Title Page

AN ARTIFICIAL INTELLIGENCE APPROACH TO PREVENTING AND PREDICTING HEART ATTACK IN THE GAMBIA USING MACHINE LEARNING PREDICTION MODELS.

BY

YUSUPHA KINTEH

Matriculation No: 22011002

yk22011002@utg.edu.gm

A Project submitted to the Department of Computer Science, School of Information Technology and Communication as part of the requirements for the award of a Bachelor of Science (B.Sc.) Degree in Computer Science from the UNIVERSITY OF THE GAMBIA.

AUTHORIZATION TO COPY

Certification

I certify that this work was carried out by Mr. Yusupha Kinteh in the Department of Computer Science, School of Information Technology and Communications, the University of the Gambia.

……
……
……………………………………………………………………………………………

Supervisor

Mr Kutobo Jaiteh, Bachelor’s Degree in Telecommunications Engineering UNEFA Maracay, Venezuela. Masters in Electronics Engineering Uludag University, (in Bursa, Turkey) and

Dedication

This work is entirely dedicated to my late father Mr. Lamin Kinteh who passed away on the 07th of September 2023. He was an inspiration and a shoulder to lean on throughout my educational career. I saw him leave this world at kanifing general hospital while I was unable to do anything. Due to lack of sufficient doctors, we were only able to confirm his death in the morning when the only doctor assigned to the ward was around.

This work won’t be completed without the special effort, understanding and support from the entire Lang Kinteh Mai Family. My mother Mrs. Sally Konateh and my siblings. They were the beacon through which this work was completed.

His only wish was for me to become a medical doctor but since I have chosen technology, I therefore want to use my knowledge gained in technology to contribute to the field of medicine.

May Allah (GOD) forgive his shortcomings and grant him Jannah.

ACKNOWLEDGEMENT

I thank Allah all mighty for giving me the strength and health to carry out this noble work.

I warm heartedly thank my Supervisor Mr.Kutobo Jaiteh, for his unwavering support throughout the course of this study

I love to thank Mr. Assan Jahateh, faculty officer at school of information technology whose valuable support in assisting and tailoring down the entire research process into easy fathomable chunks.

My special thanks goes to the dean of the school of information technology and communication, Dr. Mbemba Hydara for his insightful comments and method of teaching research methods as a course.

A blushing thanks goes to Mr. Dembo Minteh for his valuable support and dedication in ensuring that this project is a success.

I will thank my Artificial Intelligence and big data tutor, Dr. Bamfa Ceesay, HoD of the school of ICT for shaping my thoughts towards AI.

Finally, congratulations to Dr. Lamin Jaiteh, Head of the Cardiology Unit in Banjul, for his support and valuable assistance.

List of Figures

List of Acronyms

1. AI	Artificial Intelligence
2. ML	Machine Learning
3. ACS	Acute Coronary Syndrome
4. AKI	Acute Coronary Injury
5. CKD	Chronic Kidney Disease
6. GDP	Gross Domestic Product
7. ROC	Receiver Operating Characteristics
8. AUROC	Area Under Receiver Operating Characteristics
9. RRT	Renal Replacement Therapy
10. SVM	Support Vector Machine
11. BleeMAC	Bleeding Complications in a MultiCenter Registry.
12. MI	Myocardial Infarction
13. AMI	Acute Myocardial Infarction
14. XGBoost	Extreme Gradient Boosting
15. CP	Constrictive Pericardium
16. CVD	Cardiovascular Disease
17. AA	Arachidonic Acid
18. HF	Heart Failure
19. SMOTE	Synthetic Minority Oversampling Technique
20. IHD	Ischemic Heart Disease
21. DCM	Dilated Cardiomyopathy.
22. HRV	Heart Rate Variability.
23. LVEF	Left Ventricular Ejection Fraction
24. HRFLM	Hybrid Random Forest with Linear Model
25. RL	Reinforcement Learning.

Abstract

Hospitals in the Gambia and Sub-Saharan Africa have a huge gap in their doctor to patient ratio. This gap can be closed through training more medical specialist throughout the region but there is viable option which is using Artificial Intelligence. This study aims or investigates the use of machine learning algorithms as an AI tool to predict heart attack and thereby preventing heart attack in the Gambia.

The data used in this research is obtained from Kaggle which makes it a secondary data. After preprocessing the dataset which is removing missing values, duplications and outliers, different machine learning algorithms are used such as Logistic Regression, K-Nearest Neighbor, Decision Tree, and Random Forest. The dataset is divided into train and test part, each model is trained on the dataset and then get tested with new separate data of the same format to make predictions.

The accuracy of the results is test by an analysis metric from which it is found out that only K-Nearest Neighbor algorithm produce an accuracy score of 95% while the rest scored 100%. This shows that machine learning could be used as an artificial intelligence approach to predicting and preventing heart attack. Additionally, it is shown that some of heart disease conditions could increase the chances of attack such as cholesterol level and trestbps.

The results indicate that Machine learning algorithms could be used in predicting heart attack. On that note, it is recommended that hospitals in the Gambia start digitalizing their records, collaborate with AI enthusiasts and explore hybrid algorithms. The limitations set by the use of secondary data made it important for further research to be conducted to strengthen the effectiveness of this study using primary dataset.

CHAPTER ONE: INTRODUCTION

1.1 BACKGROUND TO THE STUDY

It is self-evident that quality health care services are essential in human development and crucial in any Sustainable Development Goals (SDGs). A healthy nation is a healthy working population that can strive as a driving force of a country's development. Of late, artificial intelligence technologies has played a vital role in the health care system of developed countries such as China and the USA, and emerging economies such as India, Brazil etc. It positively impacted the speed of diagnosis, prediction and prevention of diseases which marvelously set the pace for improved medical services. This research is centered on providing innovative technology solution in the cardiology sector of the Gambia using Machine Learning algorithms.

The Gambia is a small West African country bordering on the Atlantic Ocean and surrounded on three sides by Senegal. It comprises 10,689 square kilometers (about half the area of New Jersey) of land, is home to an estimated 2.3 million people (UN, 2019), and has a density of 176 people per square kilometer [1], 39.8% of its population resides in the rural area while the whopping remaining 60.2% live in the urban areas. Interestingly, it has only 6% of its GDP as total health expenditure, life expectancy at birth 62.3, and 59.6 for women and men, respectively. Regarding the doctor- to -population ratio, the density of physicians across the country is 0.11 per 10,000 populations. Similarly, the density for midwives and nurses is 0.87 per 10,000 populations. Additionally, The Gambia has an infant mortality rate of 75/1000 live birth, an estimated 2maternal mortality ratio at 75/100,000 live birth. (Jaw et al., n.d.). The Gambia been among the countries with the lowest doctor-to-population density of 0.11 per 10,000 populations, the need to mitigate these challenges is of great significance. (Jaw et al., n.d.).

This is massively challenging to the physician's efficiency du
e to the inevitable possibility of dealing with too many patients at a go.

A similar approach was in the Gambia for the prediction of mortality rate in children under the age of five with clinical pneumonia in rural Gambia and their result is stated “When we applied the final model to the test set (55 deaths), the area under the Receiver Operating Characteristic Curve was 0.88 (95% confidence interval: 0.84, 0.91), sensitivity was 0.78 and specificity was 0.77.”(Jarde et al., 2021). In their conclusion, it is indicated “Our evaluation of multiple machine learning methods combined with minimal and pragmatic feature selection led to a predictive model with very good performance. We plan further validation of our model in different populations.” (Jarde et al., 2021).

Armoring doctors with machine learning technology does not only improve the efficiency of physicians but improve their capacity of predictability. With selected key features in heart attack conditions doctors can give diagnosis or possibility of having a heart attack in the

Gambia while averting the risk of diagnostic errors because diagnostic errors have been a massive challenge for Global Health care quality and safety, and The Gambia is not an exception. Research has proven an estimated 5.08% of outpatient diagnostic errors in the United States, which is about 12 million adults yearly. Furthermore, half of these errors were estimated to be harmful to the patients [7]. (Jaw et al., n.d.).

Therefore, this research is centered on preventing and predicting heart attack in the Gambia to improve the speed, efficiency and analytical capacity of our doctors.

1.2 STATEMENT OF RESEARCH PROBLEM

1.2.1 Ideal Situation

Diagnostic errors described in section **1.1** could have been avoided if there was an availability of machine learning models that helps doctors to correlate predicates of a disease to the diseases itself. **ML** models can facilitate this process and minimize diagnostic errors; and a low doctor to patient ratio as described in section **1.1** could be reduced if Gambian doctors are using ML models to assist them in disease prediction thereby reducing the workload on doctors.

1.2.2 Reality on the Ground

Since there are diagnostic errors in the most developed country in the world, the United States then one could say without fear of err that the Gambia has recorded an even higher rate. The efficiency of doctors in the Gambia is burdened by the high number of patients assign to a single doctor.

1.2.3 The Way Forward

Heart disease is one of the top leading causes of death in the world and the second leading cause of death in the Gambia. Forty-two (42%) percent of the total deaths in the Gambia is due to ischemic heart disease. This study aims at aiding doctors in preventing and predicting heart attacks using machine learning models to improve their efficacy.

1.3 Research Questions

- ❖ How does age and chest pain type correlate to heart attack?
- ❖ What are the chances of surviving after an individual got affected by certain conditions of heart attack such as “CP”, “Thalack” etc.?
- ❖ Which heart conditions mostly result in heart attack?

1.4 Research Objectives

The aim of this research is to design a ML model that will assist doctors in preventing and predicting heart attack in the Gambia. The objectives include:

1. Determine whether different machine learning prediction models for predicting heart attack such as Logistic regression, Decision tress, Random Forest will be accurate.
2. Determine whether age, chest type results to heart attack.
3. Relate heart conditions that are most likely to cause heart attack.

1.5 Research Hypothesis

Heart diseases or conditions such as age, constrictive pericardium, thalack, cholesterol, slope, old peak etc could be used by machine learning algorithms to predict to heart attack.

1.6 Scope of Study

1.6.1 General Scope

This study focuses on using ML techniques for the prediction and prevention of heart attack in the Gambia. The study aims to provide and explore an alternative way of assisting medical personnel through AI to remedy the current situation of deaths as a result of heart attack. The research covers the geographical location of the Gambia and specifically government owns health centers. It will be conducted within a period of about 2 months, which is from June 15 to August 15. The study will be conducted through both qualitative and quantitative means following an experimental design approach.

1.6.2 Machine Learning Algorithm Scope

A range of machine learning techniques appropriate for heart attack prediction will be investigated in this study. The scope will cover well-known algorithms such as decision trees, Random forests, logistic regression, and K-nearest neighbor.

1.7 Significance / Justification of the study

Machine Learning is a new field, AI technologies in medical institutions of the Gambia are in their kindergarten stage and ischemic heart attack is the second leading cause of death in the Gambia. It is, therefore, a desideratum to assist doctors, and nurses with an **ML** technology that will aid in preventing and predicting heart disease in the Gambia.

1.8 Definition of Key terms:

1. **Machine Learning:** It is a term that describes the ability for machines or computers to learn without being explicitly programmed. Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things.
2. **Heart Attack:** this is a phenomenon that describes the heart situation when flow of blood to the heart is reduced or blocked.
3. **Data- Driven:** This is a decision-making prediction strategy based on data analysis and interpretation.
4. **Dependent Variables:** This is a variable that is dependent on other variables within a given dataset. Its result can only be attained through reliance on other data points.
5. **Independent Variables:** These are variables upon which the dependent variable rely. They determine the value of the dependent variable.

6. **Artificial Intelligence:** This is the broader field of machine learning; it is a concept to create intelligent machines that can simulate human thinking capability and behavior.
7. **Machine Learning Models/ Prediction Models:** are techniques that help to create mathematical models that, without being explicitly programmed, aid in making decisions or predictions with the assistance of sample historical data or training data.

Chapter Two: Literature Review

2.1 Conceptual Review:

Artificial intelligence and particularly machine learning has played a vital and crucial role in the development of modern states and nations ranging from medicine to military. In this time and age one can see an upward trend in the development and improvement of machine learning algorithms for the prediction of events and derivation of meaningful information from millions of terabytes of piled up data. Machine learning techniques and algorithms have been significant AI tools in many domains as ways of solving daunting challenges faced in different societies and contexts. This is clearly evident in the diverse works embedded by scholars of this field. Predictions of phenomena have been around since time immemorial, but the method or techniques of prediction adamantly changes with time and space. Before the advent of AI these techniques were empirical and statistical in nature. These traditional techniques of prediction are however with a series of drawbacks. Today machine learning prediction models have compensated for these drawbacks and catapult prediction ability of humans close to probability of one (1).

Today artificial intelligence seems unavoidable, and many African countries are on the race in implementing it in all sectors of work life. According to (Tapo et al., 2024) in Senegal Crop yield prediction, resilient agriculture, machine learning for rice detection, monitoring artisanal fisheries, predicting road accident severity, estimating electrification rates, analyzing the energy - climate - economy - population nexus and in Nigeria Extensive ML applications including diabetes prevalence detection, crude oil production modeling, flood area prediction, food insecurity prediction, entrepreneurial success prediction, mobile forensics for cybercrime detection, genre analysis of Nigerian music, terrorism activity prediction, stock market forecasting, poverty prediction using satellite imagery.

The demand for renewable energy sources from the ocean soars as the depletion of fossil energy reserves and environmental pollution increases throughout the world. Renewable energy sources from the ocean such as waves and tides because of their high potential energy. On the other hand, the large-scale deployment of ocean energy converters to meet future energy needs requires the use of large farms of these converters, which may have negative environmental impacts on the ocean ecosystem. In the meantime, a very important point is the volume of data produced by different methods of collecting data from the ocean for their analysis, which makes the use of advanced tools such as different machine learning algorithms even more colorful. (Rezaei & Javadi, 2024).

Though machine learning models are used in the health sector of developed countries as everything made by humans, artificial intelligence tends to have disparities in certain scenarios such as Bias. Bias in AI algorithms can arise from biased training data or decision-making processes, leading to disparities in health care outcomes. Addressing bias requires careful examination of the data used to train AI models and implementation of strategies to mitigate bias during algorithm development. (Grzybowski et al., 2024). Specific applications of AI in image recognition include breast histopathology analysis, skin cancer classification, ophthalmologic issues, cardiovascular disease risk prediction, and lung cancer detection (Grzybowski et al., 2024). This research in particular is based on a predictive model given a dataset instead of image.

In the pursuit of substantial and revolutionizing AI systems one must not lose sight of the fact that not all AI systems can guarantee a reliable predictive outcome for all diseases. According to (Joel et al., 2021) there is no ML algorithm that can guarantee a reliable predictive outcome for all kinds of diseases in every given problem case, the quantity of the dataset employed has a significant contribution to the performance of the predictive algorithms, and that quality time must be given to the data preparation stages because the result of the ML algorithms employed depends on the quality of the dataset used.

Chronic Kidney Disease (CKD) is increasingly recognized as a major health concern due to its rising prevalence. The average survival period without functioning kidneys is typically limited to approximately 18 days (about 2 and a half weeks), creating a significant need for kidney transplants and dialysis. Early detection of CKD is crucial, and machine learning methods have proven effective in diagnosing the condition, despite their often-opaque decision-making processes. This study utilized explainable machine learning to predict CKD, thereby overcoming the 'black box' nature of traditional machine learning predictions. (Dharmarathne et al., 2024)

The rapid growth in the use of machine learning techniques has increased in recent years resulting in the diagnosis of often life threating disease such as Parkinson’s disease and atypical parkinsonisms. Parkinson's disease is a neurodegenerative movement disorder associated with motor and non-motor symptoms causing severe disability as the disease progresses. The development of biomarkers for Parkinson's disease to diagnose patients earlier and predict disease progression is imperative. As artificial intelligence and machine learning techniques efficiently process data and can handle multiple data types, we reviewed the literature to determine the extent to which these techniques have been applied to biomarkers for Parkinson's disease and movement disorders. We determined that the most applicable machine learning techniques are support vector machines and neural networks, depending on the size and type of the data being analyzed. (Dennis & Strafella, 2024).

Machine learning is used in hypertention prediction the United States. The findings according to (Silva et al., 2022) the screening of the articles was conducted using a machine learning algorithm (ASReview). A total of 21 articles published between January 2018 and May 2021 were identified and compared according to variable selection, train-test split, data balancing, outcome definition, final algorithm, and performance metrics. Overall, the articles achieved an area under the ROC curve (AUROC) between 0.766 and 1.00. The algorithms most frequently identified as having the best performance were support vector machines (SVM), extreme gradient boosting (XGBoost), and random forest. Machine learning algorithms are a promising tool to improve preventive clinical decisions and target public health policies for hypertension.

Recently, in the Gambia sixty-five children died of acute kidney injury (AKI) due to a drug imported from India. This tragedy infuriated the Gambian population, and the government was scolded for it. This unfortunate tragedy scarred the hearts of Gambians across the globe. In the advent of machine learning AKI could have been prevented or diagnosed earlier than normal to expedite doctors the burden of dealing with too many individuals at a time. According to (Dong et al., 2021) acute kidney injury (AKI) in pediatric critical care patients is diagnosed using elevated serum creatinine, which occurs only after kidney impairment. There are no treatments other than supportive care for AKI once it has developed, so it is important to identify patients at risk to prevent injury. A study has been conducted in the United States to use machine learning algorithm for the early prediction of Acute Kidney Injury. EHR data from 16,863 pediatric critical care patients between 1 month to 21 years of age from three independent institutions, were used to develop a single machine learning model for early prediction of creatinine-based AKI using intelligently engineered predictors, such as creatinine rate of change, to automatically assess real-time AKI risk. The primary outcome is prediction of moderate to severe AKI (Stage 2/3), and secondary outcomes are prediction of any AKI (Stage 1/2/3) and requirement of renal replacement therapy (RRT). Predictions generate alerts allowing fast assessment and reduction of AKI risk, such as: "patient has 90% risk of developing AKI in the next 48 h" along with contextual information and suggested response such as "patient on aminoglycosides, suggest check level and review dose and indication". The model was successful in predicting Stage 2/3 AKI prior to detection by conventional criteria with a median lead-time of 30 h at AUROC of 0.89. The model predicted 70% of subsequent RRT episodes, 58% of Stage 2/3 episodes, and 41% of any AKI episodes. The ratio of false to true alerts of any AKI episodes was approximately one-to-one (PPV 47%). Among patients predicted, 79% received potentially nephrotoxic medication after being identified by the model but before development of AKI. (Dong et al., 2021)

In the world today, heart disease is the second largest cause of death. The need for predicting ischemic stroke outcomes using machine learning models may be useful treatment or contribute to diagnostic process. This approach is highly adapted and useful in the field of medicine due to its high predictable accuracy. In a similar study by (Heo et al., 2019) to investigate the applicability of machine learning techniques to predict long term outcomes in ischemic stroke patients concluded Machine learning algorithms, particularly the deep neural network, can improve the prediction of long-term outcomes in ischemic stroke patients.

In the Gambia the use of ML techniques in health care institutions is very rare while the use of ML algorithms in every domain of life in the 21st century is self-evident of the accuracy and dependability of using ML techniques for the purpose of prediction in every domain of life. In the health sector of the Gambia per say this covet AI technology can improve the efficacy and performance of medical physicians.

2.2 Empirical Review:

This is the part of the literature review that clearly states the opinion, findings and conclusions of researchers to tackle the problem at hand. The essence of an empirical review is to analyze various researchers’ ideas to help find answers to the research questions.

According to the study conducted by (D'Ascenzo et al., 2021) the accuracy of current prediction tools for ischemic and bleeding events after an acute coronary syndrome (ACS) remains insufficient for individualized patient management strategies. Therefore, different machine learning models for the prediction of 1-year post-discharge all-cause death, myocardial infarction, and major bleeding (defined as Bleeding Academic Research Consortium type 3 or 5) were trained on a cohort of 19 826 adult patients with ACS (split into a training cohort [80%] and internal validation cohort [20%]) from the BleeMACS and RENAMI registries, which included patients across several continents. 25 clinical features routinely assessed at discharge were used to inform the models. The best-performing model for each study outcome (the PRAISE score) was tested in an external validation cohort of 3444 patients with ACS pooled from a randomized controlled trial and three prospective registries. Model performance was assessed according to a range of learning metrics including areas under the receiver operating characteristic curve (AUC). The findings of (D'Ascenzo et al., 2021) showed that the PRAISE score showed an AUC of 0·82 (95% CI 0·78-0·85) in the internal validation cohort and 0·92 (0·90-0·93) in the external validation cohort for 1-year all-cause death; an AUC of 0·74 (0·70-0·78) in the internal validation cohort and 0·81 (0·76-0·85) in the external validation cohort for 1-year myocardial infarction; and an AUC of 0·70 (0·66-0·75) in the internal validation cohort and 0·86 (0·82-0·89) in the external validation cohort for 1-year major bleeding. (D'Ascenzo et al., 2021) concluded that a machine learning-based approach for the identification of predictors of events after an ACS is feasible and effective.

A similar study was undertaken by (Khera et al., 2021) to evaluate whether contemporary machine learning methods can facilitate risk prediction by including a larger number of variables and identifying complex relationships between predictors and outcomes. Three machine learning models were developed and validated to predict in-hospital mortality based on patient comorbidities, medical history, presentation characteristics, and initial laboratory values. Models were developed based on extreme gradient descent boosting (XGBoost, an interpretable model), a neural network, and a meta-classifier model. Their accuracy was compared against the current standard developed using a logistic regression model in a validation sample. According to (Khera et al., 2021) none of the tested machine learning models were associated with substantive improvement in the discrimination of in-hospital mortality after AMI, limiting their clinical utility. However, compared with logistic regression, XGBoost and meta-classifier models, but not the neural network, offered improved resolution of risk for high-risk individuals.

In the quest of (Segar et al., 2019) to develop and validate a novel, machine learning-derived model to predict the risk of heart failure (HF) among patients with type 2 diabetes mellitus (T2DM). Using data from 8,756 patients free at baseline of HF, with <10% missing data, and enrolled in the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial, we used random survival forest (RSF) methods, a nonparametric decision tree machine learning approach, to identify predictors of incident HF. The RSF model was externally validated in a cohort of individuals with T2DM using the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). (Segar et al., 2019) concluded that they developed and validated a novel, machine learning-derived risk score that integrates readily available clinical, laboratory, and electrocardiographic variables to predict the risk of HF among outpatients with T2DM.

A prospective cohort study by (You et al., 2023) which sought to identify predictors among a comprehensive variable space, and then employ machine learning (ML) algorithms to develop a novel CVD risk prediction model. From a longitudinal population-based cohort of UK Biobank, this study included 473 611 CVD-free participants aged between 37 and 73 years old. We implemented an ML-based data-driven pipeline to identify predictors from 645 candidate variables covering a comprehensive range of health-related factors and assessed multiple ML classifiers to establish a risk prediction model on 10-year incident CVD. The model was validated through a leave-one-center-out cross-validation. In their findings a novel UK Biobank CVD risk prediction (UKCRP) model was established that comprised 10 predictors including age, sex, medication of cholesterol and blood pressure, cholesterol ratio (total/high-density lipoprotein), systolic blood pressure, previous angina or heart disease, number of medications taken, cystatin C, chest pain and pack-years of smoking. Our model obtained satisfied discriminative performance with an area under the receiver operating characteristic curve (AUC) of 0.762±0.010 that outperformed multiple existing clinical models, and it was well-calibrated with a Brier Score of 0.057±0.006. ML-based classification models can learn expressive representations from potential high-risked CVD participants who may benefit from earlier clinical decisions. (You et al., 2023).

From the above research conducted, it is clearly manifested that the use of machine learning algorithms is an AI approach to assist doctors in diagnosis and predictions of certain life-threatening heart conditions.

Furthermore, we can understand the unmet need in patients for accurate prediction of acute decompensated heart failure or death. A study by (Ma et al., 2022) aimed at developing an AA score to accurately predict mortality in patients with acute decompensated HF and explore the causal relationship between the AA predictors and HF. The serum AA metabolites was measured in patients with acute decompensated HF (discovery cohort n=419; validation cohort n=386) by mass spectroscopy. We assessed the prognostic importance of AA metabolites for 1-year death using Cox regression and machine learning approaches. A machine learning-based AA score for predicting 1-year death was created and validated. We explored the mechanisms using transcriptome and functional experiments in a mouse model of early ischemic cardiomyopathy. According (Ma et al., 2022) in their concluding remarks, propose that the AA score predicts death in patients with acute decompensated HF and inhibiting sEH serves as a therapeutic target for treating HF.

Machine learning algorithms have been used to predict mortality due to certain heart conditions such as acute myocardial infarction. This AI approach of predicting death with the biomarkers induced my AMI is crucial in human existence. This approach of using machine learning techniques to predict mortality due to AMI per se is life saving and can improve the performance of doctors and physicians in predicting mortality. An exploratory study by (Oliveira et al., 2023) investigation created a model based on machine learning for predictive analysis of mortality in patients with AMI upon admission, using different variables to analyze their impact on predictive models. They used a discharged patients' episodes database, including administrative data, laboratory data, and cardiac and physiologic test results, whose primary diagnosis was AMI and found out that for Experiment 1, Stochastic Gradient Descent was more suitable than the other classification models, with a classification accuracy of 80%, a recall of 77%, and a discriminatory capacity with an AUC of 79%. Adding new variables to the models increased AUC in Experiment 2 to 81% for the Support Vector Machine method. In Experiment 3, we obtained an AUC, in Stochastic Gradient Descent, of 88% and a recall of 80%. These results were obtained when applying feature selection and the SMOTE technique to overcome imbalanced data. It was concluded by (Oliveira et al., 2023) that integrating Artificial Intelligence (AI) and machine learning with clinical decision-making can transform care, making clinical practice more efficient, faster, personalized, and effective. AI emerges as an alternative to traditional models since it has the potential to explore large amounts of information automatically and systematically.

2.3 Theoretical Review

Under this section of the literature review a thorough exploration of diverse theories will be conducted to provide substantial justification of this research which will help in a proper explanation of the idea or concept. The section captures the theories, the theorist, year, views, assumptions and justification for the adoption of this approach or theory. This will give a better understanding of the topic which might aid in identifying gaps in current research or suggest new directions for future study.

In article by (Ghasemieh et al., 2023) early detection of heart complications are highly effective in treating patients with cardiovascular diseases. Various machine learning methods have previously been used for the early detection of heart diseases. However, existing data-driven machine learning (ML) approaches fall short of providing efficient and accurate heart disease detection. They discuss the following theories.

Stacking Ensemble Learner (SEL): stacking is a machine learning strategy that combines the predictions of numerous base models, also known as first-level models or base learners, to obtain a final prediction. It entails training numerous base models on the same training dataset, then feeding their predictions into a higher-level model, also known as a meta-model or second-level model, to make the final prediction. The main idea behind stacking is to combine the predictions of different base models to get more extraordinary predictive performance than utilizing a single model.

Behavior Based Systems: are to create a new class label for emergency readmission of patients, which has not been previously explored in existing data-driven machine learning approaches.

Ensemble-Based Detection methods: Ensemble-based techniques are introduced to overcome the limitations of single-based methods and obtain a robust classification. An ensemble combines two or more classifiers with varying strengths/ weaknesses to build a more sustainable model with better performance.

Feature Selection: After preprocessing, there are 719 elements left in the dataset, with 113 features represented. However, to select the best set of statistically significant features in explaining the Emergency class, the features must pass through correlation elimination and nested cross-validation. In the correlation elimination step, highly correlated features are removed until a final set of features with a correlation coefficient of less than 40% is left.

With the aim of differential diagnosis between ischemic heart disease (IHD) and Dilated Cardiomyopathy (DCM), particularly in the early stages of the diseases by (Iscra et al., 2022) put forward the following theories.

Machine Learning Algorithms: Approaches such as linear/logistic regression, classification trees, and naive Bayes models are employed in several sectors of healthcare, such as toxicology [15, 16], endocrinology [18], neurology [16] and cardiology [19–21], due to their high degree of interpretability and ease of use in practice. These algorithms help in identifying the best prediction model is the naïve bays model.

Multivariable Composites: Another method for generating multivariable composites to distinguish two or more groups is to use logistic regression. In general, the sigmoid function argument of a logistic regression classifier can be a linear combination of more than one feature value or explanatory variable. The sigmoid function produces a number between 0 and 1 as its output. The middle value is used as a criterion to determine what belongs in class 1 and what belongs in class 0.

Feature Selection and Classification: The models were built considering selected HRV features together with LVEF. The features were chosen based on their correlation with the target parameter, which was computed using the information gained or the expected amount of information. The features that have information gain of at least 0.025 were considered for further modeling.

2.4 Existing Gap in Literature:

Spurred by advances in processing power, memory, storage, and an unprecedented wealth of data, computers are being asked to tackle increasingly complex learning tasks, often with astonishing success. (Deo, 2015).

According to (L. Jaiteh, personal communication, July 4, 2024) ischemic heart disease is recognized as the second most common cause of death in the Gambia by Global Burden of Diseases (GBD) study, but the information can be accurate with a margin of error because they always estimate while there is no research conducted in the Gambia. The need for machine learning in the Gambia is a necessity and crucial in the medical sector as he is the only one of two cardiologist practitioners in the Gambia. But with the aid of machine learning models, doctors in remote rural Gambia can use the tool to assess cardiologist practitioner.

This project therefore seeks to explore the medical data of the Gambia and fill in the knowledge-gap that existed in the medical sector of the Gambia using machine learning models. Machine learning is used in the medical sectors of many developed and developing countries to improve the work of medical doctors from America, Europe, Asia and Africa but the usage is quite small in Africa and particular in the Gambia. Using machine learning models for heart-Attach prediction is a knowledge-gap that generate the need for **an artificial intelligence approach to preventing and predicting heart attack in the Gambia using ML prediction models.**

2.5 Theoretical Framework:

This section will give a subtle but crucial breakdown of the key concepts of this research work that is substantially essential to the understanding of this work. It will help in explaining and presenting theories and models that other researchers have already developed. There will be evaluation, selection and comparing relevant theories. The concepts and theories that are essential to this research are highlighted below:

Machine Learning: A rapidly developing field of technology, machine learning allows computers to automatically learn from previous data. For building mathematical models and making predictions based on historical data or information, machine learning employs a variety of algorithms. It is currently being used for a variety of tasks, including speech recognition, email filtering, auto-tagging on Facebook, a recommender system, and image recognition. (*Machine Learning: What It Is, Tutorial, Definition, Types - Javatpoint*, n.d.).

It can therefore be summarized as, without being explicitly programmed, machine learning enables a machine to automatically learn from data, improve performance from experience, and predict things.

Artificial intelligence (AI) has transformed key aspects of human life. Machine learning (ML), which is a subset of AI wherein machines autonomously acquire information by extracting patterns from large databases, has been increasingly used within the medical community, and specifically within the domain of cardiovascular diseases.(Al'Aref et al., 2019)

According to (Singh et al., 2018) machine learning is defined as computer-based algorithms that can effectively learn from data to make predictions on future observations, without being explicitly programmed for a specific task or following pre-specified rules.

Main Categories of Machine Learning:

Supervised-learning: In supervised learning, the training data provided to the machines work as a supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.(*Supervised Machine Learning - Javatpoint*, n.d.)

According to (Singh et al., 2018) supervised machine learning is based on task when the model is presented with a labeled dataset also known as feature.

Supervised learning works as in real life human experts, as humans become better and better in their prediction ability based on how long they have worked in a particular field, machines also learn from pass data to draw prediction models that predict future events.

For supervised learning to work there must be a labelled data and classed data that assist the classification models such as decision tree, support vector machine, Naïve Bayes etc.

Classification:

A research article by (Khan et al., 2023) classification is the process of categorizing a given set of data into classes. Classification can be performed for both structured and unstructured data.

Classification algorithm is a type of supervised machine learning algorithm where the dependent data is categorical which is **0** or **1**, **Yes** or **No**, **Spam** or **Not Spam**, **True** or **False** etc.

The historical medical data that is collected from hospitals mostly has two parts, the predictor which is known as the possible conditions that could lead to an event, which is the event. In this study the predictors include chest pain type, cholesterol level, Exercise induce angina etc and the event is **heart attack.**

Unsupervised Machine Learning:

In the comprehensive review of (Naeem et al., 2023) Unsupervised Learning (UL) is a machine learning approach for detecting patterns in datasets with unlabeled or unstructured data points. In this learning approach, an artificial intelligence system gets just the input data and not the associated output data. Unsupervised machine learning, unlike supervised learning, does not need the presence of a person to oversee the model [4]. The data scientist enables the system to learn on its own by looking at the data and identifying patterns.

This can be painted with simple words as saying, this category of ML works on data without been supervised to produce a particular output.

According to (Naeem et al., 2023) unsupervised machine learning is important due to the following reasons:

- There is a lot of unlabeled data
- Data tagging is a time-consuming operation that necessitates human intervention.
- However, ML may be used to drive the same process, making coding easier for everyone involved.
- It can be used to investigate unknown or unprocessed data.
- It comes in handy when dealing with massive data sets and pattern detection.

Unsupervised ML has a clustering, anomalies detection, auto-encoding and association algorithms which are used to identify things which are unknown to us. Example if you are given a dog and a cat photo without label, an unsupervised ML model will generate “clusters” in the case of clustering to identify a cat from a dog based on distinguishable features such as tail, head, claw, foot etc.

Reinforcement Machine Learning:

In a survey by (Coronato et al., 2020) reinforcement Learning (RL), which is a branch of Machine Learning (ML), has received significant attention in the medical community since it has the potentiality to support the development of personalized treatments in accordance with the more general precision medicine vision. The objective of the study is to provide an overview of the applications of RL in healthcare domains, emphasizing the potentialities of this approach to support the development of personalized treatments in accordance with the more general precision medicine vision.

During this survey (Coronato et al., 2020) discern the following applications of RL healthcare sector:

- Precision Medicine
- Dynamic Treatment Regime
- Personalized Rehabilitation
- Medical Imaging
- Diagnostic Systems
- Control Systems
- Dialog Systems, Chat-bots and advanced interfaces
- Health Management Systems.

Analogically, RL is like an individual who newly bought a house but never saw or entered the house, but the moment he enters the house the brain goes with scanning and storing the names of the different facilities and their location available in the house which may be provided by taking a tour of the house or guidance from someone. Therefore, when the person visits the house for the second, he has full knowledge of facilities available and their location without any guidance from anything.
In the world of AI and ML, self-driving cars are using the same methodology. For a self-driving car to drive without the intervention of human agent it must first be taken on test driving tour and on route the self-driving car will store the information and make future decisions based on the information obtained during the self-driving test tour which was guided by human agent.

Prediction Algorithms/ Techniques:

According to recent survey by WHO organization 17.5 million people dead each year is caused by heart diseases. It will increase to 75 million in the year 2030[1]. Medical professionals working in the field of heart disease have their own limitation, they can predict chance of heart attack up to 67% accuracy [2], with the current epidemic scenario doctors need a support system for more accurate prediction of heart disease. Machine learning algorithm and deep learning opens new door opportunities for precise predication of heart attack.(Sharma & Rizvi, n.d.)

Early techniques have not been so efficient in finding it even medical professors are not so efficient enough in predicting heart disease [3]. There are various medical instruments available in the market for predicting heart disease there are two major problems in them, the first one is that they are very much expensive and second one is that they are not efficiently able to calculate the chance of heart disease in human. According to the latest survey conducted by WHO, the medical professionals are able to correctly predict only 67% of heart disease [2] so there is a vast scope of research in the area of predicating heart disease in humans. (Sharma & Rizvi, n.d.)

From the above it can be deduced that the intervention of machine learning algorithms in the medical field to predict heart disease is a desideratum, but identification of the problem is a part of the problem, and the other part is solving the problem. The question remains, what should be done? And the answer lies in machine learning prediction algorithms such as Random trees, Neural- networks, Logistic regression, decision trees, Naïve Bayes etc.

These algorithms will use historical data that have been recorded, stored and cleaned to build mathematical models to make predictions. According to (Sharma & Rizvi, n.d.) each algorithm has its peculiarity such as Naive Bayes used probability for predicating heart disease, whereas decision tree is used to provide classified report for the heart disease, whereas the Neural Network provides opportunities to minimize the error in predication of heart disease. All these techniques are using old patient record for getting predication about new patient. This predication system for heart disease helps doctors to predict heart disease in the early stage of disease resulting in saving millions of lives.

Risk Factors/Predictors:

Risk factors are the factors that act as indicators to determine heart attack. In the realm of ML, risk factors are the variables or features used as dependent variables to generate a mathematical model that will predict heart attack. The risks factors for heart attack are numerous but, in this study, we will be focusing on features such as Cholesterol level, Fasting Blood pressure, chest pain type etc.

This has been beautifully stipulated by (S. Mohan et al., 2019) in their quest to generate an effective heart disease prediction using hybrid machine learning techniques. They stated, it is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been employed to find out the severity of heart disease among humans. The severity of the disease is classified based on various methods like K -Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB) [11], [13].

Due to the reason given above, (S. Mohan et al., 2019) took a different approach to predicting heart disease by introducing HRFLM. According to (S. Mohan et al., 2019) HRFLM makes use of ANN with back propagation along with 13 clinical features as the input. The obtained results are comparatively analyzed against traditional methods [20], [23]. The risk levels become very high, and several attributes are used for accuracy in the diagnosis of the disease [24]. The nature and complexity of heart disease require an efficacious treatment plan. Data mining methods help in remedial situations in the medical field. The data mining methods are further used considering DT, NN, SVM, and KNN. Among several employed methods, the results from SVM prove to be useful in enhancing accuracy in the prediction of disease [25].

From the discussions presented above it will be substantial to give the approach that this research will take to affect its objectives. The different theories that the different researchers have explored include machine learning supervised classification, unsupervised ML and Reinforcement ML.

The Approach:

Using the above information, this research will take a supervised classification machine learning algorithm and it will be conducted in three phases and below is the framework of the paradigm:

Data Exploration:

Data exploration refers to having the heart dataset cleaned from duplicates, null values and establish correlation between features or predictors. Due to the type of ML algorithms that we will be using the dataset must be cleaned and all the entry values be converted to numerals. Duplicate values which will reduce the efficiency of the algorithms will be removed or managed in such a way that it won't hinder the application of the ML models.

The Preferred ML model:

A supervised classification algorithm will be used in contrast to unsupervised and reinforcement ML models. The nature of the dataset demands such an approach since it has two parts which fit in the classification prediction model. The dataset for this research is both labeled and has a dependent and independent part, and the values are all numerals which perfectly fit in any classification algorithm. The algorithms include Random Forest, Decision Tree, Logistic Regression, KNN algorithm, Naive Bayes classifier and Support Vector Machines (SVM).

CHAPTER THREE: METHODOLOGY

This section will explore the different methods and approaches that will be used to achieve the aims and objectives of this research.

3.1 Research Design:

This research is based on quantitative research experimental design, and due to the predictive nature of the research project. The predictive machine learning models or algorithms that will used require a particular design style and this research project will explore the same approach and they are as follow:

3.1.1 Data Source Identification:

This is where the data source is identified and extracted. The step is crucial because there are many data sources available for machine learning therefore one must choose the one that best suits your problem statement.

3.1.2 Gathering Data:

This is the first step of Machine Learning which is essential as it helps in identifying different data sources, collecting data and integrating the data obtained from different data sources. The data obtained from the above task is called a **dataset.**

3.1.3 Data Preparation:

The dataset obtained from step one two above needs to go through a refinery phase known as data preparation. The dataset will now be prepared to be used in a machine learning algorithm. It involves two basic but crucial steps, which are data exploration and data preprocessing. Data exploration is understanding the data and the features while Data preprocessing is removing abnormalities in the data.

3.1.4 Data Wrangling:

The whole purpose of this phase is to convert raw data into machine learning usable format. It is not uncommon that the data we collect is sometimes not in usable format. Therefore, it becomes a necessity to clean the data, which involves removing missing values, duplicates, invalid data and noise.

3.1.5 Data Analysis:

Taking from the above step, the data is now taking to the analysis phase. It involves three basic processes which are selection of analytical techniques, build models and review the results. The aim of the step is to build a machine learning model to analyze using various analytical techniques and review the outcome.

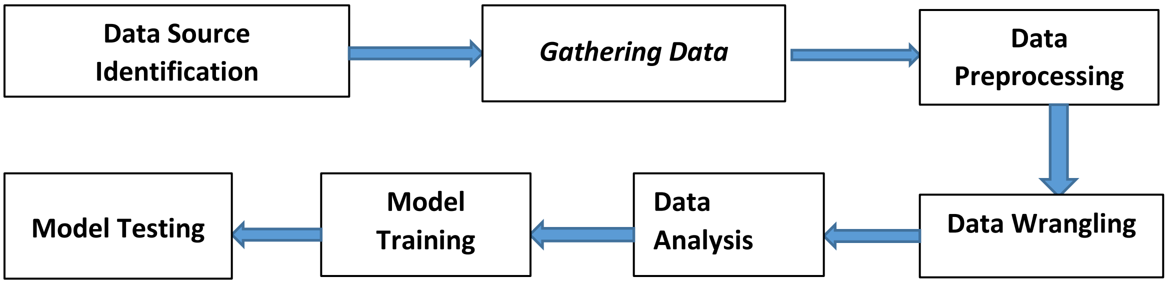
3.1.6 Train Model:

Now the next step is training our machine learning algorithms or models for better performance of the problem. They will help the machine learning algorithms to draw a mathematical model using rules, patterns and features in the dataset.

3.1.7 Test Model:

This is the final stage of the project, which involves testing our model with a new set of data to say whether the model is working properly. This can be confirmed with performance metrices such accuracy scores or precision.

Figure 1: Research Design Cycle



3.2 Area of Study:

The area under study for this research project is the Gambia generally and specifically, general hospitals of the Gambia. The quest is to develop machine learning models that will predict heart disease in the Gambia using secondary data. Since heart disease is the second most common cause of death in the Gambia and around the world.

Special focus will be giving to Banjul General Hospital (EFSTH) where the model will be first tested and after successful completion it will be deployed, and later to be used by all health institutions of the Gambia to provide support to cardiologist and improve their efficacy.

The Gambia been a country in West Africa and the smallest in mainland Africa is home to some 2.08 million people (about the population of New Mexico). The country is surrounded by Senegal on all sides except the west which opens to the Atlantic Ocean. According (L. Jaiteh, personal communication, July 4, 2024) there are only two cardiologists in the Gambia which is disastrous per se. The hospitals in the Gambia are not sufficiently equipped to handle heart disease problems in the country, for that reason most heart patient individuals are sent abroad for treatment. The Gambia compared to other sub-Saharan African countries has a low health budget. It is the 24th most polluted nation globally. The river Gambia also contains some level of Arsenic, and the temperature of the country ranges from 18 degrees to 30 degrees during the dry season and 23 degrees to 33 degrees during the wet season.

The above environmental conditions are all factors for heart disease.

Therefore, this predictive data dependent machine learning project is essential in the socio – economic development of the Gambia and can be a substantial contributing factor to the realization of the sustainable development goals of this nation.

3.3 Population of the Study:

The population of this study involves typically, people with heart disease in the major urban health centers of Gambia. The focus of the study will be focused on heart disease individuals in in the metropolitan. Most of the heart disease study diagnosis are recorded in the metropolitan and for that reason the study will be conducted within that Geographical area.

Though secondary dataset will be used for this application but in the advent of deploying the model for use, the dataset to be used will not be limited to the metropolitan Gambia but to all Government owned health institutions in the country.

There will be a degree of specification in the data collection process. In the data collection process certain features such as cholesterol level, heart bps, etc.

3.4 Sampling Technique and Sample Size:

As per the definition of sampling technique, which is the process of selecting subset of participants from a larger group. Since the population of the study is metropolitan Gambia, the sample size will be limited to the Banjul City Council (BCC) and Kanifing Municipality.

Since the research is also dealing with data, which is secondary and quantitative in nature, the sampling technique to be explored is probability sampling technique. Probability sampling technique involves selecting a group or unit of participants of interest on a statistically random basis. The selection is based on feature identification which is a process-driven approach (a predetermined process for selecting participants). In this case the research findings will help us draw general conclusions for a broader population.

This technique is essential in this research because we will be using a historical heart disease secondary dataset from a group of heart disease patients that will help us build a machine learning model.

Hence the sample size of the dataset is three (300) hundred participants with the examination of heart features such as cholesterol level, bps, chest pain type etc. It could be said that the sample size for such data-driven machine learning models, the bigger the sample size the more accurate the model becomes. Therefore, even a sample size of ten thousand entries will still function well. The entire process is about prediction based on past data mathematical model building or probability rule using historical data.

3.5 Types and Sources of Data Collection:

The type of data that will be explored in this research is secondary quantitative discrete and continues data. At the heart of the research is data. A secondary data is a type of data that is generated from indirect contact with participants, and it is used in this research because currently in the Gambia, the historical data for heart disease patients is not digitalized therefore to test its applicability, a secondary must be explored. The data is quantitative involving on numeral values which will enhance the working of the machine learning models. Since the prediction model to be explored is classification, therefore the data has "target" discrete values.

The source of data collection is kaggle and can be accessed through the link https://www.kaggle.com/ . It is a data repository website for researchers across the world. The data available are mostly incomplete and unclean, thus required by the researcher to explore and preprocess to merge the problem requirement.

3.6 Definition and Measurement of Variables

The demystification of the variables that are available in the research will give a rational and unassail understanding of the problem at hand. The variables are mostly in medical terms which will necessitate a proper description for ease of understanding. At a broader category there are two main types of variable available in this research, namely:

- **The Dependent Variable:** The dependent variable is the variable that is dependent on other variables within the dataset. In this research, the dataset contains features that affect the heart and the "target" feature in the dataset is the dependent variable which gives the information that a heart disease individual is affected. When it reads 0 then the individual is not affected and when it reads 1, it means an individual is affected.
- **The Independent Variable:** The independent variables are all the other variables that the outcome of the dependence variable depends on. They are sometimes called predictors. These variables will define and be explained thoroughly in the table below:

Figure 2: Independent Variable Features description

No	Features	Description	Data Type
1.	***Age***	***Age in year***	***Numerical***
2.	***Sex***	***Gender***	***Numerical***
3.	***CP***	***Chest pain type***	***Numerical***

4.	Trestbps	Resting blood pressure	Numerical
5.	Chol	Serum cholesterol	Numerical
6.	Fbs	Fasting blood sugar	Numerical
7.	Resteg	Resting electrographic results	Numerical
8.	Talach	Maximum heart rate achieved	Numerical
9.	Exang	Exercise induce angina	Numerical
10.	Oldpeak	ST depression induced by exercise relative to rest	Numerical
11.	Slope	The slope of the peak exercise ST segment	Numerical
12.	CA	Number of major vessels colored by fluoroscopy	Numerical
13.	Thal	Number of major vessels colored by fluoroscopy	Numerical

The above table gives a detailed definition of the variable features in the prediction of heart disease. The fact that all the variables are numeral makes the dataset a perfect fit for the supervised classification machine learning model.

3.7 Validity and Reliability of Research Instruments:

The validity and reliability of research is essential to building trust on the data used, thereby assuring reliability on the results or outcomes of the research. There are many research tools at the disposal of researchers. Different research attracts different research tools depending on the data that ought to be collected and analyzed. In this research the data used is secondary and thereby making the research take a different approach. For such data the focus is on the following:

1. **Credibility of Data Source:**

   The data source of the dataset to be used in this research is obtained from Kaggle which is the world's largest data science community with powerful tools and resources to help researchers and data scientist achieve their data science goals.

2. **Publication Date of the Data:**

   The dataset is published within the period of the advent of machine learning. The dataset is therefore within the range of usage for this research.

3. **The purpose of the Data:**

   Since the dataset is obtained from Kaggle that makes it to be typically suitable for data-driven problems such as machine learning prediction models. The preponderance given to this data is because all the identifiers are being removed by Kaggle thus making it fit the problem under review.

Hence, the validity of the research tools will be redirected to the validity of the dataset since the validity of the research tool is a process of confirming the validity of the research data. The validity of the data in this research is therefore assured through the following methods:

- **Assess Relevance of the Dataset:**

  The data to be used is purely designed for machine learning supervised classification algorithms. The dataset is all numerical categorized into a dependent variable and independent variables. Which made it super suitable for research problems.

  - **The accuracy of the Dataset:**

    The accuracy of a data is another measurement of its validity. The dataset obtained is free from null values and errors. It is precise, complete and generalizable for the problem at hand. It scores perfection in completeness and comprehensiveness thus making it a perfect fit for this research.

  - **Review the ethics:**

    Kaggle, a reputable data repository website, makes the data ethical per se. But even most trusted websites can have breaches thereby making it incumbent on a researcher to review the ethics of the dataset. Since the data is completely dependent on individuals' medical information it should be obtained through their consent, the fortunate thing about Kaggle's dataset is that the, the identifiers are removed which is names of patients' data collected.

Likewise, the reliability of the data, can be assured through the following:

- **Analyze the Data:**

  Since there is no primary data for this research, the dataset used must be tested for a certain thing such as suitability and d feasibility. It should be suitable and feasible for research design, methods and tools. By ensuring that the secondary data is suitable for the research design, methods and tools, the more reliable it can be.

3.8 Methods of Data Analysis:

Methods of data analysis in quantitative research in their essence is to measure difference between groups, to assess relationship between variables and test hypothesis in a scientific rigorous way. This research draws on the inferential statistical approach using machine learning models. The models include logistic regression, KNN model, Decision tree etc. These machine learning models will give a thorough analysis of the dataset from which will give a perfect. A descriptive analysis approach will be explored to analyze the content of the sample data. The python language offers a set of libraries that could be used to understanding a dataset of value and some of the libraries to the libraries that will be used as a method of giving descriptive analysis of the data in this research include:

A. **NumPy**:

   This is a library offered by python in Jupiter notebook to include mathematical operations in a code. It is the fundamental package for scientific calculations in python. NumPy is imported in the Jupiter notebook to analysis data using mathematical operations such as mean, average and median.

B. **Matplotlib:**

   It is a library that is used to plot any type of chart in Jupiter notebook to build to give a visual of the relationships between the different variables. It is a 2D plotting library that is imported with a sub-library called **pyplot**. This library will help in this research to give the basic correlation between the independent (predictors of heart disease) variables of the dataset and dependent variables.

C. **Pandas Library:**

   Pandas Library can be termed as the most used python library for importing and managing dataset. It's a free python library for data analysis, ranging from reading dataset to sub-setting. With Pandas Library any dataset to Jupiter Notebook.

D. Seaborn:

Seaborn is a data visualization tool in python and can provide exploratory analysis. It is like an adjustment to matplotlib. It established relation between Mutiple variables. It will be crucial in describing the dataset for this research.

To some up things we can say that the above python libraries provide descriptive analysis of the dataset which in simple terms is providing macro and micro view of the dataset, help identify anomalies in the dataset, identify correlations and inform the which ML algorithm is most suitable for inferential analysis of the dataset.

Furthermore, an inferential approach will be undertaken to give an inferential analysis of the data at hand. The inferential analysis methods that will be used in this research will be predictions on the relationship between variables. The data analysis methods to be used is discussed below:

A) Logistic Regression Algorithm:

This is a popular Machine Learning Algorithms, which comes under the Supervised Learning algorithms. LR is used for predicting categorical dependent variables using a given set of independent variables. This algorithm is essential in this research for inferential analysis of the data. This is so because the dataset is divided into two broad categories which are independent and dependent variables. Which will help in classifying whether an individual has a heart disease or not. It is a greedy algorithm

B) Decision Tree Algorithm:

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.(*Machine Learning: What It Is, Tutorial, Definition, Types - Javatpoint*, n.d.). It is called a tree because it functions as tree, it starts at the root node and down to the leaf nodes. Decision Tree being a classification algorithm will make it suitably good for this research. It will make predictions on the conditions of a patient due to certain features of a patient.

C) Random Forest Algorithm:

Random Forest is like Decision Tree Algorithm, but it is less sensitive to training data compared to it. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. (*Machine Learning: What It Is, Tutorial, Definition, Types - Javatpoint*, n.d.) This makes the algorithm a best fit for classification problems. The greater the number of trees in the forest the greater the accuracy and prevents the problem of overfitting. The model will be used in this research to build a strong relationship between variables and make predictions based on those relationships.

D) KNN Algorithm:

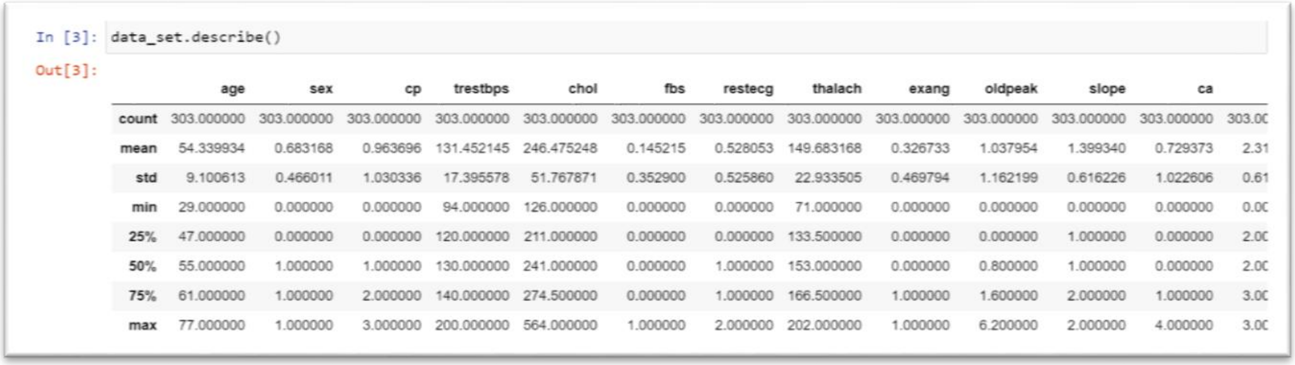
KNN algorithm means K-Nearest Neighbour Algorithm. It is one of the simplest supervised classification algorithms. The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. (*Machine Learning: What It Is, Tutorial, Definition, Types - Javatpoint*, n.d.)**.** This makes it suitable for classification problems as the one encountered in this research.

CHAPTER FOUR: DATA PRESENTATION, ANALYSIS AND DISCUSSIONS

4.1 Socio-Demographic Characteristics of Respondents:

The demographic characteristics of the respondents in this research cannot be determined directly. This is because the data is secondary. But the socio demographic characteristics of the individuals whose data is been collected can be well categorized using python. Below is a snapshot description of the socio-demographic characteristics of the respondents:

Figure 3: illustration of the socio-demographic characteristic of respondents.



```
In [3]: data_set.describe()
Out[3]:
             age         sex          cp    trestbps        chol         fbs     restecg     thalach       exang     oldpeak       slope          ca
count 303.000000  303.000000  303.000000  303.000000  303.000000  303.000000  303.000000  303.000000  303.000000  303.000000  303.000000  303.000000  303.00
 mean  54.339934    0.683168    0.963696  131.452145  246.475248    0.145215    0.528053  149.683168    0.326733    1.037954    1.399340    0.729373    2.31
  std   9.100613    0.466011    1.030336   17.395578   51.767871    0.352900    0.525860   22.933505    0.469794    1.162199    0.616226    1.022606    0.61
  min  29.000000    0.000000    0.000000   94.000000  126.000000    0.000000    0.000000   71.000000    0.000000    0.000000    0.000000    0.000000    0.00
  25%  47.000000    0.000000    0.000000  120.000000  211.000000    0.000000    0.000000  133.500000    0.000000    0.000000    1.000000    0.000000    2.00
  50%  55.000000    1.000000    1.000000  130.000000  241.000000    0.000000    1.000000  153.000000    0.000000    0.800000    1.000000    0.000000    2.00
  75%  61.000000    1.000000    2.000000  140.000000  274.500000    0.000000    1.000000  166.500000    1.000000    1.600000    2.000000    1.000000    3.00
  max  77.000000    1.000000    3.000000  200.000000  564.000000    1.000000    2.000000  202.000000    1.000000    6.200000    2.000000    4.000000    3.00
```

From the dataset the mean age of respondents is fifty-four (54) and the standard deviation of age feature is approximately nine (9). Twenty-five percent (25%) of the respondents are forty-seven (47) years old and Twenty-five percent of the respondents are female. Seventy-five percent (75%) of the respondents are about the age of sixty-one (61). The maximum age is Seventy-seven (77) and seventy-five (75) percent of the respondents are male. This is a step through the descriptive analysis of the dataset.

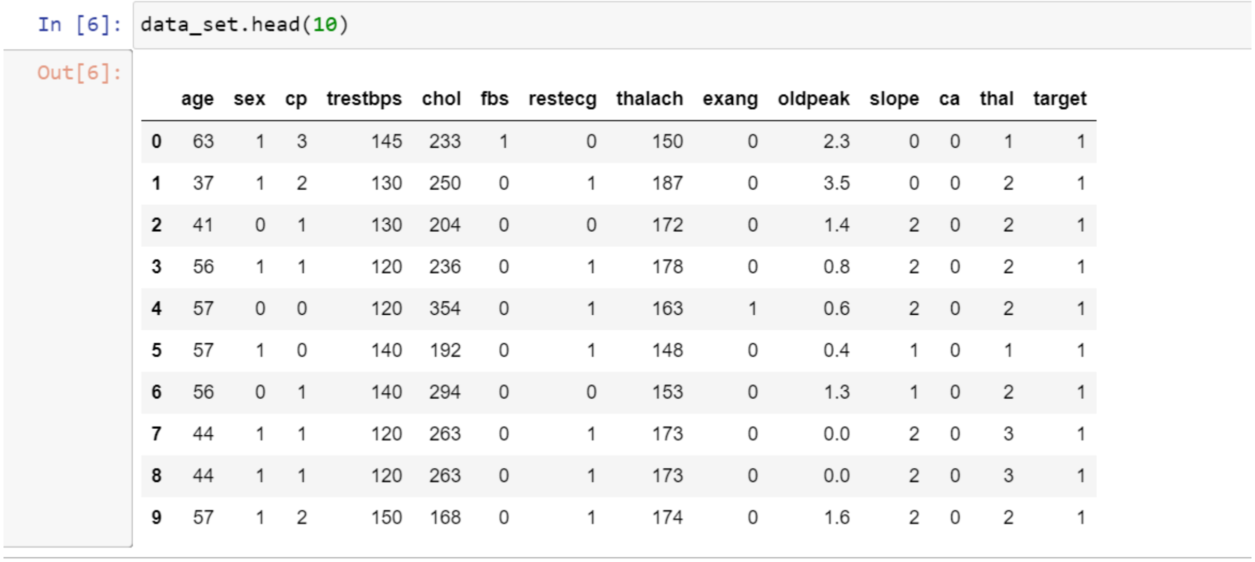
4.2 Data Presentation on Research Issues (Objective by Objective):

Before presentation of the data objective by objective, machine learning algorithms work under clean dataset. This is process in world of machine learning which is known as Data Preprocessing, which basically is a fancy term for cleaning a dataset. Below is series of steps taken to ensure that the dataset is clean and fit to be used for machine learning algorithms.

4.2.1 The Dataset

The dataset of the research that is been undertaken is a secondary dataset from Kaggle data repository. Since the dataset is secondary which makes it unfit for processing without preprocessing. The research will take a holistic approach hence seventy percent (70%) of focus is stressed on the applicability of machine learning algorithms to predict heart disease and twenty five percent (25%) of the research will focus on how efficient the algorithms will be given certain feature engineered dataset. Which will make the dataset that is not from the Gambia irrelevant to the whole research process. Figure 4 below shows a snapshot of the unprocessed dataset.

Figure 4: Snapshot of the Dataset



```
In [6]: data_set.head(10)
```

Out[6]:

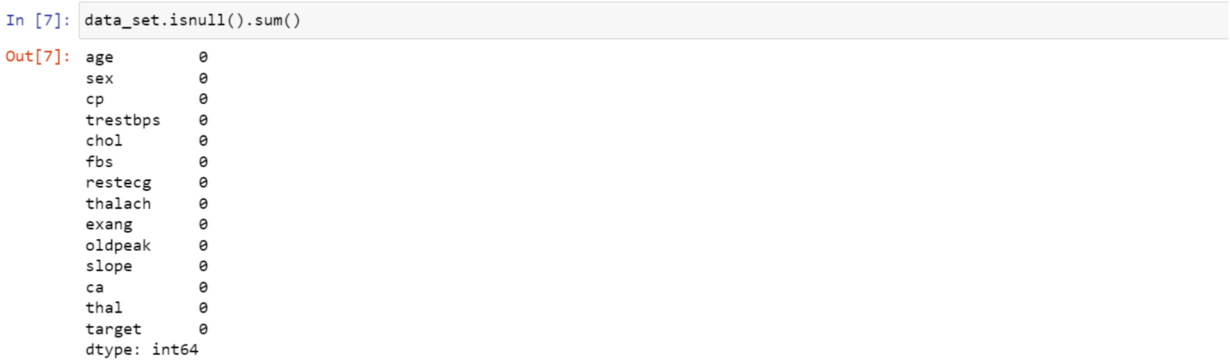
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

4.2.2 Data Preprocessing:

Data preprocessing in its broadest form consist of basically the handling of missing values, normalizing data values to fit algorithm requirements, outlier handling and lastly feature engineering which are discussed below with snapshot of the dataset after each phase:

Firstly, ***handling missing values*** is essential and subtle in the process of predicting outcome when using machine learning algorithms on particular dataset. The python Pandas library offers this feature with ease and accuracy. Below is the code snippet and output of the code in Jupiter notebook:

Figure 5: Output of the Dataset after Invoking the isNull() Method:

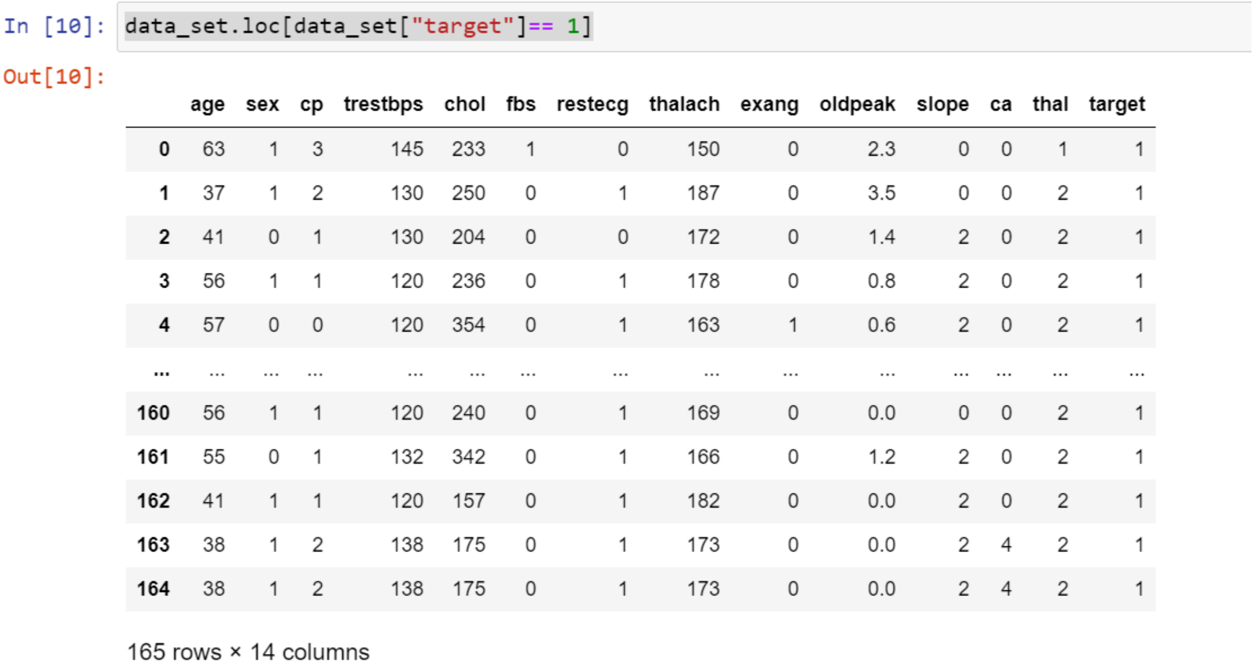


```
In [7]: data_set.isnull().sum()

Out[7]: age         0
        sex         0
        cp          0
        trestbps    0
        chol        0
        fbs         0
        restecg     0
        thalach     0
        exang       0
        oldpeak     0
        slope       0
        ca          0
        thal        0
        target      0
        dtype: int64
```

From the above result it can be seen that all columns of dataset have zero null values or missing values which makes the dataset past the first phase of cleaning and pave the way for next phase.

Secondly, ***data normalization*** which is essentially converting the values of a dataset to conform to the required nature of values for it to run in a particular machine learning algorithm, in our study classification algorithm. Since the research take a classification approach of machine learning, which is the result of a prediction be either 1 or 0, Yes or No so on and so forth. The bottom line is the result must be in a category. The dataset used in this research has dependent variable value that is either 1 or 0.

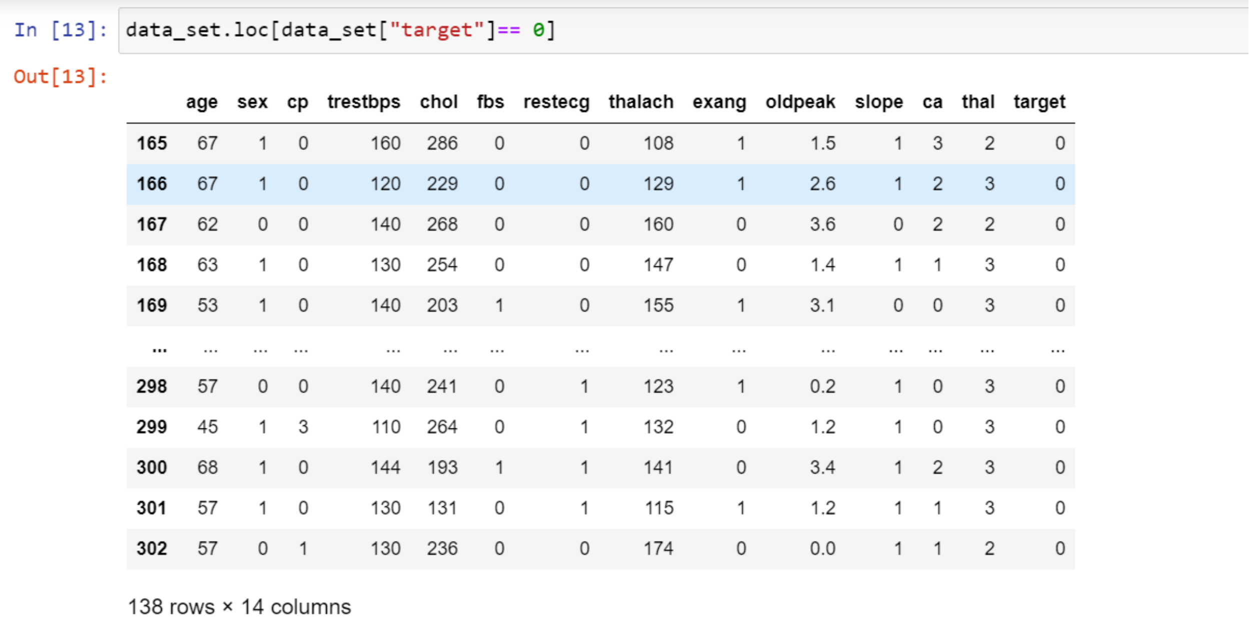
Figure 6: illustration showing the normalized state of the dependent variable:



```
In [10]: data_set.loc[data_set["target"]== 1]

Out[10]:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
160	56	1	1	120	240	0	1	169	0	0.0	0	0	2	1
161	55	0	1	132	342	0	1	166	0	1.2	2	0	2	1
162	41	1	1	120	157	0	1	182	0	0.0	2	0	2	1
163	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1
164	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1

165 rows × 14 columns



```
In [13]: data_set.loc[data_set["target"]== 0]
```

Out[13]:

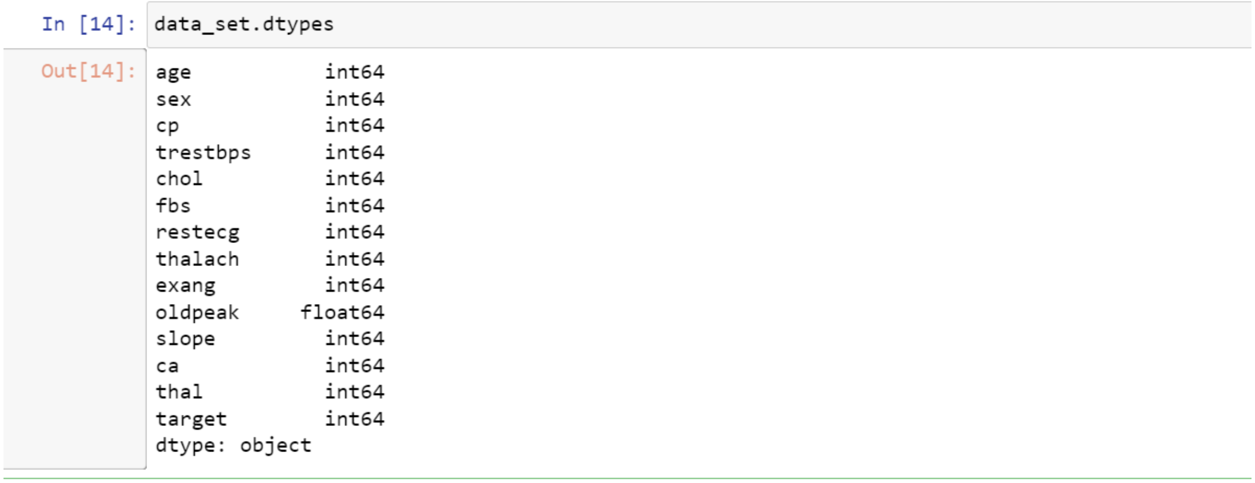
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
165	67	1	0	160	286	0	0	108	1	1.5	1	3	2	0
166	67	1	0	120	229	0	0	129	1	2.6	1	2	3	0
167	62	0	0	140	268	0	0	160	0	3.6	0	2	2	0
168	63	1	0	130	254	0	0	147	0	1.4	1	1	3	0
169	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

138 rows × 14 columns

The above snapshots show that out of the three hundred and three (303) data entries 138 results to the category 0 and 165 results to the category 1. Which shows that the data dependent value is normalized.

Another important aspect of normalization is the necessity to convert all data values to the desired datatype so that a carefully selected algorithm can process it. In this research it is essential that all data are converted to numeric datatype such as Integers and Floats for the efficient running of algorithms such as logistic regression. All datatypes of the dataset used in this research are numeric. And below is a snippet of the code and outcome:

Figure 7: code snippet to show numeric nature of dataset

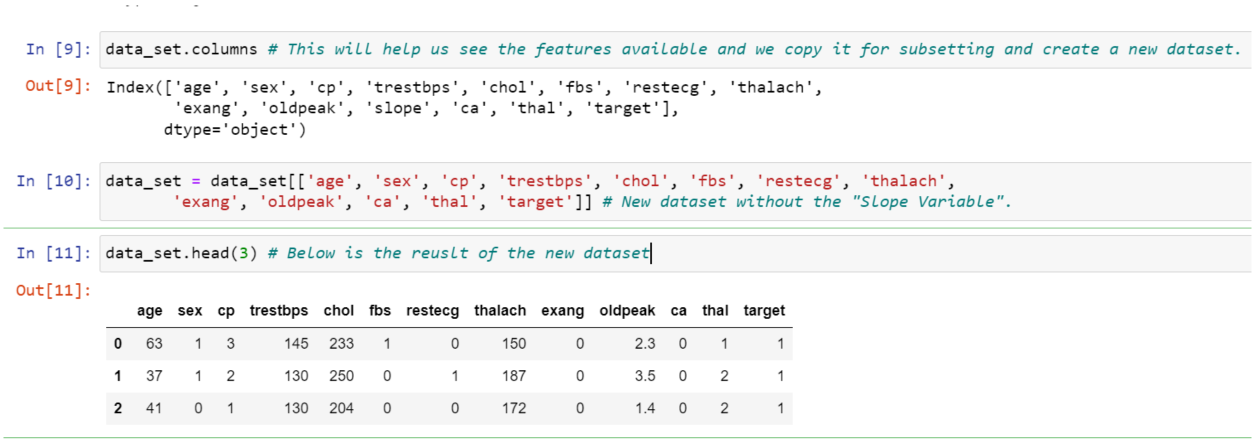


```
In [14]: data_set.dtypes

Out[14]: age           int64
         sex           int64
         cp            int64
         trestbps      int64
         chol          int64
         fbs           int64
         restecg       int64
         thalach       int64
         exang         int64
         oldpeak     float64
         slope         int64
         ca            int64
         thal          int64
         target        int64
         dtype: object
```

Moreover, handling outliers is subsumed in the preprocessing of this dataset because certain features included makes it hard for the algorithms to function properly. In simple terms, the outliers are the set of variables or values that do not conform to the general pattern of the dataset. Therefore, the research focus on minimizing this anomaly through sub-setting or feature engineering.

The last of the pre-processing process is ***feature engineering/ selection*** which is a process of ensuring only those predictors in the dataset that directly affects the outcome of the dependent variable are selected. The dataset used in this research included all essential features that are substantial in the accurate prediction of heart attack. Despite this, the feature "Slope" (***the slope of the peak exercise ST segment)*** and the feature "Oldpeak" (***ST depression induced by exercise relative to rest)*** are almost identical where "Slope" is just a graphical representation of "Oldpeak"; for that reason, the "Slope" feature will not be included in the predicting attributes of heart attack for this research. Below is the new dataset after the feature engineering process:

Figure 7: The feature engineered dataset



```
In [9]:  data_set.columns # This will help us see the features available and we copy it for subsetting and create a new dataset.

Out[9]:  Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
                'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'],
               dtype='object')

In [10]: data_set = data_set[['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
                'exang', 'oldpeak', 'ca', 'thal', 'target']] # New dataset without the "Slope Variable".

In [11]: data_set.head(3) # Below is the reuslt of the new dataset

Out[11]:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	0	2	1

Now that the dataset is clean and free from anomalies, it can be used for training our models and testing the accuracy of the models from there if the accuracy of the models is reliable then we will use the algorithm with the highest accuracy to analyze the data.

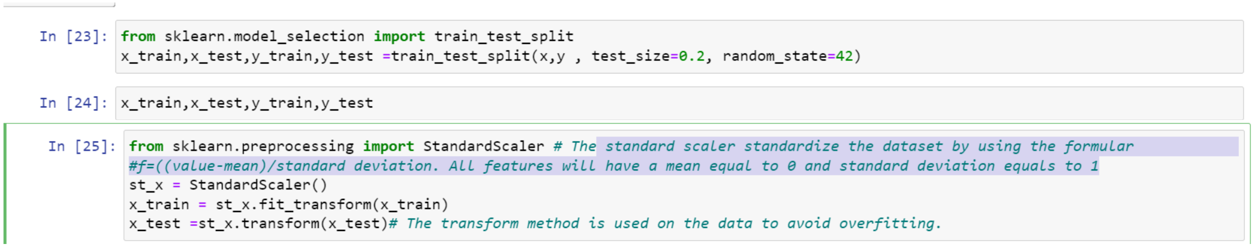
OBJECTIVE ONE: Determine whether different machine learning prediction models for predicting heart attack such as Logistic regression, Decision tress, Random Forest will be accurate in heart attack prediction.

4.2.3 Model Training and Testing:

The dataset that has been cleaned freed from anomalies that impede the proper function of the dataset are been effectively removed. Now the model training of the dataset on the different machine learning algorithms will be done. The process of training a model on a dataset is just using past data to make future predictions. There are a lot of amazing machine learning algorithm available but the problem at demands that we explore only four out of the rest. The machine learning algorithms that are been explored includes Logistic regression, Decision Tree, Random Forest and K Nearest Neighbor (KNN).

The whole process of the model training involves splitting the dataset into training and test data and also scaling the data to a format which will make the model more accurate. The splitting of the dataset into train and test data is done with the help of library called **train_test_split()** which takes three (3) arguments and that is the independent variable values, the dependent variable values and the splitting ratio. In this research project eighty (80) percent of the dataset is used for training and twenty percent is used for testing data. After a successful splitting of the dataset into training and testing; standard scaler is used to standardize the dataset by using the formula f=((value-mean)/standard deviation. All features will have a mean equal to 0 and standard deviation equals to 1. Below is the code snippet of the splitting and standardization process:

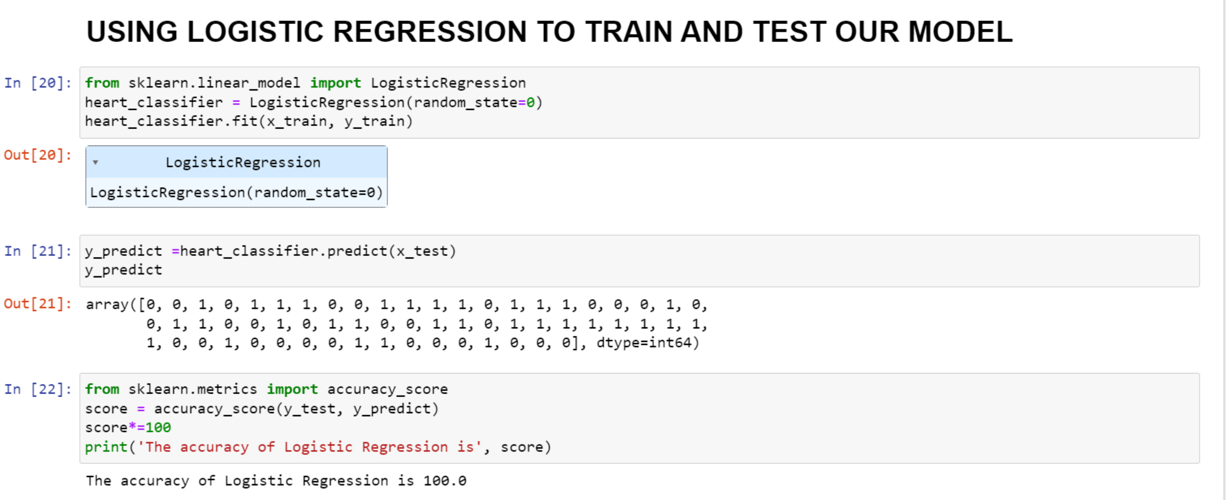
Figure 8: Illustration of splitting dataset into train and test with standardization:



```
In [23]: from sklearn.model_selection import train_test_split
         x_train,x_test,y_train,y_test =train_test_split(x,y , test_size=0.2, random_state=42)

In [24]: x_train,x_test,y_train,y_test

In [25]: from sklearn.preprocessing import StandardScaler # The standard scaler standardize the dataset by using the formular
         #f=((value-mean)/standard deviation. All features will have a mean equal to 0 and standard deviation equals to 1
         st_x = StandardScaler()
         x_train = st_x.fit_transform(x_train)
         x_test =st_x.transform(x_test)# The transform method is used on the data to avoid overfitting.
```

After a successful splitting of the dataset and standardizing it, the different machine learning models will be trained and test on the dataset to test the accuracy of the dataset. The output of the different models is shown below:

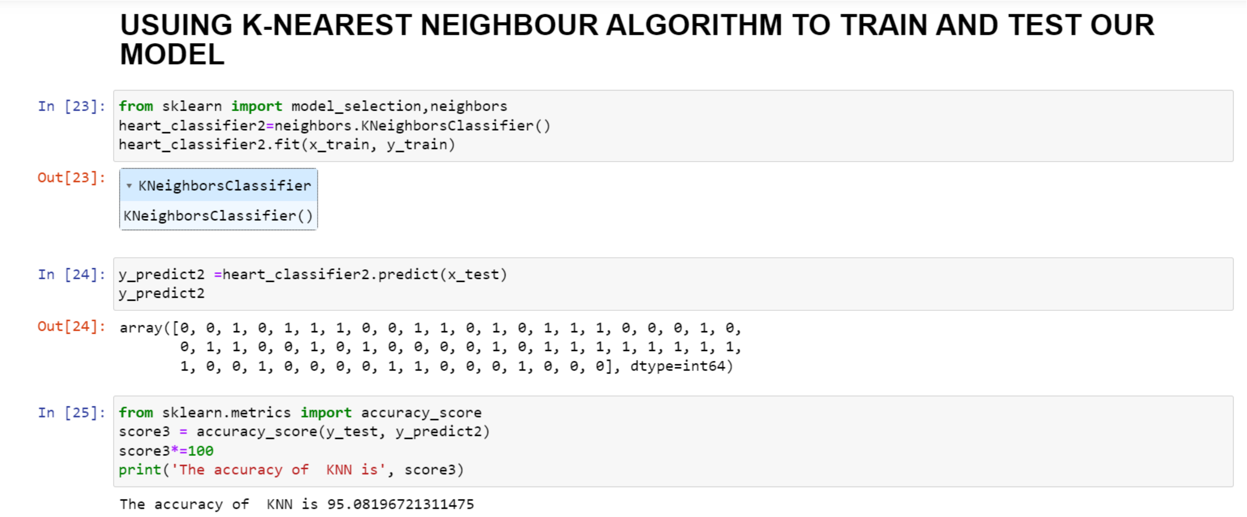
Figure 9: Output of Logistic Regression.



USING LOGISTIC REGRESSION TO TRAIN AND TEST OUR MODEL

```
In [20]: from sklearn.linear_model import LogisticRegression
         heart_classifier = LogisticRegression(random_state=0)
         heart_classifier.fit(x_train, y_train)

Out[20]: ▾          LogisticRegression
         LogisticRegression(random_state=0)

In [21]: y_predict =heart_classifier.predict(x_test)
         y_predict

Out[21]: array([0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0,
                0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0], dtype=int64)

In [22]: from sklearn.metrics import accuracy_score
         score = accuracy_score(y_test, y_predict)
         score*=100
         print('The accuracy of Logistic Regression is', score)

         The accuracy of Logistic Regression is 100.0
```

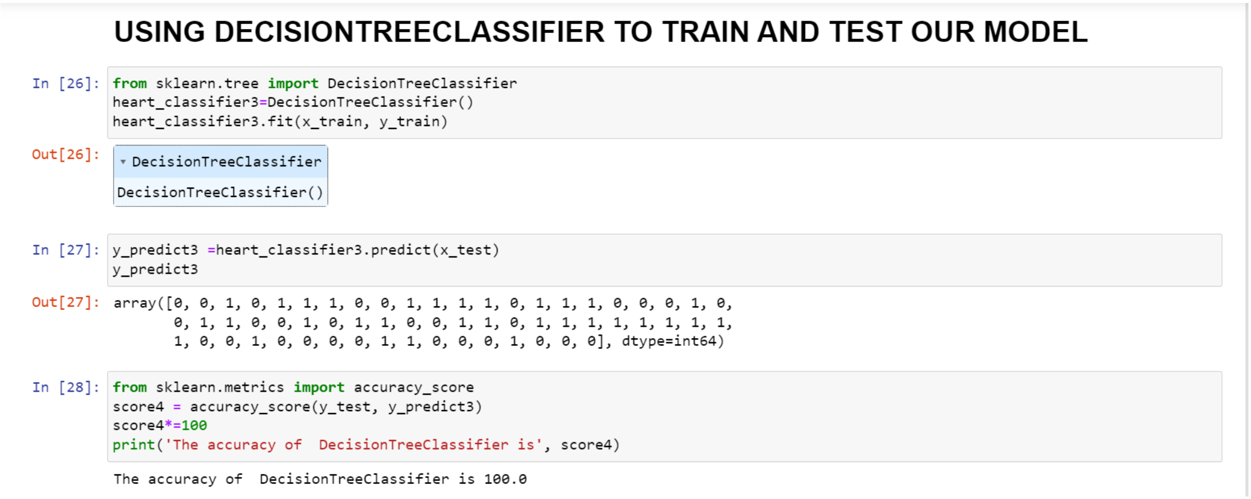
Figure 10: Output K-Nearest Neighbor.



USUING K-NEAREST NEIGHBOUR ALGORITHM TO TRAIN AND TEST OUR MODEL

```
In [23]: from sklearn import model_selection,neighbors
         heart_classifier2=neighbors.KNeighborsClassifier()
         heart_classifier2.fit(x_train, y_train)

Out[23]: ▾ KNeighborsClassifier
         KNeighborsClassifier()

In [24]: y_predict2 =heart_classifier2.predict(x_test)
         y_predict2

Out[24]: array([0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0,
                0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0], dtype=int64)

In [25]: from sklearn.metrics import accuracy_score
         score3 = accuracy_score(y_test, y_predict2)
         score3*=100
         print('The accuracy of  KNN is', score3)

         The accuracy of  KNN is 95.08196721311475
```

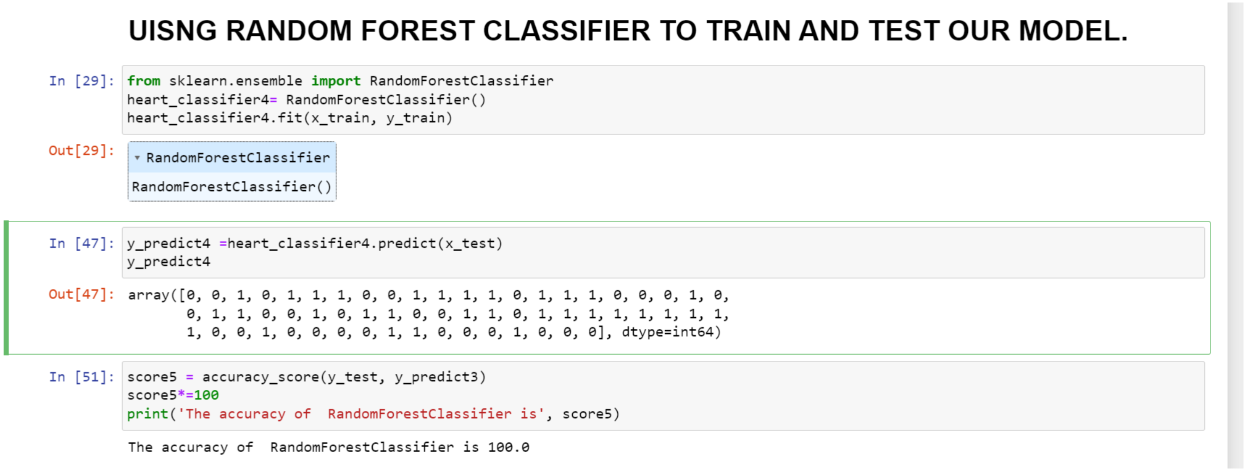
Figure 11: Output Decision Tree Classifier



USING DECISIONTREECLASSIFIER TO TRAIN AND TEST OUR MODEL

```
In [26]: from sklearn.tree import DecisionTreeClassifier
         heart_classifier3=DecisionTreeClassifier()
         heart_classifier3.fit(x_train, y_train)

Out[26]: ▾ DecisionTreeClassifier
         DecisionTreeClassifier()

In [27]: y_predict3 =heart_classifier3.predict(x_test)
         y_predict3

Out[27]: array([0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0,
                0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0], dtype=int64)

In [28]: from sklearn.metrics import accuracy_score
         score4 = accuracy_score(y_test, y_predict3)
         score4*=100
         print('The accuracy of  DecisionTreeClassifier is', score4)

         The accuracy of  DecisionTreeClassifier is 100.0
```

Figure 12: Output of Random Forest.



UISNG RANDOM FOREST CLASSIFIER TO TRAIN AND TEST OUR MODEL.

```
In [29]: from sklearn.ensemble import RandomForestClassifier
         heart_classifier4= RandomForestClassifier()
         heart_classifier4.fit(x_train, y_train)

Out[29]: ▾ RandomForestClassifier
         RandomForestClassifier()

In [47]: y_predict4 =heart_classifier4.predict(x_test)
         y_predict4

Out[47]: array([0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0,
                0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0], dtype=int64)

In [51]: score5 = accuracy_score(y_test, y_predict3)
         score5*=100
         print('The accuracy of  RandomForestClassifier is', score5)

         The accuracy of  RandomForestClassifier is 100.0
```

All the machine learning models are running excellently with prediction accuracy of nine five percent to one hundred percent (95% to 100%). The below is an accuracy table per model:

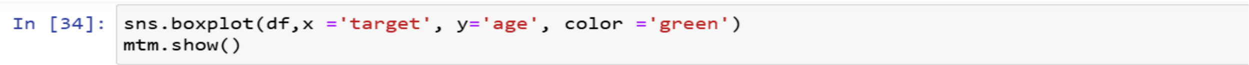
ML Model Name	**Accuracy of Model (%)**
Logistic Regression	100
Decision Tree Classifier	100
K-Nearest Neighbour	95
Random Forest Classifier	100

Objective Two (2): Relate heart conditions that are most likely to cause heart attack.

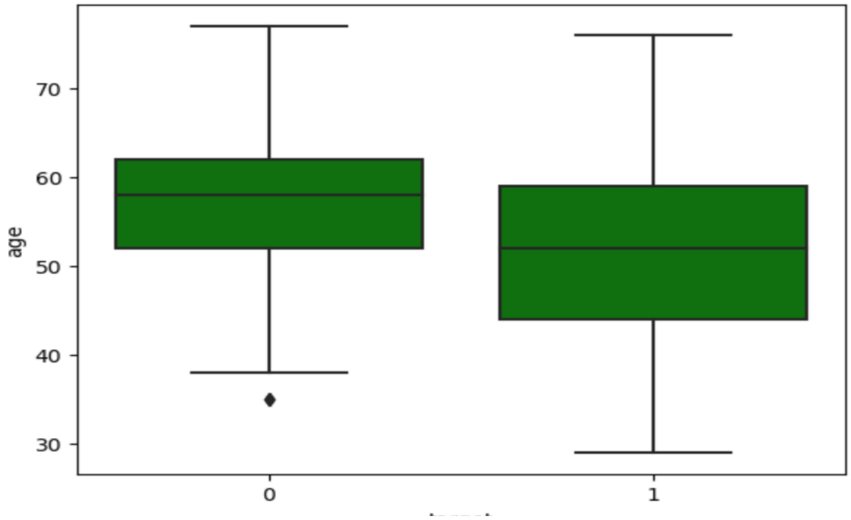
Whether an individual will have heart attack is affected by a set of heart and bodily conditions known as predictors such age, cholesterol level, fasting blood sugar etc. Therefore, understanding the relationship between heart and bodily conditions will give an insight in making firm predictions of heart patients.

Firstly, we will explore the relation between the age of a heart patient and his/her chances of surviving a heart attack. We will use Jupiter notebook to visualize the relationship between heart attack and age using a BoxPlot.

Figure 13: Illustrates the relationship between target (Heart Attack) and age



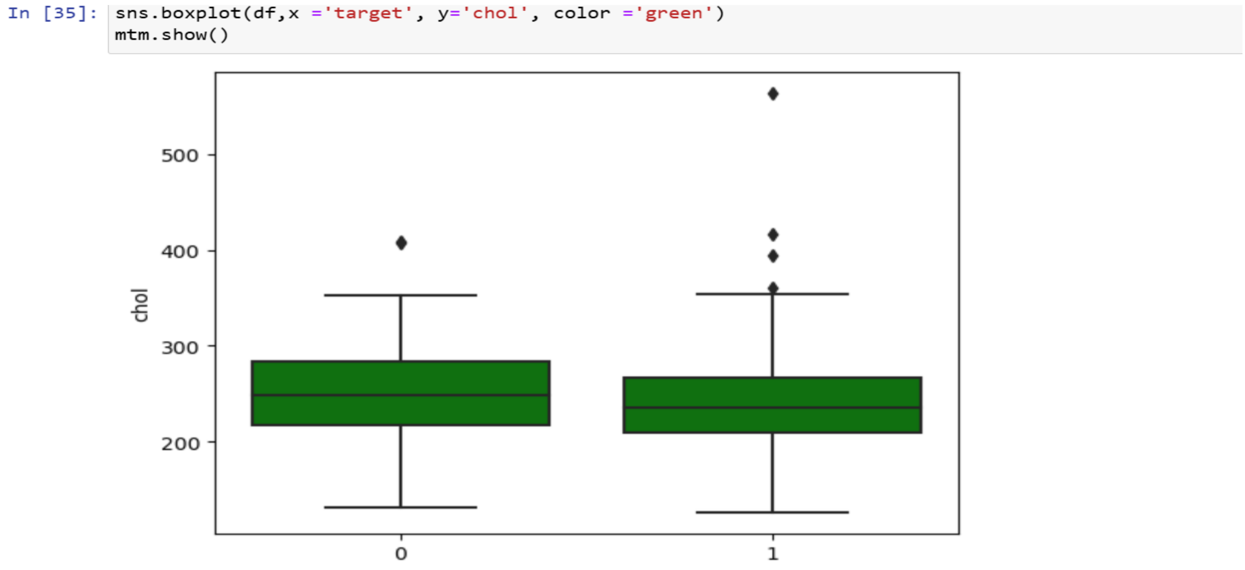
```
In [34]: sns.boxplot(df,x ='target', y='age', color ='green')
         mtm.show()
```



From the boxplot we can deduce that fifty percent of the patients that survive heart problems are approximately between the ages of 52-62 and also fifty percent of the patients that do not survive heart attack are between the ages of 45 -59 approximately.

Furthermore, the chances of heart patients with high cholesterol level could survive heart attack. Below is the relationship between heart attack patients and cholesterol levels.

Figure 14: illustrates the relationship between target (heart attack) and cholesterol



```
In [35]: sns.boxplot(df,x ='target', y='chol', color ='green')
         mtm.show()
```

4.3 Test of Hypothesis:

The accuracy score obtained from each of the machine learning algorithm could in much great degree justify that machine learning algorithms can be used as an artificial intelligence approach to predict heart conditions.

Thus justify the hypothesis of this thesis which states “**Heart diseases or conditions such as age, constrictive pericardium, thalack, cholesterol, slope, old peak etc could be used by machine learning algorithms to predict to heart attack.**”

4.4 Discussion of Findings:

Taking a genealogical approach to arrive at the findings of this research work, one will convinced that indeed machine learning can partly and efficient fill the gap of this research.

A thorough **conceptual review** gives us an understanding that, machine learning algorithms are more efficient than traditional data analysis methods, moreover the **empirical review** shows that machine learning algorithms are indeed been used by health care systems in developed and developing countries producing enormous benefit and speed in data analysis and prediction. With this known to us, the **theoretical review** of this study showed that most machine learning algorithms use the three most common theories in machine learning which are supervised learning, unsupervised learning and reinforcement learning.

From the above we explore machine learning algorithms to find solution to the **existing gap in literature** of this study which is a **knowledge gap** in the medical facilities of our beloved country which artificial intelligence and most especially machine learning could be used to support cardiologist or doctors in general.

Therefore after testing four machine learning models (algorithms) and using accuracy score to test the accuracy of the results of the various ML models, it is found that three (3) out of the four model, score an accuracy of 100%, which makes ML models a perfect fit for predicting heart diseases.

4.5 Problems encountered in the Field:

The data used in this research is secondary which is there is no physical data collection on the ground. The data used is obtained from the internet. The hurdles of physical/ primary data collections is removed with this approach which makes the entire process less problematic.

This does not remove the problem of poor and expensive internet connection to obtain data and access scholarly works of other researchers on this topic; which slows the entire process of the research project. One may argue that internet connection is provided for free at the university campus but the name itself is self-defeating, since research project could necessarily be done during odd hours when one is not within the university compound. This adds to the problems encountered in the study.

Finally, access to scholarly work relevant to once study is tremendously impeded by license fee attached to most published articles, books and journals. If the university in any way shape form could provide a subsidy to at least five articles or books of every research project for each student then this burden might be lightened.

CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Summary:

Heart disease is the second most common cause of the death in the Gambia which makes it a viable topic for study. Many methods could have been used to address this issue but technology stands out to be the only viable option for the recent situations. Studies in the Gambia with regards to heart disease is very minimal.

Thus to address the issue, implementation of predictive model is expected to be an effective and an appropriate avenue. The study aims to explore different ML models or techniques such as logistic regression, K-Nearest Neighbor, decision tree algorithm etc to make predictions using heart attack triggering parameters such as chest pain, cholesterol level, sugar level, trestbps etc to make predictions. This will effectively tackle two primary problems related to heart attacks in the Gambia which are doctor to patient ratio gab and accuracy in diagnostic speed of physicians.

Moreover, it is deduced from the literature review that machine learning techniques are more effective than traditional methods of data analysis. With machine learning algorithms predictions are more accurate and analysis. ML models have been successfully used in different fields of study to produce astonishing prediction accuracy. They are used in all spheres of work life to reduce the workload on humans. In the field of health Ml techniques are used to predict different diseases and most especially the focus of this research which is heart attack. Different theories are adopted for different problems using machine learning techniques but this research uses supervised classification modeling.

The research using machine learning predictive models is data driven thereby making the entire research centered on data collected. The data is explored or analyzed or preprocessed to get rid of missing values and duplications which will make the data fit into a supervised machine learning model. This research like any other data driven machine learning algorithm use the fundamental methodology of machine learning such as data gathering, cleaning, splitting, training and testing.

Upon using the methodology above all machine learning models used in this research work fantastically well. Most of the models produce and accuracy score one hundred percent (100%) which makes Machine learning as an artificial intelligence approach to preventing and predicting heart attack in the Gambia is essential for the socio-development of the Gambia. ML models are widely adopted by developed and developing countries throughout the globe.

Despite ML been a new technology machine learning models are used throughout the world. Some countries in West Africa have used ML models for prediction.

This research focus on the health care systems because most African countries have piles of raw data about patients of all kind of illnesses but these data could be hardly used to make predictions using traditional methods.

Exploring the problem statement of this research, the hypothesis and the discovered gap in literature an artificial intelligence approach to preventing and predicting heart attack became a necessity. Most of the researches that are been done on heart attack predictions are made in western countries.

5.2 Conclusion:

This section will conclude the study by summarizing the key research findings in relation to the research aims and research objectives, as well as the values of the finding of the research.

This research aims to design a ML model that could assist doctors in preventing and predicting heart attack in the Gambia. The findings indicate that Machine Learning models could be used to predict and thereby prevent or address heart attack in the Gambia. After feeding a preprocessed dataset into the different machine learning models and training it on the data to build probabilistic models that are later tested with new data produced accuracy scores of 75% above. Which makes ML models perfect fit for predicting heart diseases. Additionally, the investigation shows that most heart attack conditions are triggered by markers or predictors such as age, cholesterol, trestbps. It was found that half of the heart attack patients are adults above the age of 40.

Therefore, this research concludes that machine learning techniques or models could be used to predict and thereby prevent heart attacks in the Gambia.

5.3 Recommendations:

1. **Digitalization of hospital records:**
   It is strongly recommended that the medical institutions of the Gambia have their medical records digitalized. This study is limited in its ability to reflect the great need of machine learning algorithms which is due to the usage of secondary data instead of primary data. Digitalized hospital data will help future researchers to come with more accurate results that will reflect entirely the Gambian heart patients.
2. **Investing in Machine Learning Technologies:**
   Among trending technologies today is Machine learning, the global race for experts in machine learning is the craziest so far according Elon Musk. The Gambia been a tiny country with low GDP can make a step towards investing in the field of machine learning which will help the country to catch up with other West African countries. Investment here would mean training more students on machine learning, sensitizing the general public the importance machine learning and also including it in the academic curriculum.
3. **Cooperations between ML enthusiasts and health workers:**
   Machine learning with all great super human powers can only work if there is a cooperation between ML enthusiasts which are mostly computer scientist and health care workers like doctors and nurses. This cooperation will help form experts that can generate formidable models that could tackle future problems and provide predictions.
4. **University owned data repository:**
   The University of the Gambia should have standalone data repository that will help researchers within the university to access information of other researchers of the

University. This will help UTG students to have ease of access to information and to build on the studies of their peers before them. A university owned data repository will reduce the amount of money used to pay for some journals.

5. **Maintenance:**
   Technology been technology subjecting to change at any moment in time and obsoleting some software features. A continuous maintenance or updating of machine learning algorithms will help maintain a smooth functioning of the system when deployed at any hospital.

5.4 Contribution to knowledge:

The outputs of this study demonstrate its contribution to the field of cardiology. The contribution of this research is fundamental in heart attack prediction and machine learning application in the field of cardiology. This study solves the problem of the applicability of artificial intelligence in heart attack prediction through training and testing heart attack dataset on different ML models producing accuracy of above 75% and clearly demonstrating that some features of the heart directly affect heart attacks.

The existing gap in this research is machine learning applicability in the Gambia is to some extent addressed. Despite ML being a trending technology it has not been used in the field of cardiology in the Gambia. This research addresses that gap through using Gambian context.

The study conducted here with its results shows that machine learning algorithms could definitely be used in health facilities to assist cardiologist or doctors. All it requires is heart disease data with similar features which can then be feed in the algorithms which will produce probabilistic model for future predictions.

5.5 Suggestion for Further Studies:

1. Primary Data Usage:

This research uses secondary dataset which is data that has not been obtained from Gambian hospitals. It will be crucially important for any research building on this research to use primary data which is data obtained from Gambia hospitals to train and test the algorithms. This way the results will reflect Gambian heart attack conditions perfectly.

2. Hybrid Machine Learning Model for heart attack prediction:

Machine learning algorithms are all different in their unique ways and integrating these unique features might bring about a more accurate model for prediction.

3. Thorough Engagement of Cardiologist Experts and ML experts:

This study is limited by access to cardiologist and ML experts but a future study might use the expertise of these two groups to make something great. Their opinions will greatly impact the success of the research.

References

Al'Aref, S. J., Anchouche, K., Singh, G., Slomka, P. J., Kolli, K. K., Kumar, A., Pandey, M., Maliakal, G., van Rosendael, A. R., Beecy, A. N., Berman, D. S., Leipsic, J., Nieman, K., Andreini, D., Pontone, G., Schoepf, U. J., Shaw, L. J., Chang, H.-J., Narula, J., … Min, J. K. (2019). Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *European Heart Journal*, *40*(24), 1975–1986. https://doi.org/10.1093/eurheartj/ehy404

Coronato, A., Naeem, M., De Pietro, G., & Paragliola, G. (2020). Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine, 109*, 101964. https://doi.org/10.1016/j.artmed.2020.101964

D'Ascenzo, F., De Filippo, O., Gallone, G., Mittone, G., Deriu, M. A., Iannaccone, M., Ariza-Solé, A., Liebetrau, C., Manzano-Fernández, S., Quadri, G., Kinnaird, T., Campo, G., Simao Henriques, J. P., Hughes, J. M., Dominguez-Rodriguez, A., Aldinucci, M., Morbiducci, U., Patti, G., Raposeiras-Roubin, S., … De Ferrari, G. M. (2021). Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): A modelling study of pooled datasets. *Lancet (London, England), 397*(10270), 199–207. https://doi.org/10.1016/S0140-6736(20)32519-8

Dennis, A.-G. P., & Strafella, A. P. (2024). The role of AI and machine learning in the diagnosis of Parkinson's disease and atypical parkinsonisms. *Parkinsonism & Related Disorders*, 106986. https://doi.org/10.1016/j.parkreldis.2024.106986

Deo, R. C. (2015). Machine Learning in Medicine. *Circulation*, *132*(20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

Dharmarathne, G., Bogahawaththa, M., McAfee, M., Rathnayake, U., & Meddage, D. P. P. (2024). On the diagnosis of chronic kidney disease using a machine learning-based interface with explainable artificial intelligence. *Intelligent Systems with Applications*, *22*, 200397. https://doi.org/10.1016/j.iswa.2024.200397

Dong, J., Feng, T., Thapa-Chhetry, B., Cho, B. G., Shum, T., Inwald, D. P., Newth, C. J. L., & Vaidya, V. U. (2021). Machine learning model for early prediction of acute kidney injury (AKI) in pediatric critical care. *Critical Care (London, England)*, *25*(1), 288. https://doi.org/10.1186/s13054-021-03724-0

Ghasemieh, A., Lloyed, A., Bahrami, P., Vajar, P., & Kashef, R. (2023). A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients. *Decision Analytics Journal*, *7*, 100242. https://doi.org/10.1016/j.dajour.2023.100242

Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. (2019). Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. *Stroke*, *50*(5), 1263–1265. https://doi.org/10.1161/STROKEAHA.118.024293

Iscra, K., Miladinović, A., Ajčević, M., Starita, S., Restivo, L., Merlo, M., & Accardo, A. (2022). Interpretable machine learning models to support differential diagnosis between Ischemic Heart Disease and Dilated Cardiomyopathy. *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022*, *207*, 1378–1387. https://doi.org/10.1016/j.procs.2022.09.194

Jaiteh, L. (2024, July 4). *The need for heart attack prevention and prediction in the Gambia through machine learning algorithms.* [Live interview].

Jarde, A., Jeffries, D. J., & Mackenzie, G. A. (2021). Development and validation of a model for the prediction of mortality in children under five years with clinical pneumonia in rural gambia. *medRxiv*, 2021–08.

Jaw, E., Loum, W. X. M. L., & Janneh, L. L. (n.d.). *Review on the Application of Medical Artificial Intelligence in The Gambia's Health Care System*.

Khan, A., Qureshi, M., Daniyal, M., & Tawiah, K. (2023). A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction. *Health & Social Care in the Community*, *2023*(1), 1406060. https://doi.org/10.1155/2023/1406060

Khera, R., Haimovich, J., Hurley, N. C., McNamara, R., Spertus, J. A., Desai, N., Rumsfeld, J. S., Masoudi, F. A., Huang, C., Normand, S.-L., Mortazavi, B. J., & Krumholz, H. M. (2021). Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction. *JAMA Cardiology*, *6*(6), 633–641. https://doi.org/10.1001/jamacardio.2021.0122

Ma, K., Yang, J., Shao, Y., Li, P., Guo, H., Wu, J., Zhu, Y., Zhang, H., Zhang, X., Du, J., & Li, Y. (2022). Therapeutic and Prognostic Significance of Arachidonic Acid in Heart Failure. *Circulation Research*, *130*(7), 1056–1071. https://doi.org/10.1161/CIRCRESAHA.121.320548

Machine Learning: What It is, Tutorial, Definition, Types—Javatpoint. (n.d.). Retrieved July 6, 2024, from https://www.javatpoint.com/machine-learning

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An unsupervised machine learning algorithms: Comprehensive review. *International Journal of Computing and Digital Systems*.

Oliveira, M., Seringa, J., Pinto, F. J., Henriques, R., & Magalhães, T. (2023). Machine learning prediction of mortality in Acute Myocardial Infarction. *BMC Medical Informatics and Decision Making*, *23*(1), 70. https://doi.org/10.1186/s12911-023-02168-6

Rezaei, T., & Javadi, A. (2024). Environmental impact assessment of ocean energy converters using quantum machine learning. *Journal of Environmental Management*, *362*, 121275. https://doi.org/10.1016/j.jenvman.2024.121275

S. Mohan, C. Thirumalai, & G. Srivastava. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access*, *7*, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

Segar, M. W., Vaduganathan, M., Patel, K. V., McGuire, D. K., Butler, J., Fonarow, G. C., Basit, M., Kannan, V., Grodin, J. L., Everett, B., Willett, D., Berry, J., & Pandey, A. (2019). Machine Learning to Predict the Risk of Incident Heart Failure Hospitalization Among Patients With Diabetes: The WATCH-DM Risk Score. *Diabetes Care*, *42*(12), 2298–2306. https://doi.org/10.2337/dc19-0587

Sharma, H., & Rizvi, M. A. (n.d.). Prediction of Heart Disease using Machine Learning Algorithms: A Survey. *International Journal on Recent and Innovation Trends in Computing and Communication*, *5*(8).

Silva, G. F. S., Fagundes, T. P., Teixeira, B. C., & Chiavegatto Filho, A. D. P. (2022). Machine Learning for Hypertension Prediction: A Systematic Review. *Current Hypertension Reports*, *24*(11), 523–533. https://doi.org/10.1007/s11906-022-01212-6

Singh, G., Al’Aref, S. J., Van Assen, M., Kim, T. S., van Rosendael, A., Kolli, K. K., Dwivedi, A., Maliakal, G., Pandey, M., Wang, J., Do, V., Gummalla, M., De Cecco, C. N., & Min, J. K. (2018). Machine learning in cardiac CT: Basic concepts and contemporary data. *Journal of Cardiovascular Computed Tomography*, *12*(3), 192–201. https://doi.org/10.1016/j.jcct.2018.04.010

Supervised Machine learning—Javatpoint. (n.d.). Www.Javatpoint.Com. Retrieved July 6, 2024, from https://www.javatpoint.com/supervised-machine-learning

You, J., Guo, Y., Kang, J.-J., Wang, H.-F., Yang, M., Feng, J.-F., Yu, J.-T., & Cheng, W. (2023). Development of machine learning-based models to predict 10-year risk of cardiovascular disease: A prospective cohort study. *Stroke and Vascular Neurology*, *8*(6), 475–485. https://doi.org/10.1136/svn-2023-002332