additional feature selection is possible (Forward Selection and/or Backward Elimination). Will need to continually evaluate whether more attributes can be eliminated without adversely affecting predictive accuracy.

Final Step: Conclusions

- Describe the inferences that are possible using these three machine learning algorithms.
- Summarize which algorithm(s) provide the best accuracy, kappa, sensitivity, specificity precision, F1-score and Matthew's Correlation Coefficient results.
- Provide overview of threats to validity of results (such as sample size and number of attributes used) and what solutions are available to reduce the effects of such threats.
- Provide conclusions regarding whether the performance evaluation results enable a conclusion that the inferences made are appropriate (that is, that the algorithms measure what they claim to – that they provide 'construct validity').
- Summarize the conclusions in terms that are relevant to the users of such data. That is, to provide concrete, specific recommendations of which attributes should be used, and why.

Results

Correlations and Associations Between Attributes

The investigation into whether there was any correlation between any of the numeric and/or integer (continuous) variables found that only two variables had any correlation. The Pearson Correlation Coefficient calculations are shown on diagram 3 below. Results can range from 1.0 for a perfect positive correlation, to 0 for no correlation, to -1.0 for a perfect negative correlation. Diagram 3 shows that only Neutrophil and Lymphocyte attributes were negatively correlated. With regards to this inverse relationship, an abnormal increase in one kind of white blood cell can cause a decrease in another kind. Both abnormal results can be due to the same underlying condition. Regardless, later investigation shows that these two variables are not significant for Coronary Artery Disease (CAD) prediction.

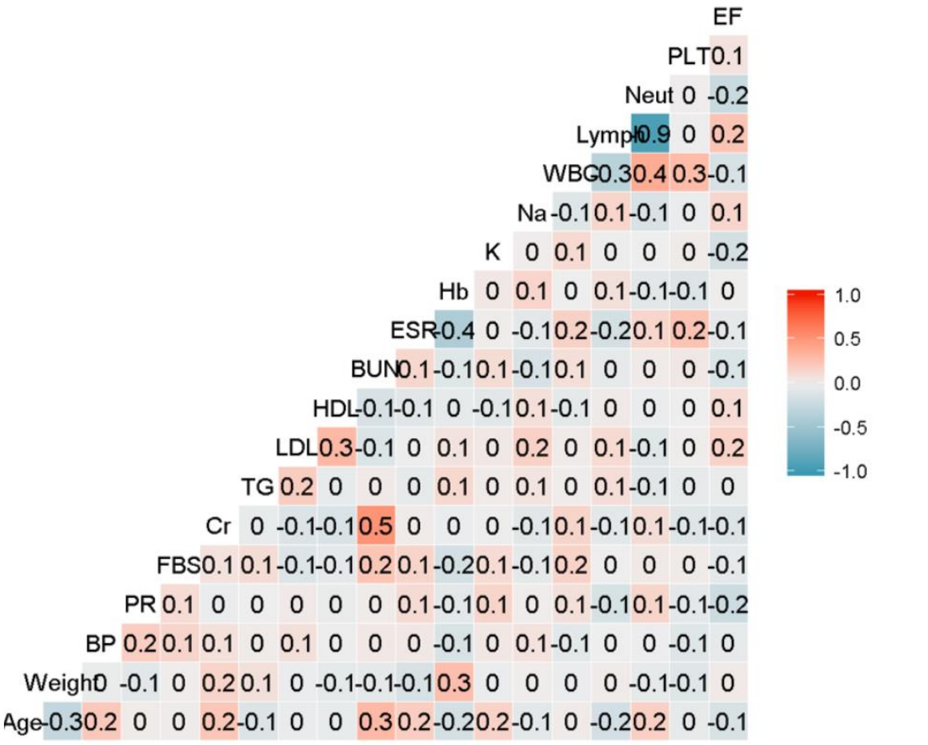
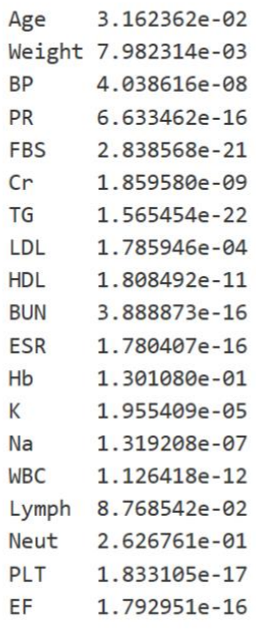


Diagram 3

Results for the Shapiro-Wilk Test of Normality

This test helps determine whether the distribution is approximately normal for the variables on the right (diagram 4). This way, we are able to decide whether parametric tests can be used for later evaluation.

The null (default) hypothesis of this test is that the population is normally distributed. If the p-value of this test is less than the chosen alpha level (0.05 in this case), then the null hypothesis is rejected. As such, p-values below 0.05 mean that the data for the variable in question is not normally distributed.
The p-value results on the right show that none of the variables with continuous data types are normally distributed. These tests were completed before these



```
Age    3.162362e-02
Weight 7.982314e-03
BP     4.038616e-08
PR     6.633462e-16
FBS    2.838568e-21
Cr     1.859580e-09
TG     1.565454e-22
LDL    1.785946e-04
HDL    1.808492e-11
BUN    3.888873e-16
ESR    1.780407e-16
Hb     1.301080e-01
K      1.955409e-05
Na     1.319208e-07
WBC    1.126418e-12
Lymph  8.768542e-02
Neut   2.626761e-01
PLT    1.833105e-17
EF     1.792951e-16
```

Diagram 4

Measures of association were completed using Cramér's V calculations – which provide a number that ranges between 0 and 1. It indicates how strongly two categorical (factor) variables are associated. The measurements show a 'small' association when the result is between 0.10 and 0.30; 'medium' between 0.30 and 0.50; and 'large' when over 0.50

Referring to diagram 5 below, the highest association result was 0.715 and was found between TCP (Typical Chest Pain) and ACP (Atypical Chest Pain). The second highest measure of association was 0.544 between SysM (Systolic Murmur) and VHD (Vascular Heart Disease). The next highest result of 0.535 was between TCP and CAD (Coronary Artery Disease – that is, the response variable). Lastly was the result of 0.498 between DiaM (Diastolic Murmur) and VHD. All of the rest of the variables showed small or medium associations. Only TCP and ACP are variables that are used with the classification algorithms. Despite this 'large' association, both variables remain in the dataset because each has an important influence for the prediction of CAD and they definitely have different medical meanings. The other variables with large associations were not found to be significant for CAD prediction.

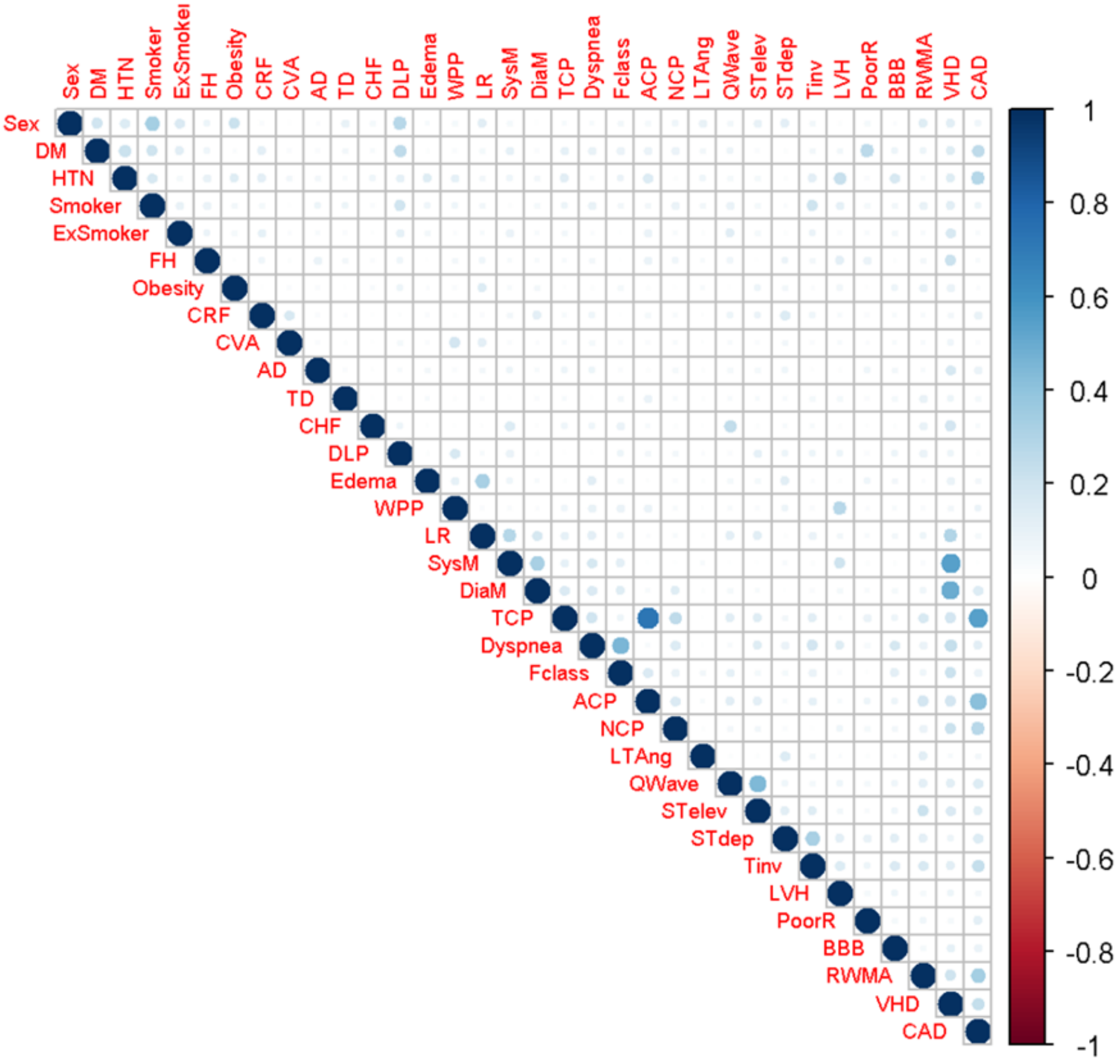


Diagram 5

Feature Selection

An important task for this study was to reduce the number of attributes from 54 to something much more manageable. Having so many variables involves a tremendous amount of time and cost to collect and is too large to effectively use on classification algorithms.

Three methods were used and compared to assist with the reduction in the number of attributes:

1. The '**FSelector**' package in R.

This is a feature ranking technique that ranks the top features according to a user-defined filter. I used an 'entropy-based' filter (that is, 'information gain') to rank the top 10 features. The top 10 features were found to be:

TCP	ACP	RWMA	EF	HTN	Age	DM	NCP	Tinv	FBS

2. The '**FSelectorRccp**' package in R.

This function also ranks the top features using the information-gain entropy-based filter. Again, it was set up to select the top 10 features. The results were as follows:

TCP	ACP	RWMA	EF	HTN	Age	DM	NCP	Tinv	FBS

*** Therefore, the results of FSelector and FSelectorRccp were identical*

3. Used the '**Recursive Feature Elimination**' function, as part of the 'Caret' package in R.

This function ranks the features using one of a number of pre-defined functions. I used the 'rfFuncs' (random forests) function, given that we are working with 'non-linear' models. As such, it is another entropy-based filter method. There was significant improvement in the accuracy of the results when choosing between 4 or 8 variable subgroups, but negligible improvement between 8 and 16 – so 8 features were chosen. The results were as follows:

TCP	EF	RWMA	NCP	ACP	HTN	Age	DM

Given these results, and that the top 8 attributes were the same for all three methods, the following final list of attributes were selected (in no particular order):

1. Typical (classic) Angina Chest Pain (**TCP**): Which is defined as [1] substernal pain (behind the sternum bone) [2] provoked by exertion or emotional stress – and [3] relieved by rest or nitroglycerine (or both). *<u>Note</u>: 'angina' is chest pain caused by reduced blood flow to the heart.*
2. Atypical (probable) Angina Chest Pain (**ACP**): Applies when two out of the three criteria of classic angina (TCP) are present.
3. Non-anginal Chest Pain (**NCP**): If only one (or none) of the three criteria of classic angina (TCP) are present.
4. History of Hypertension (**HTN**): Having a history of high blood pressure.
5. Regional Wall Motion Abnormality (**RWMA**): The motion of a region of the heart muscle is abnormal – diagnosed with an echocardiography (ultrasound imaging of the heart).
6. Ejection Fraction (**EF**): A measurement of the percentage of blood leaving the heart each time it contracts.
7. History of Diabetes Mellitus (**DM**): A disease in which the body's ability to produce or respond to the hormone insulin is impaired.
8. **Age**: Men ≤ the age of 45 and/or women ≤ the age of 55 versus men > 45 and/or women > 55.

Random Forest Algorithm Results:

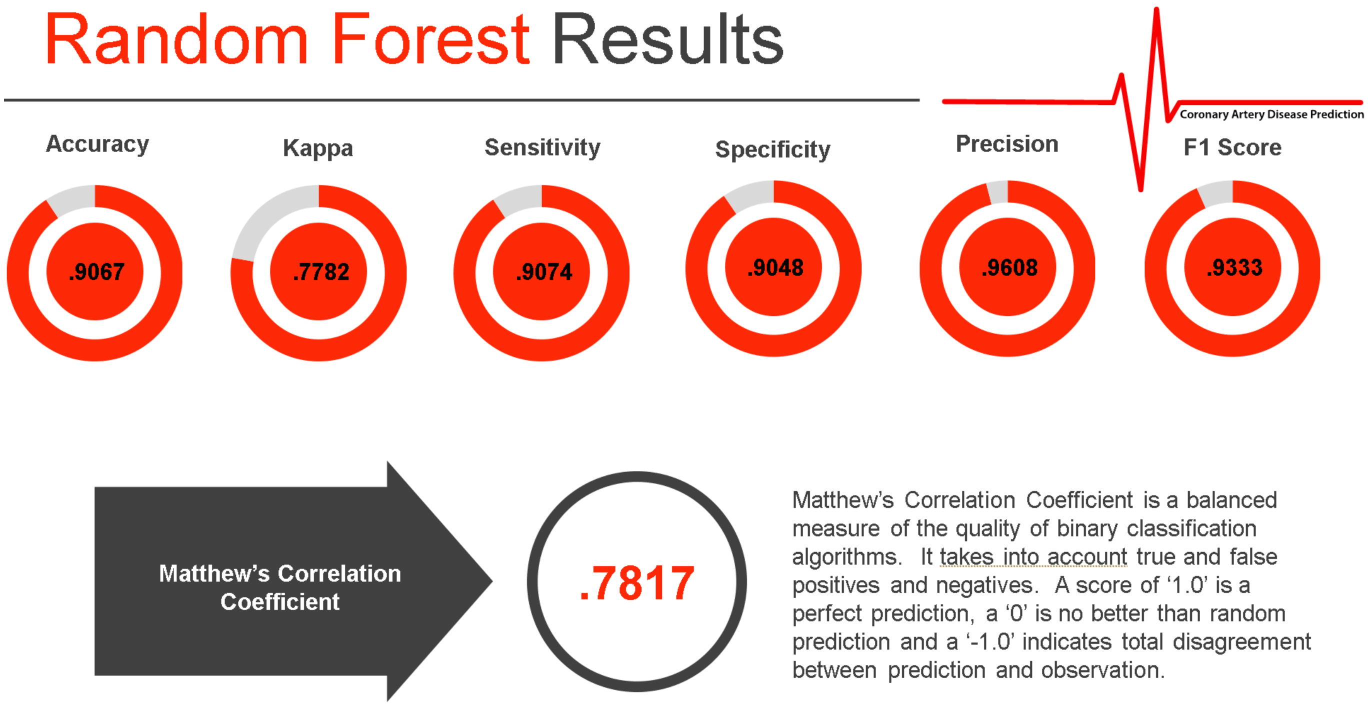
Random Forest Algorithm: Builds multiple 'decision trees' and merges them together to get a more accurate and stable prediction. (Decision Trees are constructed by repeatedly splitting feature space into smaller and smaller subsets – while at the same time a 'tree' (with decision nodes and leaf nodes) is incrementally developed. Decision nodes have two or more branches. Leaf nodes represent a classification or a decision. The topmost decision node corresponds to the best predictor, called the 'root node'.

Random Forest Algorithm - Confusion Matrix Results

		Actual		
		CAD	Normal	Row Total
Predicted	Normal	FN 5	TN 19	24
	CAD	TP 49	FP 2	51
	Column Total	54	21	75

	Results Using The Displayed Test Run	Final **Mean** Results Running Algorithm Tests with 5, 10, 15, 20, 25, 30 & 35 Folds)
Accuracy	0.9067	0.9028572
Kappa	0.7782	0.7691802
Sensitivity (or Recall)	0.9074	0.9047619
Specificity	0.9048	0.8979592
Precision	0.9608	0.9582287
F1 Score	0.9333	0.9305728
Matthew's Correlation Coefficient	**0.781740107**	

Diagram 6



Matthew's Correlation Coefficient is a balanced measure of the quality of binary classification algorithms. It takes into account true and false positives and negatives. A score of '1.0' is a perfect prediction, a '0' is no better than random prediction and a '-1.0' indicates total disagreement between prediction and observation.

Naïve Bayes Algorithm Results:

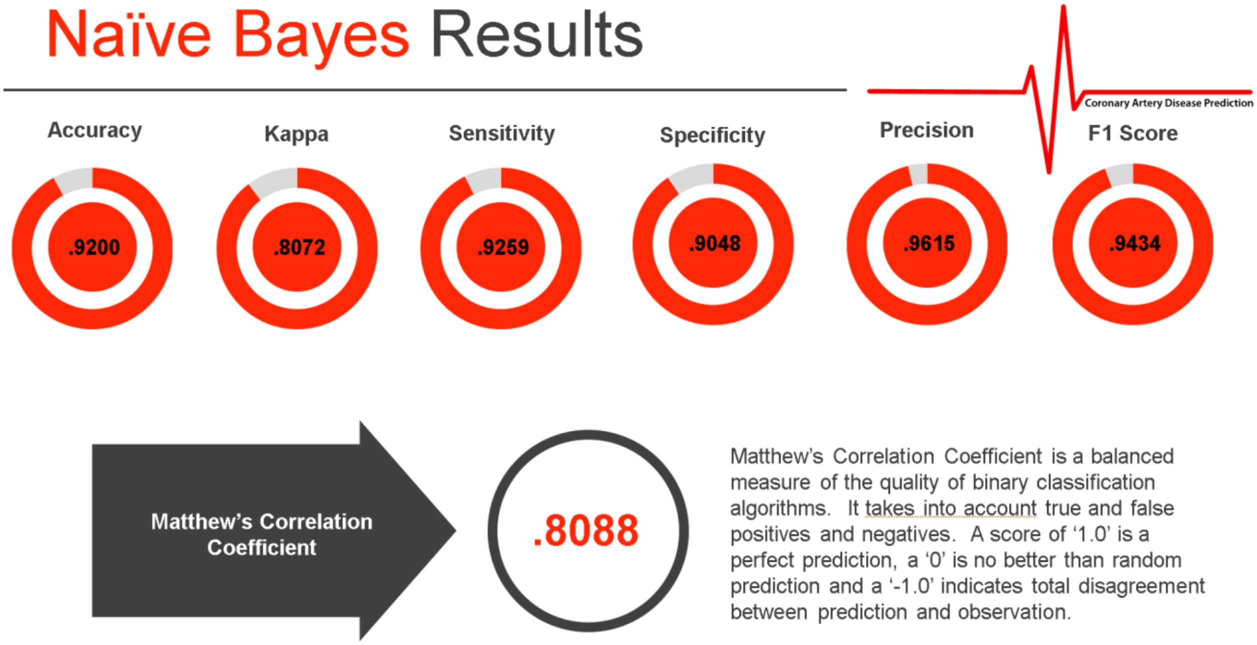
Naïve Bayes Algorithm: A classification algorithm based on 'Bayes Theorem' (calculating the probability of an event based on its association with another event). It considers that each of the independent features contribute independently to the probability that the patient has CAD or not, regardless of any correlations between features. This is why it is labelled "naïve".

Naïve Bayes Algorithm - Confusion Matrix Results

		Actual		
		CAD	Normal	Row Total
Predicted	Normal	FN 4	TN 19	23
	CAD	TP 50	FP 2	52
	Column Total	54	21	75

	Results Using The Displayed Test Run	Final **Mean** Results Running Algorithm Tests with 5, 10, 15, 20, 25, 30 & 35 Folds)
Accuracy	0.9200	0.9180952
Kappa	0.8072	0.8030554
Sensitivity (or Recall)	0.9259	0.9232804
Specificity	0.9048	0.9047619
Precision	0.9615	0.9614308
F1 Score	0.9434	0.9419586
Matthew's Correlation Coefficient	**0.808870107**	

Diagram 7



Matthew's Correlation Coefficient is a balanced measure of the quality of binary classification algorithms. It takes into account true and false positives and negatives. A score of '1.0' is a perfect prediction, a '0' is no better than random prediction and a '-1.0' indicates total disagreement between prediction and observation.

Logistic Regression Algorithm Results:

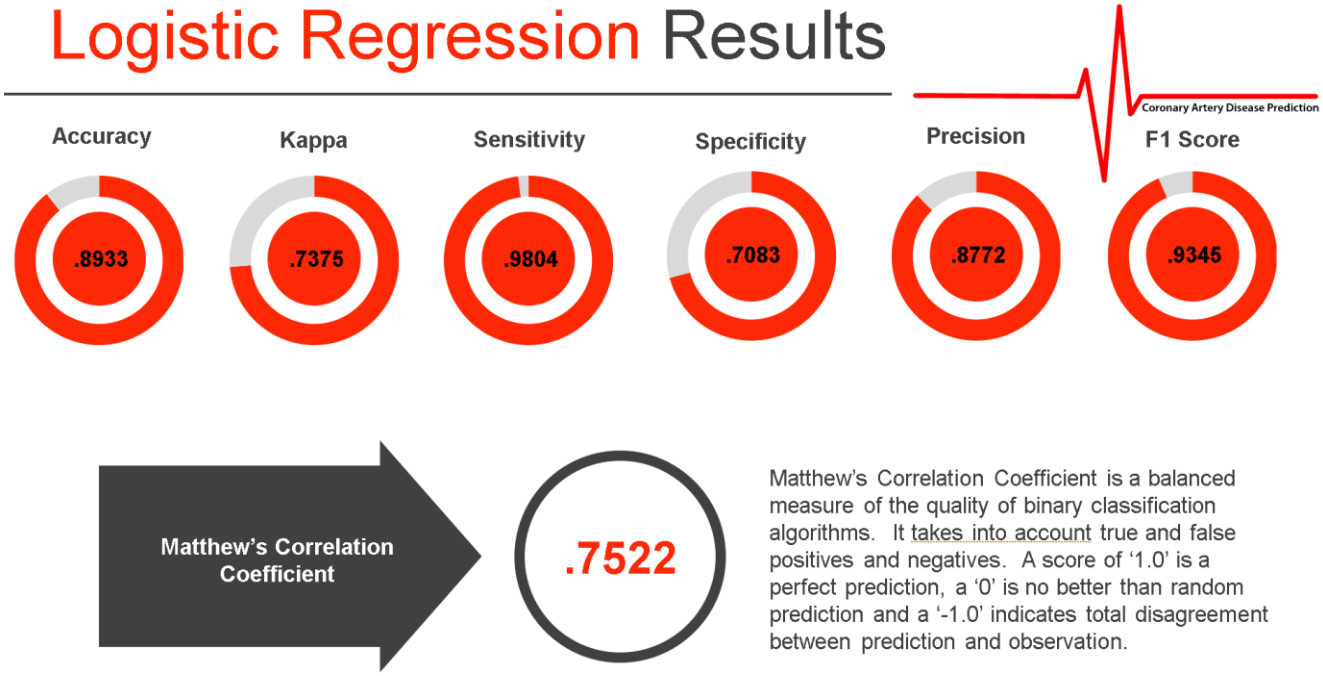
Logistic Regression Algorithm: Estimates the probability of an event occurring given some previous data. It works with binary data (where either an even happens or does not). While 'linear regression' tries to predict data by finding a linear – straight line – equation to predict future data points, logistic regression uses the natural logarithmic function to find the relationship between the variables and uses test data to find the coefficients. The function can then predict future results using these coefficients on the logistic equation.

Logistic Regression Algorithm - Confusion Matrix Results

		Actual		
		CAD	Normal	Row Total
Predicted	Normal	FN 1	TN 17	18
	CAD	TP 50	FP 7	57
	Column Total	51	24	75

	Results Using The Displayed Test Run	Final **Mean** Results Running Algorithm Tests with 5, 10, 15, 20, 25, 30 & 35 Folds)
Accuracy	0.8933	0.9009524
Kappa	0.7375	0.7329369
Sensitivity (or Recall)	0.9804	0.9814815
Specificity	0.7083	0.6938775
Precision	0.8772	0.8919588
F1 Score	0.9259	0.9345470
Matthew's Correlation Coefficient	**0.752251713**	

Diagram 8

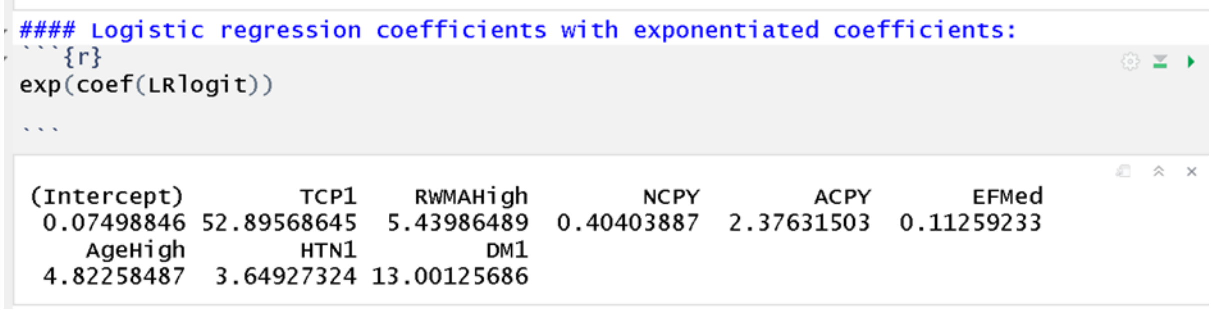


Logistic Regression Results

Matthew's Correlation Coefficient is a balanced measure of the quality of binary classification algorithms. It takes into account true and false positives and negatives. A score of '1.0' is a perfect prediction, a '0' is no better than random prediction and a '-1.0' indicates total disagreement between prediction and observation.

Note: The **Brier Score** was also calculated for the Logistic Regression algorithm results. The Brier Score indicates how far away from the prediction was. The best possible Brier Score is ‘0’ – for total accuracy, and the worst possible score is ‘1’ (wholly inaccurate). The Brier Score result for the logistic regression algorithm was 0.091.

Interpretations of the Impact of the 8 Attributes:



~~~
#### Logistic regression coefficients with exponentiated coefficients:
```{r}
exp(coef(LRlogit))

```
~~~

```
 (Intercept)        TCP1    RWMAHigh        NCPY        ACPY       EFMed 
  0.07498846 52.89568645  5.43986489  0.40403887  2.37631503  0.11259233 
     AgeHigh        HTN1         DM1 
  4.82258487  3.64927324 13.00125686 
```

Diagram 9

To provide some assistance with the interpretation of the impact of the 8 predictor variables – we can refer to the ‘exponentiated’ coefficients provided by the logistic regression algorithm (see diagram 9). You can also extrapolate these variable impacts in the same positive or negative (increasing or decreasing) directions for the other algorithms.

1. Variables which decrease the likelihood of being diagnosed with Coronary Artery Disease (CAD):

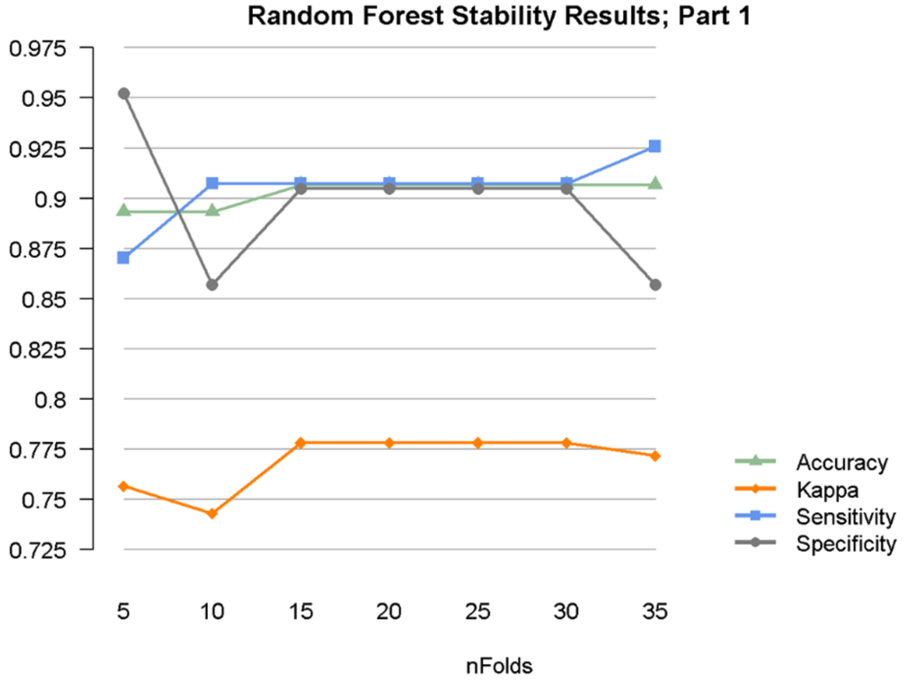
Referring to diagram 9, when there is a one unit increase in ‘Ejection Fraction (medium)’ (EFMed - versus ‘low’ which is not normal) the odds for CAD is approximately 11% lower (since the logistic regression coefficient was -2.1840). Likewise, the odds of CAD when there is a one unit increase in Non-Anginal Chest Pain (NCPY) is approximately 40% lower than when NCP is not present (exponent of -.9062 coefficient).

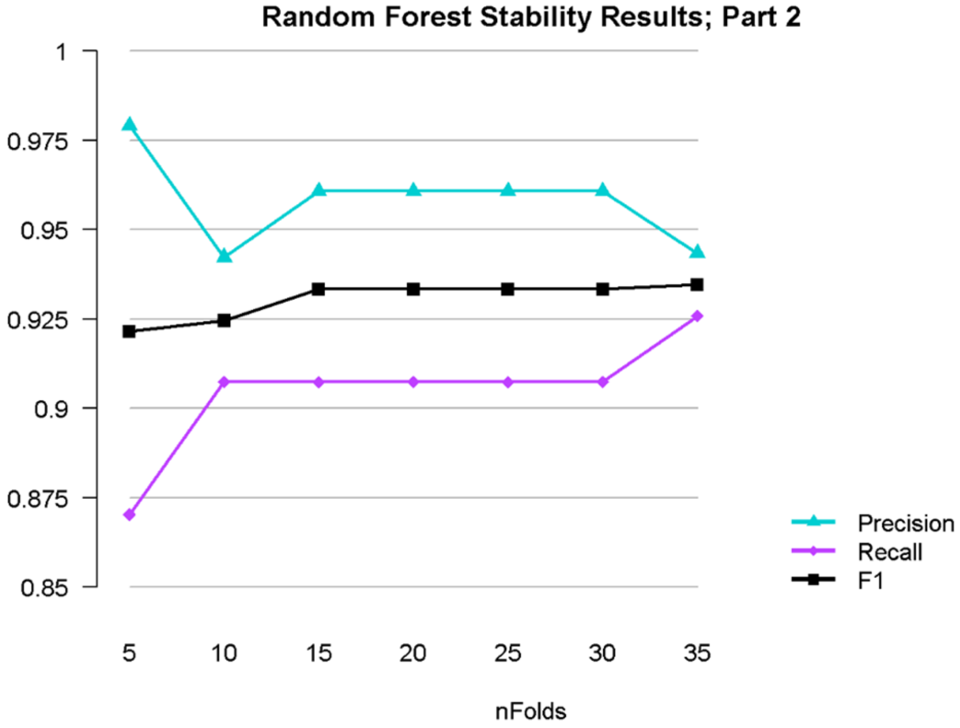
2. Variables which increase the likelihood of being diagnosed with Coronary Artery Disease (CAD):

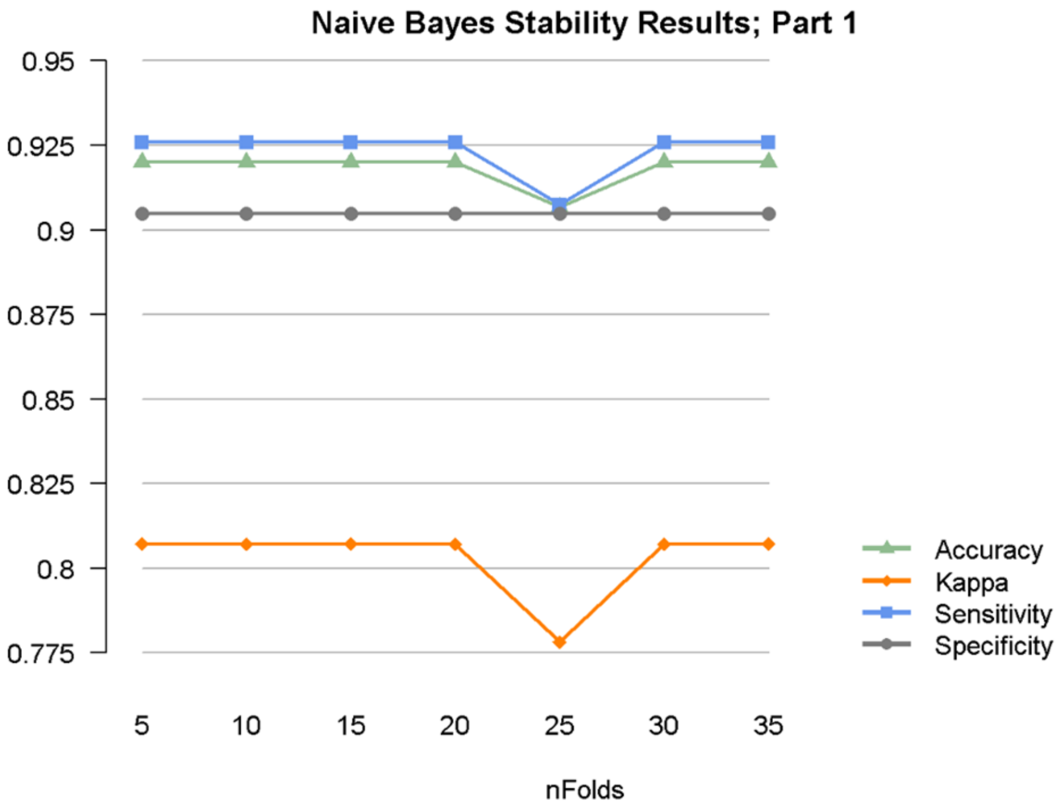
The odds of having CAD increases by a factor of 52 times when there is a one unit increase in typical angina chest pain (TCP1) versus if there is no TCP present; increases a factor of 2.37 times for a one unit increase in Atypical Chest Pain (ACPY) being present, versus not being present; increases by a factor of 5.4 times for a one unit increase in RWMA being ‘high’ (versus not being present); increases by a factor of 4.8 times for a one unit increase in Age(high) versus Age ‘medium’; increases by a factor of 3.6 times with a one unit increase in Hypertension (HTN1) versus not being present; and increases by a factor of 13 times with a one unit increase in diabetes (DM1), versus not being present.

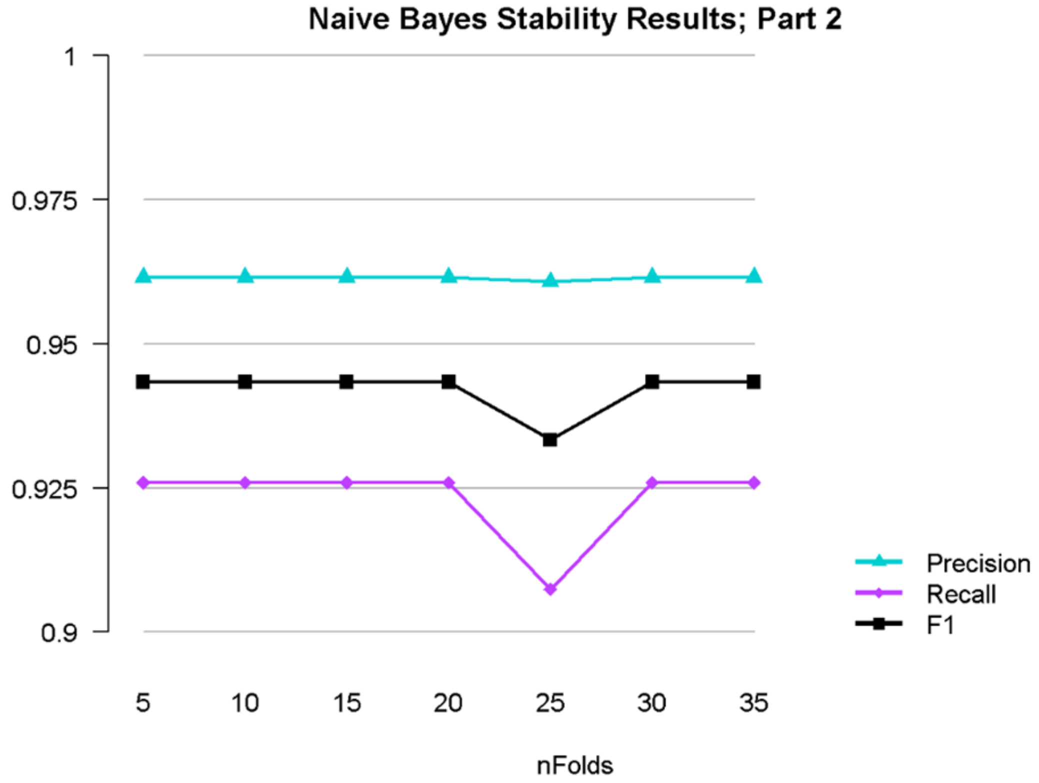
Stability of the Algorithms:

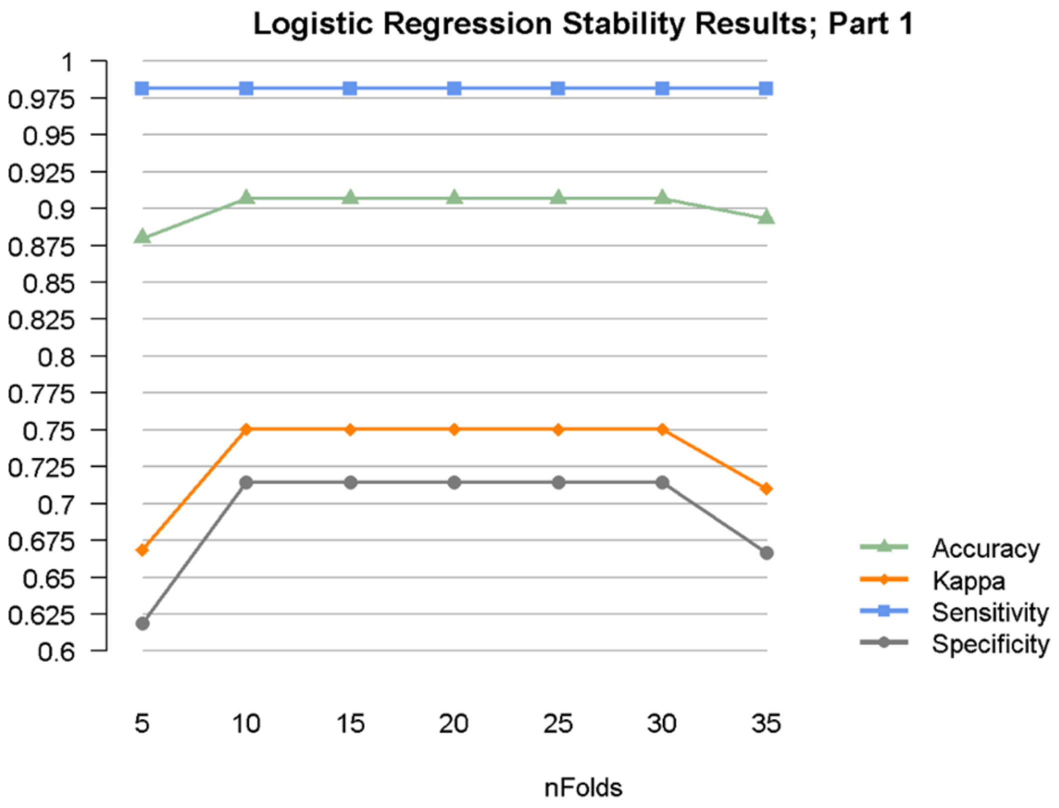
These algorithms were tested using different numbers of folds during the cross validation while training the models. They were tested using 5, 10, 15, 20, 25, 30 and 35 folds. All three of the algorithm's results were found to be stable/repeatable. There is no evidence of instability and/or 'overfitting'.

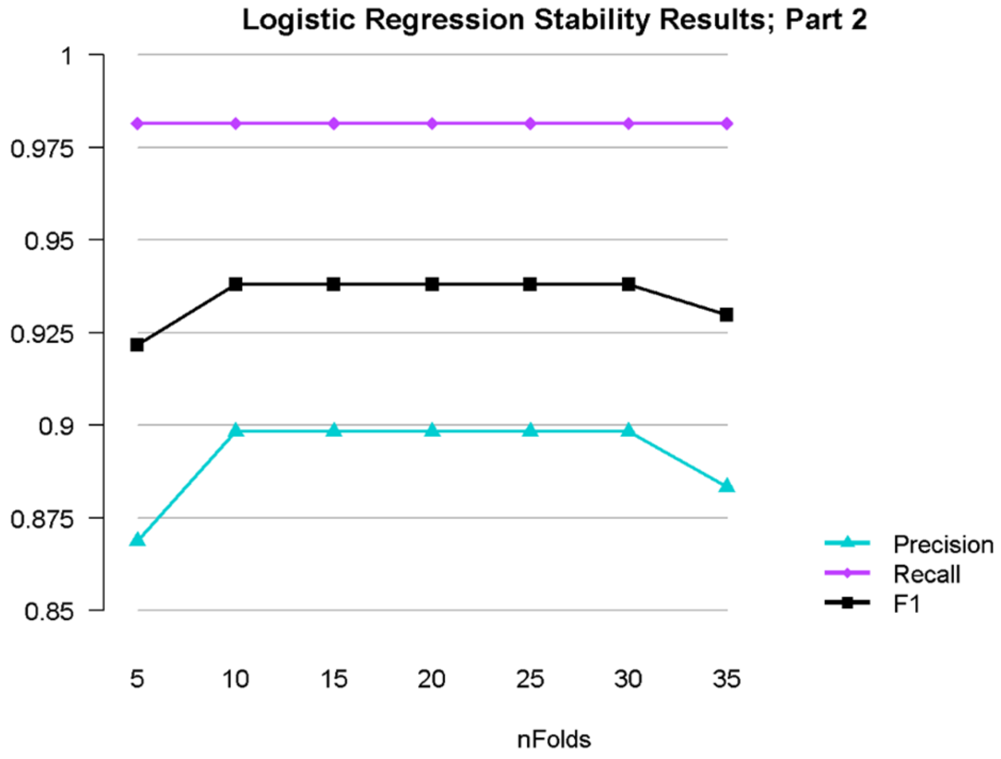





Naive Bayes Stability Results; Part 1
0.95
0.925
0.9
0.875
0.85
0.825
0.8
0.775
5
10
15
20
25
30
35
nFolds
Accuracy
Kappa
Sensitivity
Specificity

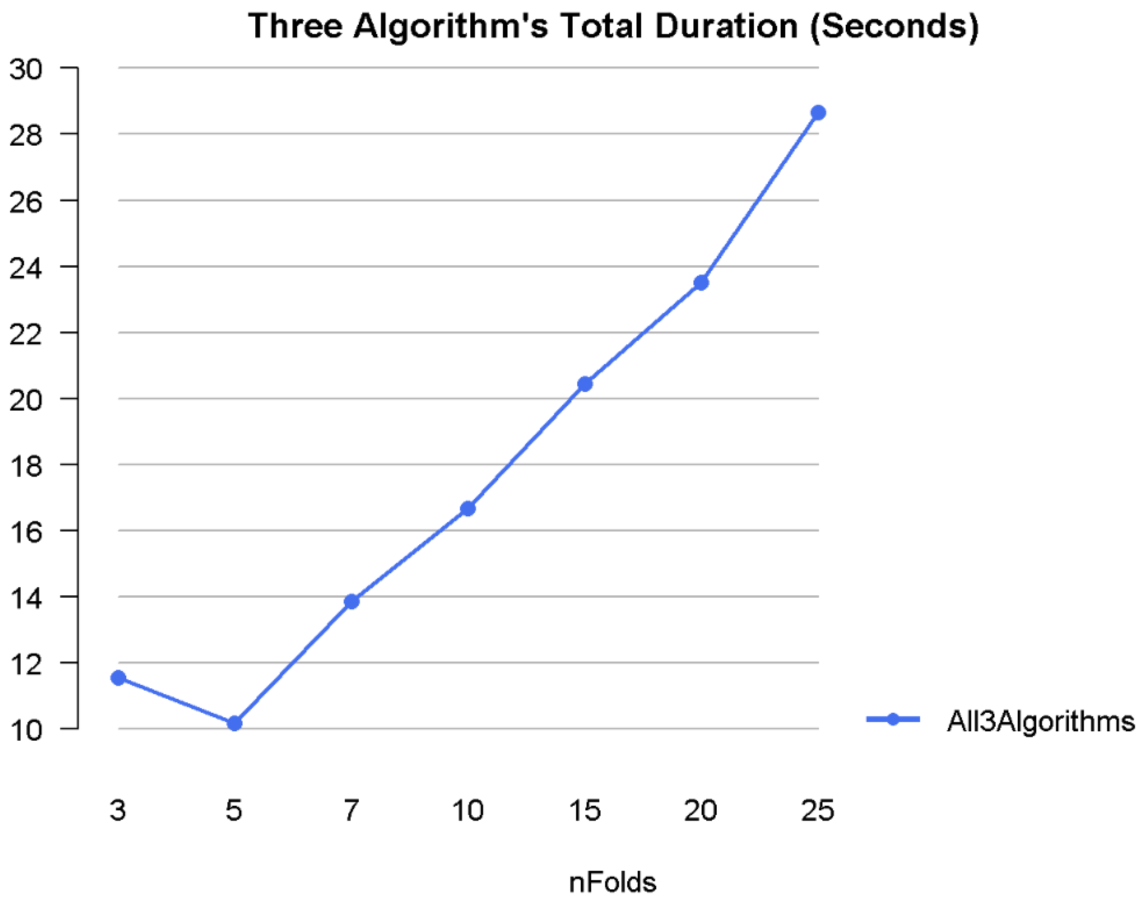

Naive Bayes Stability Results; Part 2
1
0.975
0.95
0.925
0.9
5
10
15
20
25
30
35
nFolds
Precision
Recall
F1


Logistic Regression Stability Results; Part 1
1
0.975
0.95
0.925
0.9
0.875
0.85
0.825
0.8
0.775
0.75
0.725
0.7
0.675
0.65
0.625
0.6
5
10
15
20
25
30
35
nFolds
Accuracy
Kappa
Sensitivity
Specificity


Logistic Regression Stability Results; Part 2
1
0.975
0.95
0.925
0.9
0.875
0.85
5
10
15
20
25
30
35
nFolds
Precision
Recall
F1

Timing Duration:

The algorithm timing was captured for all three algorithms combined. Timing started when the model training begins and ends when the prediction is made (and evaluation measurements are created).



Conclusions:

The 8 features that were selected to be used with all three algorithms were sufficient to produce accurate predictions of Coronary Artery Disease (CAD). The final 8 features were as follows: Typical Angina Chest Pain (TCP), Atypical Angina Chest Pain (ACP), Non-anginal Chest Pain (NCP), history of Hypertension (HTN), Regional Wall Motion Abnormality (RWMA), Ejection Fraction (EF), history of Diabetes (DM), and Age.

Although all three algorithms were capable of predicting Coronary Artery Disease using these 8 attributes, **Naïve Bayes** provided the best results. Several evaluation measures, and in particular, the Matthew's Correlation Coefficient, had results that were slightly better than the Random Forest and Logistic Regression algorithms. A possible reason for the Naïve Bayes algorithm doing better than the others is due to the fact that this algorithm assumes that all of the variables are completely independent. As it turns out, this is not far from the truth.

The objective of this study was to determine whether classification algorithms can be a viable alternative to angiography in order to get a less expensive and less invasive method of predicting Coronary Artery Disease. Rather than depending on angiography – or 54 attributes (including 17 demographic details, 13 examination results, an electrocardiogram test, 14 lab results from blood tests and an echocardiogram test) – good results were achieved using 8 variables. That is, 6 verbal questions and 1 echocardiogram test.

References

[1] UC Irvine Machine Learning Repository. "Heart Disease Data Set", retrieved 05/06/2019 from https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[2] Canada.ca. "Report from the Canadian Chronic Disease Surveillance System: Heart Disease in Canada, 2018", retrieved on 06/01/2019 from https://www.canada.ca/content/dam/phac-aspc/documents/services/publications/diseases-conditions/report-heart-disease-canada-2018/pub1-eng.pdf

[3] Healthline.com. "What Are The 12 Leading Causes of Death in the United States?" retrieved on 06/01/2019 from https://www.healthline.com/health/leading-causes-of-death.

cdc.gov. "Prevalence of Uncontrolled Risk Factors for Cardiovascular Disease: Unites States, 1999 – 2010", retrieved on 06/01/2019 from https://www.cdc.gov/nchs/data/databriefs/db103.pdf.

[4] Academicoup.com. British Medical Bulletin. "Global and Regional Causes of Death", retrieved on 06/01/2019 from https://academic.oup.com/bmb/article/92/1/7/332071.

[5] Mayoclinic.org. "Coronary Artery Disease", retrieved on 06/01/2019 from https://www.mayoclinic.org/diseases-conditions/coronary-artery-disease/diagnosis-treatment/drc-20350619.

[6] Imagingpathways.health.wa.gov.au. "Information for Consumers – Angiography (Angiogram)", retrieved on 06/01/2019 from http://www.imagingpathways.health.wa.gov.au/index.php/consumer-info/imaging-procedures/angiography.

[7] Mayoclinic.org. "Coronary Artery Disease", retrieved on 06/01/2019 from https://www.mayoclinic.org/diseases-conditions/coronary-artery-disease/diagnosis-treatment/drc-20350619.

[8] Media.neliti.com. "Prediction of Heart Disease Using Machine Learning Algorithms", retrieved on 06/13/2019 from https://media.neliti.com/media/publications/239484-prediction-of-heart-disease-using-machin-4b2e96d4.pdf .

[9] Ijiet.com. "A Comprehensive Approach to Predict Heart Disease Using Data Mining", retrieved on 06/03/2019 from http://ijiet.com/wp-content/uploads/2017/05/3.pdf.

[10] UC Irvine Machine Learning Repository. "Heart Disease Data Set – Cleveland Database", retrieved 06/03/2019 from https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[11] Pdfs.semanticscholar.org. "A Novel Approach for Heart Disease Diagnosis Using Data Mining and Fuzzy Logic", retrieved on 06/03/2019 from https://pdfs.semanticscholar.org/cd47/47f0ebabb741c3b30b83b28b71b3aeaf7b04.pdf.

[12] Ijiet.com. "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients", retrieved on 06/13/2019 from http://www.ijiet.org/papers/114-K0009.pdf

[13] Computersciencejournal.org. "Cardiovascular Disease Prediction Using Data Mining Techniques: A Review", retrieved on 06/10/2019 from http://www.computerscijournal.org/vol10no2/cardiovascular-disease-prediction-using-data-mining-techniques-a-review/ .

[14] Towardsdatascience.com. "Comparative Study on Classic Machine Learning Algorithms", retrieved on 06/03/2019 from https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222.

[15] Classeval.wordpress.com. "Basic Evaluation Measures From The Confusion Matrix", retrieved on 06/03/2019 from https://classeval.wordpress.com/introduction/basic-evaluation-measures/.

[16] Ncbi.nlm.nih.gov. "Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models", retrieved on 06/10/2019 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5394463/

[17] R. Alizadehsani et al., "A data mining approach for diagnosis of coronary artery disease," Computer Methods and Programs in Biomedicine, vol.111, no.1, pp.52-61, Jul. 2013.

[18] R. Alizadehsani, M.H. Zangooei, M.J. Hosseini, J. Habibi, A. Khosravi, M. Roshanzamir, F. Khozeimeh, N. Sarrafzadegan, S. Nahavandi, Coronary artery disease detection using computational intelligence methods, Knowledge-Based Systems, 109 (2016) 187-197.

[19] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," Computer Methods and Programs in Biomedicine, vol. 141, pp. 19-26, 2017/04/01/ 2017.

[20] Chf-solutions.com. "Heart Failure Classifications", retrieved on 06/02/2019 from https://www.chf-solutions.com/heart-failure-classifications/ .

[21] Healthline.com. "Everything you should know about Lymphocytes", retrieved on 06/21/2019 from https://www.healthline.com/health/lymphocytes#results .

[22] Healthline.com. “Understanding Neutrophils: Function, Counts and More”, retrieved on 06/21/2019 from https://www.healthline.com/health/neutrophils .

[23] ncbi.nlm.nih.gov. “Neutrophil Lymphocyte Ratio and Cardiovascular Disease Risk: A Systematic Review and Meta-Analysis”, retrieved on 06/26/2019 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6252240/ .

[24] Healthline.com. “Blood Differential Test”, retrieved on 06/26/2019 from https://www.healthline.com/health/blood-differential#test-results .

[25] ncbi.nlm.nih.gov. “What is the normal value of the neutrophil-to-lymphocyte ratio?”, retrieved on 06/26/2019 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5217256/ .

[26] ncbi.nlm.nih.gov. “The value of echocardiographic regional wall motion abnormalities in detecting coronary artery disease in patients with or without a dilated left ventricle”, retrieved on 06/29/2019 from https://www.ncbi.nlm.nih.gov/pubmed/3157304 .

[27] Mayoclinic.org. “Ejection fraction: What does it measure?”, retrieved on 06/29/2019 from https://www.mayoclinic.org/ejection-fraction/expert-answers/faq-20058286 .

[28] Researchgate.net. “Survey of Machine Learning Algorithms for Disease Diagnostic”, retrieved on 07/01/2019 from https://www.researchgate.net/publication/312629315_Survey_of_Machine_Learning_Algorithms_for_Disease_Diagnostic .

[29] Stefan Lessmann et al., “Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update”, European Journal of Operational Research (doi:10.1016/j.ejor.2015.05.030)

[30] Chakkrit Tantithamthavorn, “The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models”, retrieved on 07/15/2019 from https://arxiv.org/pdf/1801.10269.pdf .