**FACULTY OF ENGINEERING, ENVIRONMENT & COMPUTING**

7072CEM – MACHINE LEARNING

Machine Learning Algorithms for Solving Real-World Classification and Clustering Problems

**Lucas Oliviera and Martin Tay – 12357262**

MSc Data Science and Computational Intelligence

Coventry University, United Kingdom

**Abstract**

Cardiovascular disease is one of the prevalent causes of death in the world. There are several important risk factors that can lead to the high rate of deaths related to cardiovascular diseases across the globe. The Healthcare industry has a significant role to play in terms of providing timely intervention to save lives of vulnerable patients. The current number of annual cases around the world in terms of statistics is still disturbing and calls for a campaign to improve the well-being of humans with the data constantly generated in countries where the rates are terribly high.

This research paper covers the use of three classifier algorithms to examine the causes of cardiovascular disease from the heart disease indicator dataset publicly made available. Our desire is to use the historical dataset to predict outcomes. The risk factors which are likely to contribute to a cardiovascular condition is what would then become the input variables fixed into the model to make predictions. Results from the three classification models would be compared and appropriate recommendations for implementation would be made to further look at ways to overcome the burden of the disease by predicting the risk of the disease in patients who are susceptible.

The outcome of the experiments after training the three models with the heart disease dataset yielded the respective results %, %, and % models for the respective classification algorithms (Decision Tree, Support Vector Machine and Logistic Regression). Other activities like hyperparameter optimization and feature selection were explored to ensure the best effort on the predicted output variables. These results determined the best model based on the accuracy level after comparing the three models.

**Keywords:** *Decision Tree, Support Vector Machines, Logistic Regression, Confusion Matrix, Heatmap, RFE, CVD, NCD (noncommunicable disease)*

**Introduction**

Cardiovascular disease is one of the prevalent causes of death in the world and accounts for one of the most painful deaths that can be avoided under strict supervision. Cardiovascular diseases cause an average of 17.7 million deaths each year (44% of NCD fatalities) making it one of the most deserving topics for research on prevention. “Cardiovascular diseases (CVD) are a group of disorders of the heart and blood vessels which is the most significant cause of death globally.” (“Secondary Prevention of Cardiovascular Diseases and ... - Hindawi”) Despite the critical fatality rate 90% CVD can be prevented by taking necessary precautions (McGill, 2008). The impact of cardiovascular diseases can be very painful in the early years of any human being and have some ripple effects to the society due to the pain and economic effects on families and the economy of any country.

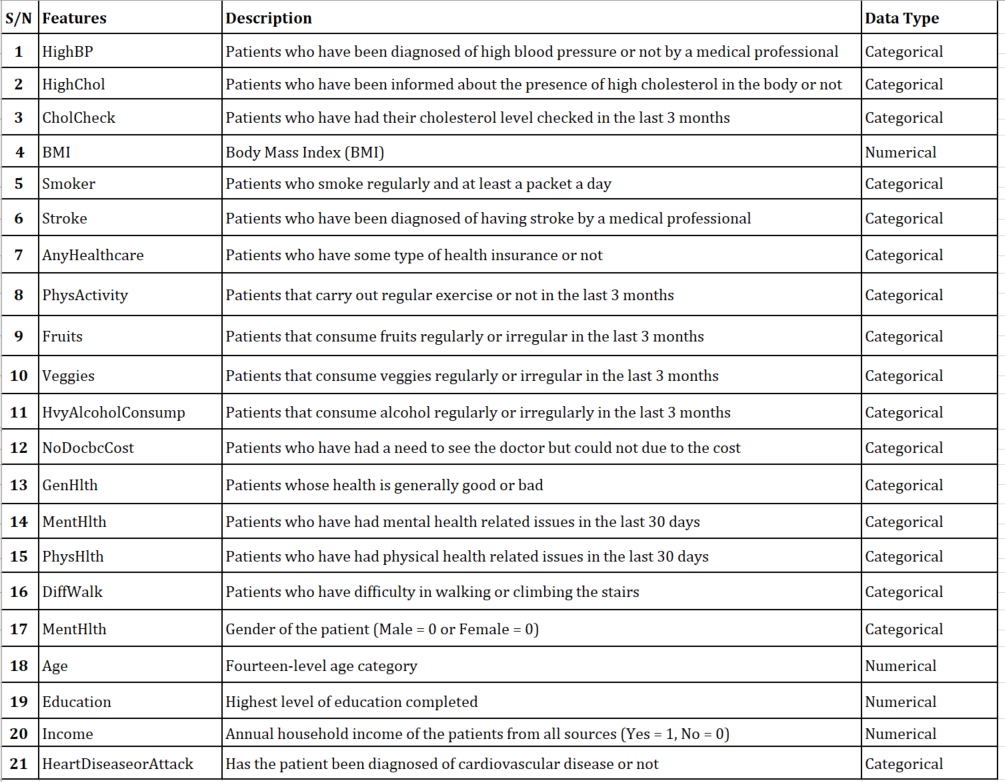
There are several risk factors that contribute to cardiovascular health conditions. It is no news that in most cases, the cause of the disease can be attributed to lifestyle and human preference of diets. Physical activities and healthy eating have been proven to help improve the health and well-being of humanity. The application of machine learning to predict cardiovascular diseases can help in the sensitization and improve the conditions of diagnosed patients. There are several case studies concerning cardiovascular diseases. According to (Karunathilake & Ganegoda, 2018), there are three types of prevention mechanisms to prevent and reduce the impacts of a disease. “Primary prevention refers to the steps taken by an individual to prevent the onset of the disease.” (“What are the primary secondary and tertiary levels of prevention?”) This is achieved by maintaining a healthy lifestyle choice such as diet and exercise. “Secondary prevention focuses on reducing the impact of the disease by early diagnosis prior to any critical and permanent damage.” (“Secondary Prevention of Cardiovascular Diseases and ... - Hindawi”) This facilitates avoiding life threatening situations and long term impairments from a disease. Tertiary prevention is used once long term effects set in, by helping the patients to manage pain, increase life expectancy, and increase the quality of life. The secondary prevention of CVD includes diagnosis and prevention. Most critical step of secondary prevention is early diagnosis which allows medical professionals to provide required care for patients and improve the quality of life. This requires identifying risk factors, criticality of risk factors, and how the variation of these factors relates to CVD. (“Secondary Prevention of Cardiovascular Diseases and ... - Hindawi”) Upon early diagnosis, patients could be directed to required treatments affording a higher quality of life (Karunathilake & Ganegoda, 2018).

The inspiration for this research paper draws strength from the use of machine learning algorithms by medical professionals to make predictions of vulnerable patients. Health improvements is tied to continuous research and implementation of state-of-the-art methodologies explored by experts to diagnose and prevent diseases like the one examined in our case study.

**Problem and Data set(s)** description (where you describe in detail the problem you want to solve and its significance)

As required, the purpose of this task is to make predictions with the independent variables and dependent variables. These independent variables contain certain information that may cause a cardiovascular condition in patients. The causes of cardiovascular disease borders around certain risk factors like blood pressure, obesity, age, sex, diet, exercise, smoking, health insurance, mental health, physical health, consumption of alcohol, rest, or sleep, and record of health check-up, etc. One major issue is the challenge with the imbalanced dataset and a resampling technique would be introduced to solve the problem. This in our case study is linked to what has been tagged as the independent features and they are outlined in the table below;

*Table 1 – Description of Independent and Dependent Features in the dataset*



The prediction of cardiovascular disease is a classification problem being that the outcome of the prediction is a categorical response variable based on certain predictor(multiple) variables, indicating whether a patient is diagnosed with a heart condition or not diagnosed. In machine learning, there are two major learning techniques, supervised learning, and unsupervised learning. Supervised learning algorithms are trained with labelled input and output data, while unsupervised learning algorithms are trained with unlabelled variables (see figure 1 and 2). The problem we have at hand is that of a supervised learning algorithm and both input and output data already labelled would be examined and evaluated when loaded unto the three classification algorithms selected for this experiment.

*Figure 1 and 2 – Graphic display of Supervised Learning and Unsupervised Learning*

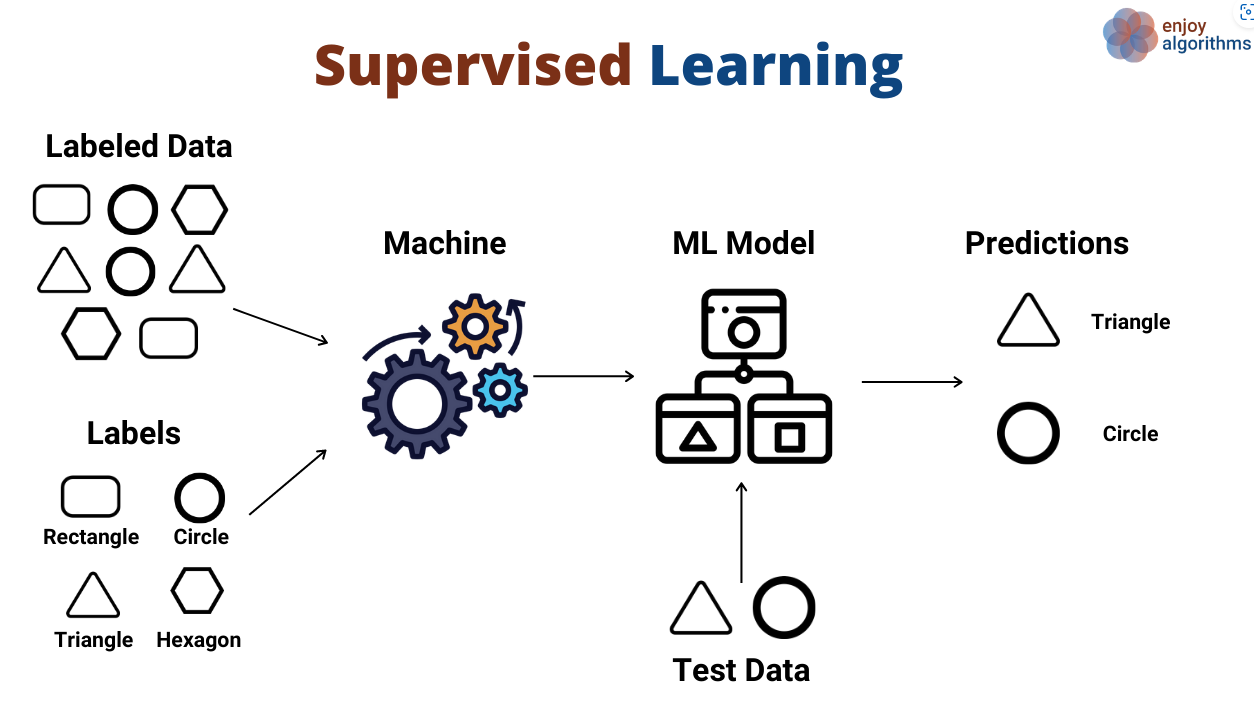
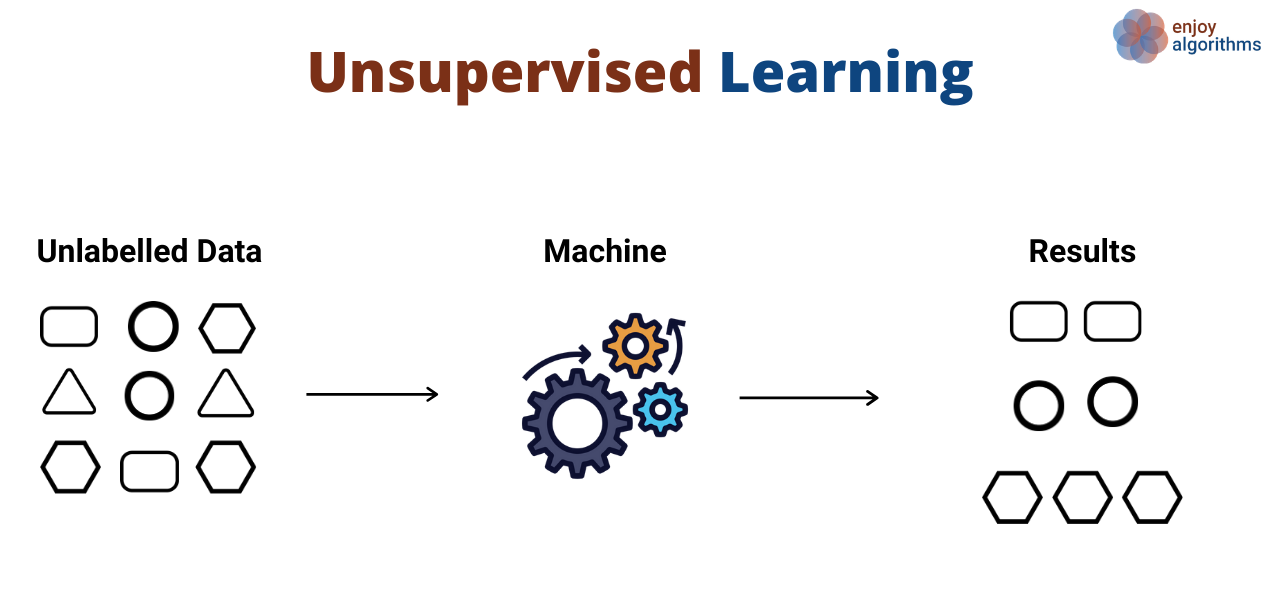
 

Figure 1 Figure 2

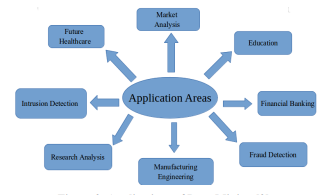
# NOTE: From *Supervised and Unsupervised Learning (an Intuitive Approach)* by Kozan, 2021, (https://medium.com/@metehankozan/supervised-and-unsupervised-learning-an-intuitive-approach-cd8f8f64b644)

As depicted in the diagram above, figure 1 is an illustration of the data is processed before insights are gleaned in supervised learning algorithms. Insights are drawn by from the correlation between the interactive features against the response variable. This is usually an iterative process, which calls for the model to be trained until optimal performance is achieved and the model can make accurate predictions. The process of model optimization is called hyperparameter tuning and would be discussed in detail in the next section. Figure 2 depicts the learning process of unsupervised learning and how it generates insights or patterns for analysis. Learning for unsupervised is done with unlabelled data to reveal insights and patterns in the dataset. Key difference is in how the data is learned and processed before predictions are done.

**Methodologies** (where you shortly describe the machine learning methods and/or other methods employed to solve the problem)

It is important to note that without data mining, machine learning is useless. The process of data mining can be applied to dataset containing human interactions from various industries.

*Figure 3 – Graphical illustration of the application areas of data mining*

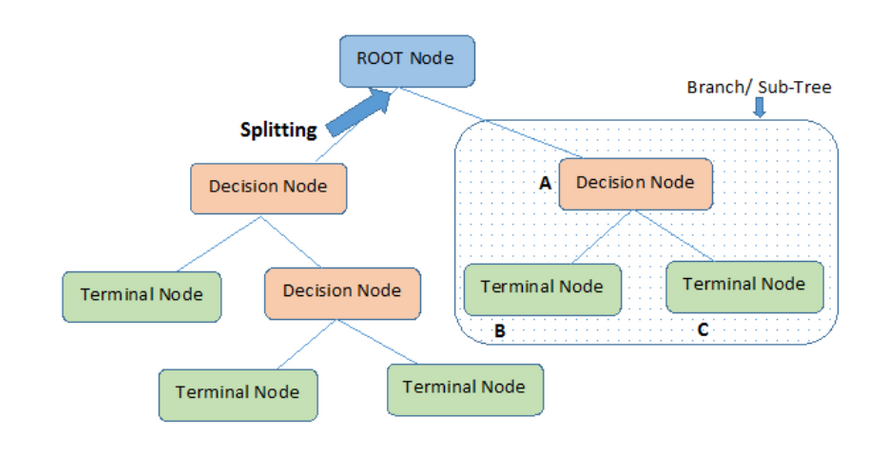


NOTE: Application of Data Mining. From Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis by Dey, 2016, (https://www.researchgate.net/publication/301335834\_Analysis\_of\_Supervised\_Machine\_Learning\_Algorithms\_for\_Heart\_Disease\_Prediction\_with\_Reduced\_Number\_of\_Attributes\_using\_Principal\_Component\_Analysis/link/583d2af708ae502a85e53634/download)

Due to the expected categorical response variable in the prediction of heart disease, the suitable machine learning methods for this would be Decision Tree, Support Vector Machine, and Logistic Regression. In binary classification algorithms, the response variable is either a one (1) or a zero (0), meaning a patient is either diagnosed of cardiovascular disease or never been diagnosed.

Decision Tree Classification algorithm is one of the simplest techniques to implement in solving classification or regression problems. The supervised learning model learns from the class labels/interactive features by making system-based rules that guides its decision to make predictions. According to (Mitchell, 1997), “Decision Tree learning method is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree”. (“Decision Tree Learning (ML 3) Flashcards | Quizlet”)

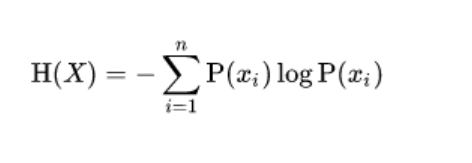
Figure 4 – A diagram of a decision learning process in Decision Tree



NOTE: From Decision Tree Algorithm – A Complete Guide by Saini, 2021

(<https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/>)

The root node is where the learning process begins with several splits into finite terminal nodes before making a prediction. To make decisions before splitting, decision tree uses the amount of information gained about the to improve the nodes. The information gained is measured by the entropy and mathematically expressed in the formula below, “where X is the random variable or process, Xi is the possible outcomes, and p(Xi) is the probability of the possible outcomes” (Wei-Meng, 2019).



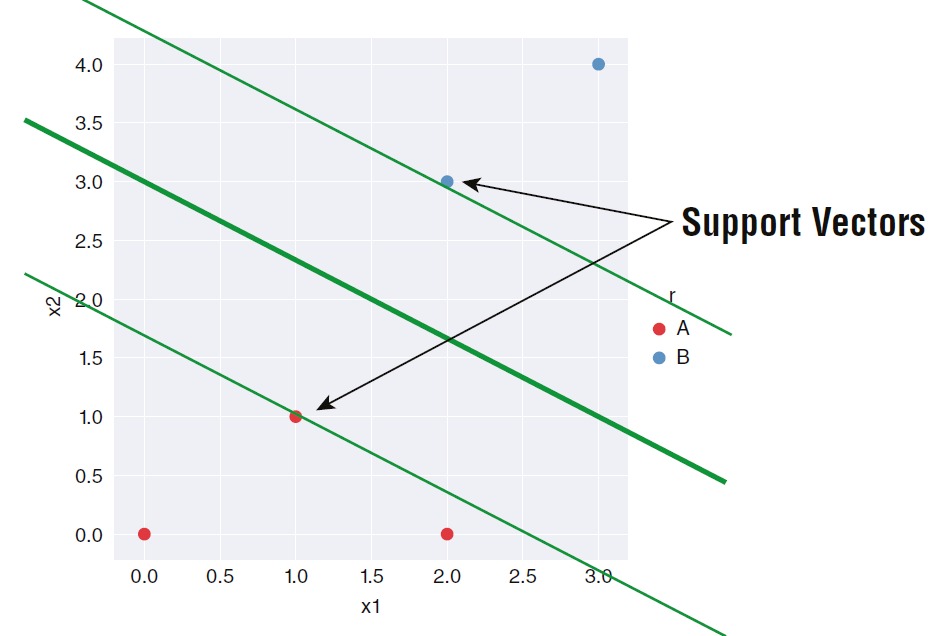
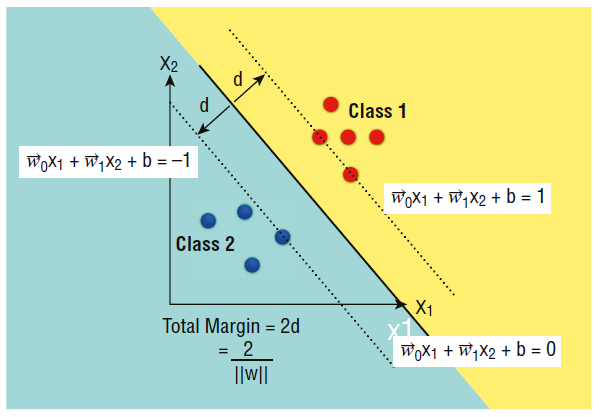
# NOTE: From A Complete Guide to Decision Tree Split using Information Gain, by Verma, 2021.

# (<https://analyticsindiamag.com/a-complete-guide-to-decision-tree-split-using-information-gain/>)

By definition, (Verma, 2021), the [entropy](https://analyticsindiamag.com/the-promise-of-maximum-entropy-reinforcement-learning/) of any random variable or random process is the average level of uncertainty involved in the possible outcome of the variable or process.

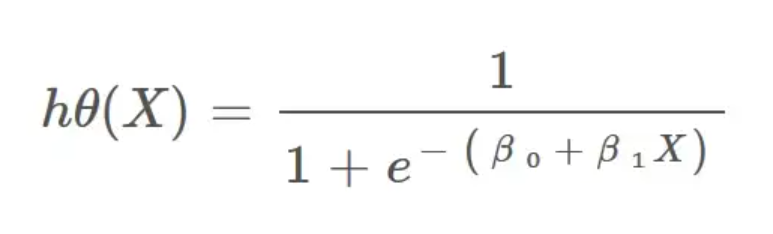
Support vector machine is a classification algorithm which performs its prediction by splitting the classes into two groups with a line mathematically positioned by two support vectors surrounding a hyperplane, typically, in a multidimensional space.

Figure 3 and 4 – Graphical display of how prediction is performed in support vector machine classification



NOTE: From Python Machine Learning, by Wei-Meng, 2019

The third classification algorithm used in this experiment is Logistic Regression. In machine learning, logistic regression uses the concept of probability to make predictions in classification problems. “The hypothesis in logistic regression tends to limit the cost function between 0 and 1” (Pant, 2019). This is mathematically expressed as;

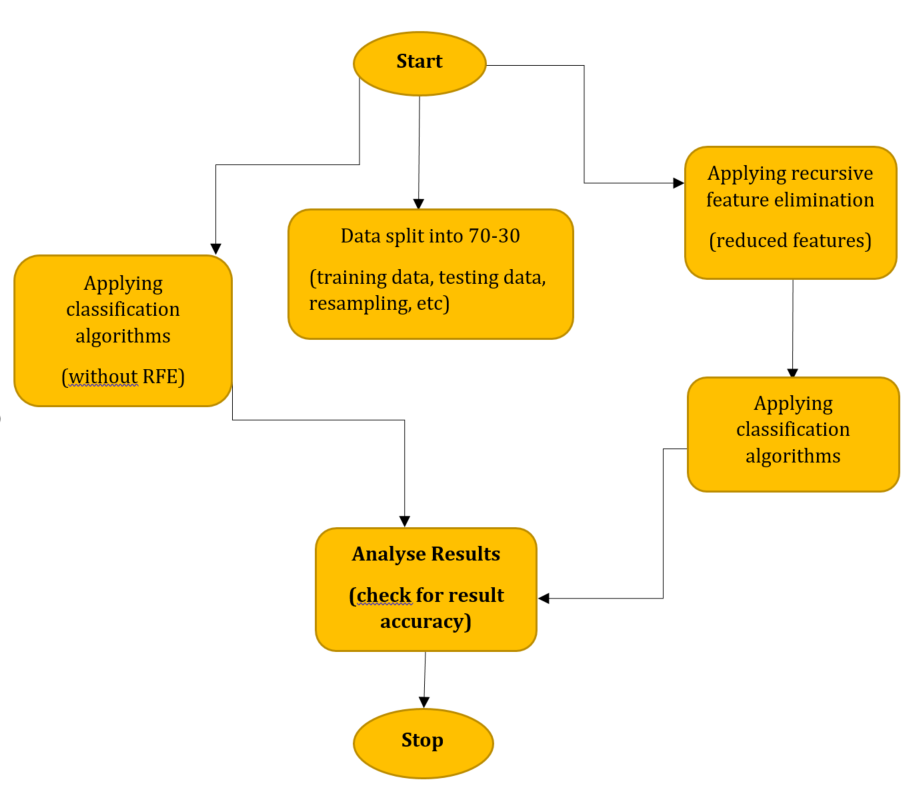


NOTE: The hypothesis of logistic regression. From: Introduction to Logistic Regression, by Pant, 2019. (<https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148>)

**Experimental** **setup** (including data pre-processing, feature selection and extraction, classification/clustering parameters)

Experimenting with data exploration and evaluation of models involves an iterative process. This process determines the uniformity or correctness of the model in the lifecycle of a machine learning workflow. There are five (5) steps in the lifecycle of a machine learning process to bear in mind. These steps include data acquisition, data cleaning, model training, testing of the model and deployment of the model. Data acquisition involves data mining process from various sources. Data cleaning takes care of removing data duplicates, and any outliers that result to incorrect predictions. The model training is one which involves feature selection and hyperparameter tuning before deployment of the model for key decision-making purposes. Our setup is illustrated in the flowchart below, the method adopted to compare the performance of the three (3) classifier algorithms before feature selection and after feature selection.

Figure 5 – Flowchart of experimental process of comparing the models



**Data Pre-processing**

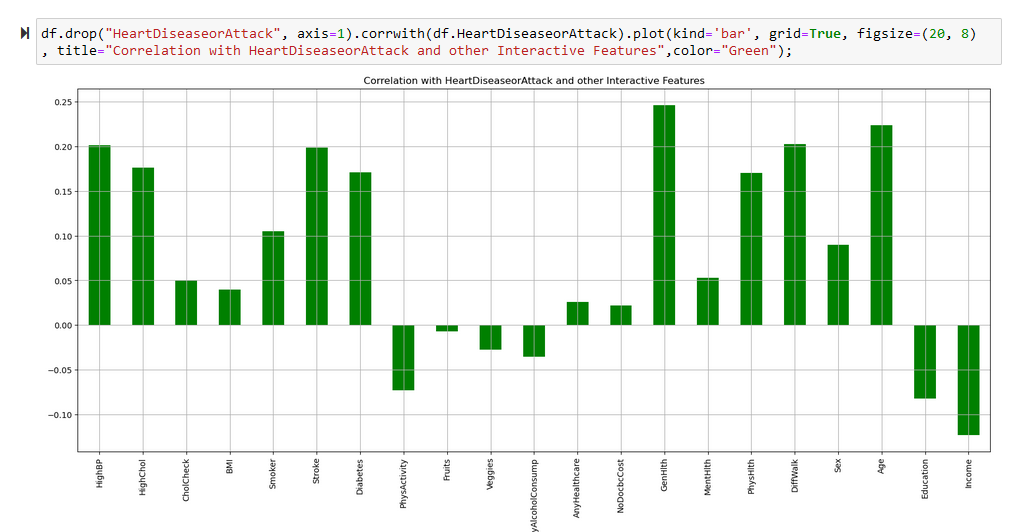
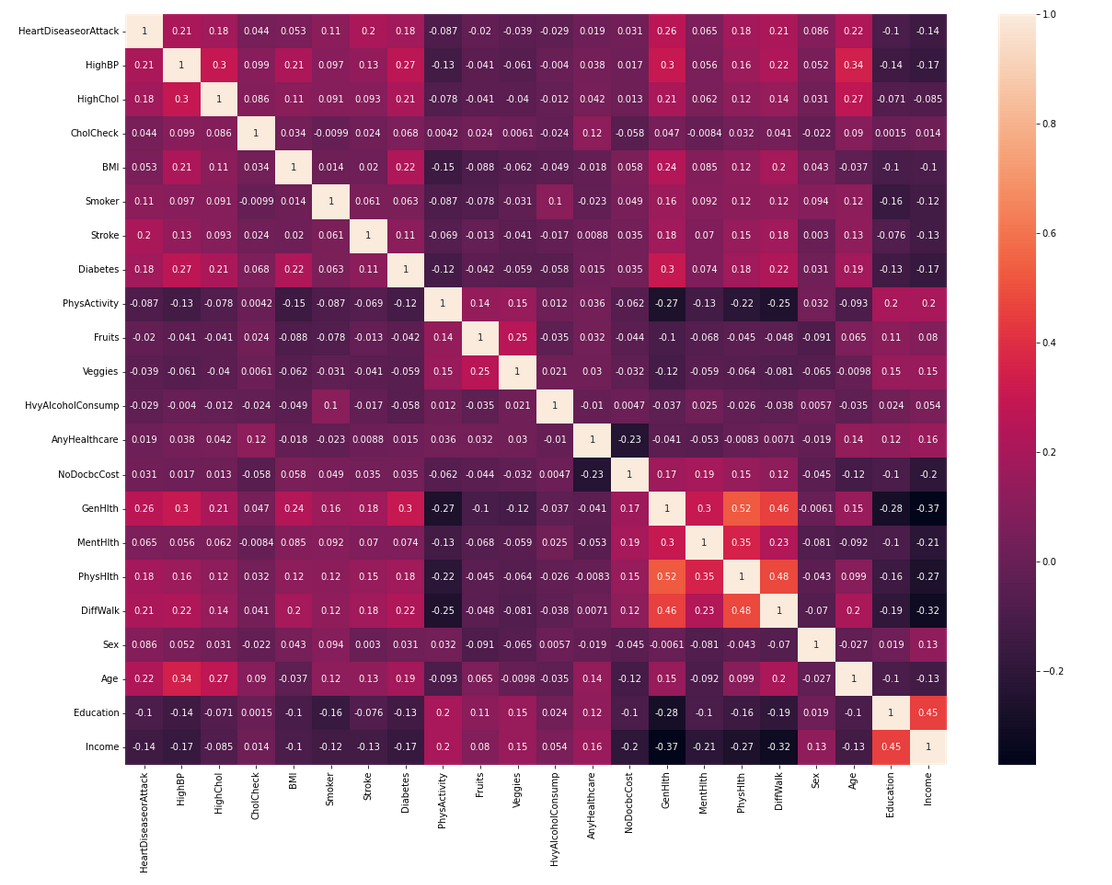
To ensure we have accurate results from the models, we must make sure the dataset is clean by examining the column data types, making the necessary conversions and removing columns and rows with missing values. In our case, there were no missing values, errors, or noise in the dataset that may cause underfitting or overfitting at the prediction stage.

**Feature Selection**

The heart disease dataset contains 20 features and not all the features have impact on the target variable. To ensure we have an accurate result of the model, the number of features must be reduced. This process is called feature selection. Feature selection is a technique in machine learning where feature variables are reduced. In this work, we have implemented a recursive feature elimination algorithm to perform the feature selection for the models and the result is depicted in the figure 5. The algorithm was used to avoid the “The Curse of Dimensionality” in the dataset by reducing the features from 20 to 7 features.

From the above plot, we can notice the test accuracy begins to decrease above 18 to 20 features ensuring the statistical performance of the model is accurate at the prediction stage.

**Results**



**Discussion and Conclusions**

In this report, the experiment was carried out with python examined and evaluated the performance accuracy of the results obtained from the three classification algorithms selected to make predictions from on the heart disease indicator dataset. The challenge of the imbalanced dataset was overcome with the SMOTEN resampling technique. From the results obtained from all three algorithms after optimizing, .While this report is for the sole purpose of comparing the selected classification models and appraising the performances, we encourage health establishments to embrace a data-driven approach to track patient health information in countries with high rate of cardiovascular diseases.

**References**

Secondary Prevention of Cardiovascular Diseases and Application of Technology for Early Diagnosis.

McGill, H. (2008), “Preventing heart disease in the 21st century implications of the pathobiological determinants of atherosclerosis in youth (PDAY) study,” Circulation, vol. 117, no. 9, (pg. 1216–1227)

Dey, A. (2016). Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of

Attributes using Principal Component Analysis. <https://www.researchgate.net/publication/301335834_Analysis_of_Supervised_Machine_Learning_Algorithms_for_Heart_Disease_Prediction_with_Reduced_Number_of_Attributes_using_Principal_Component_Analysis/link/583d2af708ae502a85e53634/download>

Pant, A. (2019). Introduction to Logistic Regression. <https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148>

Mitchell, T. (1997). Machine Learning. McGraw Hill Companies (pg. 52)

Lee, W. (2019). Python Machine Learning. John Wiley & Sons. (pg. 180)

# Mwanthi, D. (2022). Getting started with Recursive Feature Elimination algorithm in Machine Learning

<https://www.section.io/engineering-education/recursive-feature-elimination/#:~:text=the%20RFE%20algorithm.-,Implementing%20RFE%20algorithm,data%20into%20a%20decision%20tree.&text=The%20output%20above%20shows%20that%20the%20optimal%20number%20of%20features%20is%20t>

# Saini, A. (2021). Decision Tree Algorithm – A Complete Guide. <https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/>

**Appendix**