**Ryerson University**

**CKME 136 Data Analytics Capstone Course**

Submitted By: Todd Bethell

**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 K-Nearest Neighbour (kNN) 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 K-Nearest Neighbour (kNN) and the Logistic Regression algorithms. All three of these machine learning algorithms are 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 and specificity. 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.

**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 |
| Body Mass Index (kg/m²) | BMI | 18 – 41 |
| 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 using the quartiles of 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/*](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 & ECHO 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

*Note: 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 the quartiles of these attributes since they were not discretized in the original dataset.*

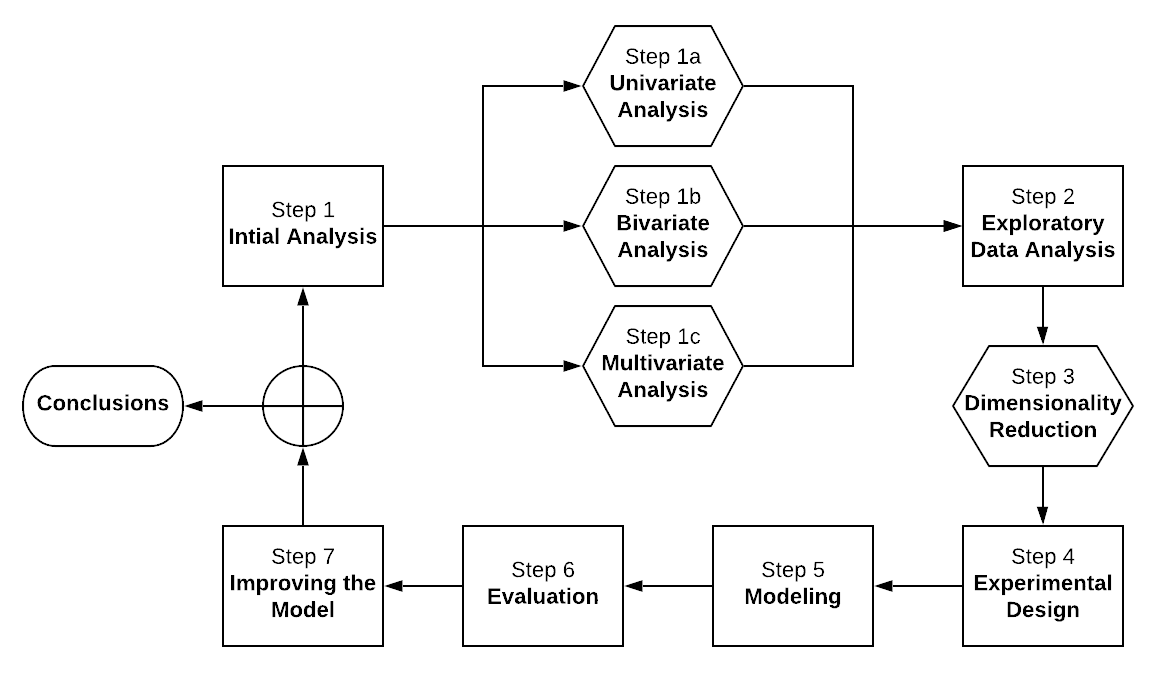
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Discretized Features and Their Range of Values** | | | |  |
|  | **Feature Name** | **Low** | **Medium** | **High** |  |
|  |  |  |  |  |  |
|  | Fasting Blood Sugar | FBS < 70 | 70 ≤ FBS ≤ 105 | FBS > 105 |  |
|  | Creatine | Cr < 0.7 | 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 |  |
|  | \* 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**



**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
* 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.
* Remove dependent attribute from data set in R – 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. \*\*
* 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. Although the algorithms used later are appropriate for non-parametric data, it is still useful to have an understanding of the distributions for interpretation of results later.
* Check for any ‘very low’ variance for individual attributes. This will help determine which attributes can be removed later (if any).
* 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 evaluated.
* Create visualizations of the attributes within each of the four attribute groups. Keep visualizations that appear to display something ‘interesting’ about the dataset. This will be updated/improved as I learn more about the relationships of the attributes.

**Step 1b: Bivariate Analysis**

* Investigate the pairwise relations between the input variables as well as between the input variables and output variable. Create scatter plots to visualize these relations.
* Complete correlation analysis and check the significance of relationships between attributes. As appropriate, use Pearson Correlation Coefficient for two quantitative variables; Chi-square for two categorical variables; and ANOVA for one categorical and one quantitative variable.

**Step 1c: Multivariate Analysis**

* Will consider the One-Way-ANOVA for the parametric attributes and the Kruskal-Wallis test for the non-parametric attributes.

**Step 2: Exploratory Data Analysis**

* Determine which attributes should be normalized, if any. 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, subset those patients who have been diagnosed with CAD versus those that do not. Subset further between men and women. Save these dataframes 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. In particular, Forward Selection and/or Backward Elimination will be used for. The feature selection will be completed within each of the four attribute groups (not by considering the entire dataset at once).
* Investigate whether PCA can be an option to use for dimensionality reduction – given that most of these attributes will ultimately be factor data types. Determine whether the ‘non-linear transformation’ of each variable can be done effectively for factor (as well as numeric) attributes. Is it a problem given that the attributes will not be normalized first? Supposedly, the R package named ‘Homals’ may be an option to do this?
* Save final dataframes for each of the four attribute groups. Also save a separate dataframe with all four attribute groups combined.

**Step 4: Experimental Design**

* Given that this data set is not very large, the splitting to be used for the training and testing of the algorithms will be the done by using the ‘k-fold cross-validation’ method. Specifically, I will use a 10-fold cross validation.
* 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. Planning to use the ‘SMOTE’ function in R to assist with this problem. SMOTE stands for “Synthetic Minority Over-Sampling Technique” and this function will synthesize new minority instances between the existing (actual) minority instances.

**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 k-Nearest Neighbour, and Logistic Regression algorithms. The effectiveness, efficiency and stability of these three algorithms will be compared.
* These three algorithms will be run on the four attribute groups separately. The results of the algorithm’s effectiveness will be evaluated within each attribute group and between all four attribute groups.
* Run the three algorithms on the ‘combined’ dataframe (that is, with the four final attribute group dataframes together).

**Step 6: Evaluation**

* 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.
* \*\* However, when it comes to the diagnosis of coronary heart disease, being falsely diagnosed with CAD is more ‘tolerable’ than being falsely diagnosed as ‘Normal’. As such, emphasis should be placed on the Sensitivity results for this data set.
* I will also generate the F-score (F1) to help evaluate the results. That is, the F1-Score considers both ‘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).

Precision provides a measure of how many of those patients who were predicted to have CAD, actually have CAD. Precision = True Positive CAD/False Positive CAD + True Negative CAD.

The F1 Score = 2\*(Precision\*Sensitivity)/Precision + Sensitivity.

*Precision, Sensitivity and the F1-Score can be generated with the function “confusionMatrix()” from the ‘Caret’ package in R.*

* For additional evaluation of the logistic regression algorithm, I will use the PREDICT function in R to predict the probabilities CAD positive. The logistic regression model accuracy will be measured as the proportion of observations that have been correctly classified. The classification error will be defined as the proportion of observations that are misclassified.
* I will measure the performance efficiency of the machine language algorithms using the ‘microbenchmark’ package in R.
* Lastly, I will evaluate the stability of the algorithms and confirm whether they generate consistent results.

**Step 7: Improving the Model**

* I will check whether building a machine learning ensemble improves my results. I will do this using the ‘SuperLearner’ R package.
* 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 concern regarding EPV, may want to determine whether bootstrapped sampling creates better results than the 10-fold cross validation for logical regression.
* 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, sensitivity and specificity 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 accuracy, sensitivity and specificity 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’).

**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> .