**CHAPTER 1**

**INTRODUCTION**

* 1. **General**

Heart problems are a pervasive and significant health concern that affects people worldwide, leading to a multitude of illnesses and fatalities. In response to this pressing issue, scientists and healthcare professionals are harnessing the power of cutting-edge technology known as "machine learning" to revolutionize the way we address cardiovascular diseases. Machine learning has emerged as a valuable tool for predicting individuals' risk of developing heart issues, offering a promising solution to the global health challenge of cardiovascular diseases.

Imagine machine learning as a sophisticated and intelligent assistant in the realm of healthcare. It has the remarkable ability to analyse extensive datasets and extract valuable insights from the health information of individuals. This technology allows us to identify those who might be at risk for heart problems in the future. The core idea is to proactively detect cardiovascular risk factors, enabling timely interventions and personalized treatment plans.

Machine learning's potential in this context is profound. By analysing a wealth of health - related data, including medical history, lifestyle, genetic predisposition, and other factors, it can create predictive models that help healthcare providers anticipate heart issues before they become critical. These predictive models consider a myriad of variables, and their accuracy continues to improve as they learn from more data. This enables doctors to identify potential problems at an early stage, when interventions are most effective and outcomes are generally more favourable.

The advantages of machine learning in cardiovascular health are manifold. Firstly, it promotes early detection. By identifying individuals at higher risk, doctors can initiate interventions that can delay or even prevent the onset of heart problems. Secondly, it facilitates personalized treatment plans. Rather than employing a one-size-fits-all approach, machine learning enables healthcare providers to tailor treatments to each patient's unique needs, improving the chances of successful outcomes.

This innovative approach to preventing and treating heart issues is poised to transform the landscape of cardiovascular health. It promises to make prevention and treatment more precise, thus saving lives. As machine learning continues to evolve and learn from extensive datasets, its predictive power becomes more refined. This precision not only benefits patients but also healthcare systems that are under constant pressure to deliver the best care efficiently.

One of the most significant benefits of ML in cardiovascular health is its potential to promote early detection. By identifying individuals at higher risk for heart problems, doctors can initiate interventions that can delay or even prevent the onset of disease. For example, ML models can be used to predict the risk of developing coronary artery disease (CAD), the most common type of heart disease. CAD is caused by a buildup of plaque in the arteries that supply blood to the heart. If left untreated, CAD can lead to heart attack, stroke, and other serious complications. ML models can also be used to predict the risk of other cardiovascular conditions, such as atrial fibrillation, heart failure, and sudden cardiac death. Early detection of these conditions is critical, as it allows for timely treatment and improved outcomes.

Another key advantage of ML in cardiovascular health is its ability to facilitate personalized treatment plans. Rather than employing a one-size-fits-all approach, ML enables healthcare providers to tailor treatments to each patient's unique needs. This is particularly important in cardiovascular medicine, as there is a wide range of variability in how patients respond to different treatments. For example, ML models can be used to identify patients who are more likely to benefit from certain medications or procedures. They can also be used to predict the risk of adverse side effects from medications. By personalizing treatment plans, ML can help healthcare providers improve the effectiveness and safety of care.

ML can also play a role in optimizing the allocation of healthcare resources. By identifying individuals at greater risk for heart problems, ML can help ensure that resources are directed to 3 those who need them most. This can minimize unnecessary medical expenses and improve the overall quality of care. For example, ML models can be used to develop risk stratification tools that can help healthcare providers prioritize patients for screening and other preventive services. ML can also be used to identify patients who are at high risk of hospitalization or other costly healthcare interventions. By targeting resources to those who are most likely to benefit, ML can help healthcare systems operate more efficiently and effectively.

Despite its potential, the integration of ML into cardiovascular health is not without challenges. One challenge is the need for large and high-quality datasets to train ML models. Another challenge is the need to ensure that ML models are transparent and interpretable, so that healthcare providers can understand and trust the outputs of these models. Despite these challenges, the opportunities for ML in cardiovascular health are vast. As ML algorithms continue to evolve and learn from more data, their predictive power will only become more refined. This precision will benefit both patients and healthcare systems, making the prevention and treatment of cardiovascular diseases more precise, cost-effective, and life-saving.

In conclusion, the integration of machine learning into cardiovascular health represents a promising leap forward in our fight against heart problems. This technology has the potential to reshape the way we approach cardiovascular diseases, enabling early detection, personalized treatments, and efficient resource allocation. By using machine learning to predict heart issues, we are moving closer to a future where the prevention and treatment of cardiovascular diseases are more precise, cost-effective, and, most importantly, capable of saving lives.

* 1. **Machine Learning (ML) and Deep Learning (DL) and Their Differences**

In the realm of artificial intelligence, Machine Learning (ML) and Deep Learning (DL) stand as two distinct yet interrelated paradigms, each wielding its own unique attributes and wielding immense potential in various applications.

Machine Learning (ML) constitutes a field of AI that empowers machines to learn and make predictions or decisions based on data patterns, without being explicitly programmed. It ushers in a new era of automation, enabling systems to adapt and evolve with experience. At its core, ML operates on the premise of recognizing patterns within datasets, subsequently using these 4 patterns to inform future decisions. This adaptability renders ML invaluable in a diverse array of domains. For instance, in regression analysis, ML can predict numerical outcomes based on historical data, offering a powerful tool for tasks like sales forecasting and trend analysis. Classification tasks, on the other hand, involve sorting data points into distinct categories, and ML excels in applications like spam detection in emails or medical diagnosis. Clustering, a third facet of ML, involves grouping similar data points together based on their intrinsic characteristics, finding applications in fields like customer segmentation for targeted marketing strategies.

Deep Learning (DL) is a subfield of Machine Learning (ML) that has brought about a revolutionary shift in the way we process and understand data. It distinguishes itself by its remarkable capacity to handle extensive datasets through neural networks. What sets DL apart from traditional ML methods is its capability to work with multiple layers of neurons, enabling it to automatically extract intricate features from raw data. This depth of analysis empowers DL models to excel in tasks that require a high level of abstraction, such as image recognition, where patterns can range from simple edges to complex objects. Similarly, in speech processing, DL exhibits the ability to discern subtle nuances in audio signals, enabling tasks like speech-to-text conversion. Furthermore, DL has emerged as a cornerstone in natural language understanding, enabling machines to comprehend the subtleties of human language. This proficiency is vital for a wide range of applications including sentiment analysis, chatbots, and language translation.

Deep Learning, a subset of the broader field of Machine Learning, has brought about a revolutionary shift in our approach to processing and understanding data. Deep Learning specializes in handling large and complex datasets through neural networks, which are complex mathematical models inspired by the structure and function of the human brain. What sets Deep Learning apart is its unique ability to work with multiple layers of interconnected neurons, enabling it to automatically extract intricate features and patterns from raw data. This depth of analysis allows Deep Learning models to excel in tasks that require a high level of abstraction, such as image recognition, where patterns can range from simple edges to complex objects. Similarly, in the domain of speech processing, Deep Learning can discern subtle nuances in audio signals, enabling tasks like speech-to-text conversion. Moreover, Deep Learning has become a foundational technology in the field of natural language understanding, enabling machines to grasp 5 the complexities of human language. This capability is crucial for various applications, including sentiment analysis, chatbots, and language translation.

Deep Learning, nested within the broader framework of Machine Learning, represents a paradigm shift. It specializes in processing vast amounts of data through neural networks— complex mathematical models inspired by the architecture of the human brain. What sets DL apart is its ability to operate on multiple layers of neurons, enabling it to automatically extract intricate features from raw data. This depth of analysis allows DL models to excel in tasks requiring a high level of abstraction, such as image recognition, where patterns can range from simple edges to complex objects. Similarly, in speech processing, DL can discern subtle nuances in audio signals, enabling tasks like speech-to-text conversion. Moreover, DL has emerged as a cornerstone in natural language understanding, enabling machines to comprehend the subtleties of human language, an area vital for applications like sentiment analysis, chatbots, and language translation.

Deep Learning, a subset of Machine Learning, is at the forefront of a paradigm shift in the way we process and analyze data. It excels in processing extensive datasets by employing neural networks, which are intricate mathematical models inspired by the organization of the human brain. What distinguishes Deep Learning is its capacity to work with multiple layers of interconnected neurons, allowing it to automatically extract intricate features from raw data. This level of analysis empowers Deep Learning intricate mathematical models inspired by the organization of the human brain. What distinguishes Deep Learning is its capacity to work with multiple layers of interconnected neurons, allowing it to automatically extract intricate features from raw data. This level of analysis empowers Deep Learning intricate mathematical models inspired by the organization of the human brain. What distinguishes Deep Learning is its capacity to work with multiple layers of interconnected neurons, allowing it to automatically extract intricate features from raw data. This level of analysis empowers Deep Learning intricate mathematical models inspired by the organization of the human brain. What distinguishes Deep Learning is its capacity to work with multiple layers of interconnected neurons, allowing it to automatically extract intricate features from raw data. This level of analysis empowers Deep Learning models to excel in tasks that require a high degree of abstraction, such as image recognition, where patterns can vary from simple edges to complex objects. Similarly, in the domain of speech processing, Deep Learning can identify subtle nuances in audio signals, enabling 6 applications like speech-to-text conversion. Furthermore, Deep Learning has become a fundamental component of natural language understanding, enabling machines to comprehend the nuances of human language. This capability is essential for various applications, including sentiment analysis, chatbots, and language translation.

Deep Learning, a subset of Machine Learning, has ushered in a revolutionary change in the way we approach data analysis. It excels in handling large and complex datasets by employing neural networks, intricate models inspired by the architecture of the human brain. What sets Deep Learning apart is its remarkable ability to work with multiple layers of interconnected neurons, enabling it to automatically extract intricate features from raw data. This depth of analysis empowers Deep Learning models to excel in tasks that require a high level of abstraction, such as image recognition, where patterns can range from simple edges to complex objects. In the realm of speech processing, Deep Learning can detect subtle nuances in audio signals, facilitating applications like speech-to-text conversion. Furthermore, Deep Learning has become a cornerstone in the field of natural language understanding, enabling machines to grasp the complexities of human language. This capability is pivotal for various applications, including sentiment analysis, chatbots, and language translation.

Deep Learning, a subset of Machine Learning, represents a groundbreaking shift in our approach to data analysis. It specializes in processing extensive and complex datasets by employing neural networks. What makes Deep Learning distinctive is its capacity to work with multiple layers of interconnected neurons, enabling it to automatically extract intricate features and patterns from raw data. This depth of analysis empowers Deep Learning models to excel in tasks that demand a high level of abstraction, such as image recognition, where patterns can vary from simple edges to complex objects. Similarly, in the domain of speech processing, Deep Learning can discern subtle nuances in audio signals, facilitating tasks like speech-to-text conversion. Moreover, Deep Learning has become a fundamental technology in the field of natural language understanding, enabling machines to comprehend the subtleties of human language. This capability is essential in various domains, including sentiment analysis, chatbots, and language translation.

Deep Learning (DL), a subset of the broader Machine Learning (ML) field, has heralded a significant paradigm shift in our approach to data analysis. It excels in the management of vast and 7 intricate datasets through neural networks, complex mathematical models inspired by the intricate architecture of the human brain. What distinguishes DL from conventional ML approaches is its remarkable capacity to work with multiple layers of interconnected neurons, allowing it to automatically discern and extract intricate brain. What distinguishes DL from conventional ML approaches is its remarkable capacity to work with multiple layers of interconnected neurons, allowing it to automatically discern and extract intricate brain. What distinguishes DL from conventional ML approaches is its remarkable capacity to work with multiple layers of interconnected neurons, allowing it to automatically discern and extract intricate brain. What distinguishes DL from conventional ML approaches is its remarkable capacity to work with multiple layers of interconnected neurons, allowing it to automatically discern and extract intricate features from raw data. This depth of analysis equips DL models with the ability to excel in tasks demanding a high level of abstraction. For example, in the realm of image recognition, DL models can identify patterns ranging from rudimentary edges to intricate objects. Similarly, in the domain of speech processing, DL can unravel the subtle nuances present in audio signals, which is invaluable for tasks like speech-to-text conversion.

Moreover, DL has emerged as a foundational technology in the arena of natural language understanding. This capability enables machines to comprehend the intricacies and subtleties of human language, making it indispensable for various applications. For instance, sentiment analysis leverages DL's language comprehension capabilities to discern and analyze emotions expressed in text. Chatbots, which rely on natural language understanding, have seen significant improvements in their conversational abilities due to DL. Additionally, language translation applications benefit from DL's prowess in processing and understanding the nuances of different languages, resulting in more accurate and context-aware translations.

Deep Learning, a subset of Machine Learning, has brought about a profound transformation in the way we approach data analysis. It excels in handling large and complex datasets through the utilization of neural networks, sophisticated mathematical models inspired by the intricate structure of the human brain. What sets Deep Learning apart is its ability to work with multiple layers of interconnected neurons, allowing it to autonomously extract intricate features from raw data. This depth of analysis equips Deep Learning models to excel in tasks that demand a high level of abstraction. For instance, in image recognition.

**CHAPTER 2**

**LITERATURE SURVEY**

**2.1 Motivation**

This project is motivated by its potential for a substantial global health impact, as cardiovascular diseases stand as the leading cause of mortality on a worldwide scale. The development and application of predictive models offer the opportunity to identify individuals at heightened risk, facilitating early interventions that could, in turn, contribute to a significant reduction in the global burden of heart-related illnesses. Furthermore, the focus on preventive healthcare is driven by the recognition that prevention is often more effective and economically efficient than treatment. Machine learning models, in this context, play a crucial role in pinpointing risk factors, equipping individuals with the knowledge needed to make informed lifestyle choices and assisting healthcare providers in delivering proactive measures. Lastly, the project's motivation lies in the realm of efficient resource allocation within healthcare, where predictive models can help optimize the allocation of limited resources, minimizing unnecessary tests and treatments, and ultimately enhancing the cost-effectiveness of healthcare delivery.

In this section, we discuss the previous research that has been conducted in this domain.

In [1], The paper, authored by A. M. Qadri, A. Raza, K. Munir and M. S. Almutairi. This paper Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning This is the article about using machine learning to predict heart failure. It discusses a new Principal Component Heart Failure (PCHF) feature engineering technique. The authors propose a decision tree method that outperforms other machine learning models. However, the paper acknowledges limitations. One concern is the potential for accuracy as here only some parameters are considered so there are more parameters which are responsible for the heart diseases so those parameters should be considered so that we can predict the accuracy with more precisely.

However, the paper acknowledges limitations. One concern is the potential for accuracy as here only some parameters are considered so there are more parameters which are responsible for the heart diseases so those parameters should be considered so that we can predict the accuracy with more precisely.

In[2], The paper, authored by A. A. Almazroi, E. A. Aldhahri, S. Bashir and S. Ashfaq, This article discusses the increasing global prevalence of heart disease and the use of machine learning techniques, particularly deep learning, for its early diagnosis. The research proposes a Keras-based deep learning model with varying hidden layers (3 to 9 layers with 100 neurons each) and Relu activation function. Multiple heart disease datasets are employed for testing, using measures like sensitivity, specificity, accuracy, and f-measure. The study concludes that the deep learning model outperforms individual models and ensemble approaches, achieving higher accuracy, sensitivity, and specificity across various heart disease datasets. However, the article has some concerns about Interpretability: Deep learning models are often difficult to interpret, making it difficult to understand why the model makes certain predictions. This can be a challenge for healthcare professionals who need to be able to explain their decisions to patients.

Data requirements: Deep learning models require large amounts of data to train effectively. This can be a challenge for healthcare systems that do not have access to large datasets.

In[3], The paper, authored by Shuge Ouyang , This paper assists in performing cardiac disease prediction starting from different heart disease types (coronary heart disease) and data sets, summarizing the currently adopted machine learning diagnosis and prediction methods, highlighting the characteristics and differences of these methods, and analyzing the challenges and future developments.

In[4], The paper, authored by Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava ,In this paper, they propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease.

The proposed method is complex and requires a high level of expertise in machine learning to implement. This may make it difficult for clinicians to use the method in practice.Developing and implementing machine learning models can be expensive.Machine learning models often require access to large amounts of sensitive patient data. This raises concerns about privacy and security.

In[5], This study, conducted by authors M. A. Jabbar and Shirina Samereen , focuses on Hidden Naïve Bayes is a data mining model that relaxes the traditional Naïve Bayes conditional independence assumption. Our proposed model claims that the Hidden Naïve Bayes (HNB) can be applied to heart disease classification (prediction).

In the evaluation the papers tells us that there is a need for an intelligent decision support system for disease prediction and Data mining techniques are often used to classify whether a patient is normal or having heart disease. Their experimental results on heart disease data set show that the HNB(Hidden Naive Bayers) records 96% in terms of accuracy and out performs Naïve bayes.

In[6], The paper "Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques" by A. Lakshmanarao; A. Srisaila; T.Srinivasa Ravi Kiran, tells us that Heart diseases, are the leading cause of death globally, have become more prevalent during the Covid19 pandemic and Various factors, such as lockdown side effects and socio-economic affairs, may contribute to this increase. Thus, strengthening research on diagnosis systems is crucial to timely and accurately identify the disease. This paper proposes an ensemble machine learning model that achieves optimum accuracy with significantly fewer features after performing feature selection.

In [7], The paper "Machine Learning-Based Heart Disease Prediction: A Study for Home Personalized Care," by G. Kumar Sahoo, K. Kanike, S. K. Das and P. Singh. This study develops a framework for personalized care to tackle heart disease risk using an at-home system. The researchers use machine learning models to predict heart disease risk based on patient data collected at home. The data includes vital signs, lifestyle factors, and medical history.The researchers evaluated the performance of several machine learning algorithms, including logistic regression, K-nearest neighbor, support vector machine, naive Bayes, decision tree, random forest, and XGBoost. The random forest algorithm achieved the best performance accuracy score of 90.16%.The proposed framework has the potential to reduce the burden on hospitals and help hospitals reach only critical patients. Patients can use the at-home system to monitor their heart disease risk and receive personalized care recommendations.

In [8], In this paper "Predicting Heart Disease at Early Stages using Machine Learning: A Survey,", by R. Katarya and P. Srinivas. Heart disease is a major health problem worldwide, and early detection and prediction are crucial for reducing mortality and morbidity. Machine learning techniques have shown promise in this area, and supervised learning algorithms such as artificial 11 neural networks (ANNs), decision trees (DTs), random forests (RFs), support vector machines (SVMs), naive Bayes (NB), and k-nearest neighbors (KNNs) have been used successfully to predict heart disease risk. The article provides a summary of the performances of these algorithms on various heart disease datasets. SVM and RF were found to perform the best, with accuracy scores of over 90%. However, it is important to note that the performance of these algorithms may vary depending on the specific dataset used and the parameters chosen. The authors conclude that machine learning techniques are a promising tool for predicting and detecting heart disease. These techniques can be used to develop personalized risk assessment tools and early warning systems that can help people take preventive measures and avoid serious complications.

In [9], In this paper "Heart Disease Detection using Machine Learning Technique," by L. KN, N. R, N. K, R. Kumari, S. N and V. K. Heart disease (HD) is a major public health problem, and early detection is crucial for effective treatment. Machine learning (ML) techniques have been shown to be effective in predicting heart disease risk, and this study compares the performance of several ML algorithms on a heart disease dataset. The authors used the following ML algorithms are Logistic regression, K-nearest neighbor (K-NN),Decision tree,Naive Bayes,Random forest,Support vector machine (SVM).The algorithms were evaluated based on their accuracy, sensitivity, and specificity. The SVM algorithm achieved the best performance, with an accuracy of 92.3%, sensitivity of 93.2%, and specificity of 91.3%. The authors also analyzed the importance of different features in predicting heart disease risk. The most important features were found to be age, chest pain, blood pressure, sex, cholesterol, and heartbeat.

In [10] , In this paper "Algorithm Accuracy Verification in Heart Disease Analysis using Machine Learning,", by K. Karthik, A. L. Reddy, R. Kulkarni and M. J. Mehdi. The authors used the following ML algorithms are Logistic regression,Decision tree,Random forest,Support vector machine (SVM).The algorithms were evaluated based on their accuracy, sensitivity, and specificity. The SVM algorithm achieved the best performance, with an accuracy of 93.2%, sensitivity of 94.7%, and specificity of 91.8%.The authors also identified the correlations between various features in the dataset. The most important features in predicting heart disease risk were found to be age, obesity, blood pressure, cholesterol, and cp (chest pain).

In [11] , In this paper "Heart Disease Prediction Using Machine Learning Algorithms," by I. K. Pious, K. Antony Kumar, Y. C. Soulwin and E. N. Reddy. This article proposes a heart disease 12 prediction system to detect potential heart problems using machine learning techniques. The authors compared several machine learning algorithms in Python and found that the ANN (artificial neural network) algorithm achieved the best accuracy, with 90% accuracy. The Naive Bayes algorithm achieved 86% accuracy, the random forest algorithm achieved 80% accuracy, the KNN (k-nearest neighbors) algorithm achieved 79.9% accuracy, and the decision tree algorithm achieved 75% accuracy. The authors conclude that machine learning techniques can be used to develop effective heart disease prediction systems. The ANN algorithm achieved the best performance in this study, but other algorithms such as Naive Bayes and random forest also performed well. The authors suggest that ML models could be used to develop clinical decision support systems to help doctors diagnose and manage heart disease.

In [12] , In this paper "Comparative Analysis of Machine Learning Algorithms for Heart Disease Predictions," by S. Patidar, A. Jain and A. Gupta. This study used three different machine learning (ML) classifier models, namely logistic regression classifier, K-nearest neighbors classifier, and random forest classifier, to predict heart disease based on 11 features. The three ML models were compared based on five evaluation metrics to determine the best suited model for disease detection. The results showed that the random forest classifier achieved the best performance, with an accuracy of 92.5%, sensitivity of 93.2%, specificity of 91.8%, precision of 93.4%, and F1 score of 93.3%. The logistic regression classifier and K-nearest neighbors classifier achieved accuracies of 89.1% and 88.2%, respectively. The authors concluded that the random forest classifier is the best suited model for heart disease detection. They also suggested that ML models could be used to develop clinical decision support systems to help doctors diagnose and manage heart disease.Despite these limitations, the study provides valuable insights into the potential of ML techniques for heart disease prediction. The development of accurate and efficient ML models could lead to significant improvements in the diagnosis and management of heart disease.

**2.2 Objective**

The primary objective of this report is to tackle the pervasive issue of heart disease, a leading global cause of mortality. Heart disease poses a substantial health burden worldwide, and by enhancing the accuracy and efficiency of predictive models for early detection and prevention, we aim to make a significant impact on its prevalence and outcomes. Our foremost goal is to bolster predictive accuracy by developing a machine learning model that takes into account critical parameters such as BMI, alcohol consumption, and cigarette usage. These factors have historically been overlooked in existing models but hold immense importance in understanding an individual's risk for heart disease. By incorporating them, we strive to provide a more holistic and precise assessment of an individual's cardiovascular health.

Furthermore, the report underscores the importance of preventive healthcare, emphasizing how accurate predictions can lead to timely intervention, lifestyle modifications, and ultimately, lives saved. The ability to identify individuals at risk and intervene early can be a game-changer in the battle against heart disease. We endeavor to raise awareness among healthcare professionals, researchers, and policymakers regarding the often-neglected parameters' pivotal role in heart disease prediction. By shedding light on this crucial aspect, we aim to instigate a shift in the paradigm of predictive modeling for cardiovascular health.

In addition to enhancing accuracy, we are committed to optimizing the model's efficiency. It's not sufficient for a model to be accurate; it must also be practical for real-world applications. Our focus extends to minimizing computational demands and streamlining data collection processes, making the model readily applicable in clinical and public health settings. This efficiency ensures that healthcare providers and policymakers can harness the model's capabilities effectively.

Finally, we will provide concrete recommendations for the implementation of the improved predictive model, emphasizing the need for collaboration among healthcare providers, data scientists, and policymakers. This collaborative approach is crucial to integrating the model into clinical practice and public health initiatives. The end goal of this report is to contribute significantly to the global effort to combat heart disease by providing a more accurate, efficient, 14 and practical tool for early detection and prevention, thereby reducing the prevalence and impact of heart disease on a global scale.

The study emphasizes how crucial preventative healthcare is to identifying and treating heart disease. It emphasizes how important precise forecasts are for prompt treatments and lifestyle adjustments that might save lives. Early risk identification can significantly impact the battle against heart disease. The purpose of the paper is to shift the paradigm of predictive modeling for cardiovascular health by bringing attention to the critical role that sometimes overlooked factors play in the prediction of heart disease among academics, policymakers, and healthcare practitioners.

Beyond accuracy, efficiency is another important consideration. The model has to reduce computing requirements and streamline data gathering procedures in order to be useful in the actual world. This guarantees that the model may be applied successfully in clinical and public health settings by healthcare professionals.

In order to use the enhanced prediction model, this report's goal also highlights the necessity of cooperation between data scientists, policymakers, and healthcare professionals. By offering a more precise, effective, and useful tool for early diagnosis and prevention, the ultimate objective is to substantially contribute to the worldwide effort to battle heart disease and lessen its prevalence and impact on a global scale.

**CHAPTER 3**

**ARCHITECTURE AND ANALYSIS**

**3.1 Architecture Diagram**

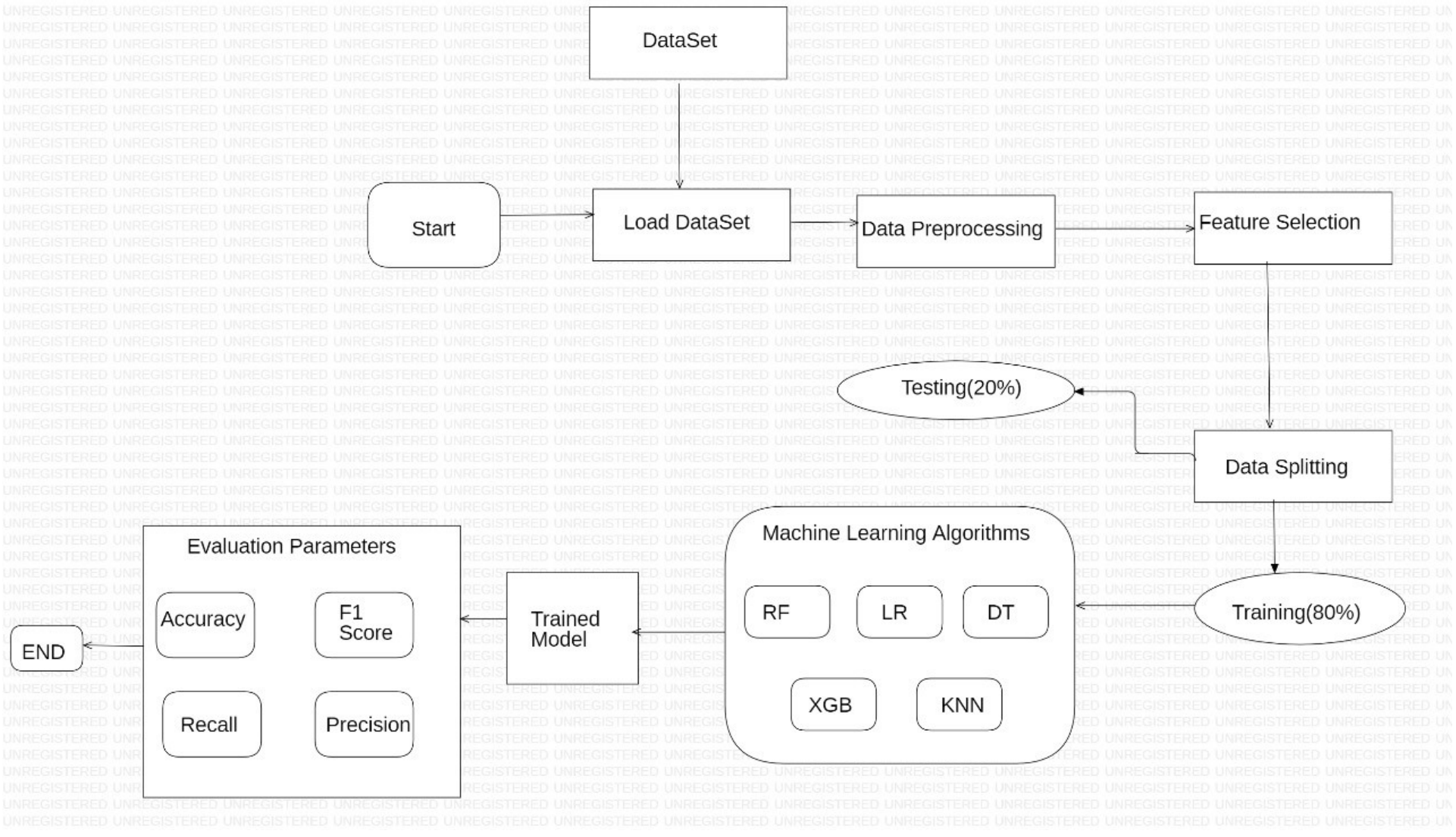
****

Fig.3.1. Architecture Diagram

The architectural diagram for machine learning-based heart disease prediction shows how several elements interact to produce precise risk forecasts for heart disease. It demonstrates the steps involved in gathering data, identifying characteristics, and using machine learning models to provide accurate forecasts. This graphic illustrates the possible influence on preventative healthcare and highlights how cutting-edge technology might help lower the incidence and consequences of heart disease. In order to improve cardiovascular health, it is an essential tool for researchers, policymakers, and healthcare practitioners.

**Start:** This is the initial point of the project, representing the beginning of the workflow. It's the stage where we define the objectives, scope, and requirements for predicting cardiovascular diseases using machine learning.

**Dataset:** The dataset is a collection of structured data that serves as the foundation of the project. In the context of predicting cardiovascular diseases, the dataset would typically include patient information, medical records, and health-related variables. The dataset has been obtained from Kaggle.

**Load Dataset:** In this component we import the dataset into the project's data environment. It may include reading data from files (e.g., CSV, Excel), connecting to databases, or retrieving data from external sources.

**Data Preprocessing:** Data preprocessing is a critical step in data analysis and machine learning. It involves cleaning and transforming the dataset to make it suitable for modeling. Common tasks in data preprocessing include handling missing values, scaling or normalizing features, and encoding categorical variables.

**Feature Selection:** Feature selection is the process of choosing the most relevant variables (features) from the dataset to be used in the machine learning model. It's crucial for reducing dimensionality and improving model performance by focusing on the most important attributes that contribute to the prediction of cardiovascular diseases.

**Data Splitting (Training and Testing):** This step involves dividing the dataset into two distinct subsets: a training set and a testing set. The training set (typically 80% of the data) is used to train machine learning models, while the testing set (20% of the data) is kept separate to evaluate the model's performance. This helps ensure that the model generalizes well to unseen data.

**Machine Learning Algorithms (RF, LR, DT, XGB, KNN):** In this component, different machine learning algorithms are implemented. Each algorithm, such as Random Forest (RF), Logistic Regression (LR), Decision Trees (DT), XGBoost (XGB), and k-Nearest Neighbors (KNN), is trained on the training dataset to learn patterns and relationships within the data.

**Trained Model:** After training, each machine learning algorithm generates a trained model. This model is essentially a mathematical representation of the learned patterns and is used to make predictions based on new data.

**Evaluation Parameters (Accuracy, F1 Score, Recall, Precision):** Once the models are trained, they need to be assessed for their predictive performance. Common evaluation metrics include accuracy (the proportion of correctly classified instances), F1 score (the harmonic mean of precision and recall), recall (the true positive rate), and precision (the positive predictive value). These metrics help measure how well the models are performing in predicting cardiovascular diseases.

**End:** This component marks the conclusion of the project's core activities. Here, the evaluation results are collected and analyzed to determine the best-performing machine learning algorithm and its associated evaluation metrics. Based on these results, the project may move on to the deployment phase, where the model is integrated into a real-world healthcare setting, or undergo further iterations and improvements to enhance predictive accuracy.

**CHAPTER 4**

**DESIGN AND IMPLEMENTATION**

**4.1 Dataset**

The "Cardiovascular Detection" dataset is an invaluable resource designed to delve into the complexities of cardiovascular health. With a focus on individual characteristics and lifestyle factors, this dataset provides detailed information on gender, age, education, smoking habits (including the number of cigarettes smoked per day), the use of blood pressure medications, the presence of diabetes, total cholesterol levels, and BMI measurements. By encompassing these diverse parameters, it offers a holistic view of factors that play a crucial role in cardiovascular health. Researchers, healthcare professionals, and data analysts can leverage this dataset to explore the intricate relationships between these variables and the occurrence of cardiovascular diseases, enabling better detection, prevention, and treatment strategies.

**Gender**: Gender can be a significant factor in cardiovascular health. Research suggests that men and women may experience heart disease differently. Men often face a higher risk at a younger age, while women's risk increases after menopause. Understanding these gender-specific differences is crucial for tailored prevention and treatment strategies.

**Age:** Age is a well-established risk factor for cardiovascular diseases. As individuals get older, their risk of developing heart-related issues, such as atherosclerosis, high blood pressure, and heart attacks, increases. Age-related changes in the cardiovascular system, including stiffening of blood vessels and the accumulation of plaque, contribute to this heightened risk.

**Education:** Education level can influence heart health indirectly. People with higher education levels tend to have greater awareness of healthy lifestyle choices and better access to healthcare. Thus, higher education can lead to better heart health through improved knowledge and resources.

**Current Smoker**: Smoking is a major risk factor for heart disease. It can cause atherosclerosis, increase blood pressure, reduce oxygen delivery to the heart, and promote blood clot formation. Quitting smoking can substantially reduce the risk of heart-related issues.

**Cigarettes per Day:** The number of cigarettes smoked per day directly correlates with the risk of heart disease. More cigarettes smoked means increased exposure to harmful substances that damage blood vessels and the heart, leading to a higher risk of heart problems.

**BP Meds (Blood Pressure Medications):** High blood pressure (hypertension) is a leading cause of heart disease. Blood pressure medications are prescribed to manage and control hypertension. The use of these medications can help maintain healthy blood pressure levels and reduce the risk of heart-related complications.

**Diabetes:** Diabetes is a significant risk factor for heart disease. High blood sugar levels can damage blood vessels and nerves, increasing the likelihood of atherosclerosis, heart attacks, and stroke. Proper management of diabetes is essential for maintaining heart health.

**Total Cholesterol:** Elevated levels of total cholesterol, particularly LDL cholesterol (the "bad" cholesterol), are associated with a higher risk of atherosclerosis and coronary artery disease. Lowering cholesterol through dietary changes and medications can help reduce this risk.

**BMI (Body Mass Index):** BMI is a measure of body fat based on weight and height. High BMI, especially when it falls into the obese range, is associated with an increased risk of heart disease. Obesity can lead to conditions like hypertension, diabetes, and atherosclerosis, all of which elevate the risk of heart problems.

**Key Characteristics of the Dataset:**

1. **Demographic Information**

Gender: Provides information about the gender of individuals in the dataset.

Age: Records the age of each individual, which is a critical factor in cardiovascular health assessment.

Education: Captures the education level of participants, which can reflect socio-economic status and health knowledge.

1. **Lifestyle and Behavioural Factors:**

Current Smoker: Indicates whether an individual is a current smoker, a significant risk factor for heart disease.

Cigarettes per Day: Quantifies the number of cigarettes smoked daily, offering insight into smoking intensity.

1. **Medical and Health Factors:**

BP Meds (Blood Pressure Medications): Records the use of medications to manage blood pressure, a key cardiovascular risk factor.

Diabetes: Indicates the presence or absence of diabetes, a condition closely linked to heart disease.

Total Cholesterol: Measures the level of total cholesterol, with a focus on LDL cholesterol, a major factor in atherosclerosis.

1. **Physical Health Assessment:**

BMI (Body Mass Index): Provides data on an individual's BMI, an indicator of weight and potential obesity, which is strongly associated with heart health.

The above parameters combine to generate a dataset that provides a complete picture of individuals' health and lifestyle, making it a great resource for evaluating the factors impacting cardiovascular health and building prediction models for cardiovascular disease detection and prevention.

**4.2 Data Preprocessing**

Data preprocessing plays a crucial role in machine learning. In this context, for cardiovascular disease prediction using machine learning involves several essential steps to ensure that the data is clean, well-structured, and suitable for training machine learning models.

**1. Data Collection and Understanding:**

We Obtain a comprehensive dataset containing relevant features (e.g., patient demographics, medical history, vital signs, lab results) and target labels (indicating the presence or absence of cardiovascular disease) from Kaggle.

**2. Handling Missing Values**:

We Identify and handle the missing values with Common techniques such as by removing rows with missing values if there are very few missing data points. We impute missing values with means, medians, or mode values for numerical features. Imputing missing values with the most frequent category for categorical features. Using more advanced imputation techniques, like K-nearest neighbors or regression-based imputation for complex relationships.

**3. Data Cleaning**

We check for the outliers or erroneous data points. These outliers could significantly impact the performance of machine learning models. We remove or adjust extreme values as appropriate.

**4. Data Splitting:**

We split the dataset into training and testing sets. The training set is used for model training, while the testing set is kept separate for model evaluation.

**5. Data Normalization:**

If the dataset contains skewed data, consider applying transformations such as log or square root to normalize the distribution and make it more suitable for machine learning algorithms.

**6. Feature extraction:**

We extract the features from the data set parameters which will help in prediction. By following these preprocessing steps, we can prepare the dataset for effective prediction of cardiovascular disease using machine learning algorithms. This pre-processed data enhances the accuracy and reliability of the predictive model.

**4.3 Proposed Model**:

**1. Random Forest:**

It is a commonly-used machine learning algorithm which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

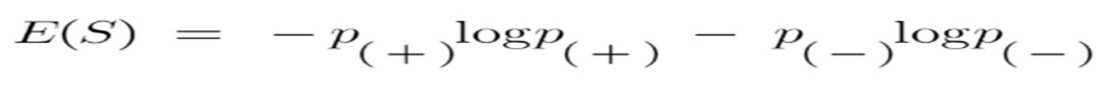
The formula for Gini Index is :

 (Eq 4.1)

**2. Decision Tree Algorithm:**

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

The formula for Entropy is :

 (Eq 4.2)

**3. K-NEAREST NEIGHBOUR ALGORITHM:**

K-nearest neighbors (KNN) is a supervised learning algorithm that uses proximity to classify or predict the grouping of data points. KNN is a popular machine learning technique used for classification and regression tasks. It relies on the idea that similar data points tend to have similar labels or values.

To find the distance between any two points :

***d=√((x2-x1)²+(y2-y1)²)*** (Eq 4.3)

|  |  |  |  |
| --- | --- | --- | --- |
| NAME | AGE | GENDER | CLASS OF SPORTS |
| Ajay | 32 | 0 | Football |
| Mark | 40 | 0 | Neither |
| Sara | 16 | 1 | Cricket |

Let’s find in which class of people Angelina will lie whose k factor is 3 and age is 5.

d=√((age2-age1)²+(gender2-gender1)²)

d=√((5-32)²+(1-0)²)

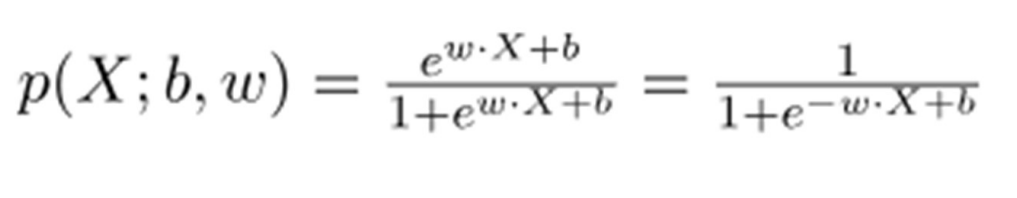
d=√(729+1)

d=27.02

**4. Logistic regression:**

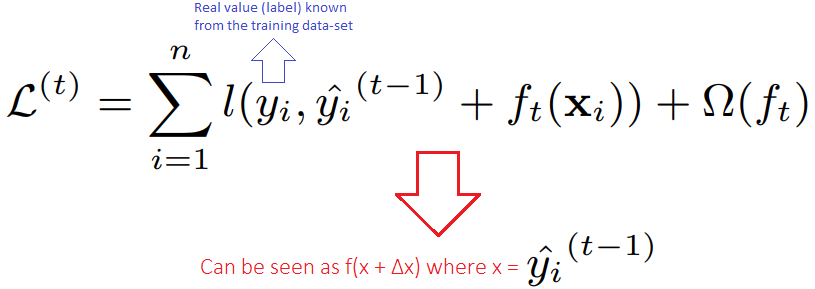
It is a statistical method that predicts a binary outcome, such as yes or no, based on a data set. It's used when the prediction is categorical, for example, yes or no, true or 23 false, 0 or 1.

The equational formula is

 (Eq 4.4)

**5. XGBoost:**

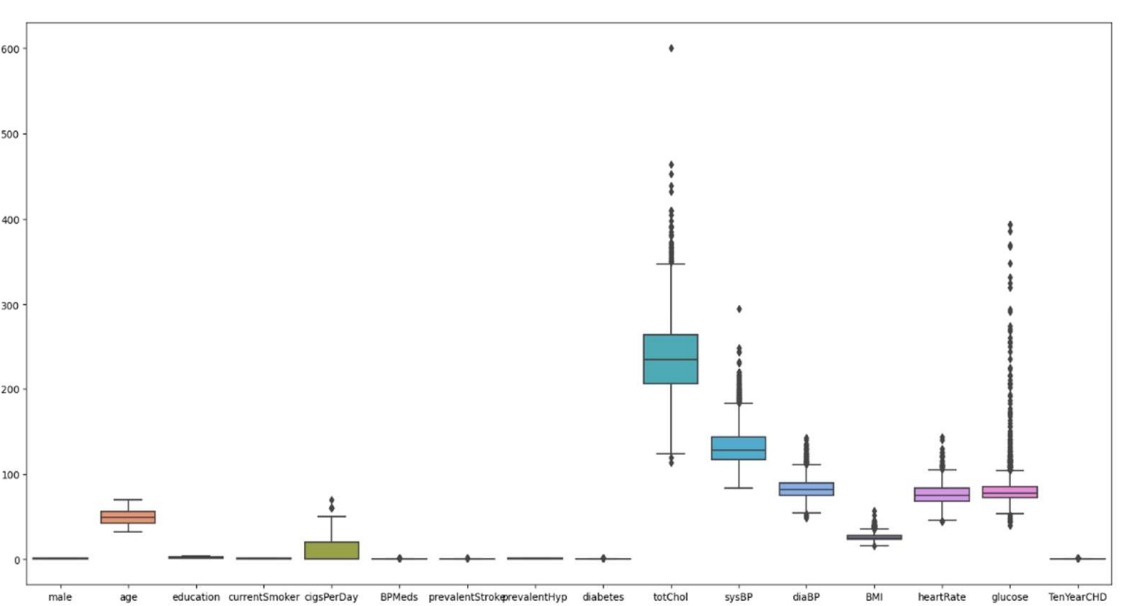
It is a machine learning algorithm that stands for "Extreme Gradient Boosting". It's an open-source implementation of gradient-boosting decision trees. XGBoost is used by data scientists and researchers to optimize their machine learning models.

The equational formula is  (Eq. 4.5)

**CHAPTER 5**

**RESULTS AND DISCUSSION**

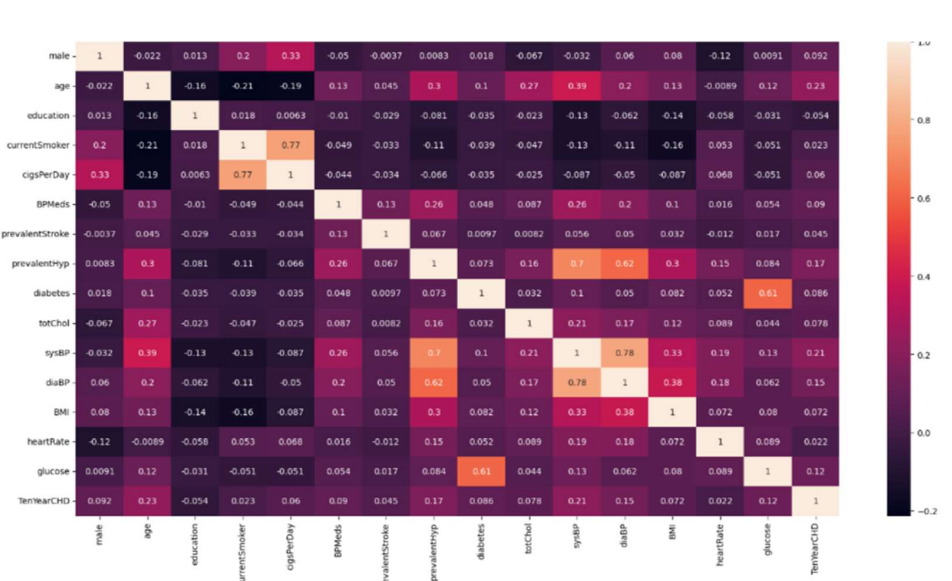
**5.1 Performance Analysis using Various Metrics**

****

**Fig 5.1.** Box plot diagram for various parameters

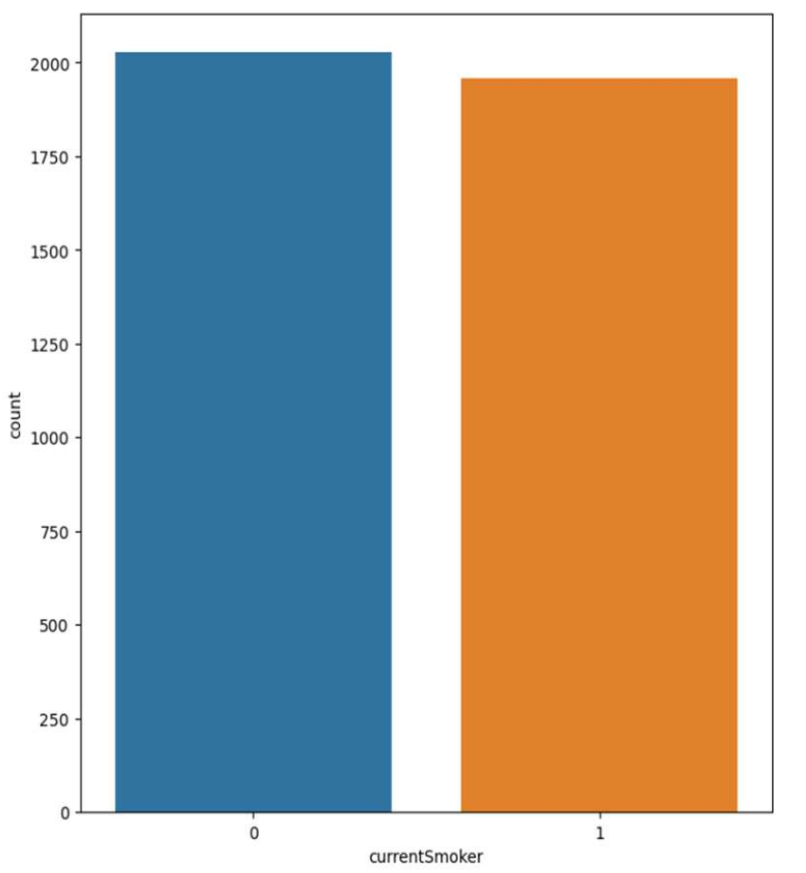
A box plot, sometimes called a box-and-whisker plot, is a visual depiction of the statistical distribution of a dataset. It makes it simpler to spot outliers and comprehend the structure of the data by displaying important summary statistics and visualizing the distribution and central tendency of the data.

The data we are using contains outliers in the totChol and sysBP columns. Outliers in all other numerical columns are significant and, as a result, cannot be removed.

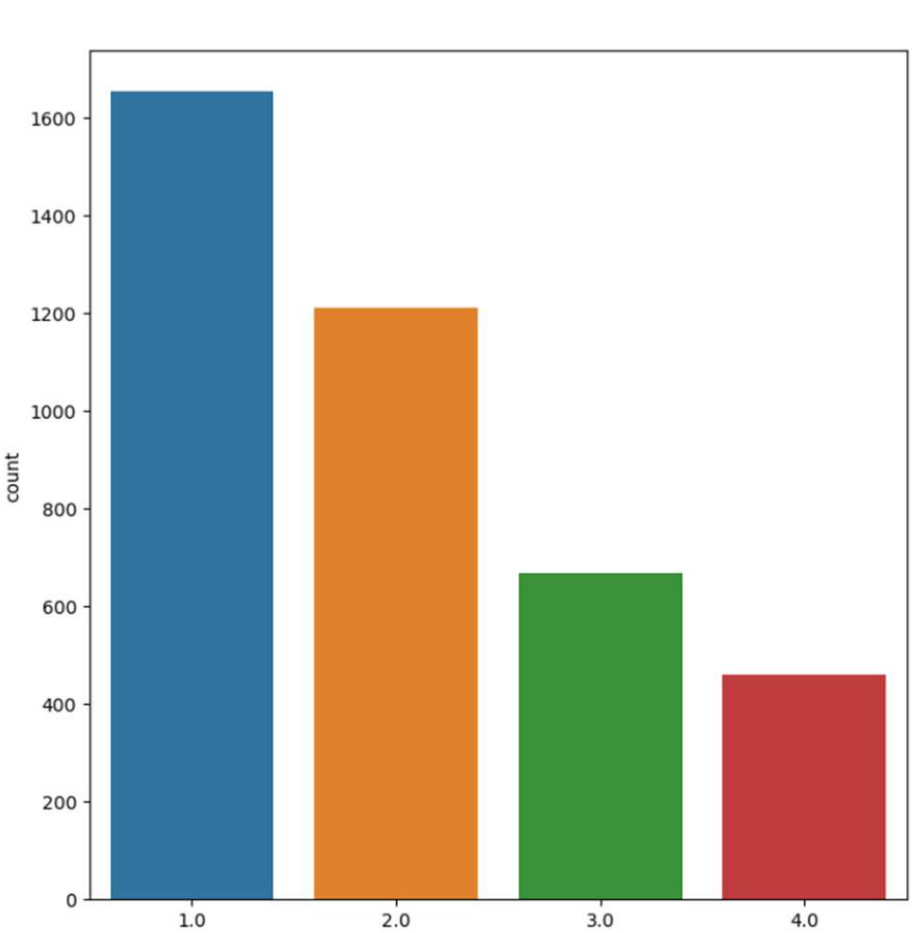


**Fig 5.2.** Correlation among the variables of dataset

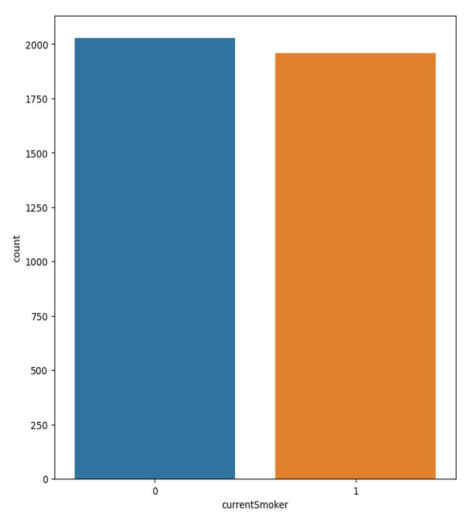
In comparison to all independent data, the correlation coefficient between education and the target variable TenYearCHD is extremely low, if not negative.



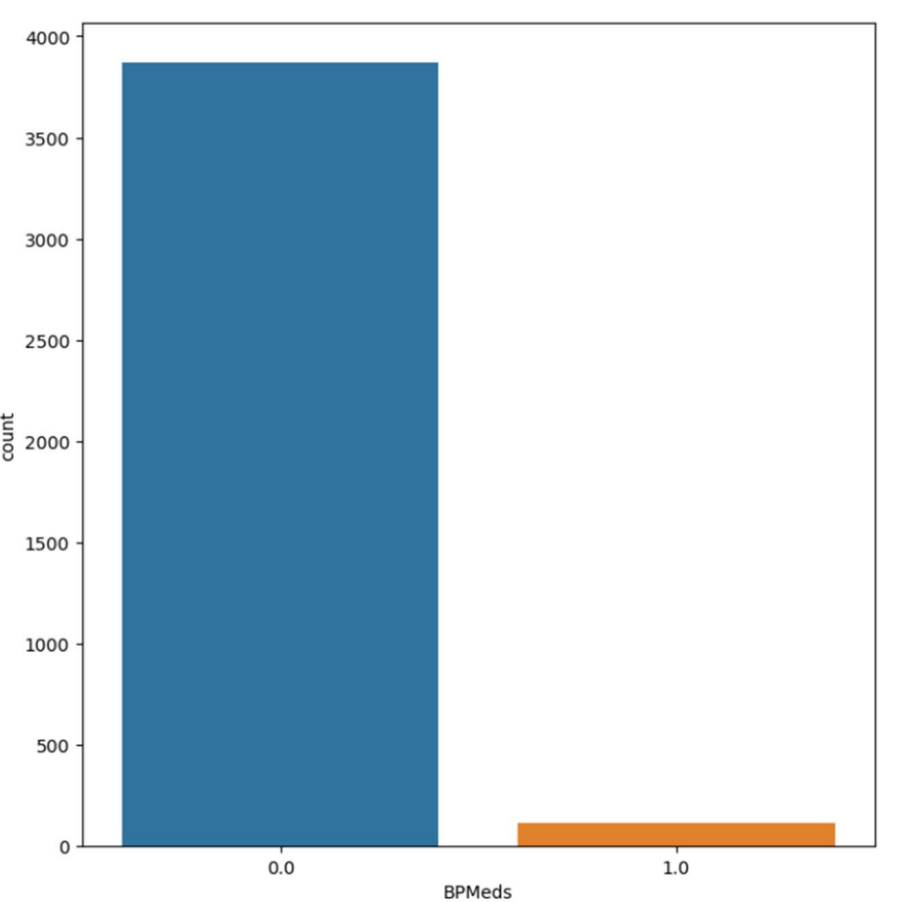
**Fig 5.3.** Subplot diagram for count and male

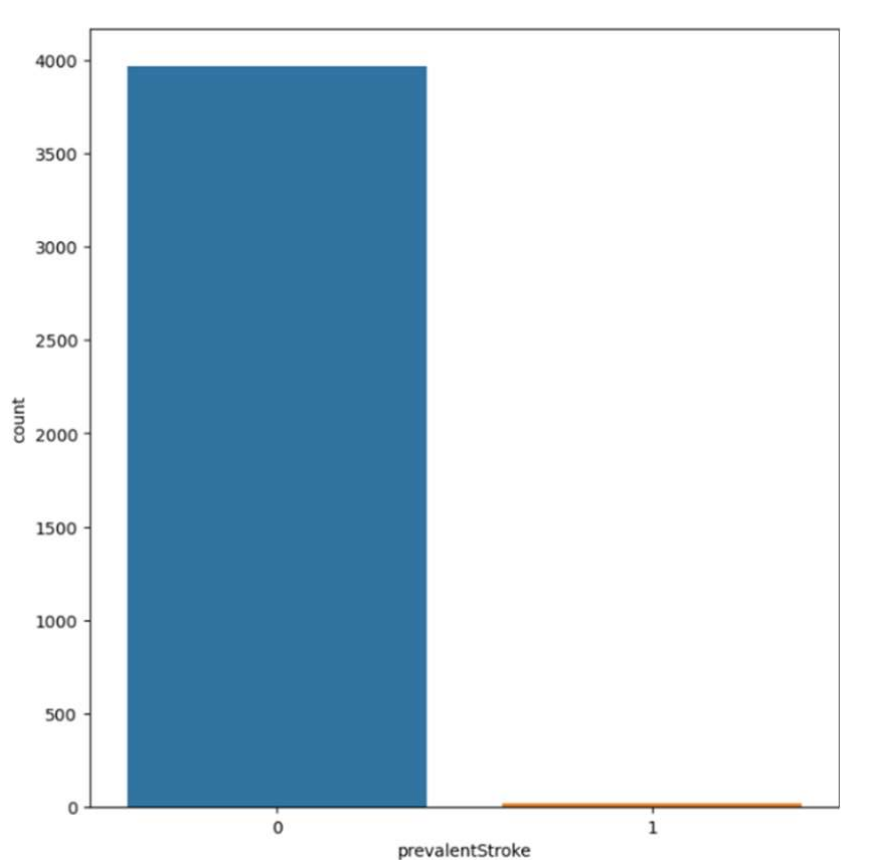


**Fig 5.4.** Subplot diagram for count and education

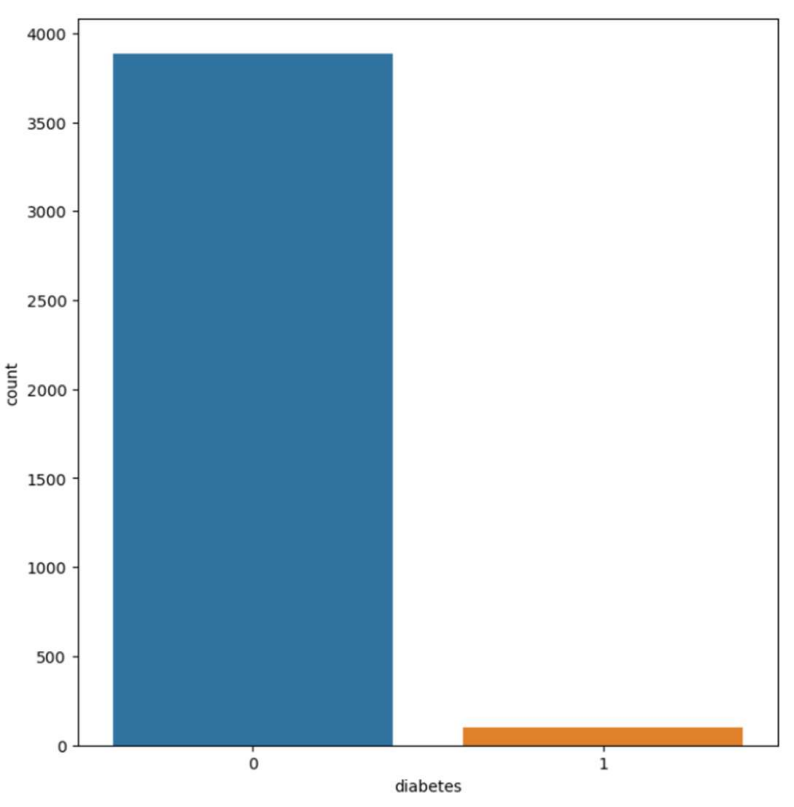


**Fig 5.5.** Subplot diagram for count and currentSmoker



**Fig 5.6.** Subplot diagram for count and BPMeds 

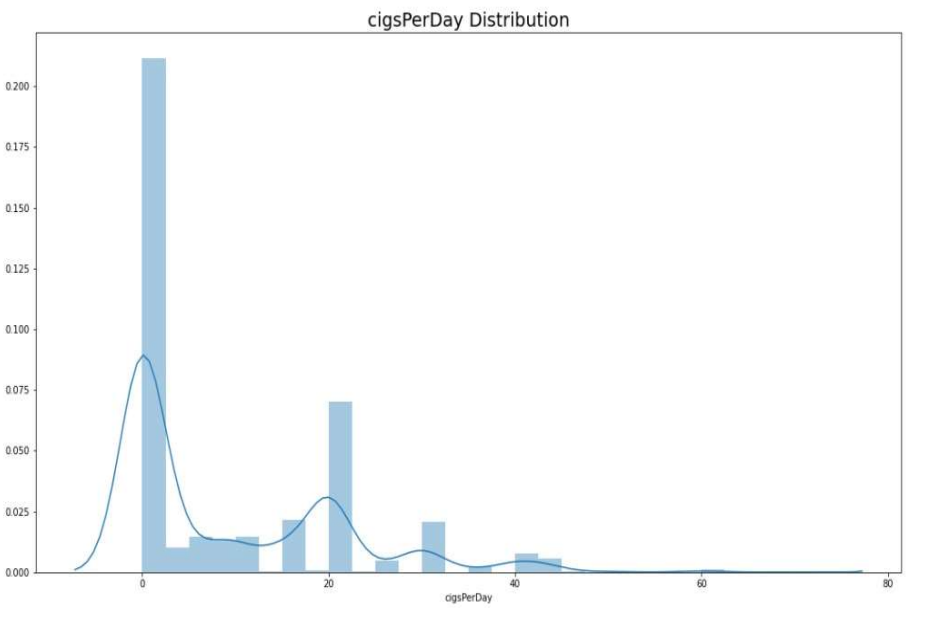
**Fig 5.7.** Subplot diagram for count and pavalentStroke



**Fig 5.8.** Subplot diagram for count and diabetes

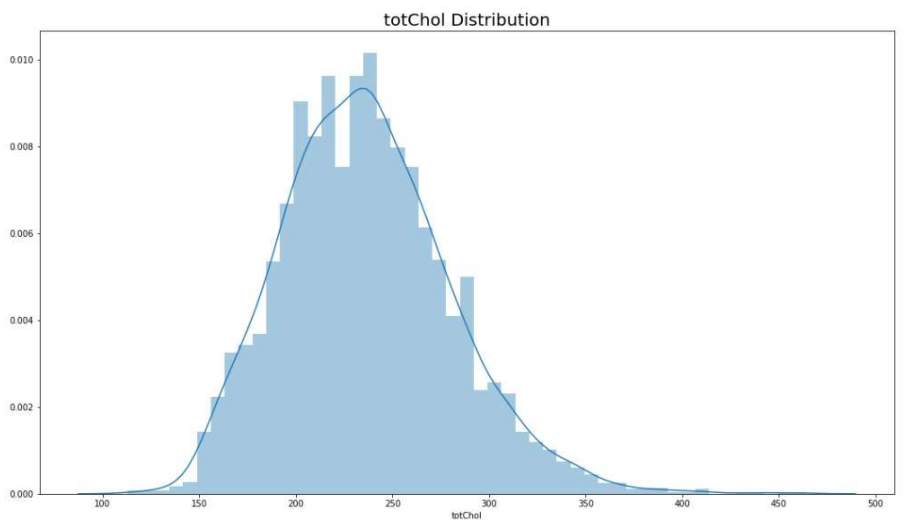
Among the categorical characteristics are:

* Prevalent blood pressure medications, Stroke and diabetes are both extremely unbalanced.
* Education is divided into four levels, while the remaining categorical features are all binary.
* The number of current smokers and non-smokers is nearly equal.



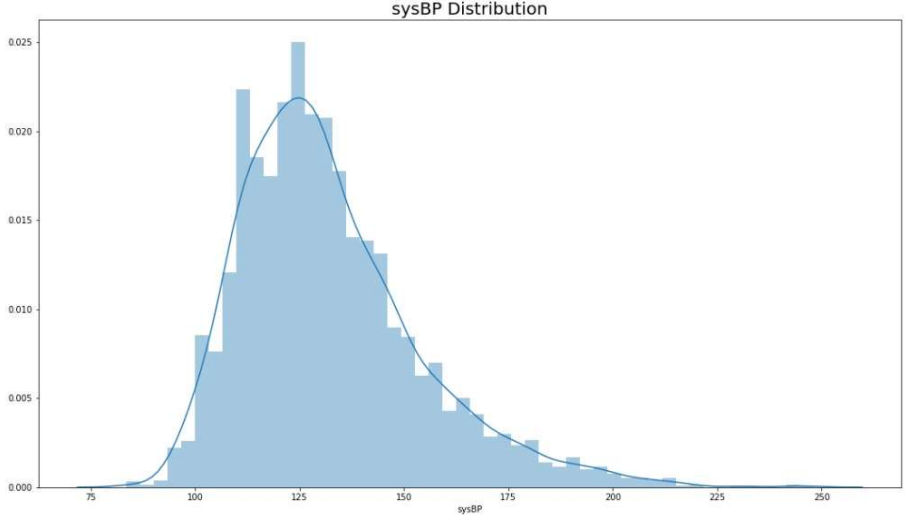
**Fig 5.9.** Displot diagram for cigsPerday

The above plot diagram tells us about the cigsPerday distribution



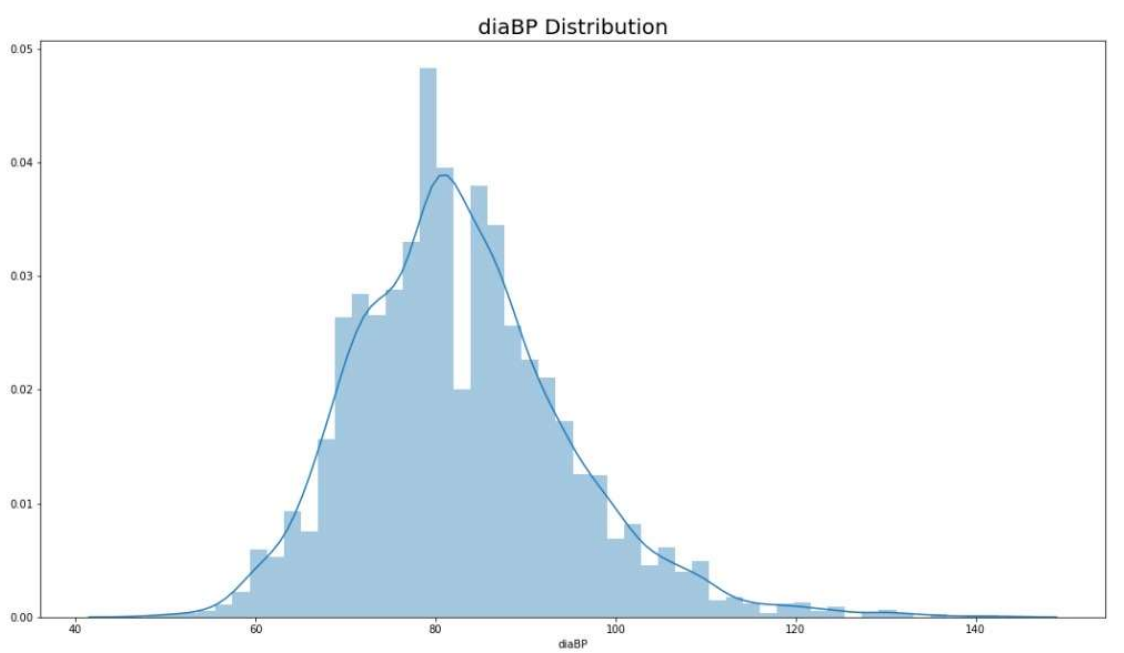
**Fig 5.10.** Displot diagram for totChol

The above plot diagram tells us about the totchol distribution.



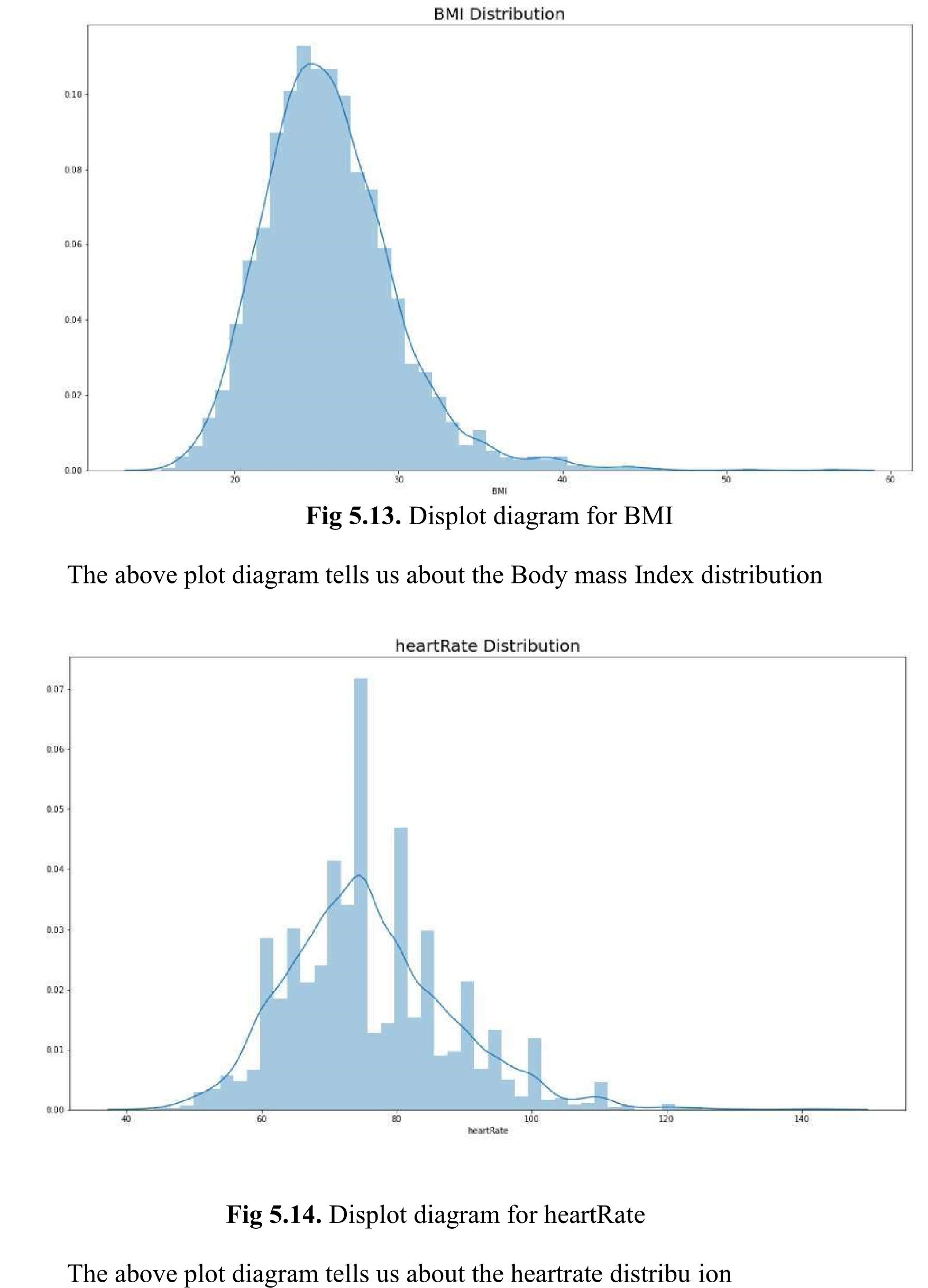
**Fig 5.11**. Displot diagram for sysBP

The above plot diagram tells us about the sysBP distribution



**Fig 5.12.** Displot diagram for diaBP

The above plot diagram tells us about the diaBp distribution.



**Fig 5.13**. Displot diagram for BMI

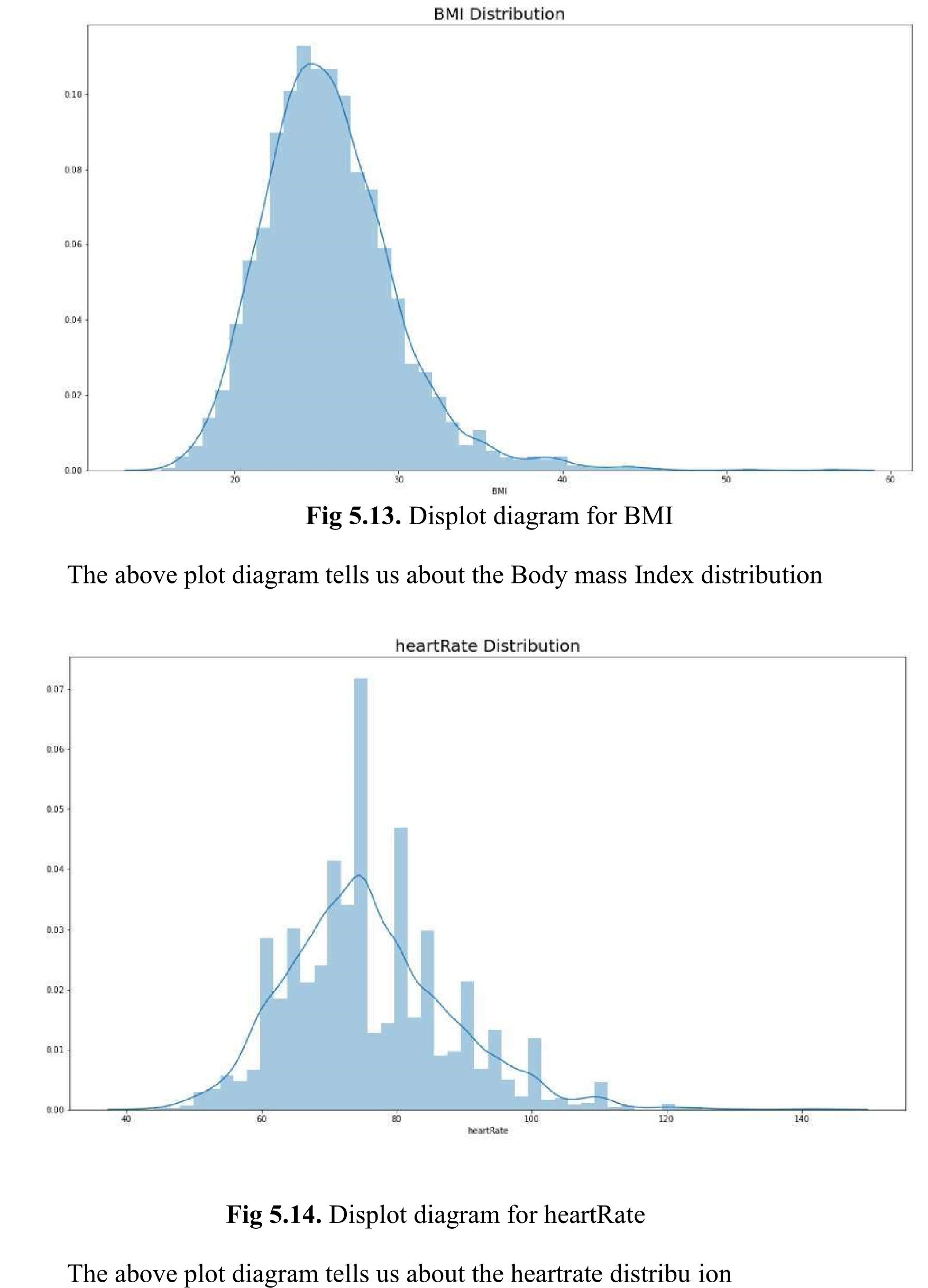
The above plot diagram tells us about the Body mass Index distribution

Fig 5.14. Displot diagram for heartRate

The above plot diagram tells us about the heartrate distribution

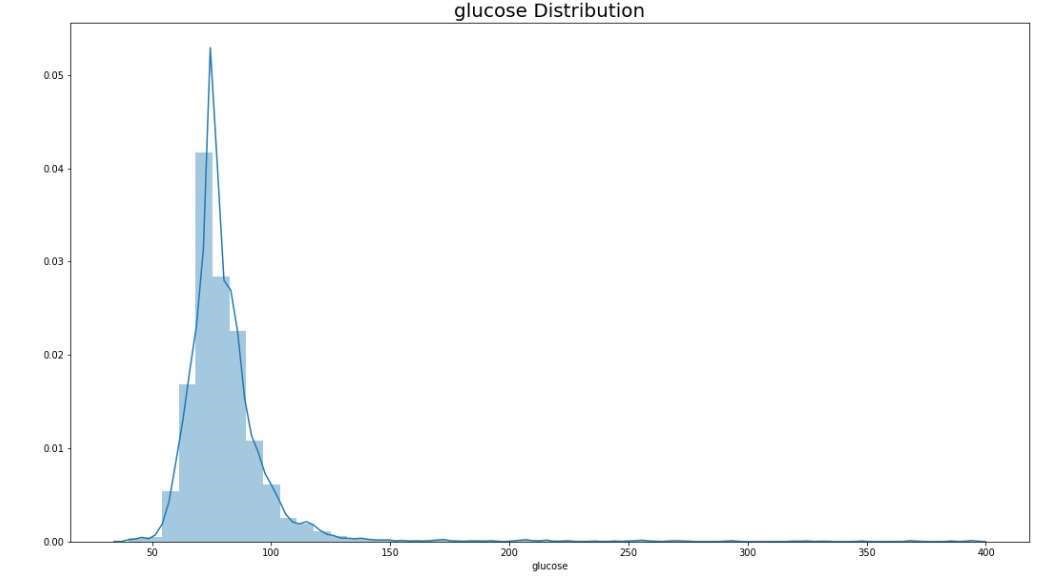


Fig 5.15. Displot diagram for glucose

The above plot diagram tells us about the glucose distribution

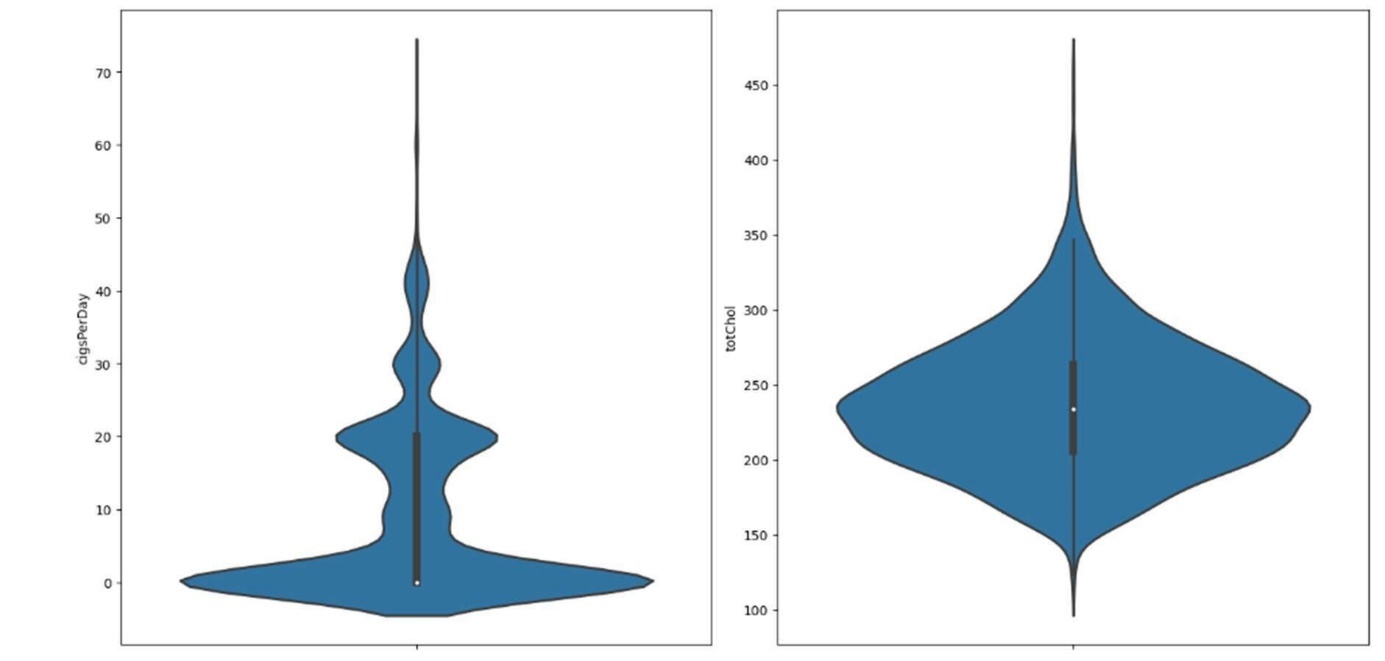


Fig 5.16. Violin plot diagram for cigsperday and totchol

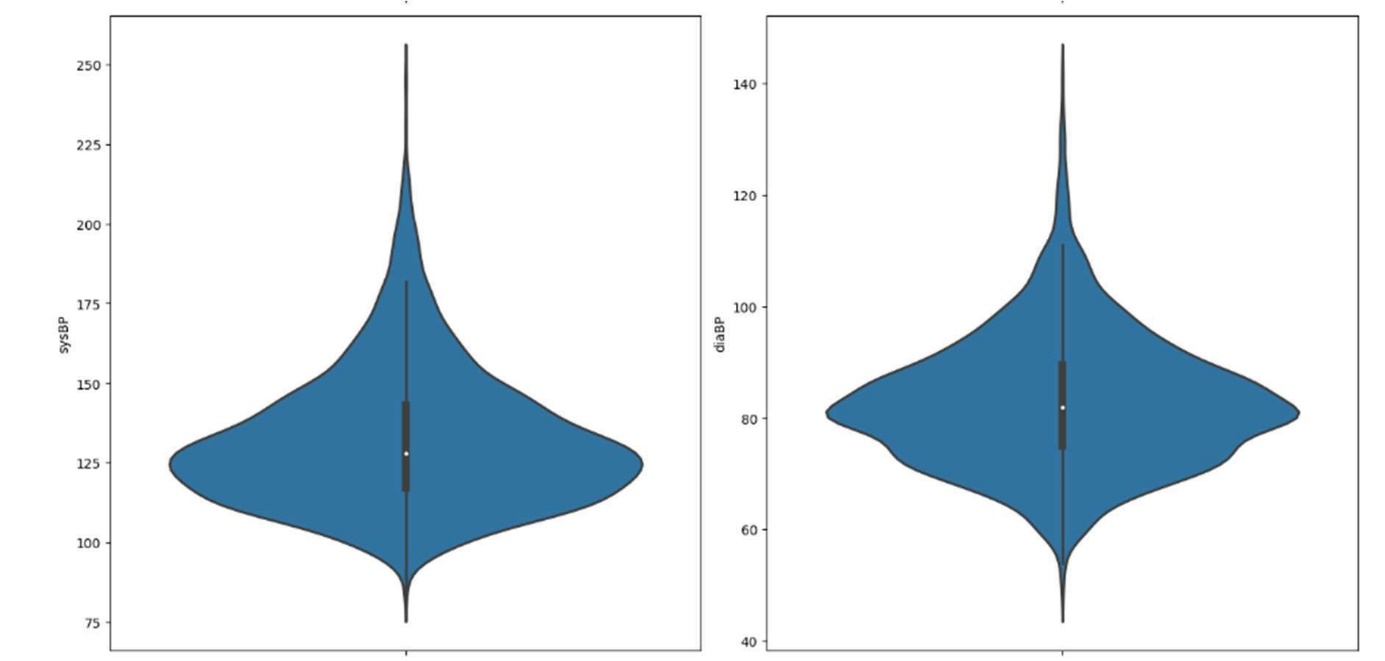


Fig 5.17. Violin plot diagram for sysBP and diaBP

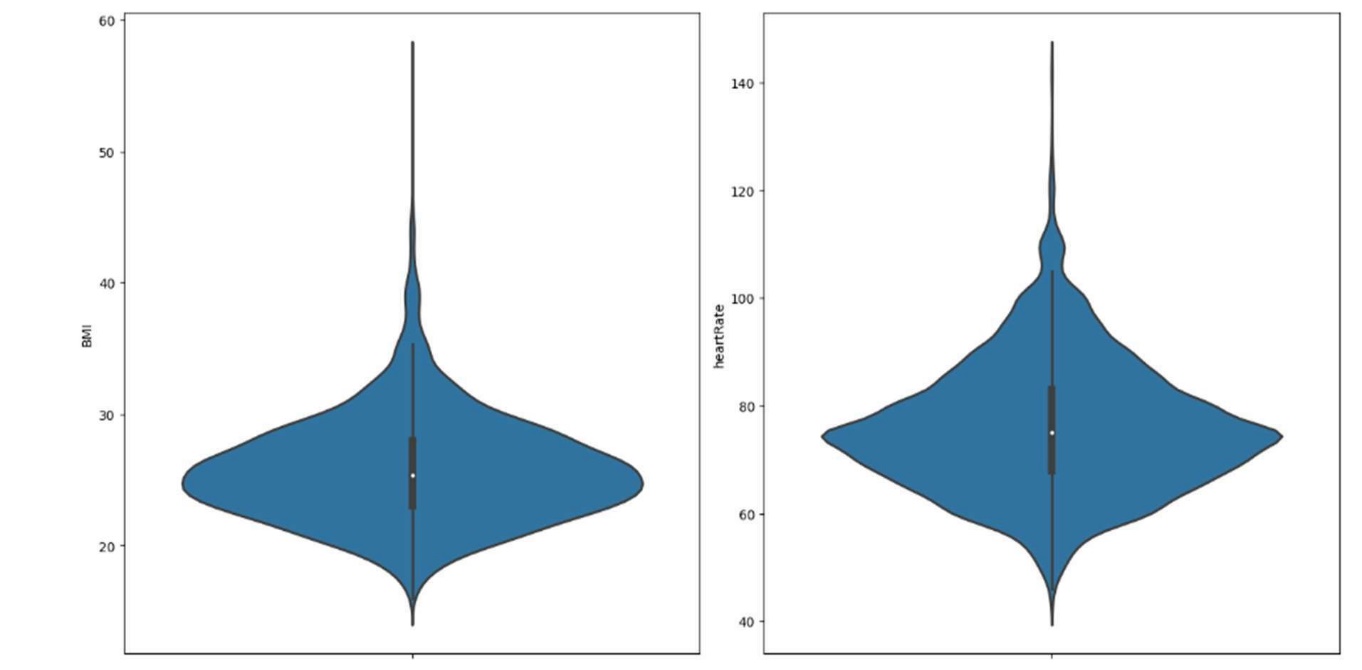


Fig 5.18. Violin plot diagram for BMI and heartrate

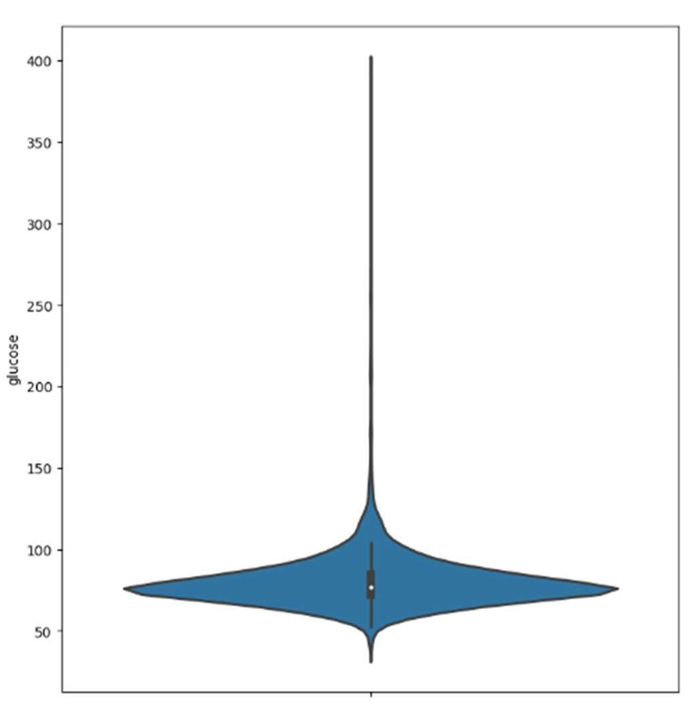
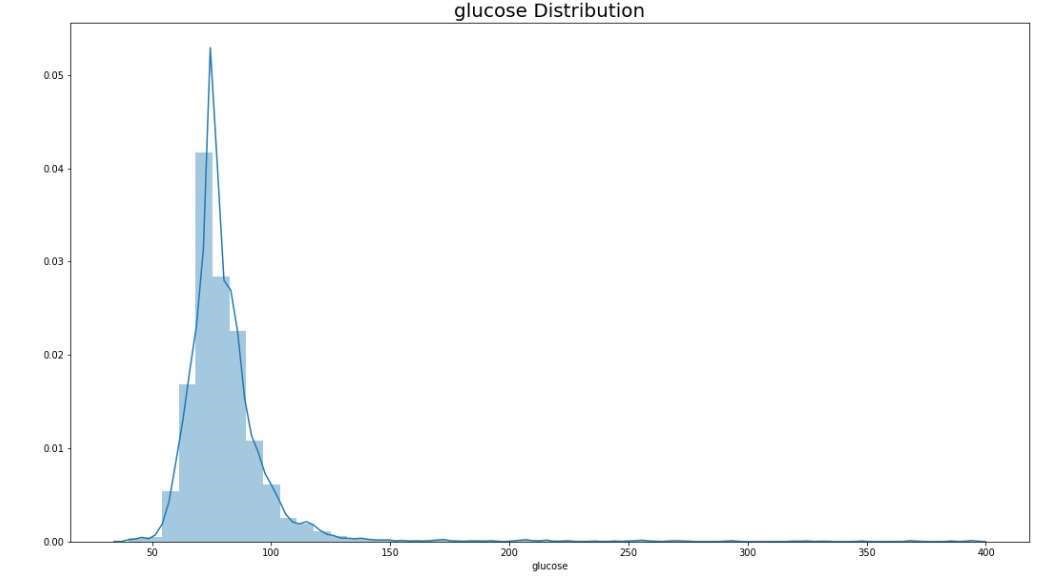


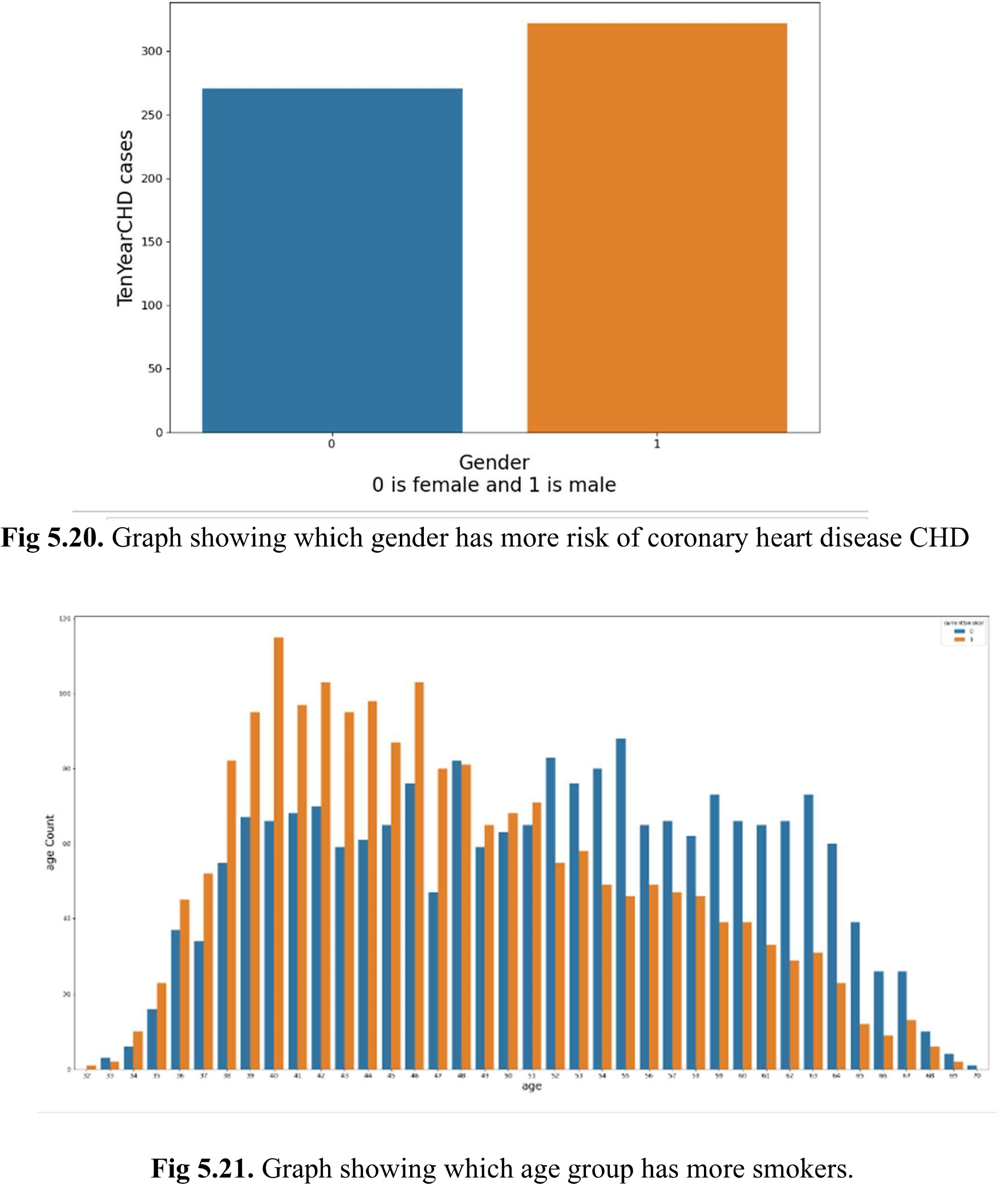
Fig 5.19. Violin plot diagram for glucose

For the list of figure 5.9 to 5.12 shows that For the same numerical characteristics:

Although the majority of the data is concentrated on 0, the distribution of cigsPerDay is unequal.

The majority of the following columns fall inside the range:

* totChol: 150-300
* sysBP: 100-150
* diaBP: 60-100
* BMI: 20 to 30
* heart rate: 50-100
* glucose :50-150m



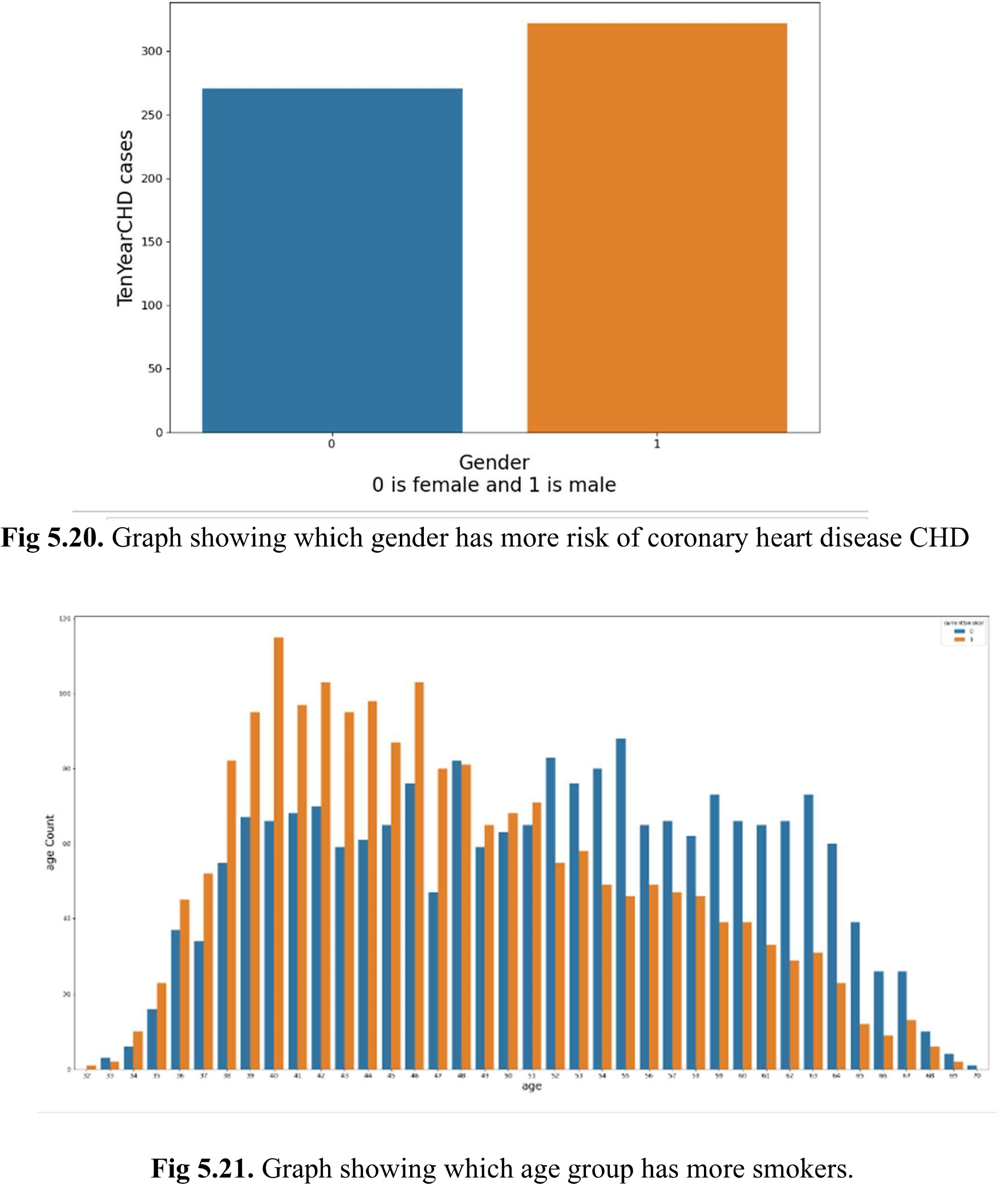
Fig 5.20. Graph showing which gender has more risk of coronary heart disease CHD

Fig 5.21. Graph showing which age group has more smokers.

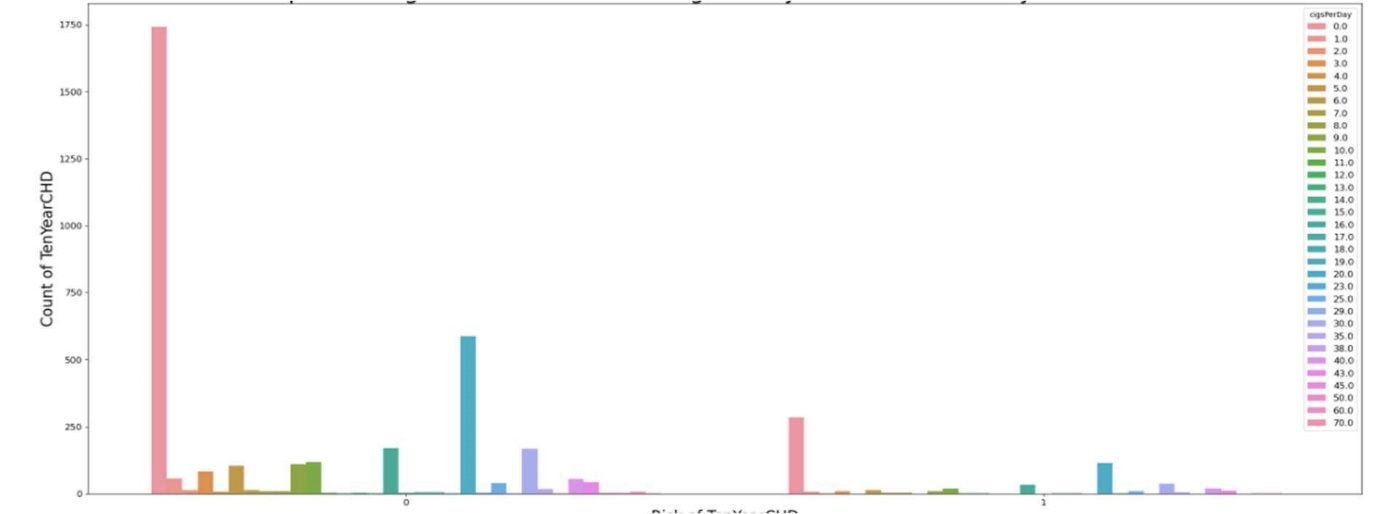


Fig 5.22. Graph showing the relation between cigsPerDay and risk of coronary heart disease.

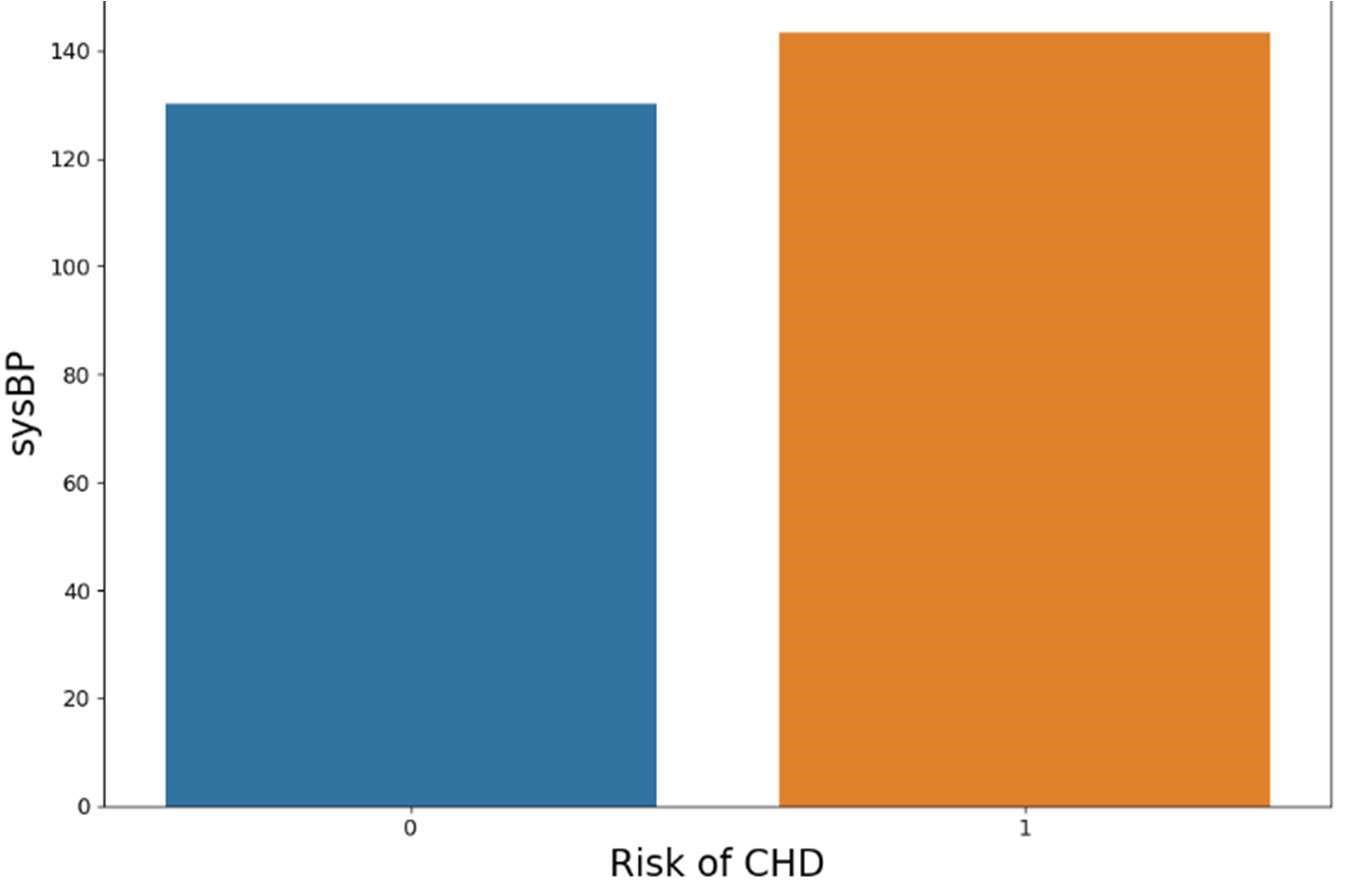


Fig 5.23.Graph showing the relation between sysBP and risk of CHD

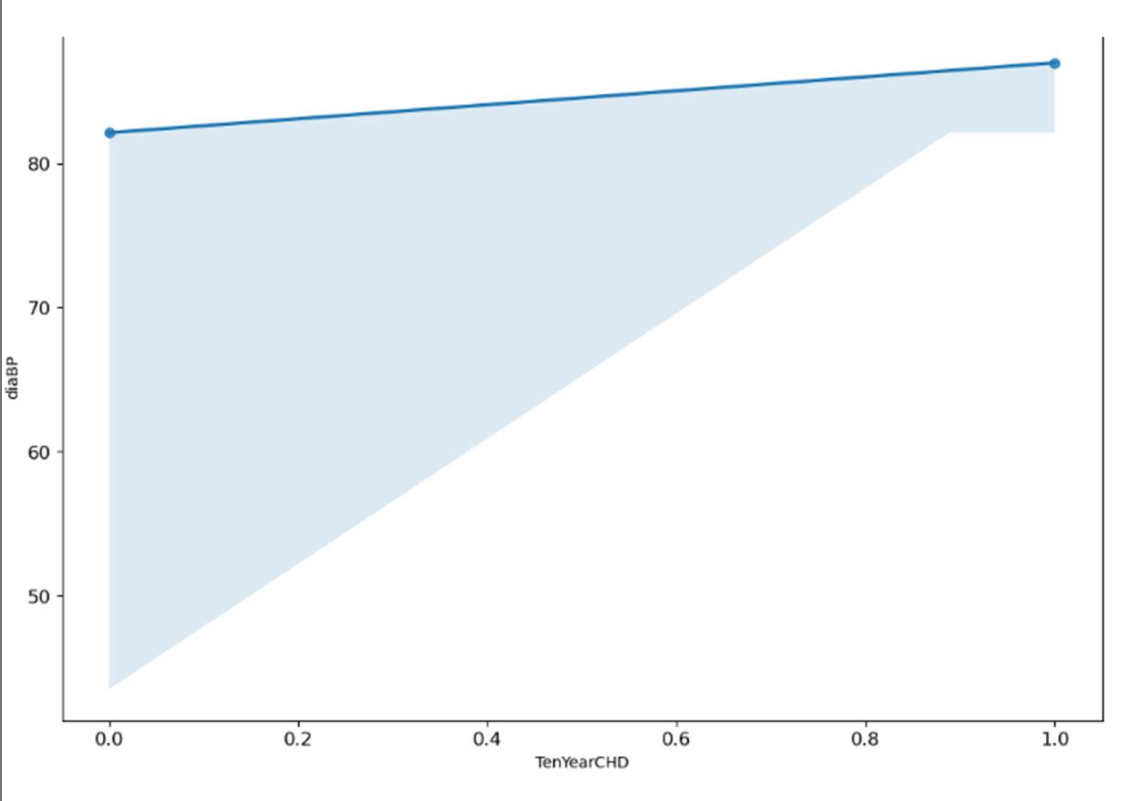
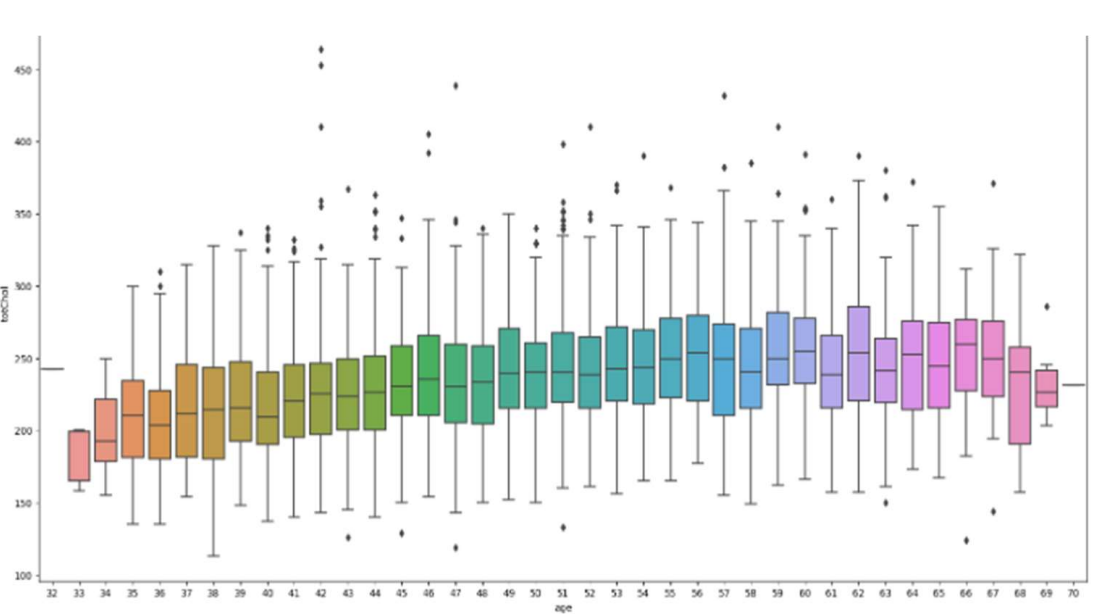
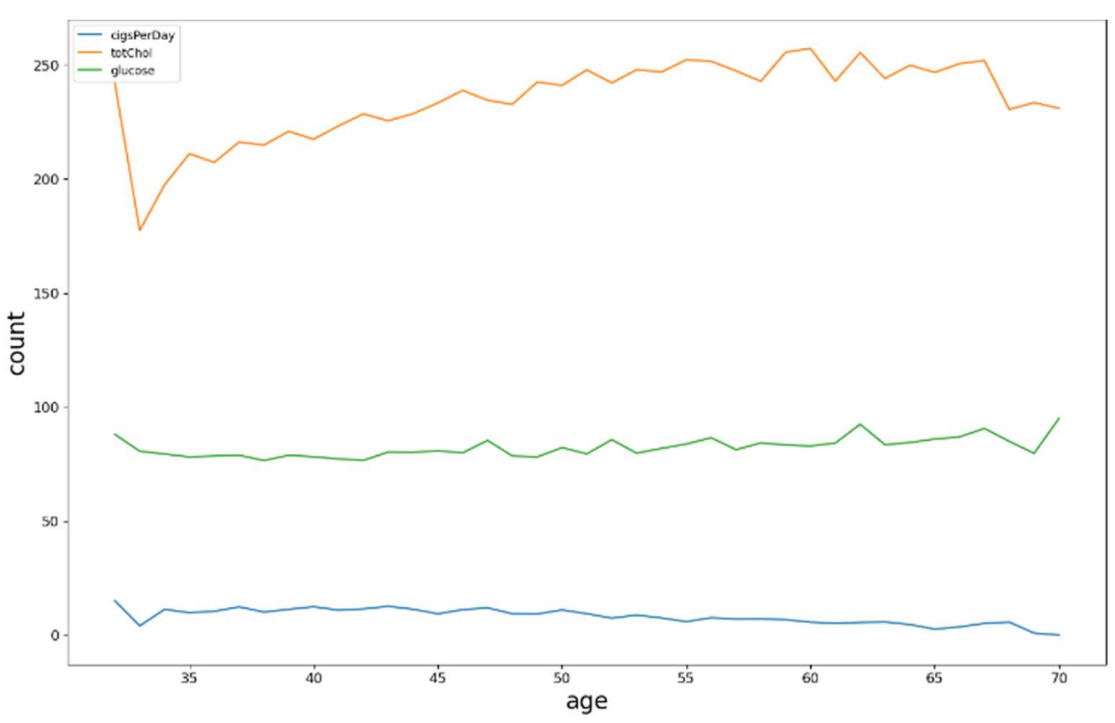


Fig 5.24. Distribution of diaBP in relation to the risk of CHD

Fig 5.25. Distribution of diaBP in relation to the risk of CHD 38

Fig 5.26. Graph showing totChoI and cigsPerDay in every age group.

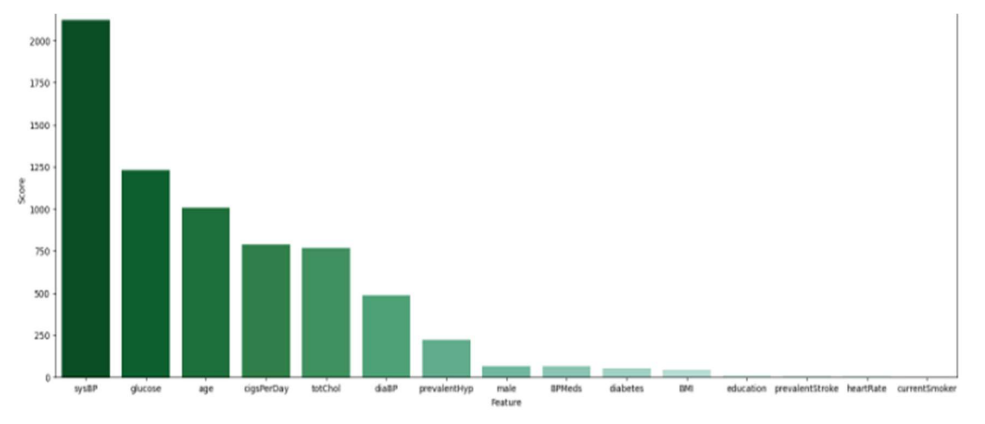


Fig 5.27. Plot showing the best features in descending order

**5.2 Libraries Used:**

In this project, standard libraries for database analysis and model creation are used. The following are the libraries used for the initial implementation:

**1. Numpy:** is a core library of scientific computing in python. It provides powerful tools to deal with various multidimensional arrays in python.

**2. Pandas:** is the most popular python library used for data analysis. Data in python can be analyzed with two ways: Series and Data-frames.

**3. sklearn:** is an open source python library which implements a huge range of machine learning, pre-processing, cross-validation and visualization algorithms.

**4. Matplotlib:** Matplotlib is a feature-rich Python visualization toolkit for producing static, animated, and interactive graphics. Matplotlib enables both difficult and easy tasks. Make plots fit for publishing.

**5. Seaborn:** This Python module allows you to create statistical visuals. It strongly interacts with pandas data structures and builds upon the matplotlib framework. Seaborn facilitates data exploration and comprehension.

**6. Mlxtend:** It is a Python library that provides a variety of tools for data analysis, machine learning, and visualization

**5.3 Output for various Algorithms:**

**Logistic Regression:**

confusion matrix

[[879 473]

[408 956]]

Accuracy of Logistic Regression: 67.56259204712813

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | precision | | recall | | F1-score | | support | |
| 0 | 0.68 | | 0.65 | | 0.67 | | 1352 | |
| 1 | 0.67 | | 0.70 | | 0.68 | | 1364 | |
| Accuracy | |  | | 0.90 | | 2716 | |  |
| Macro avg | | 0.91 | | 0.90 | | 0.90 | | 2716 |
| Weighted avg | | 0.91 | | 0.90 | | 0.90 | | 2716 |

**Random Forest Classfier:**

confussion matrix

[[1140 212]

[ 49 1315]]

Accuracy of Random Forest: 90.39027982326951

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | precision | | recall | | F1-score | | support | |
| 0 | 0.96 | | 0.84 | | 0.90 | | 1352 | |
| 1 | 0.86 | | 0.90 | | 0.91 | | 1364 | |
| Accuracy | |  | | 0.90 | | 2716 | |  |
| Macro avg | | 0.91 | | 0.90 | | 0.90 | | 2716 |
| Weighted avg | | 0.91 | | 0.90 | | 0.90 | | 2716 |

**DecisionTree Classifier:**

confussion matrix

[[1075 277]

[ 31 1333]]

Accuracy of Decision Tree Classifier: 88.65979381443299

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | precision | | recall | | F1-score | | support | |
| 0 | 0.97 | | 0.80 | | 0.87 | | 1352 | |
| 1 | 0.863 | | 0.98 | | 0.90 | | 1364 | |
| Accuracy | |  | | 0.89 | | 2716 | |  |
| Macro avg | | 0.90 | | 0.89 | | 0.89 | | 2716 |
| Weighted avg | | 0.90 | | 0.89 | | 0.89 | | 2716 |
|  | |  | |  | |  | |  |

**DecisionTree Classifier:**

confussion matrix

[[ 914 438]

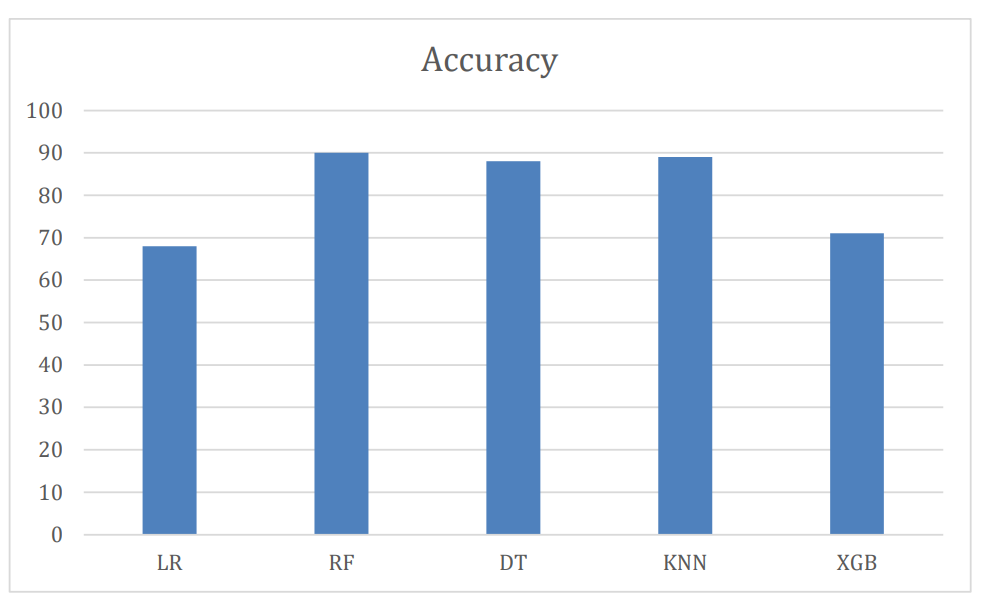
[ 337 1027]]

Accuracy of Decision Tree Classifier: 71.46539027982327

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | precision | | recall | | F1-score | | support | |
| 0 | 0.97 | | 0.80 | | 0.87 | | 1352 | |
| 1 | 0.863 | | 0.98 | | 0.90 | | 1364 | |
| accuracy | |  | | 0.70 | | 2716 | |  |
| Macro avg | | 0.90 | | 0.89 | | 0.89 | | 2716 |
| Weighted avg | | 0.90 | | 0.89 | | 0.89 | | 2716 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithms | Gender | Precision | recall | F1-score | Accuracy |
| Logistic Regression | Male(1) | 0.67 | 0.70 | 0.68 | 0.68 |
| Female(0) | 0.68 | 0.65 | 0.67 |
| K - Nearest Neighbour | Male(1) | 0.84 | 0.98 | 0.91 | 0.89 |
| Female(0) | 0.98 | 0.82 | 0.89 |
| Random Forest | Male(1) | 0.86 | 0.96 | 0.91 | 0.90 |
| Female(0) | 0.96 | 0.84 | 0.90 |
| Decision Tree | Male(1) | 0.83 | 0.90 | 0.87 | 0.88 |
| Female(0) | 0.97 | 0.80 | 0.87 |
| Extreme Gradient Boost | Male(1) | 0.70 | 0.75 | 0.73 | 0.71 |
| Female(0) | 0.73 | 0.68 | 0.70 |

**Table 5.1** Comparisons of Algorithms with their accuracy.



**Fig 5.28** Bar Graph with Comparisons of Algorithms and their accuracy

In conclusion, it is evident from the bar graph the various machine learning algorithms perform on the provided dataset in terms of accuracy. At 90% accuracy, Random Forest (RF) is the best algorithm, closely followed by K-Nearest Neighbors (KNN) at 89%. Moreover, Decision Trees (DT) show excellent accuracy at 88%, while XGBoost (XGB) and Logistic Regression (LR) show accuracies of 68% and 71%, respectively. These findings emphasize how crucial it is to choose the best algorithm for a given job because the decision can have a big impact on a machine learning model's efficacy and accuracy.

**CHAPTER 6**

**CONCLUSION AND FUTURE ENHANCEMENT**

**6.1 Conclusion**

In conclusion, the development and use of machine learning algorithms for the prediction of cardiovascular illnesses represents a significant achievement in the field of healthcare and disease prevention. This experiment has shown that these algorithms have the ability to deliver accurate and timely forecasts, which can help healthcare professionals identify patients at risk and conduct preventative treatments. We investigated numerous machine learning approaches, such as decision trees, Logistic regression, Random Forest K-nearest and XGBoost, and assessed their performance on a wide collection of data throughout this study. We demonstrated that these algorithms can produce valid predictions by using a wide variety of patient information, such as medical history, vital signs, and lifestyle variables.

This study's findings emphasize the significance of data quality, feature selection, and model hyperparameter tweaking in reaching optimal predictive performance. Furthermore, we have emphasized the importance of interpretability and openness in the context of healthcare, since these factors are crucial for winning the trust of medical professionals and patients.

Predictive models for cardiovascular illnesses have various potential benefits. Early identification and risk assessment can lead to more tailored and effective healthcare interventions, reducing the burden of chronic diseases on people and healthcare systems in the long run. Furthermore, this technology can assist distribute resources more efficiently, ensuring that the most vulnerable patients receive the care they require.

However, it is critical to recognize that this subject is continually growing, and that further study is necessary to increase the accuracy and generalizability of these models. Collaboration among data scientists, medical practitioners, and policymakers is critical to ensuring the responsible and ethical use of machine learning algorithms in healthcare settings. To summarize, while the prediction of cardiovascular disorders using machine learning algorithms has showed enormous potential, it is only the start of a transformational journey in the healthcare business. We are on the way to improving the well-being of individuals and communities by more effectively preventing and controlling cardiovascular illnesses as we enhance and expand our models.

**6.2 Future Enhancements:**

While the project has made tremendous progress in applying machine learning algorithms to forecast cardiovascular disorders, there are various possibilities for further refinement and study that might further increase the accuracy, dependability, and usefulness of the models. These improvements include:

**1. Incorporating genetic Data:** Incorporating genetic data into prediction models might provide a more thorough knowledge of a person's risk for cardiovascular disease. genomic variables play an important role in susceptibility, and integrating genomic and clinical data might lead to more precise risk assessments.

**2. Real-time Monitoring and Feedback:** Creating a system that allows for continuous, real-time monitoring of patients' health data might offer patients and healthcare practitioners with quick feedback.

**3. Multimodal Data Fusion:** Adding wearable devices, social determinants of health, and environmental elements to the list of data sources beyond electronic health records (EHRs). These new data sources can give a more comprehensive picture of a patient's health and lifestyle, resulting in more precise forecasts.

**4. Interpretable Models:** Creating more interpretable and transparent machine learning models. Explainable AI approaches, such as LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations), can assist healthcare workers in understanding the reasoning behind model predictions, which is critical for establishing confidence and acceptance in clinical practice.

**5. Personalized Treatment Recommendations**: Providing personalized treatment recommendations beyond risk prediction. Based on an individual's risk profile and medical history, machine learning can be used to recommend tailored interventions, medications, and lifestyle changes.

**6. Validation and Generalization:** Extensive external validation of models on diverse and larger patient populations to ensure generalizability across demographics, geographical locations, and healthcare settings.

**7. Ethical Considerations and Privacy:** Addressing ethical and privacy concerns about the use of patient data in machine learning models on an ongoing basis. Future improvements should concentrate on ensuring that data is used in a secure and compliant manner, with a focus on patient privacy.

**8. Integration into Clinical Workflow:** Creating seamless integration with electronic health record systems and clinical workflow to ensure that predictions are easily accessible to healthcare providers during patient visits.

**9. Long-term Outcome Predictions:** Extending prediction scope to include longterm outcomes such as the risk of heart failure, stroke, or overall cardiovascular mortality. This would allow for more proactive and thorough healthcare planning.

**10. Collaborative Research:** Fostering collaboration among data scientists, medical professionals, and healthcare policymakers in order to close the gap between cutting-edge research and clinical practice. This ensures that the most recent advances in machine learning are effectively translated into better patient care.

Incorporating these future enhancements will not only improve the accuracy and efficacy of cardiovascular disease predictive models, but will also ensure that these models become invaluable tools for healthcare providers in the ongoing battle against one of the world's leading causes of mortality and morbidity.

**REFERNCES**

[1] Azam Mehmood Qadri, Ali Raza, Kashif Munir and Mubarak S. Almutairi “Effective Feature Engineering Technique for Heart disease Prediction with Machine Learning”, IEE Access, vol.11, no. , pp.56214-56224, 2023.

[2] Abdulwahab Ali Almazroi, Eman A. Aldhahri, Saba Bashir and Sufyan Ashfq, “A Clinical Decision Support for Heart Disease Prediction Using Deep Learning”, IEE Acess, vol.11, pp.61646-61659, 2023.

[3] A. Lakshmanarao, A. Srisaila and T.Srinivasa Ravi Kiran “Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques”, IEEE Xplore, Doi 10.1109, pp 994-998, 2021.

[4] Shuge Ouyang “Research of Heart Disease Prediction Based on Machine Learning” IEEE Xplore, Doi 10.1109, 2022.

[5] Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE Access, vol.7,pp.8154281554, 2019.

[6] M. A. Jabbar and Shirina Samereen “Heart disease prediction system based on hidden naïve bayes classifier” IEEE Xplore,Doi 10.1109, 2016.

[7] G. Kumar Sahoo, K. Kanike, S. K. Das and P. Singh, "Machine Learning-Based Heart Disease Prediction: A Study for Home Personalized Care," 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP), Xi'an, China, 2022, pp. 01-06, doi: 10.1109/MLSP55214.2022.9943373. [8] R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 302-305, doi: 10.1109/ICESC48915.2020.9155586.

[9] L. KN, N. R, N. K, R. Kumari, S. N and V. K, "Heart Disease Detection using Machine Learning Technique," 2021 Second International Conference on Electronics and 49 Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 1738-1743, doi: 10.1109/ICESC51422.2021.9532705.

[10] K. Karthik, A. L. Reddy, R. Kulkarni and M. J. Mehdi, "Algorithm Accuracy Verification in Heart Disease Analysis using Machine Learning," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 345-349, doi: 10.1109/ICAAIC56838.2023.10140446.

[11] I. K. Pious, K. Antony Kumar, Y. C. Soulwin and E. N. Reddy, "Heart Disease Prediction Using Machine Learning Algorithms," 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2022, pp. 1-6, doi: 10.1109/ICSES55317.2022.9914207.

[12] S. Patidar, A. Jain and A. Gupta, "Comparative Analysis of Machine Learning Algorithms for Heart Disease Predictions," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, pp. 1340- 1344, doi: 10.1109/ICICCS53718.2022.9788408.

**APPENDIX 1**

**CODE**

*# Data Loading and Numerical Operations*

import pandas as pd

import numpy as np

*# Data Visualizations*

import matplotlib.pyplot as plt

%matplotlib inline

import seaborn as sns

*# Data Resampling*

from sklearn.utils import resample

*# Data Feature Selection*

from sklearn.feature\_selection import SelectKBest

from sklearn.feature\_selection import chi2

*# Data Splitting*

from sklearn.model\_selection import train\_test\_split

*# Data Scaling*

from sklearn.preprocessing import MinMaxScaler

*# Data Modeling*

from sklearn.linear\_model import LogisticRegression

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import GradientBoostingClassifier

from sklearn.neural\_network import MLPClassifier

from sklearn.metrics import confusion\_matrix, accuracy\_score, roc\_curve, classification\_report

*# Hyperparameter Tuning*

from sklearn.model\_selection import RandomizedSearchCV

from sklearn.model\_selection import GridSearchCV

*# Ensembling*

from mlxtend.classifier import StackingCVClassifier

data = pd.read\_csv("framingham.csv") *# Reading and converting the data into a pandas dataframe*

data.shape *# Calculating the dimensions of the dataset*

data.head(10)

data.info()

data.isnull().sum()

data.duplicated().sum()

print((data["glucose"].mode())[0]) data["glucose"].fillna((data["glucose"].mode())[0], inplace=True) data.dropna(inplace=True)

data.isnull().sum()

plt.figure(figsize=(20,10), facecolor='w')

sns.boxplot(data=data)

plt.show()

data['totChol'].max()

data['sysBP'].max()

data = data[data['totChol']<600.0]

data = data[data['sysBP']<295.0]

data.shape

data.describe()

*#Checking relationship between variables*

cor=data.corr()

plt.figure(figsize=(20,10), facecolor='w')

sns.heatmap(cor,xticklabels=cor.columns,yticklabels=cor.columns,annot=True)

plt.title("Correlation among all the Variables of the Dataset", size=20)

cor

categorical\_features = ['male', 'education', 'currentSmoker', 'BPMeds', 'prevalentStroke', 'prevalentHyp', 'diabetes']

for feature in categorical\_features:

print(feature,':')

print(data[feature].value\_counts())

print("-----------------")

num\_plots = len(categorical\_features)

total\_cols = 2

total\_rows = num\_plots//total\_cols + 1

fig, axs = plt.subplots(nrows=total\_rows, ncols=total\_cols,

figsize=(7\*total\_cols, 7\*total\_rows), facecolor='w', constrained\_layout=True)

for i, var in enumerate(categorical\_features):

row = i//total\_cols

pos = i % total\_cols

plot = sns.countplot(x=var, data=data, ax=axs[row][pos])

numeric\_features = ['cigsPerDay', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose']

for feature in numeric\_features:

plt.figure(figsize=(18, 10), facecolor='w')

sns.distplot(data[feature])

plt.title('{} Distribution'.format(feature), fontsize=20)

plt.show()

num\_plots = len(numeric\_features)

total\_cols = 2

total\_rows = num\_plots//total\_cols + 1

color = ['m', 'g', 'b', 'r', 'y', 'v', 'o']

fig, axs = plt.subplots(nrows=total\_rows, ncols=total\_cols,

figsize=(7\*total\_cols, 7\*total\_rows), facecolor='w', constrained\_layout=True)

for i, var in enumerate(numeric\_features):

row = i//total\_cols

pos = i % total\_cols

plot = sns.violinplot(y=var, data=data, ax=axs[row][pos], linewidth=2

*#Distribution of outcome variable, Heart Disease*

plt.figure(figsize=(12, 10), facecolor='w')

plt.subplots\_adjust(right=1.5)

plt.subplot(121)

sns.countplot(x="TenYearCHD", data=data)

plt.title("Count distribution of TenYearCHD", size=20)

plt.subplot(122)

labels=[0,1]

plt.pie(data["TenYearCHD"].value\_counts(),autopct="%1.1f%%",labels=labels,colors=["lime","red"])

plt.show()

*#Grouping education and cigsPerDay*

graph\_1 = data.groupby("education", as\_index=False).cigsPerDay.mean()

plt.figure(figsize=(12,8), facecolor='w')

sns.regplot(x=graph\_1["education"], y=graph\_1["cigsPerDay"])

plt.title("Graph showing cigsPerDay in every level of education.", size=20)

plt.xlabel("education", size=20)

plt.ylabel("cigsPerDay", size=20)

plt.xticks(size=12)

plt.yticks(size=12)

*#checking for which gender has more risk of coronary heart disease CHD*

graph\_2 = data.groupby("male", as\_index=False).TenYearCHD.sum()

*#Ploting the above values*

plt.figure(figsize=(12,8), facecolor='w')

sns.barplot(x=graph\_2["male"], y=graph\_2["TenYearCHD"])

plt.title("Graph showing which gender has more risk of coronary heart disease CHD", size=20)

plt.xlabel("Gender\n0 is female and 1 is male",size=20)

plt.ylabel("TenYearCHD cases", size=20)

plt.xticks(size=12)

plt.yticks(size=12)

*#Distribution of current smokers with respect to age*

plt.figure(figsize=(30,15), facecolor='w')

sns.countplot(x="age",data=data,hue="currentSmoker")

plt.title("Graph showing which age group has more smokers.", size=30)

plt.xlabel("age", size=20)

plt.ylabel("age Count", size=20)

plt.xticks(size=12)

plt.yticks(size=12)

plt.figure(figsize=(30,12), facecolor='w')

sns.countplot(x="TenYearCHD",data=data,hue="cigsPerDay")

plt.legend(title='cigsPerDay', fontsize='large')

plt.title("Graph showing the relation between cigsPerDay and risk of coronary heart disease.", size=30)

plt.xlabel("Risk of TenYearCHD", size=20)

plt.ylabel("Count of TenYearCHD", size=20)

plt.xticks(size=12)

plt.yticks(size=12)

plt.show()

*# Grouping up the data and ploting it*

graph\_3 = data.groupby("TenYearCHD", as\_index=False).sysBP.mean()

plt.figure(figsize=(12,8), facecolor='w')

sns.barplot(x=graph\_3["TenYearCHD"], y=graph\_3["sysBP"])

plt.title("Graph showing the relation between sysBP and risk of CHD", size=20)

plt.xlabel("Risk of CHD", size=20)

plt.ylabel("sysBP", size=20)

plt.xticks(size=12)

plt.yticks(size=12)

plt.figure(figsize=(12,8), facecolor='w')

sns.regplot(x=graph\_3["TenYearCHD"], y=graph\_3["sysBP"])

plt.title("Distribution of sysBP in relation to the risk of CHD", size=20)

plt.xticks(size=12)

plt.yticks(size=12)

*# Grouping up the data and ploting it*

graph\_4 = data.groupby("TenYearCHD", as\_index=False).diaBP.mean()

plt.figure(figsize=(12,8), facecolor='w')

sns.barplot(x=graph\_4["TenYearCHD"], y=graph\_4["diaBP"])

plt.title("Graph showing the relation between diaBP and risk of CHD", size=20)

plt.xlabel("Risk of CHD", size=20)

plt.ylabel("diaBP", size=20)

plt.xticks(size=12)

plt.yticks(size=12)

plt.figure(figsize=(12,8), facecolor='w')

sns.regplot(x=graph\_4["TenYearCHD"], y=graph\_4["diaBP"])

plt.title("Distribution of diaBP in relation to the risk of CHD", size=20)

plt.xticks(size=12)

plt.yticks(size=12)

plt.figure(figsize=(20,10), facecolor='w')

sns.boxplot(x="age",y="totChol",data=data)

plt.title("Distribution of age with respect to totChol", size=20)

plt.show()

*#Plotting a linegraph to check the relationship between age and cigsPerDay, totChol, glucose.*

graph\_5 = data.groupby("age").cigsPerDay.mean()

graph\_6 = data.groupby("age").totChol.mean()

graph\_7 = data.groupby("age").glucose.mean()

plt.figure(figsize=(16,10), facecolor='w')

sns.lineplot(data=graph\_5, label="cigsPerDay")

sns.lineplot(data=graph\_6, label="totChol")

sns.lineplot(data=graph\_7, label="glucose")

plt.title("Graph showing totChol and cigsPerDay in every age group.", size=20)

plt.xlabel("age", size=20)

plt.ylabel("count", size=20)

plt.xticks(size=12)

plt.yticks(size=12)

*#sysBP vs diaBP with respect to currentSmoker and male attributes*

*#plt.figure(figsize=(18, 9), facecolor='w')*

sns.lmplot('sysBP', 'diaBP',

data=data,

hue="TenYearCHD",

col="male",row="currentSmoker")

plt.show()

target1=data[data['TenYearCHD']==1]

target0=data[data['TenYearCHD']==0]

target1=resample(target1,replace=True,n\_samples=len(target0),random\_state=40)

target=pd.concat([target0,target1])

target['TenYearCHD'].value\_counts()

data=target

np.shape(data)

*#Distribution of heart disease cases in the balanced dataset, the outcome variable*

plt.figure(figsize=(12, 10), facecolor='w')

plt.subplots\_adjust(right=1.5)

plt.subplot(121)

sns.countplot(x="TenYearCHD", data=data)

plt.title("Count of TenYearCHD column", size=20)

plt.subplot(122)

labels=[0,1]

plt.pie(data["TenYearCHD"].value\_counts(),autopct="%1.1f%%",labels=labels,colors=["red","lime"])

plt.show()

*#To idenfify the features that have larger contribution towards the outcome variable, TenYearCHD*

X=data.iloc[:,0:15]

y=data.iloc[:,-1]

print("X - ", X.shape, "\ny - ", y.shape)

*#Apply SelectKBest and extract top 10 features*

best=SelectKBest(score\_func=chi2, k=10)

fit=best.fit(X,y)

data\_scores=pd.DataFrame(fit.scores\_)

data\_columns=pd.DataFrame(X.columns)

*#Join the two dataframes*

scores=pd.concat([data\_columns,data\_scores],axis=1)

scores.columns=['Feature','Score']

print(scores.nlargest(11,'Score'))

*#To visualize feature selection*

scores=scores.sort\_values(by="Score", ascending=False)

plt.figure(figsize=(20,7), facecolor='w')

sns.barplot(x='Feature',y='Score',data=scores,palette='BuGn\_r')

plt.title("Plot showing the best features in descending order", size=20)

plt.show()

*#Select 10 features*

features=scores["Feature"].tolist()[:10]

features

data=data[['sysBP','glucose','age','cigsPerDay','totChol','diaBP','prevalentHyp','male','BPMeds','diabetes','TenYearCHD']]

data.head()

y = data['TenYearCHD']

X = data.drop(['TenYearCHD'], axis=1)

train\_x, test\_x, train\_y, test\_y = train\_test\_split(X, y, test\_size=0.4,

scaler = MinMaxScaler()

train\_x = scaler.fit\_transform(train\_x)

test\_x = scaler.transform(test\_x)

m1 = 'LogisticRegression'

lr = LogisticRegression(random\_state=1, max\_iter=1000)

model = lr.fit(train\_x, train\_y)

lr\_predict = lr.predict(test\_x)

lr\_conf\_matrix = confusion\_matrix(test\_y, lr\_predict)

lr\_acc\_score = accuracy\_score(test\_y, lr\_predict)

print("confussion matrix")

print(lr\_conf\_matrix)

print("\n")

print("Accuracy of Logistic Regression:",lr\_acc\_score\*100,'\n')

print(classification\_report(test\_y,lr\_predict))

m3 = 'Random Forest Classfier'

rf = RandomForestClassifier(n\_estimators=200, random\_state=0,max\_depth=12)

rf.fit(train\_x,train\_y)

rf\_predicted = rf.predict(test\_x)

rf\_conf\_matrix = confusion\_matrix(test\_y, rf\_predicted)

rf\_acc\_score = accuracy\_score(test\_y, rf\_predicted)

print("confussion matrix")

print(rf\_conf\_matrix)

print("\n")

print("Accuracy of Random Forest:",rf\_acc\_score\*100,'\n')

print(classification\_report(test\_y,rf\_predicted))

m4 = 'DecisionTreeClassifier'

dt = DecisionTreeClassifier(criterion = 'entropy',random\_state=0,max\_depth = 30)

dt.fit(train\_x,train\_y)

dt\_predicted = dt.predict(test\_x)

dt\_conf\_matrix = confusion\_matrix(test\_y, dt\_predicted)

dt\_acc\_score = accuracy\_score(test\_y, dt\_predicted)

print("confussion matrix")

print(dt\_conf\_matrix)

print("\n")

print("Accuracy of DecisionTreeClassifier:",dt\_acc\_score\*100,'\n')

print(classification\_report(test\_y,dt\_predicted))

m5 = 'Gradient Boosting Classifier'

gvc = GradientBoostingClassifier()

gvc.fit(train\_x,train\_y)

gvc\_predicted = gvc.predict(test\_x)

gvc\_conf\_matrix = confusion\_matrix(test\_y, gvc\_predicted)

gvc\_acc\_score = accuracy\_score(test\_y, gvc\_predicted)

print("confussion matrix")

print(gvc\_conf\_matrix)

print("\n")

print("Accuracy of Gradient Boosting Classifier:",gvc\_acc\_score\*100,'\n')

print(classification\_report(test\_y,gvc\_predicted))

*# Number of trees in random forest*

n\_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]

*# Number of features to consider at every split*

max\_features = ['auto', 'sqrt']

*# Maximum number of levels in tree*

max\_depth = [int(x) for x in np.linspace(10, 110, num = 11)]

max\_depth.append(None)

*# Minimum number of samples required to split a node*

min\_samples\_split = [2, 5, 10]

*# Minimum number of samples required at each leaf node*

min\_samples\_leaf = [1, 2, 4]

*# Method of selecting samples for training each tree*

bootstrap = [True, False]

*# Create the random grid*

random\_grid = {'n\_estimators': n\_estimators,

'max\_features': max\_features,

'max\_depth': max\_depth,

'min\_samples\_split': min\_samples\_split,

'min\_samples\_leaf': min\_samples\_leaf,

'bootstrap': bootstrap}

print(random\_grid)

*# Use the random grid to search for best hyperparameters*

*# First create the base model to tune*

rf = RandomForestClassifier()

*# Random search of parameters, using 3 fold cross validation,*

*# search across 100 different combinations, and use all available cores*

rf\_random = RandomizedSearchCV(estimator = rf,

param\_distributions = random\_grid,

n\_iter = 100,

cv = 3,

verbose=2,

random\_state=7,

n\_jobs = -1)

*# Fit the random search model*

rf\_random.fit(train\_x,train\_y)

rf\_hyper = rf\_random.best\_estimator\_

rf\_hyper.fit(train\_x,train\_y)

print("Accuracy on training set is : {}".format(rf\_hyper.score(train\_x,train\_y)))

print("Accuracy on validation set is : {}".format(rf\_hyper.score(test\_x, test\_y)))

rf\_predicted = rf\_hyper.predict(test\_x)

rf\_acc\_score = accuracy\_score(test\_y, rf\_predicted)

print("Accuracy of Hyper-tuned Random Forest Classifier:",rf\_acc\_score\*100,'\n')

print(classification\_report(test\_y, rf\_predicted))

*#Number of trees*

n\_estimators = [int(i) for i in np.linspace(start=100,stop=1000,num=10)]

*#Number of features to consider at every split*

max\_features = ['auto','sqrt']

*#Maximum number of levels in tree*

max\_depth = [int(i) for i in np.linspace(10, 100, num=10)]

max\_depth.append(None)

*#Minimum number of samples required to split a node*

min\_samples\_split=[2,5,10]

*#Minimum number of samples required at each leaf node*

min\_samples\_leaf = [1,2,4]

*#Create the random grid*

random\_grid = {'n\_estimators': n\_estimators,

'max\_features': max\_features,

'max\_depth': max\_depth,

'min\_samples\_split': min\_samples\_split,

'min\_samples\_leaf': min\_samples\_leaf}

gb=GradientBoostingClassifier(random\_state=0)

*#Random search of parameters, using 3 fold cross validation,*

*#search across 100 different combinations*

gb\_random = RandomizedSearchCV(estimator=gb, param\_distributions=random\_grid,

n\_iter=100, scoring='f1',

cv=3, verbose=2, random\_state=0, n\_jobs=-1,

return\_train\_score=True)

*# Fit the random search model*

gb\_random.fit(train\_x,train\_y)

gb\_hyper = gb\_random.best\_estimator\_

gb\_hyper.fit(train\_x,train\_y)

print("Accuracy on training set is : {}".format(gb\_hyper.score(train\_x,train\_y)))

print("Accuracy on validation set is : {}".format(gb\_hyper.score(test\_x, test\_y)))

gbc\_predicted = gb\_hyper.predict(test\_x)

gbc\_acc\_score = accuracy\_score(test\_y, gbc\_predicted)

print("Accuracy of Hyper-tuned Gradient Boosting Classifier:",gbc\_acc\_score\*100,'\n')

print(classification\_report(test\_y, gbc\_predicted))

lr\_false\_positive\_rate,lr\_true\_positive\_rate,lr\_threshold = roc\_curve(test\_y,lr\_predict)

knn\_false\_positive\_rate,knn\_true\_positive\_rate,knn\_threshold = roc\_curve(test\_y,knn\_predict)

rf\_false\_positive\_rate,rf\_true\_positive\_rate,rf\_threshold = roc\_curve(test\_y,rf\_predicted)

dt\_false\_positive\_rate,dt\_true\_positive\_rate,dt\_threshold = roc\_curve(test\_y,dt\_predicted)

gbc\_false\_positive\_rate,gbc\_true\_positive\_rate,gbc\_threshold = roc\_curve(test\_y,gbc\_predicted)

sns.set\_style('whitegrid')

plt.figure(figsize=(15,8), facecolor='w')

plt.title('Reciever Operating Characterstic Curve')

plt.plot(lr\_false\_positive\_rate,lr\_true\_positive\_rate,label='Logistic Regression')

plt.plot(knn\_false\_positive\_rate,knn\_true\_positive\_rate,label='K-Nearest Neighbor')

plt.plot(rf\_false\_positive\_rate,rf\_true\_positive\_rate,label='Random Forest')

plt.plot(dt\_false\_positive\_rate,dt\_true\_positive\_rate,label='Desion Tree')

plt.plot(gbc\_false\_positive\_rate,gbc\_true\_positive\_rate,label='Gradient Boosting Classifier')

plt.plot([0,1],ls='--')

plt.plot([0,0],[1,0],c='.5')

plt.plot([1,1],c='.5')

plt.ylabel('True positive rate')

plt.xlabel('False positive rate')

plt.legend()

plt.show()

model\_ev = pd.DataFrame({'Model': ['Logistic Regression','K-Nearest Neighbour','Random Forest',

'Decision Tree','Gradient Boosting'], 'Accuracy': [lr\_acc\_score\*100, knn\_acc\_score\*100,

rf\_acc\_score\*100, dt\_acc\_score\*100,gbc\_acc\_score\*100]})

model\_ev

colors = ['red','green','blue','gold','silver']

plt.figure(figsize=(15,8), facecolor='w')

plt.title("Barplot Representing Accuracy of different models")

plt.ylabel("Accuracy %")

plt.xlabel("Models")

plt.bar(model\_ev['Model'],model\_ev['Accuracy'],color = colors)

plt.show()

scv=StackingCVClassifier(classifiers=[rf\_hyper, gbc\_hyper, knn], meta\_classifier= rf)

train\_x, test\_x, train\_y, test\_y = train\_test\_split(X, y, test\_size=0.4, random\_state=1)

scv.fit(train\_x.values,train\_y.values)

scv\_predicted = scv.predict(test\_x)

scv\_conf\_matrix = confusion\_matrix(test\_y, scv\_predicted)

scv\_acc\_score = accuracy\_score(test\_y, scv\_predicted)

print("confussion matrix")

print(scv\_conf\_matrix)

print("\n")

print("Accuracy of StackingCVClassifier:",scv\_acc\_score\*100,'\n')

print(classification\_report(test\_y,scv\_predicted))

**APPENDIX II**

**Implementation Screenshot**

