**Heart Failure Prediction Project**

**Student Name:**

**Student ID:**

**Date:**

**Declaration**

I the author of the Heart Failure Research declare the following,

That all the included data and information in this report was uniquely gathered, scrutinized, and composed by myself. This means that there are absolutely no submissions made earlier on in this discipline of field research as part or whole of the information in the entire report. Any relevant secondary materials discerned are appropriately recognized via paperwork references. The entire project is completely owned by myself.

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Acknowledgments**

I would like to appreciate the entire support as well as the tremendous recommendations I gained from a large group of people during my entire dissertation period. My dissertation instructor has been there providing splendid guidance rhythm as well as encouraging feedback which was, which I took it positively during my dissertation period. However, I would like to thank my fellow students who were involved in the testing phase. It is through the identified errors that I made immediate corrections as well as outstanding adjustments towards the development of the entire web application.

# **Abstract**

This project outlined various challenges faced by the health sector in fighting heart failure-related problems. It evaluated how different studies show that heart failure has been linked with coronary artery disease. This type of disease develops as a result of fatty deposits in the heart arteries, this causes a reduction in the blood flow leading to a heart attack. Heart attacks generally occur suddenly when the coronary artery becomes blocked thoroughly. The research was based on a heart failure dataset that has all the relevant attributes that are required to provide insights on defining all the required patterns. This means that only pre-existing open-source dataset was collected, verified, and analyzed thoroughly.

This research was based on machine learning models which were used to impact patients’ care delivery schemes. It is through the application of machine learning classification models to the existing data where the detection of patterns associated with heart failure diseases as well as patient records will be analyzed appropriately. The implementation phase of the entire Heart Failure research included carrying out several tasks. These tasks included data reading, data cleaning, feature selection, data descriptions, data correlation, data plotting as well as data prediction using various machine learning models. The machine learning models gave various results. Jupyter Notebook Application is a client-server-based application that allowed the editing as well as the running of notebook documents via different web browsers. The Jupyter Notebook application was executed on the local machine with no requirement of internet access. However, the notebook application gave the dashboard alias the control panel which was used for data as well as results editing, display, and running of the notebook documents.

Jupyter Notebook application contains notebook kernel which is referred to as the computational engine that was used for executing the code written for the heart failure research. Through the use of the Jupyter Notebook, the application was discovered to be powerful in showcasing one’s work in that I was able to see the code and its relative results. However, the running of the application’s code was facilitated by python version 3.9 libraries. It was noticed that Gradient Boosting Model topped to be the best in predicting the model for the heart failure task after yielding an accuracy average of 88.3% in all sampling use cases.

Table of Contents

[Abstract 4](#_Toc92934316)

[Introduction 8](#_Toc92934317)

[Background Information 8](#_Toc92934318)

[Research Aim 9](#_Toc92934319)

[Problem Statement 9](#_Toc92934320)

[Scope, Objectives, Approach, and Project Plan 9](#_Toc92934321)

[Scope 9](#_Toc92934322)

[Objectives 10](#_Toc92934323)

[Project Plan 11](#_Toc92934324)

[Ethics, Legal, Social, Security and Professional Issues 11](#_Toc92934325)

[i. Ethical Issues 11](#_Toc92934326)

[ii. Legal Issues 11](#_Toc92934327)

[iii. Social Issues 11](#_Toc92934328)

[iv. Professional Issues 12](#_Toc92934329)

[v. Security 12](#_Toc92934330)

[Literature Review 12](#_Toc92934331)

[Introduction 12](#_Toc92934332)

[Related Work 13](#_Toc92934333)

[Machine Learning Models 18](#_Toc92934334)

[a. Logistic Regression model 18](#_Toc92934335)

[Types of Logistic Regression 19](#_Toc92934336)

[Assumptions 20](#_Toc92934337)

[b. K-Nearest Neighbor model 21](#_Toc92934338)

[How does K-NN Model Work? 21](#_Toc92934339)

[c. Support Vector Machine model 22](#_Toc92934340)

[Methodology 23](#_Toc92934341)

[Classification 23](#_Toc92934342)

[a. Binary Classification 25](#_Toc92934343)

[b. Multi-class Classification 25](#_Toc92934344)

[Classification Tasks Learners 25](#_Toc92934345)

[a. Lazy Classification Learners 25](#_Toc92934346)

[a. Eager Classification Learners 25](#_Toc92934347)

[Types of Machine Learning Classification Algorithms 26](#_Toc92934348)

[Design and Implementation 26](#_Toc92934349)

[Introduction 26](#_Toc92934350)

[a. Data Reading 27](#_Toc92934351)

[b. Data Cleaning 31](#_Toc92934352)

[c. Data Correlation 33](#_Toc92934353)

[d. Feature Selection 35](#_Toc92934354)

[e. Dataset Outlier Removal 36](#_Toc92934355)

[f. Data Model’s Building 37](#_Toc92934356)

[i. K- Neighbors Model 37](#_Toc92934357)

[ii. Logistic Regression Model 37](#_Toc92934358)

[iii. Support Vector Machines (SVM) Model 38](#_Toc92934359)

[iv. Gradient Boosting 38](#_Toc92934360)

[Data Model Building without Sampling Technique 38](#_Toc92934361)

[Data Model Building with Sampling Technique 44](#_Toc92934362)

[g. Prediction 49](#_Toc92934363)

[Project Evaluation and Discussion 55](#_Toc92934364)

[Conclusion and Recommendations 57](#_Toc92934365)

[Conclusion 57](#_Toc92934366)

[Recommendations 58](#_Toc92934367)

[References 59](#_Toc92934368)

**Tables**

[Table 1: Literature Summary 16](#_Toc89988850)

[Table 2: Model's without sampling Analysis Table 41](#_Toc89988851)

[Table 3: Prediction on correlated dataset features 44](#_Toc89988852)

[Table 4: Prediction with sampling Analysis 46](#_Toc89988853)

[Table 5: Prediction with sampling on Selected dataset features Analysis 49](#_Toc89988854)

[Table 6: K-Fold sampling Analysis 54](#_Toc89988855)

[Table 7: Model's Overall Sampling Accuracy 54](#_Toc89988856)

**Table of Figures**

[Figure 1: Sigmoid function 19](#_Toc89988879)

[Figure 2: K-NN Model 22](#_Toc89988880)

[Figure 3: SVM Notation 23](#_Toc89988881)

[Figure 4: Classification Diagram 25](#_Toc89988882)

[Figure 5: Data Reading 28](#_Toc89988883)

[Figure 6: Head data in the dataset 28](#_Toc89988884)

[Figure 7: Data Tail 29](#_Toc89988885)

[Figure 8: Features 29](#_Toc89988886)

[Figure 9: Discrete and Continuous Datatype 30](#_Toc89988887)

[Figure 10: Discrete Features 30](#_Toc89988888)

[Figure 11: Continous Features 31](#_Toc89988889)

[Figure 12: Dataset information 32](#_Toc89988890)

[Figure 13: Data is a null sum 32](#_Toc89988891)

[Figure 14: Drop duplicates method 32](#_Toc89988892)

[Figure 15: Drop Missing Values 33](#_Toc89988893)

[Figure 16: Data Correlation 34](#_Toc89988894)

[Figure 17: Correlation Heat Map 34](#_Toc89988895)

[Figure 18: Feature Selection 35](#_Toc89988896)

[Figure 19: Outlier Removal 36](#_Toc89988897)

[Figure 20: Model building without sampling 38](#_Toc89988898)

[Figure 21: Model without sampling (Logistic Regression) 39](#_Toc89988899)

[Figure 22: Model without sampling (K-NN) 39](#_Toc89988900)

[Figure 23: Model without sampling (Random Forest) 40](#_Toc89988901)

[Figure 24: Model without sampling (Support Vector Machines) 40](#_Toc89988902)

[Figure 25:Model building without sampling (with few features) 41](#_Toc89988903)

[Figure 26: Prediction using Logistic Regression 42](#_Toc89988904)

[Figure 27: Prediction using Gradient Boosting 42](#_Toc89988905)

[Figure 28: Prediction using Random Forest 43](#_Toc89988906)

[Figure 29: Prediction using Support Vector Machines 43](#_Toc89988907)

[Figure 30: Model building with Sampling Technique 44](#_Toc89988908)

[Figure 31: Prediction with sampling (Logistic Regression) 45](#_Toc89988909)

[Figure 32: Prediction with sampling (Random Forest) 45](#_Toc89988910)

[Figure 33: Prediction with sampling (Naive Bayes) 46](#_Toc89988911)

[Figure 34:Prediction with sampling on Selected dataset features 47](#_Toc89988912)

[Figure 35: Prediction with sampling on Selected dataset features (Logistic Regression) 47](#_Toc89988913)

[Figure 36: Prediction with sampling on Selected dataset features (Random Forest) 48](#_Toc89988914)

[Figure 37: Prediction with sampling on Selected dataset features (Gradient Boosting) 48](#_Toc89988915)

[Figure 38: Prediction with sampling on Selected dataset features (Naive Bayes) 49](#_Toc89988916)

[Figure 39: Repeated K-Fold Sampling/Prediction 50](#_Toc89988917)

[Figure 40: Repeated K-Fold Prediction (Random Forest) 50](#_Toc89988918)

[Figure 41: Repeated K-Fold Prediction (Gradient Boosting) 50](#_Toc89988919)

[Figure 42: Repeated K-Fold Prediction (Support Vector Machines) 51](#_Toc89988920)

[Figure 43: Repeated K-Fold Prediction (Decision Tree) 51](#_Toc89988921)

[Figure 44: Repeated K-Fold Prediction (Naive Bayes) 51](#_Toc89988922)

[Figure 45: Death prediction based on features correlation in K-Fold sampling 52](#_Toc89988923)

[Figure 46: Death prediction based on correlation in K-Fold sampling (Random Forest) 52](#_Toc89988924)

[Figure 47: Death prediction based on correlation in K-Fold sampling (Gradient Boosting) 52](#_Toc89988925)

[Figure 48: Death prediction based on correlation in K-Fold sampling (Decision Tree) 53](#_Toc89988926)

[Figure 49: Death prediction based on correlation in K-Fold sampling (Support Vector Machines) 53](#_Toc89988927)

[Figure 50: Death prediction based on correlation in K-Fold sampling (K-Neighbors) 53](#_Toc89988928)

# **Introduction**

In chapter one of the research, it gives a detailed description of heart failure research. Discussion on the background information required for the proposed project to be feasible economically, technically, operationally as well as legally. However, the background information gives a clear focus as well as precise of the research.

## **Background Information**

Technology evolution has affected society over the last decade in that it has improved the way people live across various parts of the world. It has affected the ways that different individuals communicate, learn as well as think while carrying out their day-to-day activities, (Biagi & Falk, 2017). Through this, technology has helped people to learn new modes of interacting as well as collaborating. This has helped people to have better and clear engagement.

However, technology has brought a huge and welcome change to the health sector. Through this, different patients can access the crème diagnostic relative tools as well as new cutting-edge based treatments. Machine learning and artificial intelligence dominance in the healthcare systems have resulted in accurate statistics unlike in the earlier days where statistics were extremely limited.

The application of machine learning has positively impacted patient care delivery plans, (Etu, 2018). Given an example of this, different physicians are now aided in the identification, diagnosis, and treatments of diseases. The proposed research be based on machine learning models which will be used to impact patients’ care delivery schemes. It is through the application of machine learning classification models to the existing data where the detection of patterns associated with heart failure diseases as well as patient records will be analyzed appropriately.

## **Research Aim**

The main aim of this research paper was to predict heart failure using machine learning models.It therefore utilized popular models for classification in machine learning like the Logistic Regression model, K-Nearest Neighbor model, Support Vector Machines model, Decision Trees model, Bagging and Boosting models to predict heart failure. For this case, a dataset containing different attributes of patients to predict heart failures will be used.

## **Problem Statement**

For a long time, the health sector has been experiencing several challenges on how to deal with heart failure problems. According to the WHO statistics on heart failure, at least 1.4 million people living with heart failure conditions are under 60 years of age, only 2 percent of people of ages 40 to 59 live with this critical condition across the whole world (WHO, 2021). The proposed research be based on machine learning models which will be used to impact patients’ care delivery schemes. It is through the application of machine learning classification models to the existing data where the detection of patterns associated with heart failure diseases as well as patient records will be analyzed appropriately.

## **Scope, Objectives, Approach, and Project Plan**

### **Scope**

The proposed research will be based on a heart failure dataset that has all the relevant attributes that are required to provide insights on defining all the required patterns. This means that only pre-existing open-source datasets will be collected, verified, and analyzed thoroughly. Relevant consultations will be made from the domain knowledge experts.

### **Objectives**

Since the proposed research is a data science-driven research, the following are the defined objectives to be achieved;

1. To collect data required: Online surveys to collect open-source heart failure datasets will be carried out respectively.
2. To process the collected data: This will include data cleaning process activities such as describing statistically the data attributes, removing data outliers, duplicates as well as missing values.
3. To carry out feature engineering: This will involve scaling numeric features, encoding categorical features, changing back dummy variables to numeric.
4. To explore and visualize data: This will be done by the application of data visualizing tools used in data science like heatmaps, histograms, etc.
5. To model the data: This will involve splitting the dataset into two parts that are, training and testing parts respectively.
6. To analyze and apply to learn: After splitting the dataset into two, the application of various classification machine learning models will be applied.
7. To make decisions based upon insights generated: This will be made based on the analyzed machine learning model's output.

This research will ease the workflow of predicting patients’ heart failures thus reducing the risks of treatment failure. However, the applied machine learning models will be utilized countless times giving accurate results depending on the type of dataset applied to them. This means they can learn by themselves and give results based on the data accurately. However, the following are the techniques to be used all through;

1. Classification,
2. Outlier detection,
3. Prediction.

### **Project Plan**

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Task Name** | **Planned Days** | **Deliverables** |
|  | Data collection | 14days | Relevant datasets in the domain of heart failure. |
|  | Data processing | 7 days | Redefined dataset, error-free dataset. |
|  | Feature engineering, visualization, and data modeling | 42 days | Heat maps, histograms, Model results |
|  | Review, Feedback &  Decision Making | 7 days | Predicted results. |

## **Ethics, Legal, Social, Security and Professional Issues**

### **Ethical Issues**

Ethical acting in any given in field simply refers to building a society/workplace based on integrity, application of regulations as well as trust among the people, (Daniel et al, 2019). The ethical issues to deal with in any project or research may include diversity and acceptance as well as carrying out tasks in due respect to the predetermined goals and expectations.

### **Legal Issues**

Legal issues can be referred to as the legal questions which are fundamental to any court case, (George et al, 2016). Heart failure project is based on data mining field of study, this doesn’t mean through it there will be discrimination of people datasets, mostly in regards with race, sex as well as religion orientation. With consideration of people’s origin is unethical and illegal.

### **Social Issues**

Being social is the act of seeking or providing great relationships with others. The heart failure proposed project will benefit society by building trust through which medical attention links will be created as a public service.

### **Professional Issues**

Professional behavior is highly required during the research period. This may include maintaining data privacy as well as data confidentiality for the datasets used since they might be collected from various sources.

### **Security**

Data security can be termed as the procedure of protecting data from various unauthorized access gates as well as corruption of data throughout the data lifecycle, (Scassa &Taylor, 2017). Heart failure is data-driven research that applies machine learning models and this will provide enough security. Backing up the data used will be implemented to strengthen data security.

# **Literature Review**

### **Introduction**

Heart failure is a general term used to provide a precise description of a heart that cannot keep up well with its tasks (Bytyçi & Bajraktari, 2015). This may be as a result of the body not getting enough oxygen as it requires generally. This has been a serious and awful condition because there is no accurate cure that has been discovered to mitigate its seriousness, (Toback & Clark, 2017). However, people living with this type of condition live happily when the entire condition is managed with various heart failure medications or prescriptions from different doctors in different hospitals, (Kaminsky &Tuttle, 2015). It has been discovered that family support towards their peers with heart failure conditions is an important thing to be maintained.

According to the WHO statistics on heart failure, at least 1.4 million people living with heart failure conditions are under 60 years of age, only 2 percent of people of ages 40 to 59 live with this critical condition across the whole world, (WHO, 2021). Different studies show that heart failure has been linked with coronary artery disease. This type of disease develops as a result of fatty deposits in the heart arteries, this causes a reduction in the blood flow leading to a heart attack, (Nikam, 2015). Heart attacks generally occur suddenly when the coronary artery becomes blocked thoroughly.

Different new technologies including Heart flow have been invented to help physicians in determining easily if their patients require invasive procedures to clear their clogged arteries or test if non-invasive treatment such as medications would work, (Coeckelbergh, 2019). An implantable cardiac defibrillator alias ICD is a developed electronic device that is put down inside the patient’s body, (Kunwar et al, 2016). It is designed to keep track of the patients’ heart rhythm as well as send minimal revelation to the heart muscle in a condition where the heart rhythm becomes unusual, (Perveen et al, 2016). This state can be referred to as arrhythmia. It is required for one to note that if the revelation is required, it can be very disturbing but its effects end quickly.

### **Related Work**

Several types of research have outlined how heart failure can be predicted via machine learning using different machine learning algorithms. According to (Ross et al, 2008) they carried out a structured exploration of different studies assessing different patient attributes that could be related to patients’ hospital re-admission due to heart failure. (Rahimi et al, 2016) performed a thorough review on the literature regarding risk prediction machine learning models for the patients facing heart failure issues and the literature discovered the highest consistently communicated unrestrained predictors of risk across machine learning models. Assessment of the performance regarding the prevailing set of models that were used to forecast hospital readmission was carried out by, (Kansagara et al, 2011).

Later on, information from Telemonitoring was used to improve heart failure outcomes trial and through this activity, comparing the effectiveness of different machine learning techniques in predicting about 30- and 180- day all cause-related readmissions due to heart failure problems were carried out by, (Mortazavi et al, 2016). According to (Choi et al, 2016), they carried an initial diagnosis of heart failure-related problems in different hospitals having different patients, and this was done through implementing recurrent neural network alias as RNN model.

Through the RNN model, temporal exploitation of relations between the used 3884 heart failure records of different primary care patients and the electronic health records alias EHRs was carried out. (Choi et al, 2016) also carried out a comparison of the RNN model with other machine learning models such as the K-nearest neighbor model classifier, support vector machine (SVM), the neural network as well as the logistic regression model.

There were many RNN architecture variants. Unidirectional RNN tends to be designed so that it can only draw inferences from previous inputs when making predictions about the present. Bidirectional is another RNN architecture variant which plus future data for improved accuracy. LSTM is another critical RNN architecture and is perceived to be the solution to the vanishing gradient. LSTM has enhanced the applicability of RNN, thus making it more known in the technological community. Due to the internal memory capability, RNN has the potential to remember the input values. It is through remembering the input values that the RNN is efficient in predicting what may come next, (Yu et al., 2019).

RNN, due to efficiency in predicting what may come next, they are the preferred algorithm for sequential data such as video streaming, text, and many. Gated recurrent unit is also among the architecture variant and works similarly to LSTM since it focuses on resolving the related issues to short-term memory. The Gated recurrent unit uses a hidden state for the regulation of information in the RNN model.

RNN was designed to take prior input to influence the current output and input. In some traditional deep neural networks, it is assumed that both outputs and inputs are independent of one another (Yin et al., 2017). In the case of RNN, the output relied on previous factors or elements integrated within the sequence.

Unidirectional RNN could not account for future events in the predictions made when determining the output. NNs algorithm is ideal because it offers great flexibility when it is being modeled (Li et al., 2018). The architecture in terms of layers and hidden nodes can be easily adjusted to meet the desired requirements. NNs are more likely to show quality and accurate forecasting than linear regression when utilized in heart failure prediction. RNN has recent past and present as inputs. Two inputs make the RNN ideal since sequence data may contain vital information regarding what may come next. The trained model can be developed based on the dataset fed into the algorithm. According to Chan et al. (2019, p. 03), high accuracy is achieved when the algorithm is trained through a supervised learning approach.

The use of structured as well as the unstructured data from different electronic health records by (Wang et al, 2015) to forecast the start of heart failure as well as the variation of the prediction window estimated from 60 to 720 days before heart failure diagnosis. (Wang et al, 2015) at least used an aggregate of 1684 heart failure records of various primary care hospital patients. However, (Dai et al, 2015) utilized at least five machine learning models to forecast heart failure-related hospitalizations. According to, (Dai et al, 2015) they utilized the EHR patient's data experiencing heart failure and diseases from a large urban hospital based in Boston.

System state prognostics data model fusion framework was proposed by, (Liu et al, 2012). Investigation on the problem of evaluating the remaining useful life through the use of stochastic lifetime machine learning models was done by, (Nystad et al, 2012) in consideration of randomly distributed failure thresholds. According to (Gola & Nystad, 2012), a combination of a condition monitoring system that provided reliable computations of the definite erosion situation of choke valves during its performance with a lifetime remaining useful life model. (Casoetto et al, 2003) constructed a model of the normal behavior of a given health system by fusing multiple sensor signals when in its normal performance.

The implementation of this system monitored the behaviors of the entire health system for purposes of degradation which were expressed by a drift of various signals away from the normal state. However, (Jiang et al, 2012) proposed a new method to be used for calibration of a predictive model for different individual patients, it used a similar group of hospitalized patients for the calibration process.

The decision tree is among the ML algorithm that has been applied in various trading operations. The algorithm is in the manner that it does not require parameter setting or domain knowledge. Decisions tree as algorithm offers visually interpretable models compared to the Neural networks (Deng et al., 2019). Improved visual interpretation in the decision tree is because of the easy-to-understand nodes. Different models can be produced based on the decision tree concept. The simplest model that can be generated based on the decision tree concept is classification and regression tree. When applied in trading operations, classification and regression trees perform far better than the single-factor models when the same stock portfolio variables are picked. When the alternating decision tree is boosted, it is more likely to result in an abnormal return for the company.

#### **Summary**

Table 1: Literature Summary

|  |  |  |
| --- | --- | --- |
|  | **Authors** | **Achievements** |
|  | Ross et al, 2008 | * Carried out a structured exploration of different studies assessing different patient attributes that could be related to patients’ hospital re-admission due to heart failure. |
|  | Rahimi et al, 2016 | * Performed a thorough review on the literature regarding risk prediction machine learning models for the patients facing heart failure issues and the literature discovered the highest consistently communicated unrestrained predictors of risk across machine learning models. |
|  | Kansagara et al, 2011 | * Assessment of the performance regarding the prevailing set of models that were used to forecast hospital readmission. |
|  | Mortazavi et al, 2016 | * Used Telemonitoring to improve heart failure outcomes trial and through this activity, comparing the effectiveness of different machine learning techniques in predicting about 30- and 180- day all cause-related readmissions due to heart failure problems. |
|  | Choi et al, 2016 | * Carried an initial diagnosis of heart failure-related problems in different hospitals having different patients, and this was done through implementing recurrent neural network alias as RNN model. * Also carried out a comparison of the RNN model with other machine learning models such as the K-nearest neighbor model classifier, support vector machine (SVM), the neural network as well as the logistic regression model. |
|  | Yin et al., 2017 | * Designed RNN to take prior input to influence the current output and input. In some traditional deep neural networks, it is assumed that both outputs and inputs are independent of one another. |
|  | Wang et al, 2015 | * Used structured as well as the unstructured data from different electronic health records to forecast the start of heart failure as well as the variation of the prediction window estimated from 60 to 720 days before heart failure diagnosis, * At least used an aggregate of 1684 heart failure records of various primary care hospital patients. |
|  | Dai et al, 2015 | * Utilized at least five machine learning models to forecast heart failure-related hospitalizations, * They utilized the EHR patient's data experiencing heart failure and diseases from a large urban hospital based in Boston. |
|  | Liu et al, 2012 | * Proposed system state prognostics data model fusion framework. |
|  | Nystad et al, 2012 | * Investigation on the problem of evaluating the remaining useful life through the use of stochastic lifetime machine learning models in consideration of randomly distributed failure thresholds. |
|  | Gola & Nystad, 2012 | * Proposed for a combination of a condition monitoring system that provided reliable computations of the definite erosion situation of choke valves during its performance with a lifetime remaining useful life model. |
|  | Cassette et al, 2003 | * Constructed a model of the normal behavior of a given health system by fusing multiple sensor signals when in its normal performance. |
|  | Jiang et al, 2012 | * Proposed a new method to be used for calibration of a predictive model for different individual patients, it used a similar group of hospitalized patients for the calibration process. |

## **Machine Learning Models**

### **Logistic Regression model**

A logistic regression model is considered to be a statistical model that uses a defined logistic to model a binary dependent variable, (Dietz, 1987). Mostly in regression type of analysis, the logistic regression alias logit regression is usually used to estimate the logistic model parameters.

This type of machine learning model analysis is used in examining the association which could be either continuous or categorical, independent variables with the dichotomous stable variable. In this, it is a contrast to the linear type machine learning regression analysis in which the stable variable could be used as a continuous variable rather than being categorical. This type of machine learning model learns a given linear connection from the dataset given to it and afterward introduces a non-linearity in the form of a sigmoid function, (Cheon et al, 2017).

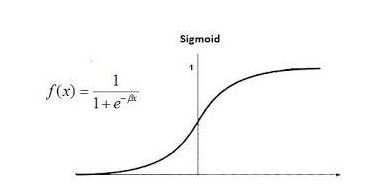


Figure 1: Sigmoid function

The logistic regression model representation uses an equation and therefore the input values (x) are linearly combined with the use of weights or the preferred coefficient values (Beta) as this predicts the output value of the model (y). By ensuring the determination of the likelihood factors as expectation, it can be termed as a procedure that explores the connection of the resulting variables together with the self-reliant variables in the binary or else multiple phases.

Logistic regression is always termed as a supervised learning classification machine learning model used in forecasting the expectation of any given target variable. The self-reliant variable nature is said to be dichotomous, meaning that there would be the possibility of two classes, (Lowrie & Lew, 1990). The key objective of performing regression analysis at any given point is to explain the variability depending on the type of a variable by the mechanism of one or more of the self-reliant or control variables.

Scientifically, the said dependent variable is binary having the data coded as either 0 which stands for failure/no, or 1 which stands for success/yes. Mathematically, it can be said that the logistic regression model forecasts P(Y=1) as a general function of X, thus making the logistic regression model to be considered as one of the simplest machine learning algorithms that can be used for classification tasks such as spam text detection, diabetes diseases forecasting, heart failure prediction as well as cancer prediction, etc.

### **Types of Logistic Regression**

Logistic regression can be grouped into two types, which are based on the number of categories of different target variables that can be used to forecast a solution. The logistic regression, therefore, means binary logistic regression that has a binary target variable.

#### Binomial or Binary Logistic Regression

In this type of classification, the self-reliant termed as dependent variable always has the possible types of either 0 or 1. In general terms, these types of variables could represent different real-life situations like success or failure, no or yes, loss or win, etc.

#### Multinomial Logistic Regression

In this type of classification, the self-reliant variable can exist and have 3 or more possible scattered types, or else these types could have no computational significance. Given an example of this, variables may be used to represent “Type X” or “Type Y” or “Type Z”.

#### Ordinal Logistic Regression

In this type of classification, the self-reliant variable could have at least 3 or more viable structured types, or else these types have a computational significance. Given an example in this, these variables may be used to represent different real-life situations like “poor” or “good”, “superb”, “excellent” and in categories can be represented with different scores like 0,1,2,3, etc.

### **Assumptions**

Before implementing logistic regression division, it has been recommended for one to consider the assumptions listed below;

1. The logistic regression model must have meaningful variables,
2. Choosing large size sample for the logistic regression model is highly recommended,
3. The models’ independent variables need to be self-reliant of each other meaning that there should be no multi-collinearity in the prediction model,
4. The desired outcome of the model is always represented by a factor level of 1 and this means that the target variable must always be in binary format.

The logistic Regression machine learning model has the following advantages that will be considered in the implementation of the heart failure prediction project.

1. Its implementation, interpretation is easy as well as being efficient to train,
2. While using logistic regression, no assumptions are made on the distribution of classes in attribute space,
3. Logistic regression can easily extend multiple classes and this gives it a natural expectation view on the class forecasts.
4. Implementation of logistic regression model gives not only the direction of association termed to be either positive or negative but also it gives an appropriate measure of a predictor.
5. Given a linearly separable dataset, it results in a good accuracy level for many simple datasets,
6. The logistic regression model can interpret its coefficients as indicators of attribute importance.

### **K-Nearest Neighbor model**

K-nearest neighbor is a classification as well as a regression solving machine learning algorithm that is said to be simple in its performance as well as implementation. It is easy to understand its results simply because it is a supervised machine learning model, (Kuang et al, 2019). For one to make a prediction using this type of model for any given new observation, he/she needs to keep all the training data in a dataset since the model is always an instance-based type of a model. This type of model assumes any resemblance between the new data and all the available data and through this, it keeps the new data in the most resembling category available.

### **How does K-NN Model Work?**

The following steps explain the basic functionality of the K-NN machine learning model;

1. Selection of the k number of neighbors,
2. Calculation of the Euclidean distance of the selected k number of neighbors,
3. Taking the k nearest neighbors with due respect to the Euclidean distance calculated above,
4. Counting the number of data points in each data category among the k neighbors,
5. Assignment of new data points to categories that have the maximum number of neighbors.
6. Deployment of the developed model.

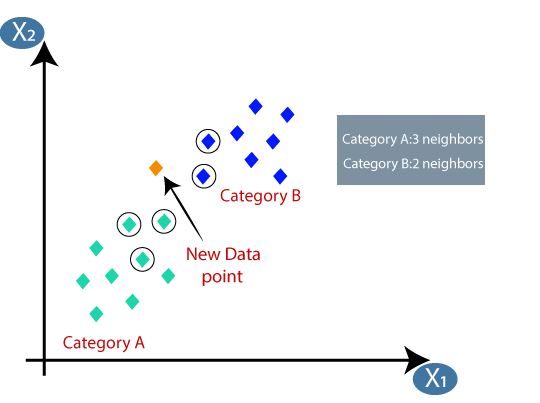


Figure 2: K-NN Model

KNN implementation has proven its special capabilities in that it has quick calculation time which is versatile in both regressions as well as classification purposes. Its versatile ability makes it suitable for imputing missing data values as well as datasets resampling. However, it has high accuracy level meaning that one doesn’t require comparisons with better-supervised machine learning models.

### **Support Vector Machine model**

SVM has been one of the most popularly known supervised machine learning models generally used for both regression and classification tasks. Its main implementation goal has been to create the best line alias as a decision boundary that can be used for n-dimensional space segregation into different classes thus making it easy for someone to put new data points into the correct categories soon. The best line alias decision boundary is generally referred to as hyperplane, (Jakkula, 2006.). The hyperplane lines are created by choosing the extreme points or vectors. The extreme points are known as support vectors, therefore, making the algorithm to be known as the Support Vector Machine model.

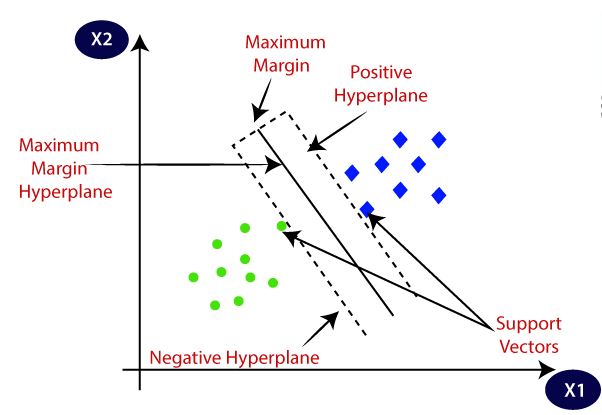


Figure 3: SVM Notation

The proposed research be based on machine learning models which will be used to impact patients’ care delivery schemes. It is through the application of machine learning classification models to the existing data where the detection of patterns associated with heart failure diseases as well as patient records will be analyzed appropriately.

# **Methodology**

Chapter one of the research gives a powerful description of the heart failure research methodologies. Discussion on the background information of every methodology to be applied is required for the proposed project to be fully functioning concerning the research objectives.

## **Classification**

Classification algorithms are under supervised machine learning methodology that is widely used in the identification of categories of new data observations with due respect to the training data supplied. While using the classification methodology, the program is ought to learn from the supplied data in a dataset or the observations then required to make classifications of new observations into different numbers of data classes or data groups. For example, a class can be “Yes” or No” depending on the data applied, (Mishra et al, 2020). These classes can be referred to as targets or labels or data categories. Unlike the regression methodology, the output variable of classification methodology is a category but not a value such as “Red or Black”, “human or animal”, etc. Classification being a supervised machine learning methodology, it operates by taking labeled input data from a dataset and this means that it possesses input with the corresponding data output.

While using classification type of machine learning methodology, it can be noted that a discrete output function (Y) is always mapped to input variable (X), therefore; Y = f (X), where Y is considered to be the categorical data output, (Templin & Bradshaw, 2013). The main objective of the classification model in machine learning tasks is always to provide identification of categories of any supplied dataset. Mostly, the classification models are used for prediction purposes on the output for the dataset categorical data. The heart failure research will use this type of machine learning methodology to achieve the predefined research objectives.

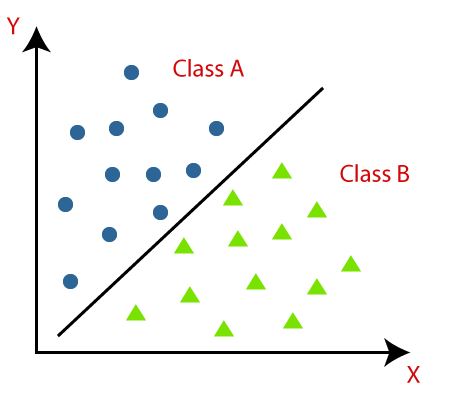


Figure 4: Classification Diagram

The above classification diagram can be used to describe the operation of the classification technique given the two classes “A and B”. The classes have attributes that resemble each other as well as dissimilar to other classes in the supplied dataset.

In any given classification task, the model that is used for classification implementation on the dataset is referred to as a classifier. Classification can be divided into two main types as described below;

### **Binary Classification**

This is always applicable where the classification task has only two possible outcomes. For example, yes or no, horse or donkey, bat or bird, etc.

### **Multi-class Classification**

This type of classification is always applied where the classification task has more than two outcomes. For example, in the classification of types of electronic gadgets, classification of types of heart diseases, classification of types of foods, etc.

## **Classification Tasks Learners**

In any given classification task, there are always two types of learners;

### **Lazy Classification Learners**

Lazy learner models operate by first storing the training dataset as they wait until they receive the testing data. In this type of classification learning, the classification exercise is carried out based on the most relative data stored in the training dataset. It has been discovered that it takes less time in the training phase but much more time for predictions. Examples of lazy learners include the K-NN model and Case-based reasoning.

### **Eager Classification Learners**

Eager learner algorithms develop classification models based on the supplied training dataset before receiving the testing dataset. This is opposite to the normal operation of lazy learners’ algorithms. An eager learner is said to take more time in the learning phase and much less time in the prediction phase. Examples of these classifiers include; Decision trees, naïve Bayes, ANN.

## **Types of Machine Learning Classification Algorithms**

The classification models can be further be classified into two categories;

1. Linear Machine learning Models: This includes; Logistic Regression, Support Vector Machines alias (SVM),
2. Non-linear Machine learning Models: These include; K-Nearest Neighbours alias (K-NN), Kernel SVM, Naïve Bayes, Decision Tree Classification, Random Forest Classification.

# **Design and Implementation**

## **Introduction**

The implementation phase of the entire Heart Failure research included carrying out several tasks. These tasks included data reading, data cleaning, feature selection, data descriptions, data correlation, data plotting as well as data prediction using various machine learning models. The machine learning models gave various results. Jupyter Notebook Application is a client-server-based application that allowed the editing as well as the running of notebook documents via different web browsers. The Jupyter Notebook application was executed on the local machine with no requirement of internet access. However, the notebook application gave the dashboard alias the control panel which was used for data as well as results editing, display, and running of the notebook documents. Jupyter Notebook application contains notebook kernel which is referred to as the computational engine that was used for executing the code written for the heart failure research. Through the use of the Jupyter Notebook, the application was discovered to be powerful in showcasing one’s work in that I was able to see the code and its relative results. However, the running of the application’s code was facilitated by python version 3.9 libraries.

## Data Reading

The dataset used for the research was in the form of CSV. It was discovered that the data was stored using the CSV format since it was discovered that it was the simplest manner that would accommodate the storage of big data sets. However, using the CSV file format was considered to be easy was to store data which was easy for human understanding for it was not encoded or did not require binary conversions before its storage. Also, it is through the research where I discovered that the CSV data format was easy in parsing, fast to manipulate as well as the data could be read with a wide range of text editors. The figure below shows how the data was read using Jupyter Notebook.

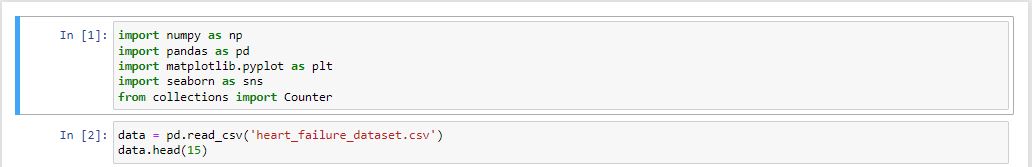


Figure : Data Reading

The following are the results of the read data using the data. head (15) command method which returned only the first 15 rows of the entire dataset. The command method was used to get a quick overview of the data and it returned the respective data.

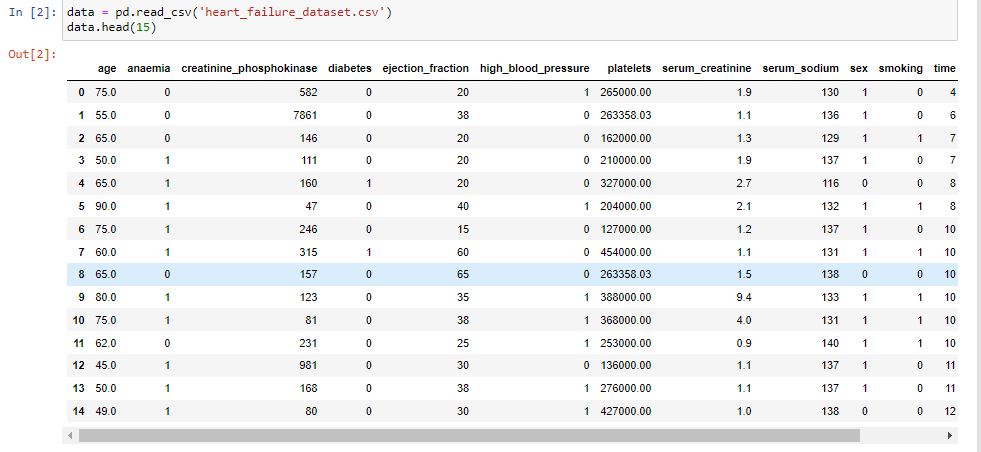


Figure : Head data in the dataset

However, I used the data. tail (10) command method to read the last 10 rows of the heart failure dataset as follows.

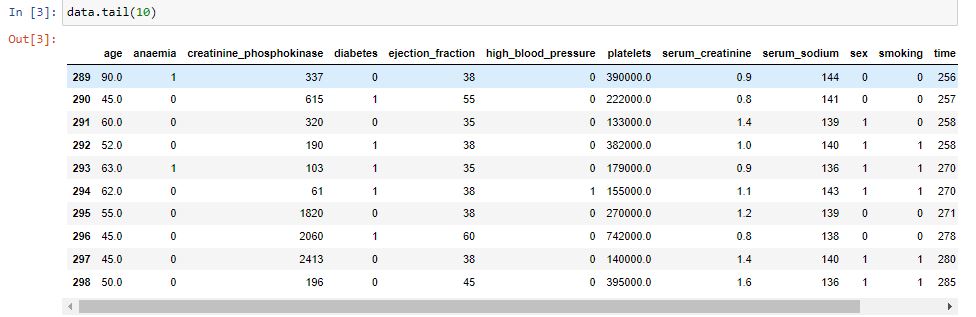


Figure : Data Tail

However, I had to carry out the feature importance identification process which was found to be the score given to the attributes of machine learning models that defined how useful the attributes of the dataset alias features were to the model’s prediction. This was considered to be a useful process later on in feature selection as it provided useful data insights about the applied heart failure dataset.

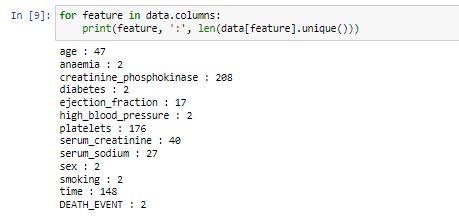


Figure : Features

Identification of attributes was also carried out, the attributes identification meant that the data could be stored either as characters, integers. Being the author of the report, I identified there was the existence of discrete as well as continuous attributes in the dataset and recorded the observations as follows;

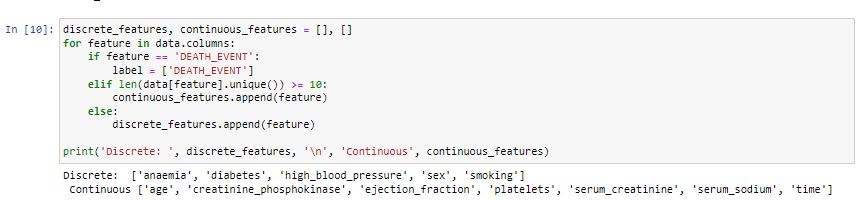


Figure : Discrete and Continuous Datatype

I carried out extra efforts in plotting the discrete and continuous data features domination and the effects towards Death Event of heart failure related issues using the plot function and recorded the respective results as shown below;

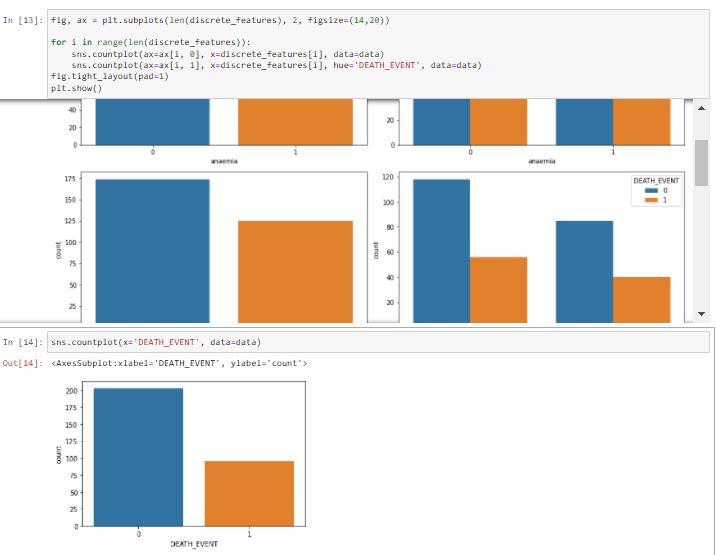


Figure : Discrete Features

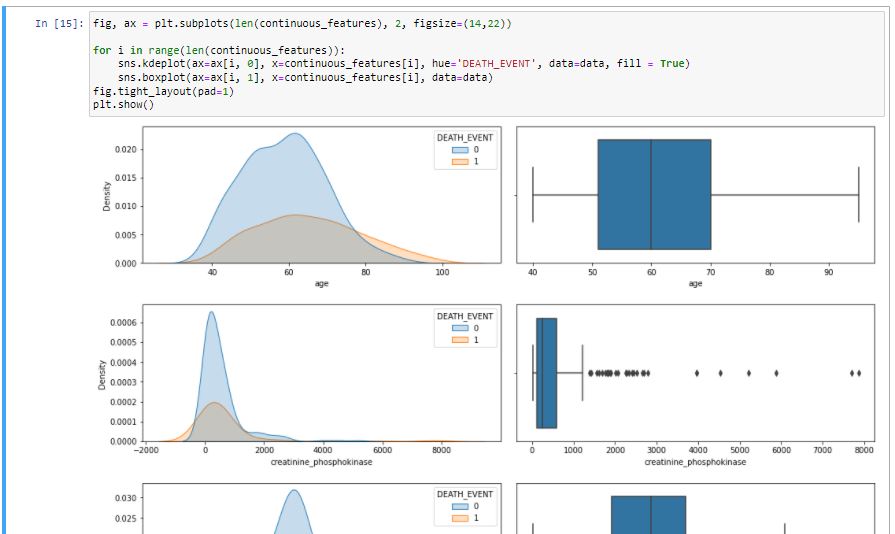


Figure : Continous Features

## Data Cleaning

The heart failure dataset data cleaning was considered as the procedure which was used to provide preparation of raw data for the analysis phase by removing all the bad data as well as organizing the raw data in the dataset. The data cleaning phase prepared the entire dataset for the process of data mining which provided useful dataset insights as well as allowed clear data predictions using various machine learning models. It is through the data cleansing phase where I identified that this procedure is always useful in removing major errors as well as inconsistencies which could later be considered to be inevitable when several sources of data could get pulled into the heart failure dataset. The entire data cleaning procedure was made successful by executing the data.info () command method first to get the information as follows;

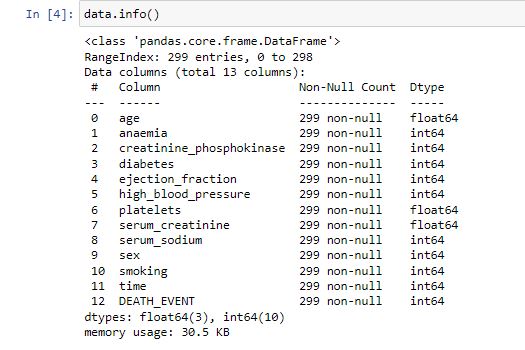


Figure : Dataset information

However, I carried out extensive research on how to remove duplicate data values by applying the data. IsNull (). sum () command to get the total summation of all dataset columns with null data values and the results were recorded as follows.

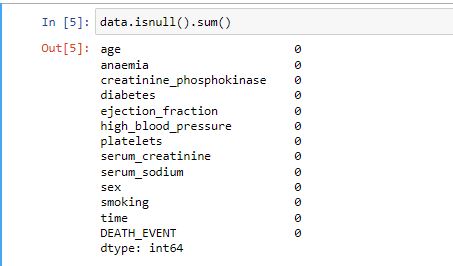


Figure : Data is a null sum

I also had to check for data duplicate values in the heart failure dataset by executing the data. drop\_duplicates (in place = True) command method and there were no results recorded since the dataset had no duplicates from the observations made.



Figure : Drop duplicates method

I also dropped all the missing data values by applying the data. dropna () command method. By dropping the missing dataset values, the following results were recorded.

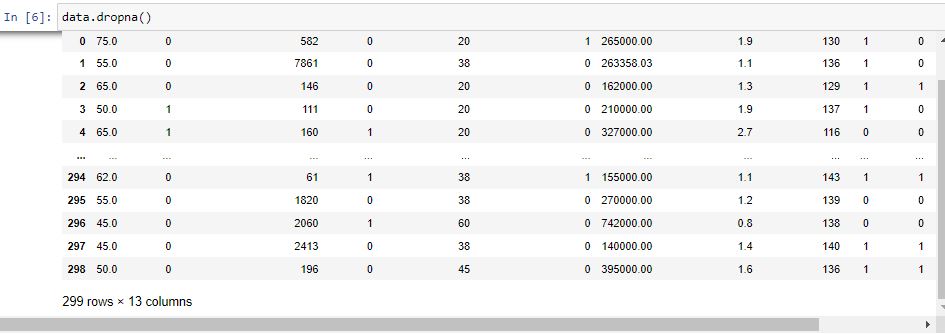


Figure : Drop Missing Values

## Data Correlation

Dataset columns correlation procedure was carried out as a statistical measure that expressed the level to which two or more dataset variables alias as the attributes were observed to be linearly related (this meant that from the observations, the attributes could change together at a constant rate). It was discovered that data correlation is a common method which one could use to describe simple data attribute relationships without making any word statements about the causes as well as effects. However, I identified that the data correlation process speeds up the machine learning model's training time, and therefore, it is always recommended to model the data correlation between the data attributes as it would save one from wasting valuable period.

Data correlation results are always recorded in a table with many numbers that are used to represent how good the attributes relationships are in the entire dataset. The recording numbers in the table vary from – (negative)1 to + (positive)1. The + 1 means that there is a one-to-one relationship which is said to be a perfect data correlation. The data correlation observations were recorded as follows;

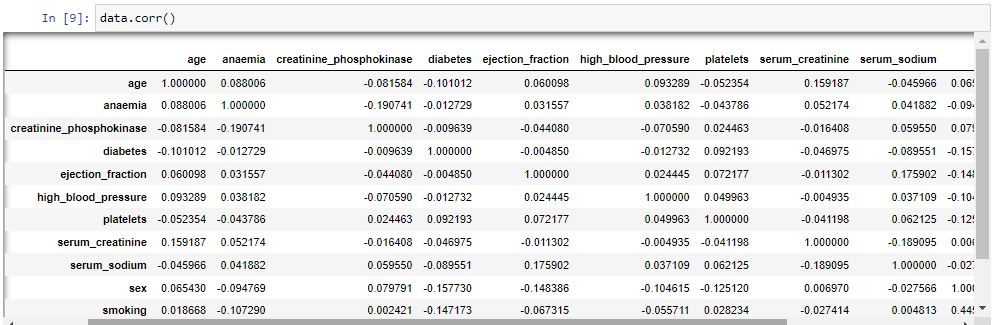


Figure : Data Correlation

However, a correlation heatmap was used to show the 2-Dimensional correlation matrix between two attributes discrete dimensions in the entire heart failure dataset. The heatmap used colored cells that were used to demonstrate data from a monochromatic scale. It was observed that the values of the heatmap’s first dimension appeared as the data rows of the heart failure dataset table while the heatmap’s second dimension represented the columns. The recorded observations were as follows;

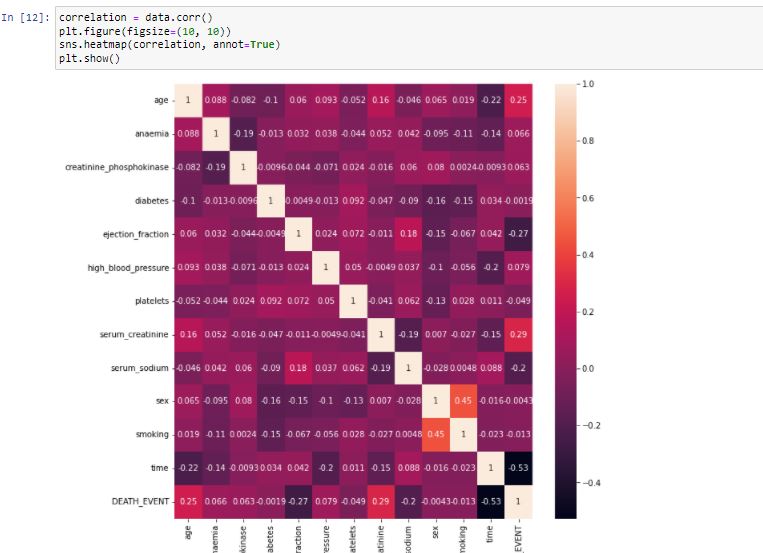


Figure : Correlation Heat Map

## Feature Selection

Feature selection can be defined as the procedure used in reducing the data inputs for processing as well as analysis purposes. Feature selection involves the extraction of useful information or attributes from an already existing dataset. I had to apply feature selection because it offered an effective and efficient way of overcoming the challenge of irrelevant as well as the elimination of redundant data. It is through the removal of irrelevant data from the heart failure dataset where I identified that there was an improvement in the model’s learning accuracy as well as reduction of computations time as this facilitated the enhanced understanding for the machine learning models. From the implemented feature selection procedure carried out by myself, it was discovered that some features in the dataset such as anemia, age, smoking as well as diabetes, and age were less leading to heart failure as compared to the other features. The feature selection was implemented as follows;

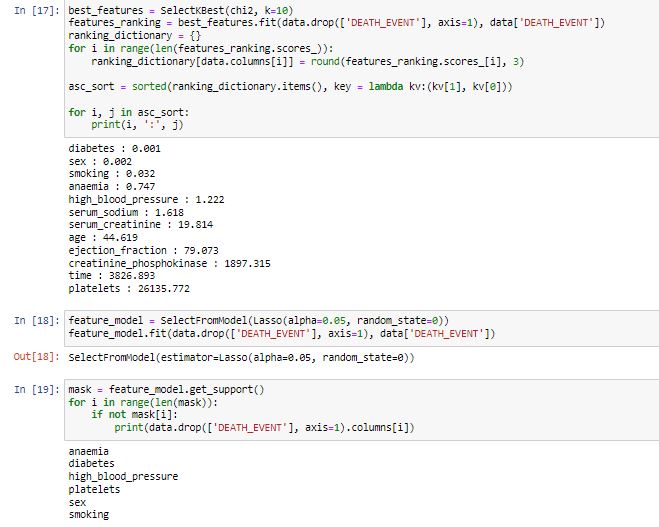


Figure : Feature Selection

## Dataset Outlier Removal

Outliers are considered to be data objects which generally deviate significantly from the rest of the data in a given dataset and they usually behave in various manners. The presence of data outliers in any given dataset can be caused by data measurements or dataset execution errors. Removal of outliers in the dataset was considered to increase the dataset's statistical significance. However, the outlier removal was done on different dataset columns as shown below;

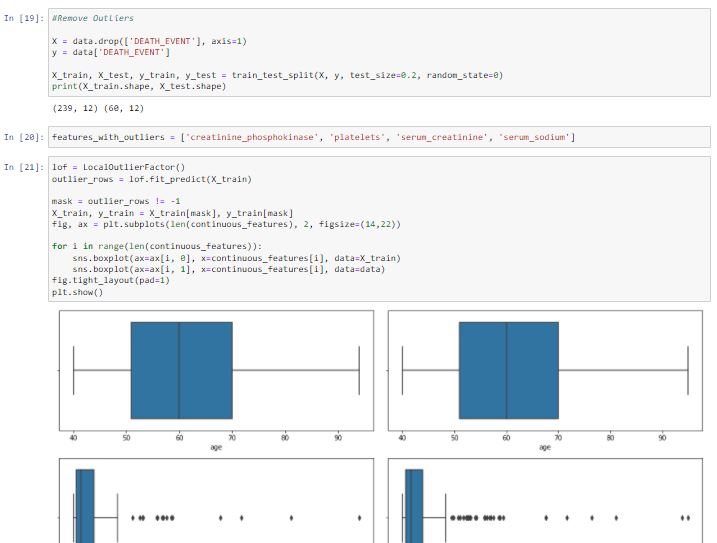


Figure : Outlier Removal

## Data Model’s Building

Data mining models are referred to be created when one applies algorithms to any given date in a dataset though it is more than just an algorithm or a container made up of metadata. However, they can be considered to be sets of data in a dataset, statistics as well as patterns that can be applied to given new data in different datasets to facilitate the generation of data predictions as well as making inferences about data relationships. The heart failure research used both the data model’s building without sampling and the data model’s building with sampling. The entire data models building was made successful by applying various machine learning models such as K-NN, Support vector machines, Gradient Boosting, Logistic Regression, Naïve Bayes, Random Forest, and Decision Trees.

### K- Neighbors Model

K-nearest neighbor is a classification as well as a regression solving machine learning algorithm that is said to be simple in its performance as well as implementation. It is easy to understand its results simply because it is a supervised machine learning model. For one to make a prediction using this type of model for any given new observation, he/she needs to keep all the training data in a dataset since the model is always an instance-based type of a model. This type of model assumes any resemblance between the new data and all the available data and through this, it keeps the new data in the most resembling category available.

### Logistic Regression Model

The logistic regression model representation uses an equation and therefore the input values (x) are linearly combined with the use of weights or the preferred coefficient values (Beta) as this predicts the output value of the model (y). By ensuring the determination of the likelihood factors as expectation, it can be termed as a procedure that explores the connection of the resulting variables together with the self-reliant variables in the binary or else multiple phases. Logistic regression is always termed as a supervised learning classification machine learning model used in forecasting the expectation of any given target variable. The self-reliant variable nature is said to be dichotomous, meaning that there would be the possibility of two classes. The key objective of performing regression analysis at any given point is to explain the variability depending on the type of a variable by the mechanism of one or more of the self-reliant or control variables.

### Support Vector Machines (SVM) Model

SVM has been one of the most popularly known supervised machine learning models generally used for both regression and classification tasks. Its main implementation goal has been to create the best line alias as a decision boundary that can be used for n-dimensional space segregation into different classes thus making it easy for someone to put new data points into the correct categories soon. The best line alias decision boundary is generally referred to as hyperplane. The hyperplane lines are created by choosing the extreme points or vectors. The extreme points are known as support vectors, therefore, making the algorithm to be known as the Support Vector Machine model.

### Gradient Boosting

Gradient boosting is considered to be a machine learning method used for both regressions as well as classification purposes in any given task. I identified that the gradient boosting model relies on the immediate cognition without the use of conscious rational processes that the best as well as possible next model when the combination of the previous models occurs to minimize the general model prediction error. However, the discovered main goal of that gradient boosting model was to set a clear target outcome for the next models, and in this prediction, an error was minimized.

### Data Model Building without Sampling Technique

This procedure was carried out considering all dataset features as well as a few dataset features.

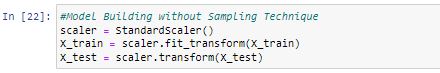


Figure : Model building without sampling

#### Logistic Regression Model

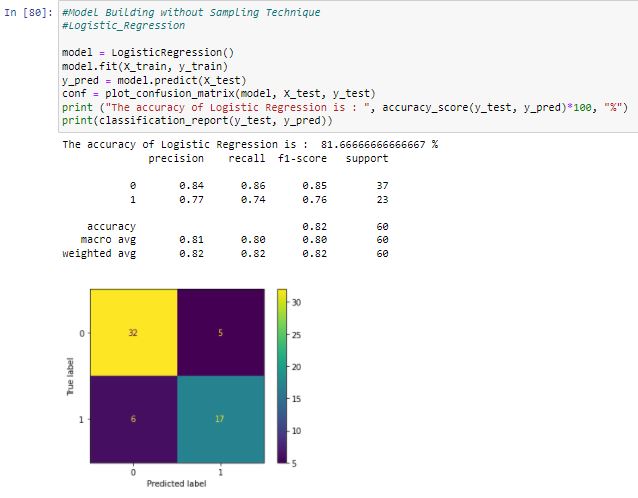


Figure : Model without sampling (Logistic Regression)

#### K-NN

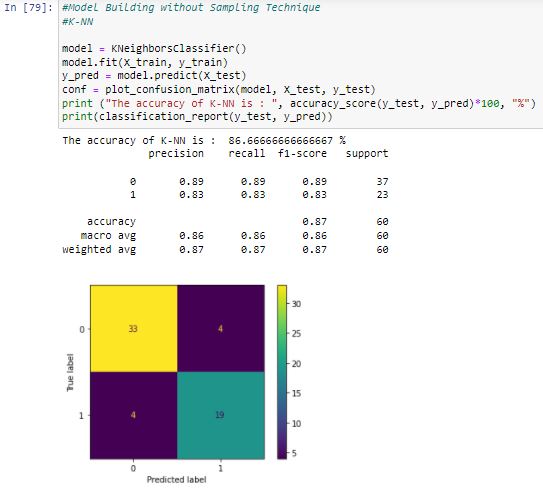


Figure : Model without sampling (K-NN)

#### Random Forest

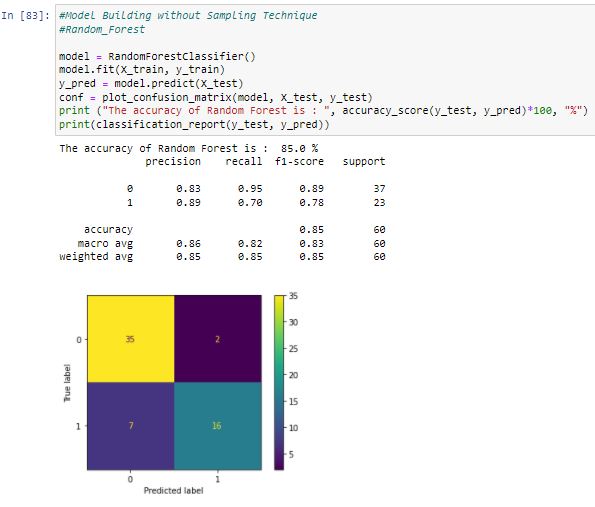


Figure : Model without sampling (Random Forest)

#### Support Vector Machines (SVM)



Figure : Model without sampling (Support Vector Machines)

After carrying out the model building without sampling technique, I recorded the general analysis of the above observations in the table below as Support Vector Machines and Logistic Regression produced promising results of 88% and 87% respectively considering all dataset features;

Table : Model's without sampling Analysis Table

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Model** | **Accuracy** |
|  | K-Nearest Neighbors | 82% |
|  | Logistic Regression | 87% |
|  | Random Forest | 85% |
|  | Support Vector Machines | 88% |

However, I had also to carry out the model building without sampling and performed data prediction considering some of the dataset features. These features were creatinine phosphokinase, ejection fraction, platelets, serum creatinine, and time.

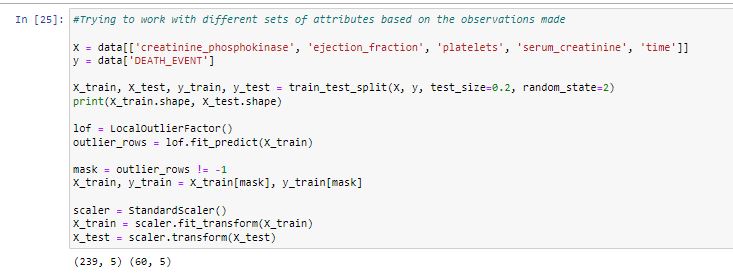


Figure :Model building without sampling (with few features)

The resulting observations after applying different machine learning models were represented below;

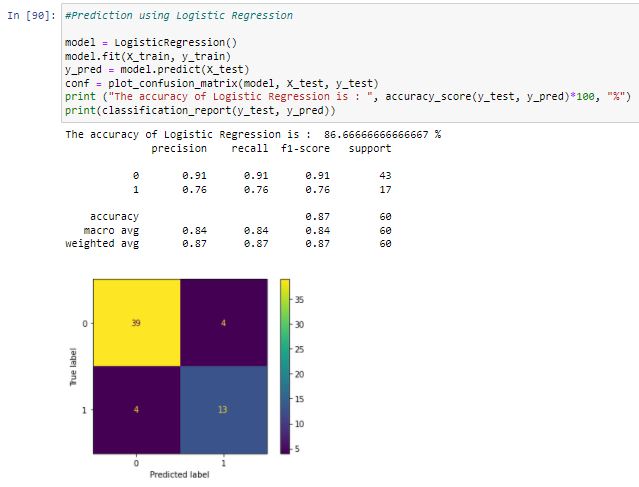


Figure : Prediction using Logistic Regression

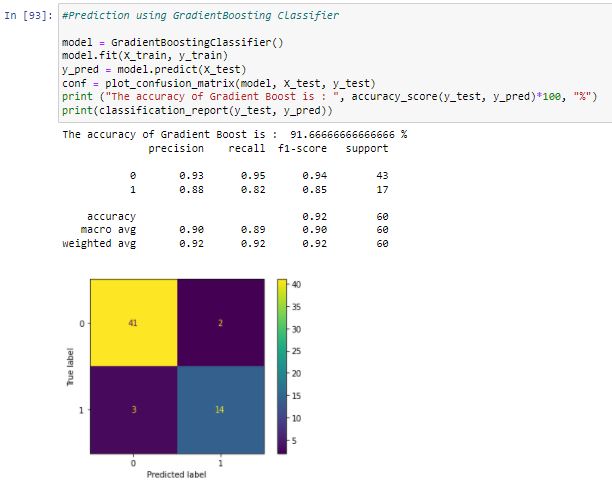


Figure : Prediction using Gradient Boosting

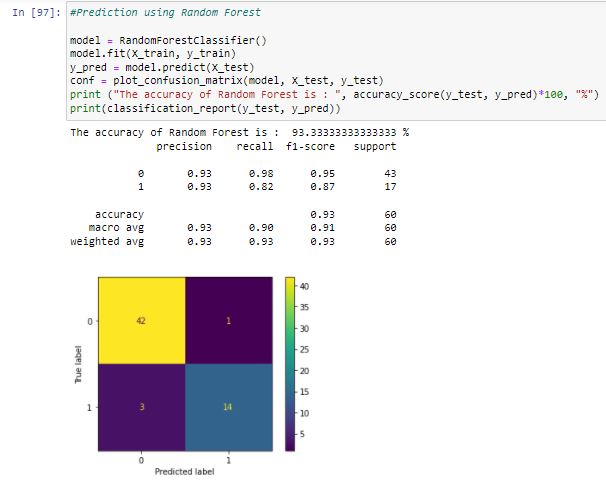


Figure : Prediction using Random Forest

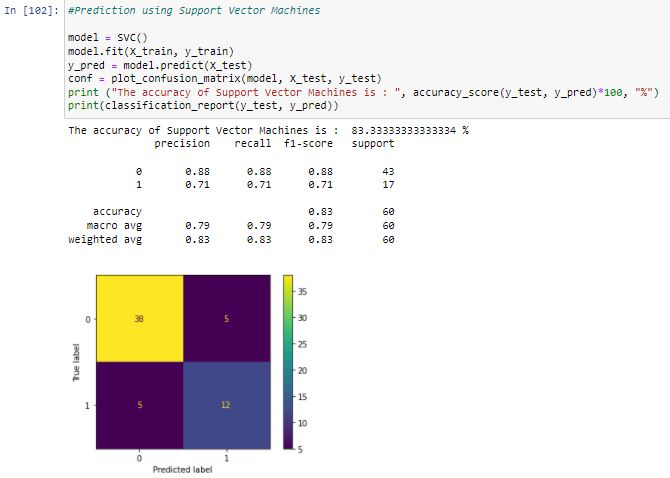


Figure : Prediction using Support Vector Machines

The analysis of the models' output after considering the few dataset features was recorded in the table below as Gradient Boosting and Random Forest models produced promising results in consideration of creatinine phosphokinase, ejection fraction, platelets, serum creatinine, and time features of the data used;

Table : Prediction on correlated dataset features

|  |  |  |
| --- | --- | --- |
| **S. No** | **Model** | **Accuracy** |
|  | Logistic Regression | 87% |
|  | Gradient Boosting | 92% |
|  | Random Forest | 93% |
|  | Support Vector Machines | 83% |

### Data Model Building with Sampling Technique

This procedure was carried out considering all dataset features as well as a few dataset features.

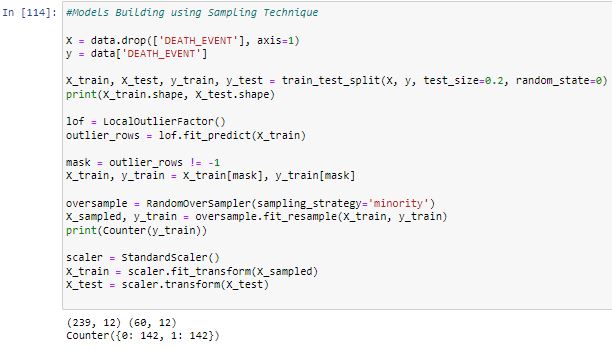


Figure : Model building with Sampling Technique

#### Logistic Regression

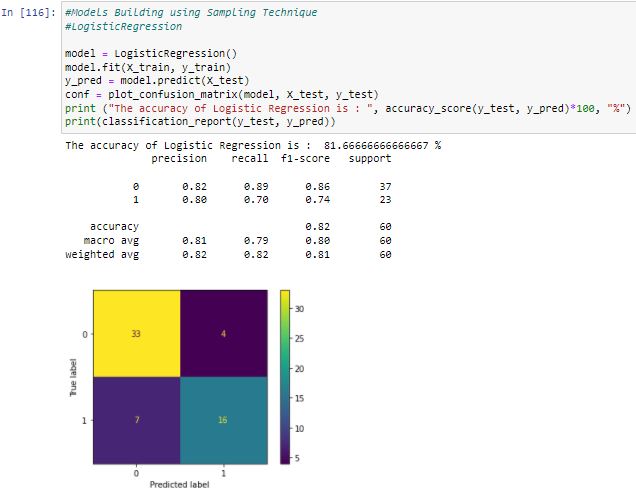


Figure : Prediction with sampling (Logistic Regression)

#### Random Forest

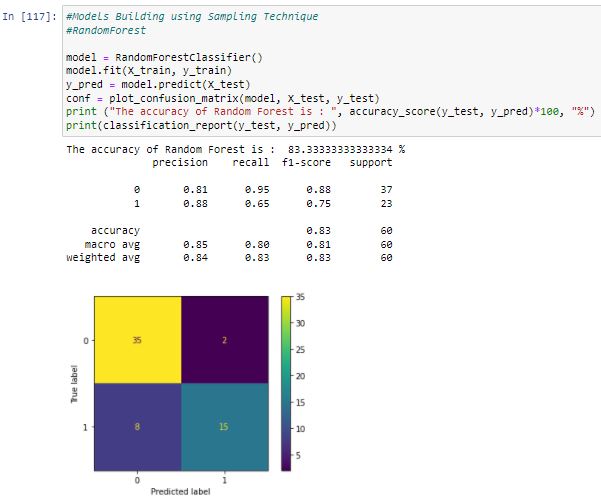


Figure : Prediction with sampling (Random Forest)

#### Naïve Bayes



Figure : Prediction with sampling (Naive Bayes)

Considering all the dataset features as well as the machine learning models applied, the Random Forest model emerged to provide the highest accuracy level. The results were summarized in the table below;

Table : Prediction with sampling Analysis

|  |  |  |
| --- | --- | --- |
| **S. No** | **Model** | **Accuracy** |
|  | Logistic Regression | 82% |
|  | Random Forest | 83% |
|  | Naïve Bayes | 80% |

However, I also had to carry out model building with sampling technique and performed data prediction considering some of the dataset features. These features were; ejection fraction, serum creatinine, and time.

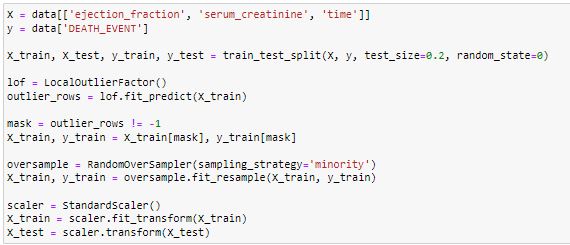


Figure :Prediction with sampling on Selected dataset features

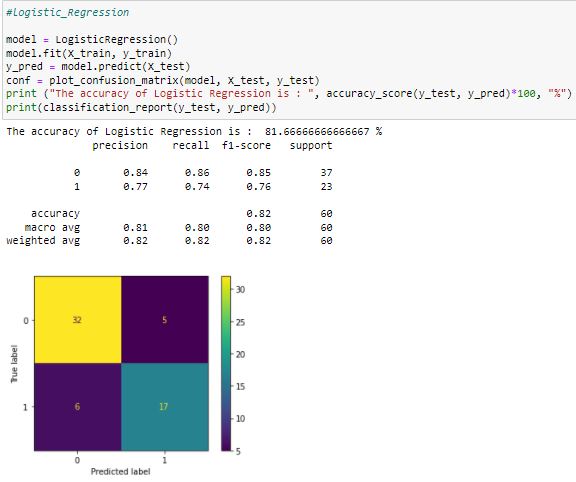


Figure : Prediction with sampling on Selected dataset features (Logistic Regression)

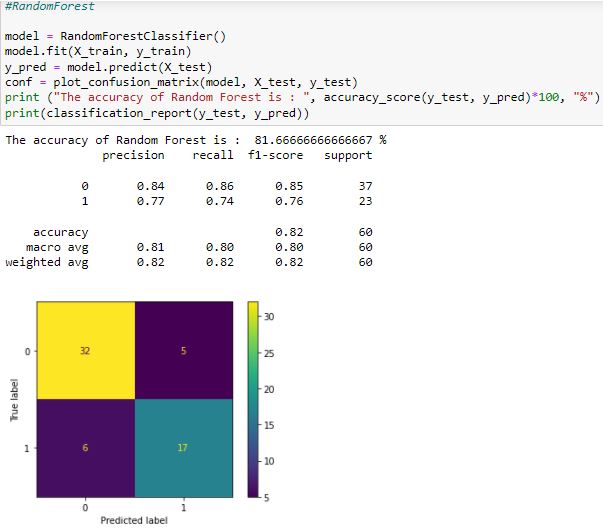


Figure : Prediction with sampling on Selected dataset features (Random Forest)

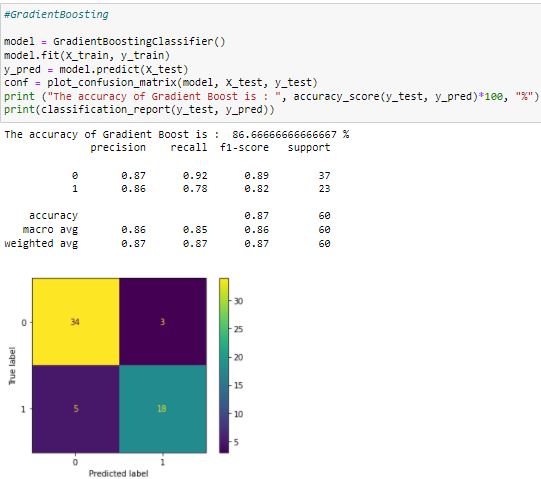


Figure : Prediction with sampling on Selected dataset features (Gradient Boosting)

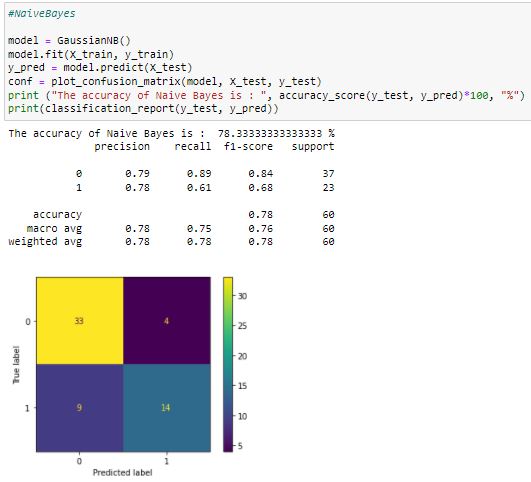


Figure : Prediction with sampling on Selected dataset features (Naive Bayes)

Considering all the dataset features as well as the machine learning models applied, the Gradient Boosting model emerged to provide the highest accuracy level. The results were summarized in the table below;

Table : Prediction with sampling on Selected dataset features Analysis

|  |  |  |
| --- | --- | --- |
| **S. No** | **Model** | **Accuracy** |
|  | Logistic Regression | 82% |
|  | Random Forest | 82% |
|  | Naïve Bayes | 78% |
|  | Gradient Boosting | 87% |

## Prediction

Data prediction was considered to be the algorithm output after several training sessions have been carried out on any given user dataset as well as applied to new datasets. I, therefore, carried out the data prediction procedure using the repeated k-fold sampling on various models and recorded the results as follows;

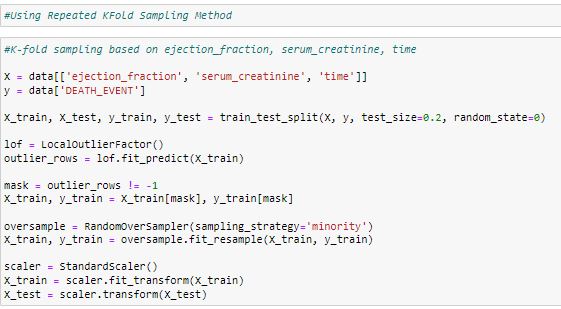


Figure : Repeated K-Fold Sampling/Prediction

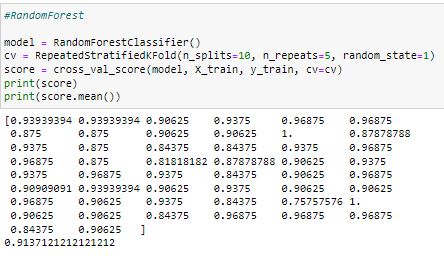


Figure : Repeated K-Fold Prediction (Random Forest)

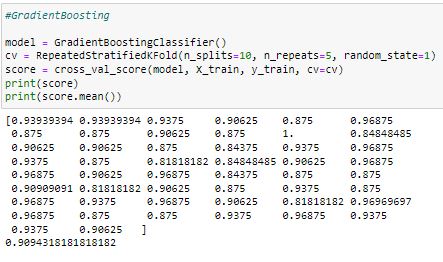


Figure : Repeated K-Fold Prediction (Gradient Boosting)

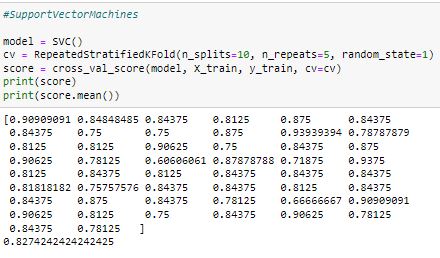


Figure : Repeated K-Fold Prediction (Support Vector Machines)

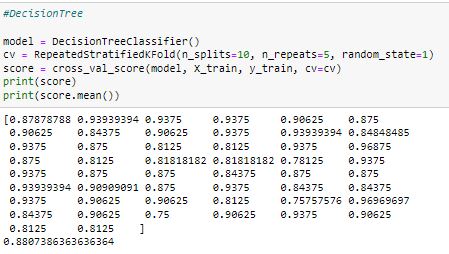


Figure : Repeated K-Fold Prediction (Decision Tree)

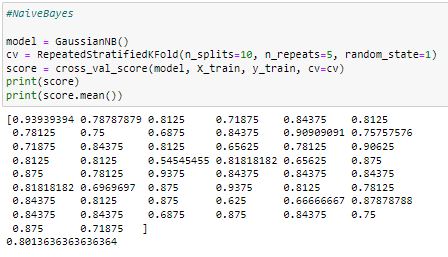


Figure : Repeated K-Fold Prediction (Naive Bayes)

I also carried out K-fold sampling based on ejection fraction, serum creatinine, time to predict Death and recorded the results as follows;

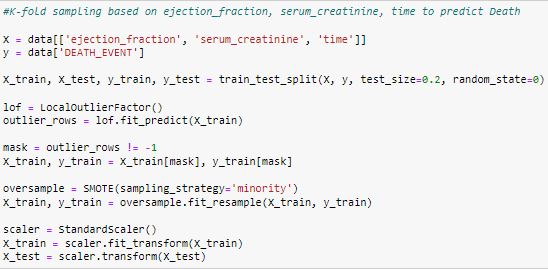


Figure : Death prediction based on features correlation in K-Fold sampling

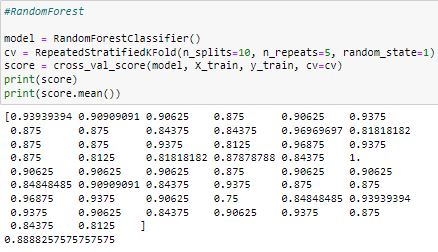


Figure : Death prediction based on correlation in K-Fold sampling (Random Forest)

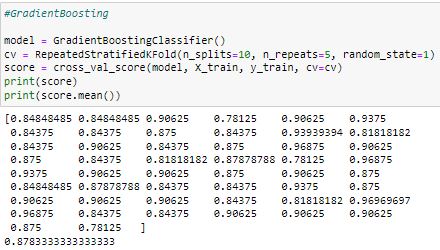


Figure : Death prediction based on correlation in K-Fold sampling (Gradient Boosting)

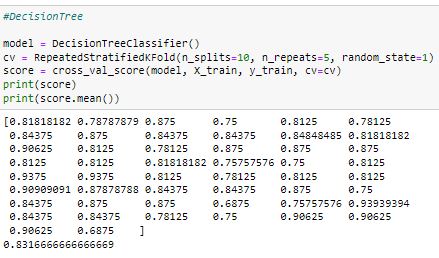


Figure : Death prediction based on correlation in K-Fold sampling (Decision Tree)

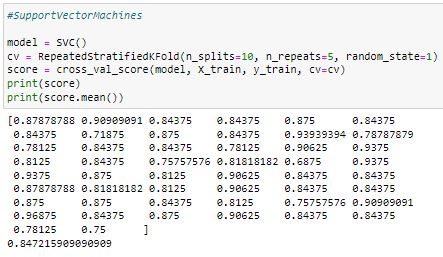


Figure : Death prediction based on correlation in K-Fold sampling (Support Vector Machines)

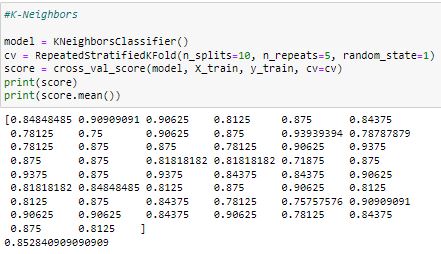


Figure : Death prediction based on correlation in K-Fold sampling (K-Neighbors)

After considering the heart failure models’ results, the following table was designed to summarize all the observations made in K-Fold sampling.

Table 6: K-Fold sampling Analysis

|  |  |  |
| --- | --- | --- |
| **S. No** | **Model** | **Accuracy** |
|  | K-Neighbors | 85.3% |
|  | Random Forest | 88.9% |
|  | Gradient Boosting | 87.8% |
|  | Support Vector Machines | 84.7% |
|  | Decision Tree | 83.2% |

From the above models which were used for the entire heart failure prediction research, it is noticed that Gradient Boosting Model topped to be the best in predicting the model for the heart failure task after yielding an accuracy average of 88.3% in all sampling use cases. The table below summarizes the model’s average accuracy in percentage.

Table 7: Model's Overall Sampling Accuracy

|  |  |  |
| --- | --- | --- |
| **S. No** | **Model** | **Overall Average Accuracy** |
|  | K-Neighbors | 78.0% |
|  | Logistic Regression | 83.0% |
|  | Random Forest | 86.6% |
|  | Gradient Boosting | 88.3% |
|  | Support Vector Machines | 84.4% |
|  | Decision Tree | 83.2% |

## **Project Evaluation and Discussion**

I enumerated the possible characteristics of the heart failure dataset and investigated their effect on the data's predictive accuracy using the machine learning models. While a substantial amount of information is known about the characteristics of the classification as well as regression methods, little was done to determine the performance of the different classifiers applied for the research. It was noted that both the classification and regression methods had varying performances in the heart failure dataset, the classifiers on average were noted to have performed differently and this could have been as a result of the number of folds each classifier took to give out the results. I used overall average accuracy as the outcome of the general accuracy of the entire analysis process, although it was well-known to be a rough measure for the performance of the machine learning models. Accuracy was widely used in practice because of its straightforward interpretation. However, I observed that in a given highly imbalanced dataset, the accuracy percentage may result in over-optimistic results in cases where classification models are used because they might easily send all dataset samples to the majority class. The majority class imbalance was therefore noted to be a key consideration when one is carrying out data interpretation on prediction accuracy. The classification models used in the research, therefore, gave a higher accuracy as compared to the proportion of the majority class of the dataset. However, to deal with the class imbalance problem when measuring the accuracy percentage in the prediction exercise I corrected all random effects data models for the class imbalance levels.

The literature review helped me in, carrying out feature selection which involved the extraction of useful information or attributes from an already existing dataset. I had to apply feature selection because it offered an effective and efficient way of overcoming the challenge of irrelevant as well as the elimination of redundant data. It is through the removal of irrelevant data from the heart failure dataset where I identified that there was an improvement in the model’s learning accuracy as well as reduction of computations time as this facilitated the enhanced understanding for the machine learning models. From the implemented feature selection procedure carried out by myself, it was discovered that some features in the dataset such as anemia, age, smoking as well as diabetes, and age were less leading to heart failure as compared to the other features. However, the dataset columns correlation procedure was carried out as a statistical measure that expressed the level to which two or more dataset variables alias as the attributes were observed to be linearly related (this meant that from the observations, the attributes could change together at a constant rate). It was discovered that data correlation is a common method which one could use to describe simple data attribute relationships without making any word statements about the causes as well as effects. However, the research helped me to note that the data correlation process speeds up the machine learning model's training time, and therefore, it is always recommended to model the data correlation between the data attributes as it would save one from wasting valuable period.

# **Conclusion and Recommendations**

## Conclusion

For a long time, human health has been experiencing several challenges on how to deal with heart failure problems. According to the WHO statistics on heart failure, at least 1.4 million people living with heart failure conditions are under 60 years of age, only 2 percent of people of ages 40 to 59 live with this critical condition across the whole world, (WHO, 2021). The research is based on machine learning models which will be used to impact patients’ care delivery schemes. It is through the application of machine learning classification models to the existing data where the detection of patterns associated with heart failure diseases as well as patient records have been analyzed appropriately and gave out the appropriate insights about the collected data from different health centers. This research was based on most of the popular models for classification in machine learning like the Logistic Regression model, K-Nearest Neighbor model, Support Vector Machines model, Decision Trees model, Bagging and Boosting models. For this case a dataset containing different attributes of patients to predict heart failures was used.

The research was a data science-driven research and the following were the defined and achieved research objectives;

1. To collect data required: Online surveys to collect open-source heart failure datasets were carried out respectively.
2. To process the collected data: This included data cleaning process activities such as describing statistically the data attributes, removing data outliers, duplicates as well as missing values.
3. To carry out feature engineering: This involved scaling numeric features, encoding categorical features, changing back dummy variables to numeric.
4. To explore and visualize data: This was done by the application of data visualizing tools used in data science like heatmaps, histograms, etc.
5. To model the data: This involved splitting the dataset into two parts that are, training and testing parts respectively.
6. To analyze and apply to learn: After splitting the dataset into two, the application of various classification machine learning models was applied using different machine learning models.
7. To make decisions based upon insights generated: This was made based on the analyzed machine learning model's output.

From the applied models which were used for the entire heart failure prediction research, it is noticed that Gradient Boosting Model topped to be the best in predicting the model for the heart failure task after yielding an accuracy average of 88.3% in all sampling use cases.

## Recommendations

This research helped in relieving the workflow of predicting patients’ heart failures thus reducing the risks of treatment failure. However, the applied machine learning models will be utilized countless times giving accurate results depending on the type of dataset applied to them. This means they can learn by themselves and give results based on the data accurately. However, the following are the recommended techniques to be used all through;

1. Classification,
2. Outlier detection,
3. Prediction.

However, before implementing logistic regression division, it has been recommended for one to consider the assumptions listed below;

1. The logistic regression model must have meaningful variables,
2. Choosing large size sample for the logistic regression model is highly recommended,
3. The models’ independent variables need to be self-reliant of each other meaning that there should be no multi-collinearity in the prediction model,
4. The desired outcome of the model is always represented by a factor level of 1 and this means that the target variable must always be in binary format.

This research can be taken further by applying other classification models as well as improving their accuracy levels through performing algorithm tuning which is always based on finding the optimum value for every parameter to increase the models’ accuracy. However, addition of more heart failure related dataset will allow the models’ accuracy level to amplify since the presence of data will make the models to rely on the data rather than assumptions.

# **References**

Biagi, F., and Falk, M., 2017. The impact of ICT and e-commerce on employment in Europe. *Journal of Policy Modeling*, *39*(1), pp.1-18.

Bytyçi, I. and Bajraktari, G., 2015. Mortality in heart failure patients. *Anatolian Journal of cardiology*, *15*(1), p.63.

Casoetto N, Djurdjanovic D, Mayor R, Ni J, and Lee J. (2013)Multisensor process performance assessment through the use of autoregressive modeling and feature maps. J Manuf Syst. 2013;22(1):64.

Chan, J., Rea, T., Gollakota, S., and Sunshine, J. E. (2019). Contactless cardiac arrest detection using smart devices. *NPJ digital medicine*, *2*(1), 1-8.

Cheon JH, Kim A, Kim M, and, Song, Y. (2017). Homomorphic encryption for the arithmetic of approximate numbers. In: Advances in Cryptology–ASIACRYPT 2017: 23rd International Conference on the Theory and Application of Cryptology and Information Security. Springer: p. 409–37.

Choi, E., Schuetz, A., Stewart, W., and Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc. 2016;24(2):361–70.

Coeckelbergh, M. (2019). Artificial intelligence: some ethical issues and regulatory challenges. *Technology and regulation*, pp.31-34.

Dai, W., Brisimi, T., Adams, W., Mela, T., Saligrama, V., and Paschalidis, C. (2015). Prediction of hospitalization due to heart diseases by supervised learning methods. Int J Med Inform. 2015;84(3):189–97.

Daniel, B. (2019). Big Data and data science: A critical review of issues for educational research. *British Journal of Educational Technology*, *50*(1), pp.101-113.

Deng, S., Wang, C., Wang, M., & Sun, Z. (2019). A gradient boosting decision tree approach for insider trading identification: An empirical model evaluation of China stock market. *Applied Soft Computing*, *83*, 105652.

Dietz, E. Application of logistic regression and logistic discrimination in medical decision making. Biom J. 1987; 29(6):747–51.

Etu, E. E. E. (2018). *The impact of machine learning algorithms on benchmarking process in healthcare service delivery*. Wayne State University.

George, G., Osinga, E., Lavie, D. and Scott, B.A., (2016). Big data and data science methods for management research.

Jakkula, V. (2016). Tutorial on support vector machine (SVM). *School of EECS, Washington State University*, *37*.

Jiang, X., Osl, M., Kim, J., Ohno-Machado, L. (2012). Calibrating predictive model estimates to support personalized medicine. J Am Med Inform Assoc. 2012;19(2):263–74

Kaminsky, L., and Tuttle, M. (2015). Functional assessment of heart failure patients. *Heart failure clinics*, *11*(1), pp.29-36.

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., Kripalani, S. (2011) Risk prediction models for hospital readmission: a systematic review. JAMA. 2011;306(15):1688–98.

Kuang, L., Yan, H., Zhu, Y., Tu, S. and Fan, X. (2019). Predicting the duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor. *Journal of Intelligent Transportation Systems*, *23*(2), pp.161-174.

Kunwar, V., Chandel, K., Sabitha, A.S. and Bansal, A. (2016) Chronic Kidney Disease analysis using data mining classification techniques. In *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)* (pp. 300-305). IEEE.

Li, S., Li, W., Cook, C., Zhu, C., and Gao, Y. (2018). Independently recurrent neural network (indrnn): Building a longer and deeper RNN. Proceedings of the IEEE conference on computer vision and pattern recognition

Liu, J., Wang, W., Ma, F., Yang, Y., and Yang, C. (2012) A data-model-fusion prognostic framework for dynamic system state forecasting. Eng Appl Artif Intell. 2012;25(4):814–23.

Lowrie, E., and Lew, N. (2015). Death risk in hemodialysis patients: the predictive value of commonly measured variables and an evaluation of death rate differences between facilities. Am J Kidney Dis. 2015; 15(5):458–82.

Mishra, S., Tripathy, H., Mallick, P., Bhoi, A., and Barsocchi, P. (2020). EAGA-MLP—an enhanced and adaptive hybrid classification model for diabetes diagnosis. *Sensors*, *20*(14), p.4036.

Mortazavi, B., Downing, N., Bucholz, E., Dharmarajan, K., Manhapra, A., and Li, S. (2016) Analysis of machine learning techniques for heart failure readmissions. Circ Cardiovasc Qual Outcomes. 2016;9(6):629–40.

Nikam, S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental journal of computer science & technology*, *8*(1), pp.13-19

Nystad, B., Gola, G., and Hulsund, J. (2012) Lifetime models for remaining useful life estimation with randomly distributed failure thresholds. In: First European conference of the prognostics and health management society, vol. 3; 2012.

Perveen, S., Shahbaz, M., Guergachi, A. and Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*, *82*, pp.115-121.

Ross, J., Mulvey, G., Stauffer, B., Patlolla, V., Bernheim, S., Keenan, P., and Krumholz, H. (2018) Statistical models and patient predictors of readmission for heart failure: a systematic review. Arch Intern Med. 2018;168(13):1371–86.

Scassa, T., and Taylor, F. (2017). Legal and ethical issues around incorporating traditional knowledge in polar data infrastructures. *Data Science Journal*, *16*.

Templin, J., and Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), pp.251-275.

Toback, M., and Clark, N. (2017). Strategies to improve self-management in heart failure patients. *Contemporary Nurse*, *53*(1), pp.105-120.

Wang, Y., Ng, K., Byrd, R., Hu, J., Ebadollahi, S., and Daar Z. (2015). Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records. In: 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC): IEEE; 2015. p. 2530–3.

Yu, Y., Si, X., Hu, C., and Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*, *31*(7), 1235-1270.

<https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1>