**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**

The gathering and exploration of data generated by human interactions has been a major factor to consider in various industries including the health care industry. Data generation in the era of internet 2.0 is humongous and diverse in size and nature, delineating and grouping them into three major categories; structured, unstructured, and semi-structured data. The compelling all organizations to implement big data processing with the use of advanced technology like machine learning for predictions. The gathering and exploration of data generated by human interactions today has become particularly important for business consideration in terms of the improvements its insights provide to various industries including the health care industry. Upon acquisition, exploration and analysis of the data, the gleaned insights enable relevant experts from the respective case studies take key decisions, which also includes the healthcare sector.

Cardiovascular disease is one of the prevalent causes of death in the world. Health condition calls for continuous research in the health industry. Certain health-related information about patients is an important ingredient in the exploration of ways to improve on the treatment or prediction heart conditions. The healthcare industry is one of the many industries that requires a continuous improvement around tackling health challenges by being able to predict heart conditions in patients for research and development purposes. This information gathered contain information or features that indicate the risk of heart disease or conditions in patients. There are various data gathering techniques that may be used to obtain patient information for experimentation and analysis to make informed decisions in the health care industry. This research paper is focused on the comparison of three classification algorithms and interpretation of their performances with respect to the results obtained after applying the relevant optimization techniques.

The experiment would be focused on using machine learning to understand the causes of heart disease and further look at ways to overcome the burden of the disease by predicting the risk of the disease in patients who are susceptible. The three classification algorithms chosen for comparison are the Decision Tree, Support Vector Machine and K – Nearest Neighbors were applied to a heart disease dataset containing 253680 rows, 21 variables, and a target or response variable. In the context of this experiment, 21 of these variables are regarded as the interactive features or variables that determine the outcome of the binary classification or target variable. These 21 variables are regarded as the interactive feature variables due to its importance of determining the predictions. They were reduced to 19 feature variables after extraction by the feature extraction tool (PCA) implemented during the experiment.

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 K – Nearest Neighbors). These results determined the best model based on the accuracy level after the implementation of the applicable hyperparameter optimization. For instance, the XGBoost ensemble machine learning algorithm was used to improve the results of ….

**Keywords:** *Decision Tree, Support Vector Machines, Logistic Regression, Confusion Matrix, Heatmap, PCA . 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. 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.

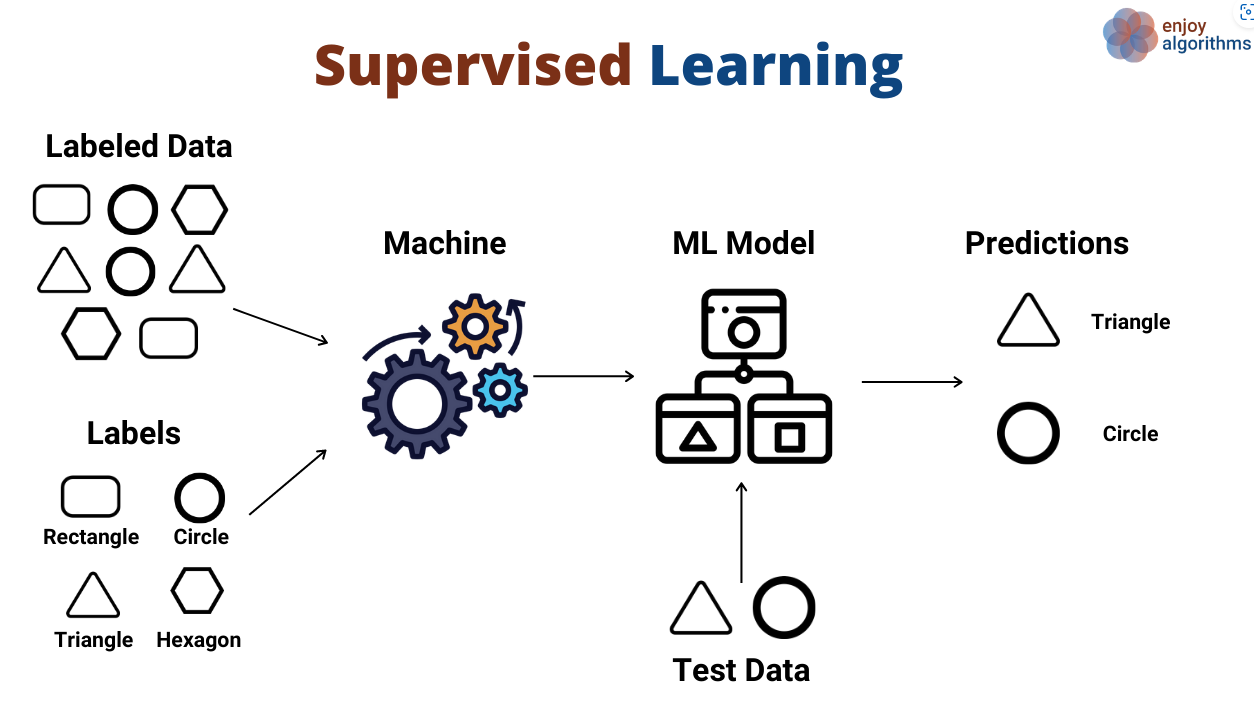
There are several risk factors that contribute to cardiovascular health conditions. It is no news that in most cases, the cause of the disease can be attributed to lifestyle and human preference of diets. Physical activities or routine exercises and healthy eating has been proven to help improve the 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. 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. 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. 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)

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. The difference is in based on how they handle data. 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.

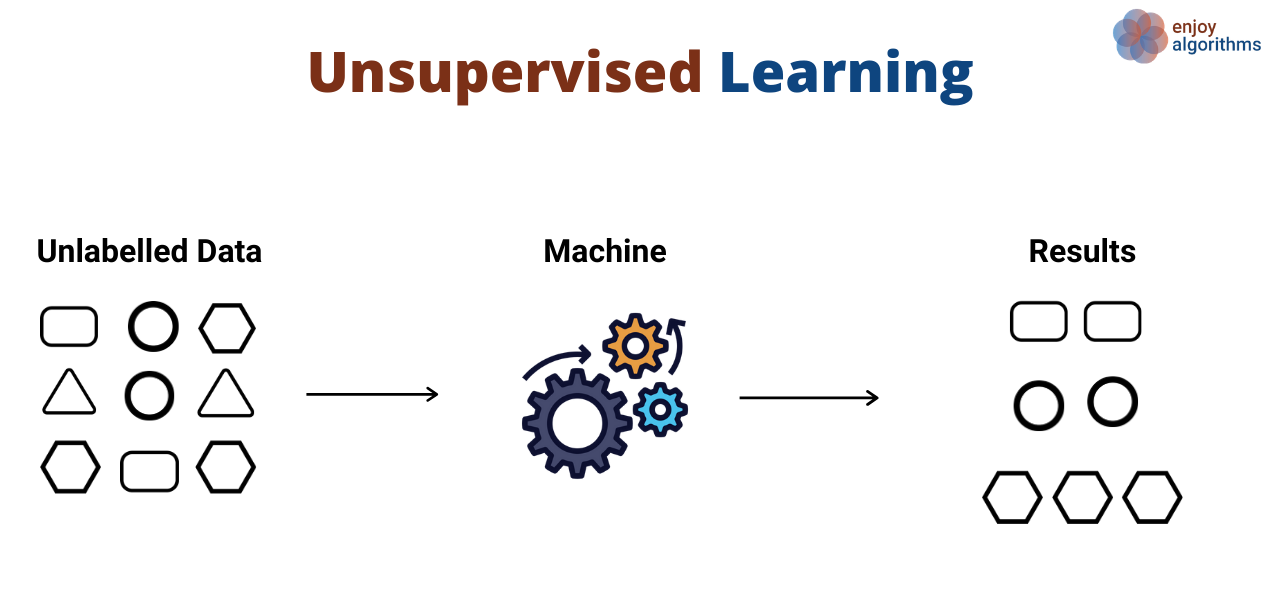
*Figure 1 - Supervised Learning*



# 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 – Unsupervised Learning*

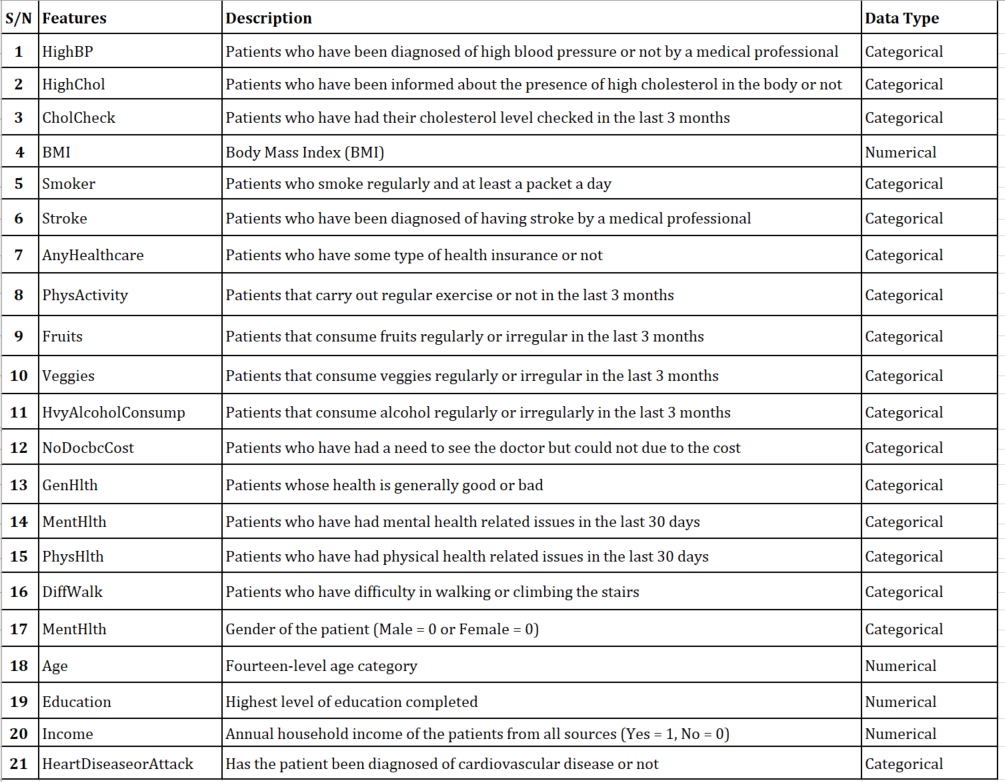


NOTE: From *Supervised and Unsupervised Learning (an Intuitive Approach)* by Kozan, 2021,  ( https://medium.com/@metehankozan/supervised-and-unsupervised-learning-an-intuitive-approach-cd8f8f64b644)

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.

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. This in our case study is linked to what has been tagged as the independent features and they are outlined in the table below;

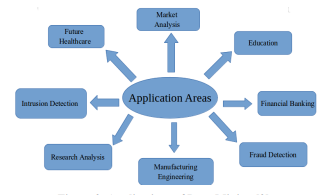
*Table 1 – Description of Independent and Dependent Features in the dataset*



**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*

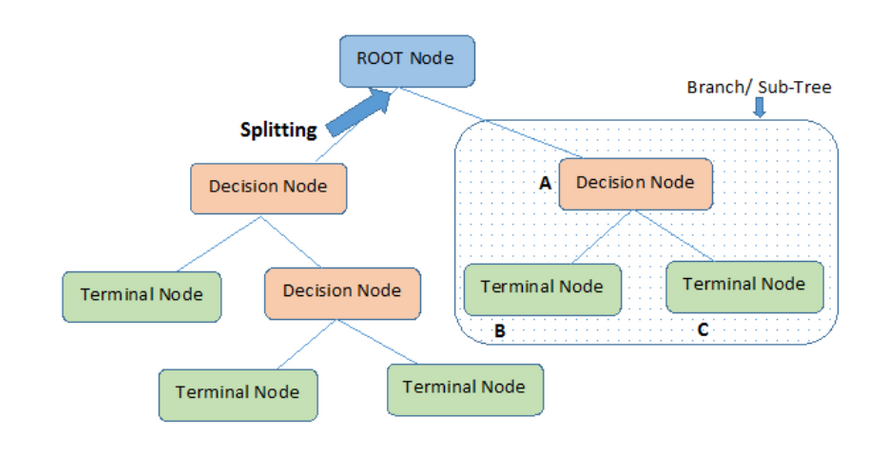


NOTE: Application of Data Mining. From Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis by Dey, 2016, (https://www.researchgate.net/publication/301335834\_Analysis\_of\_Supervised\_Machine\_Learning\_Algorithms\_for\_Heart\_Disease\_Prediction\_with\_Reduced\_Numb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

The machine learning methods adopted for the prediction of cardiovascular disease are Decision Tree, K-Nearest Neighbours, and Logistic Regression. Support Vector Machine. This is due to the expected categorical response variable for binary classification algorithms. In our case, the response variable is either a one (1) or a zero (0), meaning a patient is either diagnosed of cardiovascular disease or not.

Decision Tree Classification algorithm is one of the simplest techniques to implement in solving classification and 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”.

Figure 4 – A diagram of a decision learning process in theory



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.

**Support Vector Machine**

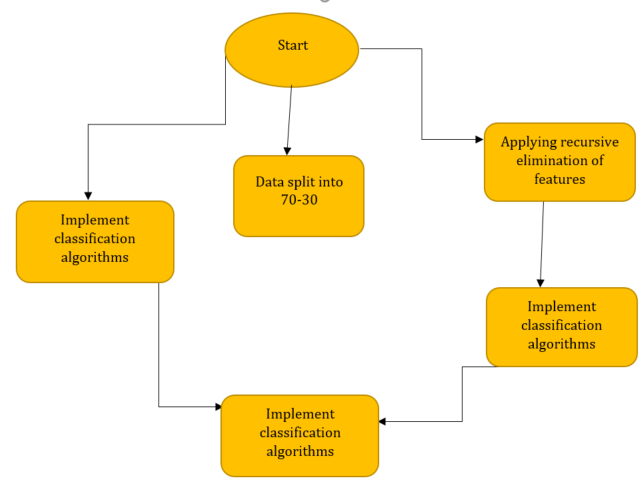
**Logistic Regression**

**K-Nearest Neighbors**

**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.

Figure 5 – Flowchart of experimental process of comparing the models

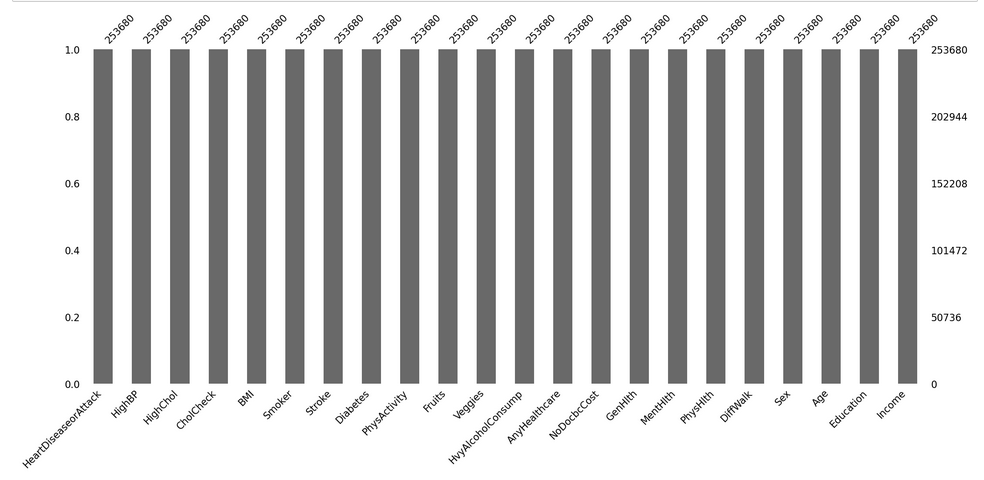


To begin with the setup, it is important to demonstrate in the flowchart below, the method adopted to compare the performance of the three (3) classifier algorithms before feature selection and after feature selection.

**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. The figure below is graphic illustration of the content of the dataset.

Figure 4 – Graphical display of the dataset without missing values

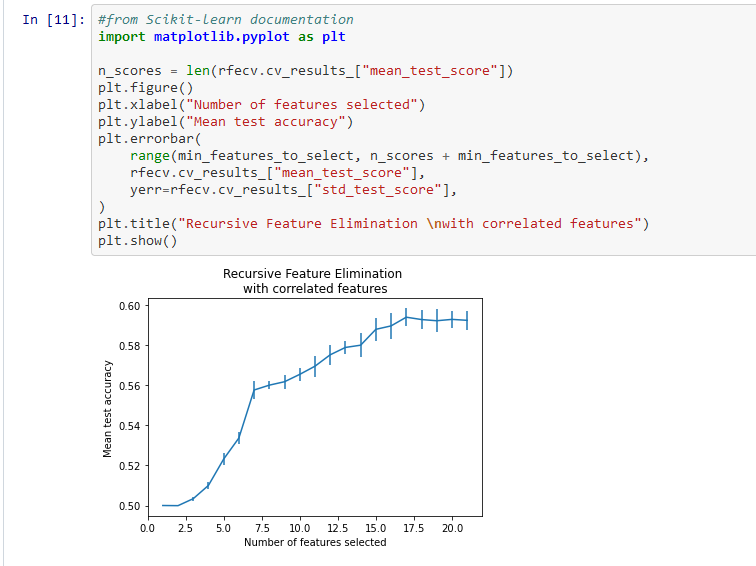


Discuss underfitting and overfitting

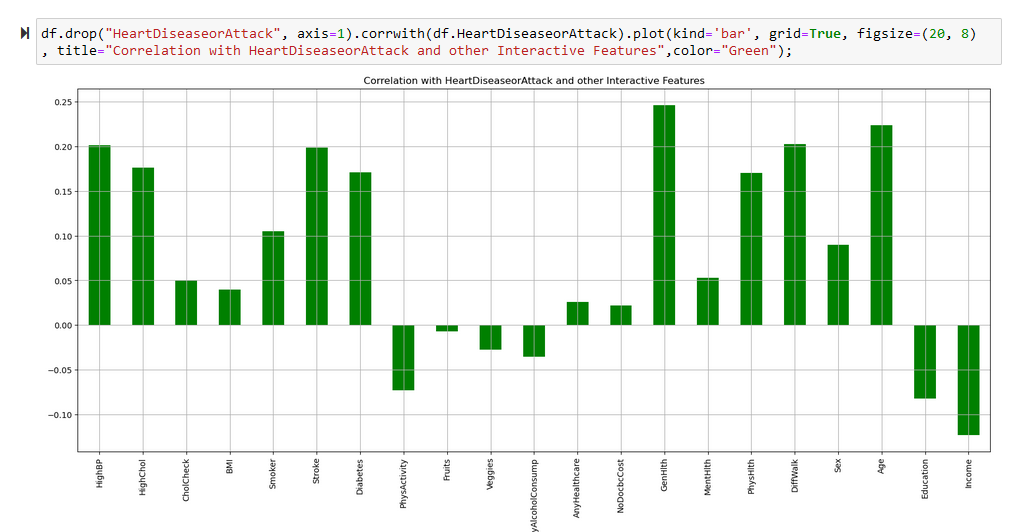
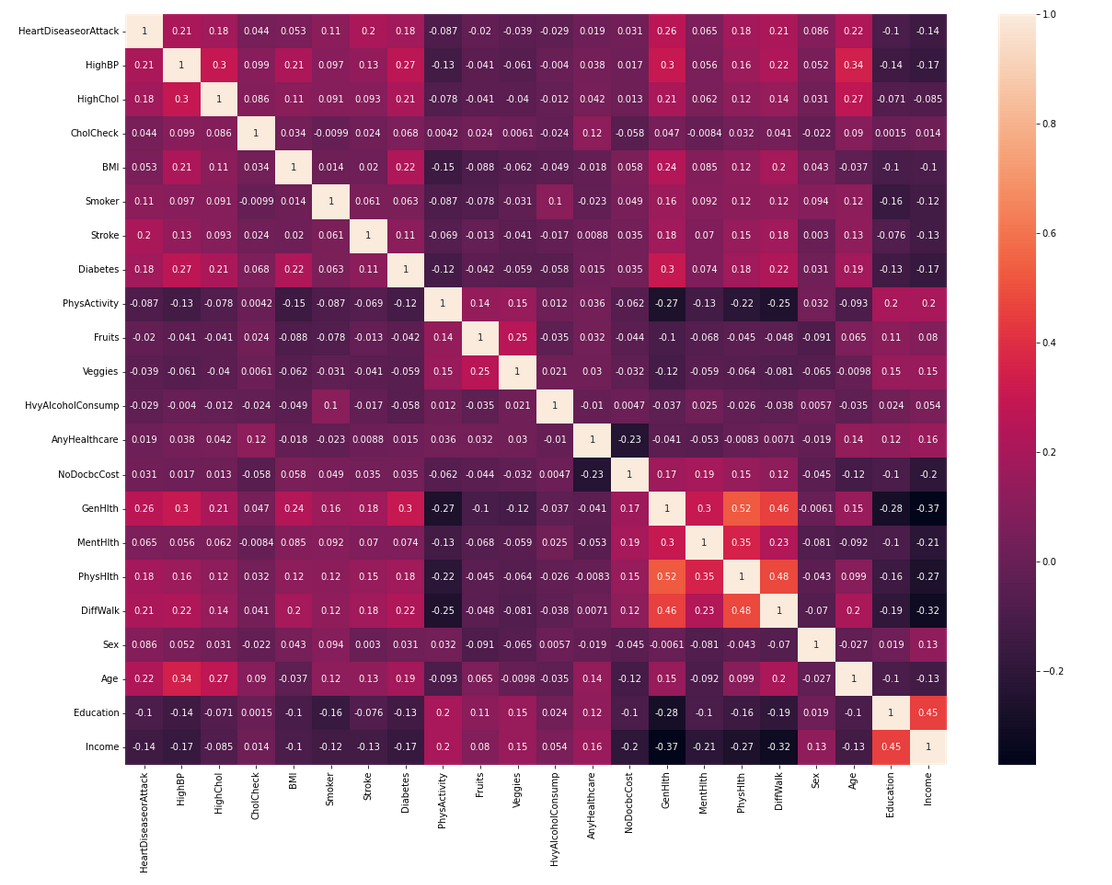
**Feature Selection**

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.

Figure 5 – Graphical display of feature selection by Recursive Feature Elimination Algorithm



**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. 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, pp. 1216–1227, 2008

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

# 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**