A project report on

AI BASED AUTOMATED SYSTEM FOR CORONARY ARTERY DISEASE PREDICTION

Submitted in partial fulfillment for the award of the degree of

M.Tech Computer Science and Engineering [Integrated]

by

HARSHITA SOROUT (19MIC0047)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

April, 2024

AI BASED AUTOMATED SYSTEM FOR CORONARY ARTERY DISEASE PREDICTION

Submitted in partial fulfillment for the award of the degree of

M.Tech Computer Science and Engineering [Integrated]

by

HARSHITA SOROUT (19MIC0047)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

April, 2024

DECLARATION

I here by declare that the thesis entitled “AI BASED AUTOMATED SYSTEM FOR CORONARY ARTERY DISEASE PREDICTION” submitted by me, for the award of the degree of M.Tech Computer Science and Engineering [Integrated] is a record of bonafide work carried out by me under the supervision of Prof. GAYATHRI S.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore

Date: **Signature of the Candidate**

CERTIFICATE

This is to certify that the thesis entitled "AI BASED AUTOMATED SYSTEM FOR CORONARY ARTERY DISEASE PREDICTION" submitted by HARSHITA SOROUT (19MIC0047), School of Computer Science and Engineering, Vellore Institute of Technology, Vellore for the award of the degree M.Tech Computer Science and Engineering [Integrated] is a record of bonafide work carried out by him/her under my supervision.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The Project report fulfils the requirements and regulations of VELLORE INSTITUTE OF TECHNOLOGY, VELLORE and in my opinion meets the necessary standards for submission.

Signature of the Guide **Signature of the HoD**

Internal Examiner **External Examiner**

ABSTRACT

Coronary artery disease (CAD) is a major global cause of death that is frequently made worse by a delay in diagnosis brought on by a lack of knowledge about its symptoms, particularly in rural regions. In order to solve this problem, this project will create an AI-based automated system for CAD prediction that is especially designed for rural communities with limited access to medical care and knowledge of the illness. The system provides a personalized risk assessment based on the user's health data. By applying machine learning methods including logistic regression, decision trees, support vector machines (SVM), and k-nearest neighbors (KNN), the system predicts CAD risk levels, which range from no risk to severe risk, by extracting relevant information from the available data. This project aims to enable people living in rural areas to take charge of their cardiovascular health and seek early medical assistance when necessary by developing an easy-to-use program application. In the end, this program aims to lessen the impact of CAD by assisting underprivileged areas with early detection and intervention.

ACKNOWLEDGEMENT

The project "AI BASED AUTOMATED SYSTEM FOR CORONARY ARTERY DISEASE PREDICTION" was made possible because of inestimable inputs from everyone involved, directly or indirectly. First, I would like to thank and express my sincere gratitude to my guide, Prof. GAYATHRI S, who was highly instrumental in providing an innovative base with constructive inputs for the completion of the project

I would like to express my gratitude to DR.G.VISWANATHAN, Chancellor VELLORE INSTITUTE OF TECHNOLOGY, VELLORE, MR. SANKAR VISWANATHAN, DR. SEKAR VISWANATHAN, DR.G V SELVAM, Vice – Presidents VELLORE INSTITUTE OF TECHNOLOGY, VELLORE, Dr. V. S. Kanchana Bhaaskaran, Vice – Chancellor, Dr. Partha Sharathi Mallick, Pro-Vice Chancellor and Dean of SCOPE, Dr. Ramesh Babu K, for all the support provided at the school for successful completion of the project.

I would also like to acknowledge the role of HOD of the Dept. of Computational Intelligence Dr. Swathi J N, who was instrumental in keeping me, updated with all necessary formalities and helped me in all aspects for the successful completion of the project.

Finally, I would like to thank Vellore Institute of Technology, for providing me with a flexible choice and for supporting my project execution in a smooth manner.

Place: Vellore

Date: **HARSHITA SOROUT**

CONTENTS

Title **Page No**

LIST OF FIGURES

Title **Page No.**

LIST OF ABBREVIATIONS

Abbreviation	Expansion
CAD	Coronary artery disease
SVM	support vector machine
KNN	k-nearest neighbors
LDL	Low-density lipoprotein cholesterol
HDL	High- density lipoprotein cholesterol
AHI	Apneas hypopneas index

CHAPTER -1

INTODUCTION

1.1 THEORETICAL BACKGROUND

Coronary artery disease (CAD), also known as coronary heart disease or ischemic heart disease, is a medical condition characterized by the narrowing or blockage of the coronary arteries, the blood vessels that supply oxygen and nutrients to the heart muscle. This narrowing is usually caused by the buildup of plaque—a combination of fat, cholesterol, and other substances—that forms within the walls of the arteries.

Over time, the plaque buildup can restrict blood flow to the heart muscle, leading to various symptoms such as chest pain (angina), shortness of breath, or, in severe cases, heart attack (myocardial infarction). CAD is a major cause of heart-related morbidity and mortality worldwide.

The project "AI Based Automated System for Coronary Artery Disease Prediction" aims to address the lack of awareness and limited access to healthcare resources for individuals in rural areas who may be unaware of the symptoms and risks associated with coronary artery disease (CAD). CAD, a prevalent cardiovascular condition characterized by the narrowing or blockage of coronary arteries, can lead to severe health complications if left untreated. Early detection and intervention are crucial for mitigating its impact.

Leveraging machine learning algorithms such as logistic regression, decision tree, support vector machine (SVM), and k-nearest neighbors (KNN), the project seeks to extract valuable insights from existing health data and predict CAD risk levels (ranging from no risk to severe risk) for individuals in rural communities. By creating an intuitive program application, this initiative aims to empower individuals with personalized risk assessments, enabling them to make informed decisions about their cardiovascular health and seek appropriate medical attention when necessary.

Through the integration of medical knowledge and machine learning techniques, the project endeavors to bridge the gap in healthcare accessibility and promote early detection and prevention of CAD in underserved populations.

1.2 MOTIVATION

The "AI Based Automated System for Coronary Artery Disease Prediction" project was motivated by the urgent need to address the healthcare inequalities that rural residents experience, especially with regard to coronary artery disease (CAD).

Rural communities frequently lack access to healthcare resources and may not be aware of the signs and dangers of CAD, which can cause a delay in the diagnosis and course of therapy.

According to a 2021 study in India found that there is a lack of awareness about risk factors for coronary artery disease (CAD).Some Studies also indicate that many individuals in India are unaware of the major risk factors for CAD, such as hypertension, diabetes, high cholesterol, obesity, and smoking. This lack of awareness contributes to a higher prevalence of these risk factors, which in turn increases the burden of CAD.

Research also suggests that there is a lack of awareness about the symptoms of CAD, particularly among individuals in rural areas and those with lower socioeconomic status. Many people may not recognize symptoms such as chest pain, shortness of breath, or fatigue as potential signs of a heart problem, leading to delays in seeking medical help.

Overall, research studies underscore the importance of addressing CAD causes and symptoms unawareness in India through comprehensive public health strategies, including targeted education, improved access to healthcare services, and culturally sensitive interventions.

By raising awareness and promoting early detection and treatment, it is possible to reduce the burden of CAD and improve cardiovascular health outcomes in the Indian population.

Hence by offering a user-friendly command in line application that uses cutting-edge machine learning techniques, such as logistic regression, decision trees, support vector machines (SVM), and k-nearest neighbors (KNN), to predict CAD risk levels based on health data, this project aims to empower people living in rural communities.

The initiative intends to encourage early diagnosis and intervention, thereby lowering the burden of CAD and improving health outcomes in rural regions. It

does this by providing users with tailored risk assessments and teaching them about CAD symptoms and prevention techniques.

1.3 AIM OF THE PROPOSED WORK

The aim of the proposed work is to develop an AI-based automated system specifically designed for rural areas, targeting individuals who lack awareness of coronary artery disease (CAD) symptoms and associated risks.

In order to estimate CAD risk levels, the main goal is to develop an intuitive application that makes use of sophisticated machine learning models like logistic regression, decision trees, support vector machines (SVM), and k-nearest neighbors (KNN). Through the integration of these models into the application, the project hopes to derive pertinent information from current health data and offer users customized risk assessments for CAD.

In the end, the intention is to equip people in rural areas with the information and resources needed to determine their risk of developing CAD, allowing them to take proactive steps toward early detection and prevention. The project aims to enhance cardiovascular health outcomes and increase healthcare accessibility for marginalized groups through this initiative.

1.4 OBJECTIVE(S) OF THE PROPOSED WORK

- The objective of this project is to featuring new data from current data by extracting information and predicting CAD (in range of - no risk, low risk, high risk, severe risk).
- Dataset features are: -
    - Weight (kg)
    - Height (m)

- BMI
- Blood Sugar (mg/dL)
- Systolic Pressure (mmHg)
- Diastolic Pressure (mmHg)
- Total Cholesterol (mg/dL)
- LDL Cholesterol (mg/dL)
- HDL Cholesterol (mg/dL)
- Triglycerides (mg/dL)
- Apneas
- Hypopneas
- Sleep Hours
- AHI
- Calcium Scoring
- Patient Status (No risk, Low risk, High risk, Severe risk)

- New featured data such as : -
  - Low-density lipoprotein (LDL) cholesterol
  - Triglycerides
  - Apneas
  - Hypopneas
  - AHI index
  - calcium scoring.

- Creating an application using the machine learning models shown below to anticipate CAD (No risk, Low risk, High risk, Severe risk) :
    - Logistic regression
    - Decision tree
    - Support vector machine (SVM)
    - K-nearest Neighbors (KNN)
- Creating a command in line application.

1.5 REPORT ORGANIZATION

The report for the project "AI Based Automated System for Coronary Artery Disease Prediction" is organized into several key sections to effectively communicate the project's objectives, methodology, findings, and implications.

An introduction of coronary artery disease (CAD) and the justification for the project's emphasis on rural people is given at the outset of the paper. This is followed by a discussion of the project's motivation, literature reviews, and gaps.

After that, the report goes into detail about the methodology used for gathering data, preprocessing it, and developing the model, emphasising the use of machine learning models like logistic regression, decision trees, support vector machines (SVM), and k-nearest neighbours (KNN).

The results section presents the predictive performance of each model in categorizing CAD risk levels, emphasizing the system's ability to accurately classify individuals into risk categories.

Discussion and analysis sections delve into the implications of the findings, considering the potential impact on healthcare accessibility and early CAD detection in rural areas.

Finally, the report concludes with recommendations for future research and implementation, aiming to further enhance the effectiveness and accessibility of the AI-based automated system for CAD prediction.

CHAPTER -2

LITERATURE SURVEY

2.1. SURVEY OF THE EXISTING MODELS/ WORK

In this [1], the authors Alaa Khaleel Faieq, Maad M. Mijwil, introduce two methods for early diagnosis of heart disease, the support vector machine and artificial neural network (ANN). The medical data is taken from the University of California Irvine (UCI) Machine Learning Repository database, and it contains reports of 170 people. The investigation results confirm that the optimal execution is the support vector machine technique. It gives high-accuracy prediction results. As for the performance of the forward propagation artificial neural networks technique is acceptable.

In this research Girish S. Bhavekar & Agam Das Goswami [2] , a hybrid deep learning methodology for the categorization of cardiac disease have been developed. As a result, deep hybrid learning is more accurate than either classic deep learning or machine learning techniques used alone

In this paper [3] ,proposed a lazy associative classification for prediction of heart disease in Andhra Pradesh and present some experimental results which will help physicians to take accurate decisions.

In research paper [4], [12] different machine learning algorithms and deep learning are applied to compare the results and analysis of the UCI Machine Learning Heart Disease dataset. In research paper [5] ,includes a review of the classification methods for machine learning and image fusion that have been demonstrated to help healthcare professionals identify heart disease.

With the machine learning brief and summarize descriptions of the mainly used classification techniques for diagnosing diseases of heart. Then, review and demonstrate

some work on the use of classification techniques for machine learning and image fusion in this area. It also provides an overview of the working algorithm, and provides a description of the current work.

This project [6], gives us significant knowledge that can help us predict the patients with heart disease ;the strength of the proposed model [13] was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naive bayes, random forest etc.

This research paper [7], the system is designed using machine learning classifiers such as Support Vector Machine (SVM), Nave Bayes (NB), and K-Nearest Neighbor (KNN). The proposed work depends on the UCI database, for the diagnosis of heart diseases.

In this work [8], numerous readings have been carried out to produce a prediction model using not only distinct techniques but also by relating two or more techniques. These amalgamated new techniques are commonly known as hybrid methods.

The main concept in [9] is to identify the age group and heart rate using the Random forest algorithm. This project tells how the heart rate and condition is estimated based on the inputs such as blood pressure and many more being provided by the user to a system. This is being much better way when it comes with others algorithms the implementation of RFA gives the better experience and provide accurate result. This helps in early prediction of the disease and is used in many ways, where as it is being provided with the input, in order to find the heart rate based on the health condition.

In this paper [10], an efficient and accurate system to diagnose heart disease is proposed and the system is based on Machine learning techniques resulting in improving the accuracy in the prediction of heart disease. A cardiovascular dataset is classified by using several state of the art Machine Learning algorithms that are precisely used for disease prediction. The prediction model is introduced with the several classification techniques and the different combinations of features. And tried to produce an enhanced performance with high accuracy level through the prediction model for cardiovascular disease with the use of Machine Learning techniques like Random Forest, Naïve Bayes and SVM.

The stored data can be useful for source of predicting the occurrence of future disease. Some of the data mining and machine learning techniques are used to predict the heart disease, such as Artificial Neural Network (ANN), Decision tree, Fuzzy Logic, K-Nearest Neighbour(KNN), Naïve Bayes and Support Vector Machine (SVM). This paper provides an insight of the existing algorithm and it gives an overall summary of the existing work [11].

This project [13] proposes a prediction model to predict whether a people have a heart disease or not and to provide an awareness or diagnosis on that. This is done by comparing the accuracies of applying rules to the individual results of Support Vector Machine, Gradient Boosting, Random forest, Naive Bayes classifier and logistic regression on the dataset taken in a region to present an accurate model of predicting cardiovascular disease.

In this study[14], 29930 subjects with high-risk of CVD were selected from 101056 people in 2014, regular follow-up was conducted using electronic health record system. Logistic regression analysis showed that nearly 30 indicators were related to CVD, including male, old age, family income, smoking, drinking, obesity, excessive waist circumference, abnormal cholesterol, abnormal low-density lipoprotein, abnormal fasting blood glucose and else.

Several methods were used to build prediction model including multivariate regression model, classification and regression tree (CART), Naïve Bayes, Bagged trees, Ada Boost and Random Forest.

In this study [15] sought to construct a genomic risk score for CAD and to estimate its potential as a screening tool for primary prevention.

In this work [16], three data mining classification algorithms like Random Forest, Decision Tree and Naïve Bayes are addressed and used to develop a prediction system in order to analyse and predict the possibility of heart disease. The main objective of this significant research work is to identify the best classification algorithm suitable for providing maximum accuracy when classification of normal and abnormal person is carried out.

Thus prevention of the loss of lives at an earlier stage is possible. The experimental setup has been made for the evaluation of the performance of algorithms with the help of heart disease benchmark dataset retrieved from UCI machine learning repository. It is that Random Forest algorithm performs best with 81% precision when compared to other algorithms for heart disease prediction.

2.2. GAPS IDENTIFIED IN THE SURVEY

In [1] research paper, the dataset used in the study contains reports from only 170 people. While this may be suitable for preliminary research, it may not be representative

of the diversity of patients with heart disease. A larger and more diverse dataset would provide more robust results and validate the model's generalization capabilities. Support Vector Machines (SVM) and Artificial Neural Networks (ANN) are considered "black-box" models, which means they lack interpretability.

In [2] research paper, ,synthetic data have been used, The use of synthetic data can introduce biases or inaccuracies if it does not accurately represent real patient data. The effectiveness of the model should be validated on real-world clinical data to ensure its practical utility.

To establish the significance of the research, it's crucial to compare the proposed hybrid deep learning approach to existing methods for cardiac disease diagnosis, such as traditional diagnostic tests or other machine learning techniques.

In [3] paper, The paper mentions that it was predicted that cardiovascular disease (CVD) would be the most important cause of mortality in India by 2015. However, the research context and data may not account for changes in healthcare trends or disease prevalence beyond that year. Including more recent data or accounting for temporal changes would be important for the model's accuracy. Associative classification methods, while potentially effective, can generate complex rules that are difficult to interpret by healthcare professionals.

In [4] paper, the dataset's sample size is not large, which can limit the generalizability of the results. Larger datasets are often preferred for training machine learning models as they can capture a broader range of patterns and variations in the data. The assumption that the dataset should have a Gaussian distribution may not always hold in real-world healthcare data. Many medical datasets exhibit non-Gaussian or skewed distributions. Relying on this assumption may limit the model's applicability to diverse healthcare datasets.

In research paper [5] , The passage mentions the potential of machine learning classification methods for reliable and instant disease diagnosis. However, it does not discuss the challenges of translating these methods into real-world clinical practice, where variations in patient data and conditions may affect performance.

Availability of a comprehensive and diverse dataset for training and testing machine learning models can be a challenge, especially in healthcare applications [6]. It's important to acknowledge the potential limitations of the dataset, including its representativeness of the broader population.

ML models heavily rely on the quality and quantity of data. If the data used for training the model is incomplete, inconsistent, or biased, it can significantly impact the model's accuracy and generalizability [7].

ML models, especially complex ones, can overfit the training data, capturing noise rather than meaningful patterns [8]. Overfit models perform well on the training data but fail to generalize to new, unseen data, leading to poor performance in real-world scenarios.

[9] The given proposed work does not mention specific methods for ensuring the quality of the data obtained from the UCI repository. Poor quality data, such as missing values or outliers, can significantly impact the accuracy and reliability of the classification model. And their prediction of cardiovascular disease results is not accurate. Data mining techniques does not help to provide effective decision making. Cannot handle enormous datasets for patient records.Biases present in the training data can be learned by ML models, leading to biased predictions. For instance, if certain demographic groups are underrepresented in the data, the model might not perform well for those groups, leading to healthcare disparities [10].

In [11] , Selecting the right set of features (variables) is a critical step in building accurate ML models. Choosing irrelevant or redundant features can degrade the model's performance. However, identifying the most relevant features can be challenging, especially in medical datasets with numerous variables.

[12] paper, assumes the availability of high-quality medical data without addressing the challenges related to data accuracy, completeness, and consistency. Real-world healthcare data often contains errors and missing values, which can significantly impact the effectiveness of predictive models.

[13] Heart disease prevalence and outcomes can be influenced by various socio-economic factors such as access to healthcare, education, and lifestyle choices. Ignoring these factors limits the holistic understanding of heart disease prediction and prevention.

In [14] several limitations should be addressed. The main limitation of the study was that it lacked external validation. ML could be deemed as internal validation to some extent since it consisted of multiple data-oriented analyses through randomly splitting the data repeatedly. And the validation and optimization of current model needed to be performed in future study.

In [15], Coronary artery disease (CAD) has substantial heritability and a polygenic architecture. However, the potential of genomic risk scores to help predict

CAD outcomes has not been evaluated comprehensively, because available studies have involved limited genomic scope and limited sample sizes.

In real-world scenarios, achieving the highest accuracy might lead to overfitting, where the model performs well on the training data but fails to generalize to unseen data [16].

2.3. PROBLEM STATEMENT

Coronary Artery Disease (CAD) poses a significant public health challenge globally, including in India, where it is a leading cause of morbidity and mortality. Despite advances in medical science and interventions, CAD remains a major cause of concern due to its high prevalence, associated healthcare burden, and often asymptomatic progression until advanced stages.

In India, awareness of CAD is significantly influenced by access to healthcare facilities.

Research has indicated that people who live in rural or isolated regions and have limited access to healthcare facilities may not be as aware of CAD and may not be as likely to obtain a timely diagnosis and treatment.

In order to educate patients and increase public awareness about CAD, healthcare providers are essential. Research indicates that in order to successfully communicate with patients about the prevention, diagnosis, and management of coronary artery disease (CAD), healthcare workers may require additional training and resources.

The problem statement revolves around the need for effective CAD prediction and prevention strategies to mitigate its impact on individuals and healthcare systems.

Many people suffer from heart problems, and often, these issues are detected too late, leading to serious consequences. The challenge is to finding a means to identify coronary artery disease early on, before it manifests and becomes critical.

Traditional methods aren't always fast or accurate enough. That's why there's a need for advanced Artificial Intelligence (AI) tools.

The challenge is to take information from demographic data that already exists and create new information. Developing effective AI solutions for early prediction of coronary artery disease is crucial for improving healthcare outcomes and saving lives.

CHAPTER-3

OVERVIEW OF THE PROPOSED SYSTEM WITH REQUIRMENTS ANALYSIS AND DESIGN

The proposed system for "AI Based Automated System for Coronary Artery Disease Prediction" aims to address the lack of awareness and accessibility to healthcare resources for individuals residing in rural areas who may be unaware of coronary artery disease (CAD) symptoms and risks.

The system is designed to feature an intuitive program application that leverages advanced machine learning models, including logistic regression, decision tree, support vector machine (SVM), and k-nearest neighbors (KNN), to predict CAD risk levels ranging from no risk to severe risk.

By utilizing these models, the system extracts valuable insights from existing health data and provides personalized risk assessments for users.

3.1. REQUIRMENTS ANALYSIS

The requirements analysis for the project "AI Based Automated System for Coronary Artery Disease Prediction" involves identifying the necessary components and functionalities to achieve its objectives.

Firstly, the system must be designed with simplicity and accessibility in mind, catering to individuals residing in rural and remote areas who may have limited access to healthcare resources.

The system should include an intuitive program application developed in Python language, allowing users to input their health data easily. Additionally, the system should incorporate machine learning models such as logistic regression, decision tree, support vector machine (SVM), and k-nearest neighbors (KNN) to predict CAD risk levels based on the provided data.

The data set for training and testing these models should include features such as weight, height, body mass index, blood sugar level, blood pressure readings, cholesterol levels, sleep patterns, and calcium scoring, along with the corresponding CAD risk status.

Furthermore, the system should compute and display performance metrics such as confusion matrix, F1-score, precision, recall, and accuracy to evaluate the predictive performance of each model.

By fulfilling these requirements, the system aims to provide accurate CAD risk assessments and empower individuals with the knowledge to manage their cardiovascular health effectively.

3.1.1. FUNCTIONAL REQUIREMENTS

Data Processing and Integration: The system must be capable of processing both fresh data and existing datasets containing essential health parameters such as weight, height, body mass index, blood sugar level, blood pressure, cholesterol levels, sleep patterns, apnea occurrences, hypopnea occurrences, apnea hypopnea index (AHI), calcium scoring, and patient status. These datasets will be utilized for training and testing the machine learning models.

Information Extraction and Prediction: Employing machine learning models including Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-nearest Neighbors (KNN), the system should extract relevant information from the datasets and predict CAD risk categories spanning from "no risk" to "severe risk".

In-Program Application Development: An intuitive in-program application must be developed using Python programming language to facilitate user interaction. The application should allow users to input their health data easily.

Model Evaluation: The system should utilize evaluation metrics such as confusion matrix, F1-score, precision, recall, and accuracy to assess the performance of the machine learning models in predicting CAD risk levels.

Feature Selection: Only eight essential features from the dataset will be utilized for training and testing within the specified machine learning algorithms.

User Interface: The in-program application should have a user-friendly interface, allowing users to input their health data and promptly receive their CAD risk level. This feature aims to facilitate proactive health management for individuals in rural and remote areas and those unaware of CAD symptoms.

3.1.1.1. PRODUCT PERSPECTIVE

From a product standpoint, the "AI Based Automated System for Coronary Artery Disease Prediction" is a helpful and easily navigable tool for those living in rural and isolated locations and for people who aren't aware that they have coronary artery disease (CAD).

By offering predicted insights about CAD risk levels, this technology is an essential tool for preventive health management. The system uses sophisticated information extraction techniques to process both new and pre-existing datasets that contain critical health characteristics in order to predict CAD risk categories that range from "no risk" to "severe risk."

3.1.1.2. PRODUCT FEATURES

The AI-based automated system for coronary artery disease (CAD) prediction offers several key features to provide accurate and accessible risk assessments for users:

Data Input: Users can easily input their health data, including weight, height, body mass index, blood sugar level, blood pressure, cholesterol levels, sleep patterns, apnea occurrences, hypopnea occurrences, apnea hypopnea index (AHI), calcium scoring, and patient status.

Machine Learning Models: The system employs advanced machine learning algorithms such as Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-nearest Neighbors (KNN) to process the input data and predict CAD risk levels.

Personalized Risk Assessment: Based on the input data, the system generates personalized CAD risk assessments for users, categorizing their risk levels from no risk to severe risk.

Performance Evaluation: The system evaluates the performance of the machine learning models using metrics such as confusion matrix, F1-score, precision, recall, and accuracy, ensuring reliability and effectiveness.

User-Friendly Interface: The system features an intuitive in-program application with a user-friendly interface, allowing users to input their health data effortlessly and receive prompt risk assessments.

Accessibility: Designed with individuals in rural and remote areas in mind, the system is accessible to users with limited access to healthcare resources, aiding in early detection and intervention for CAD.

Proactive Health Management: By providing timely and accurate CAD risk assessments, the system enables users to proactively manage their cardiovascular health and seek appropriate medical attention when necessary.

3.1.1.3. USER CHARACTERISTICS

The target users of the "AI Based Automated System for Coronary Artery Disease Prediction" exhibit specific characteristics aligned with the project's objectives. Primarily, the system caters to individuals residing in rural and remote areas, where access to healthcare resources and awareness of coronary artery disease (CAD) symptoms may be limited.

These users may lack easy access to medical facilities and expertise, making early detection and intervention for CAD challenging.

Additionally, the system is intended for individuals who may be unaware of CAD symptoms or risk factors, highlighting the need for proactive health management and education.

Moreover, the users of this system are likely to value simplicity and accessibility in healthcare solutions, necessitating an intuitive in-program application interface. Furthermore, users may vary in their level of comfort with technology, emphasizing the importance of a user-friendly design and straightforward input process.

Overall, the system targets individuals who require accessible, accurate, and timely CAD risk assessments to empower them with the knowledge and tools for proactive health management, especially in underserved communities.

3.1.1.4. ASSUMPTIONS & DEPENDECIES

Data Quality: The system assumes that the input data provided by users is accurate and reliable. Any inaccuracies or inconsistencies in the data may lead to erroneous predictions and affect the performance of the machine learning models.

Feature Importance: The system assumes that the selected features for CAD prediction, including weight, height, body mass index, blood sugar level, blood pressure, cholesterol levels, sleep patterns, apnea occurrences, hypopnea occurrences, apnea hypopnea index (AHI), calcium scoring, and patient status, are sufficient and relevant for accurate risk assessment. The effectiveness of the system relies on the importance of these features in predicting CAD risk levels.

Model Generalization: The system assumes that the machine learning models trained on the provided datasets generalize well to unseen data. However, the performance of the models may be influenced by factors such as dataset size, distribution, and representativeness.

Algorithm Suitability: The system assumes that the selected machine learning algorithms, including Logistic Regression, Decision tree, Support Vector Machine (SVM), and K-nearest Neighbors (KNN), are suitable for CAD prediction based on the provided data. The effectiveness of these algorithms may depend on the characteristics of the data and the underlying assumptions of each algorithm.

Programming Language and Tools: The system depends on the Python programming language for implementation and utilizes libraries such as scikit-learn, seaborn, numpy, pandas, matplotlib.pyplot for machine learning modeling and evaluation.

Any changes or limitations in the functionality of these tools may impact the development and performance of the system.

User Engagement: The system assumes user engagement and cooperation in providing accurate health data for CAD risk assessment. User input is essential for the system to generate personalized risk assessments and facilitate proactive health management. Any reluctance or inconsistency in user input may affect the accuracy and reliability of the predictions.

3.1.1.5. DOMAIN REQUIREMENTS

The domain requirements for the "AI Based Automated System for Coronary Artery Disease Prediction" are deeply rooted in the healthcare domain, specifically addressing the challenges faced by individuals in rural and remote areas with limited access to healthcare resources.

The system must encompass a comprehensive understanding of coronary artery disease (CAD) and its associated risk factors, including weight, height, body mass index, blood sugar level, blood pressure, cholesterol levels, sleep patterns, apnea occurrences, hypopnea occurrences, apnea hypopnea index (AHI), calcium scoring, and patient status.

Furthermore, the system must incorporate knowledge of machine learning algorithms such as Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-nearest Neighbors (KNN) to effectively process and analyze health data. For patient information to be accurate, dependable, and private, the system must strictly abide by healthcare laws and standards.

Moreover, the system needs to give top priority to user accessibility and usability, especially when creating the in-program application, so that people may take an active role in managing their healthcare. The system seeks to close the gap in healthcare accessibility and enable proactive management of CAD risk in marginalised populations by addressing these domain criteria.

3.1.2. NON-FUNCTIONAL REQUIREMENTS

- Performance: The system must be able to process and analyze health data efficiently to provide timely CAD risk assessments. Response times for user inputs and risk predictions should be minimized to enhance user experience.

- Accuracy: The CAD risk assessments generated by the system must be highly accurate and reliable. Machine learning models should be trained and evaluated using high-quality datasets to ensure the validity of predictions.

- Privacy and Security: The system must adhere to strict privacy regulations to safeguard the confidentiality of user health data. Measures such as encryption, access controls, and data anonymization should be implemented to protect sensitive information.

- Scalability: The system should be scalable to accommodate a growing number of users and increasing volumes of health data. It should be able to handle concurrent requests without compromising performance or accuracy.

- Usability: The in-program application must have an intuitive and user-friendly interface, allowing users to input their health data easily and understand the CAD risk assessments provided by the system. Clear instructions and visual aids should be provided to enhance usability, especially for users with limited technical expertise.

- Reliability: The system must be reliable and available for use at all times. It should have built-in mechanisms for error handling and recovery to minimize downtime and ensure uninterrupted service.

- Compliance: The system must comply with relevant regulatory standards and healthcare industry guidelines. This includes adherence to data protection regulations such as HIPAA and GDPR, as well as compliance with ethical standards for handling sensitive health information.

- Maintainability: The system should be easy to maintain and update to incorporate new features, address bugs, and improve performance over time. Modular design principles and well-documented code should be employed to facilitate maintenance by developers.

- Accessibility: The in-program application should be accessible to users with disabilities, adhering to accessibility standards such as WCAG (Web Content Accessibility Guidelines). This includes support for screen readers, keyboard navigation, and alternative input methods.

- Compatibility: The system should be compatible with a wide range of devices and operating systems to ensure accessibility for users across different platforms. Compatibility testing should be conducted to verify the system's functionality on various devices and browsers.

3.1.3. SYSTEM MODELING

One common mathematical model used in machine learning for classification tasks like coronary artery disease (CAD) prediction is logistic regression. In logistic regression, the probability that a given input (set of health parameters) belongs to a certain category (CAD risk level) is modeled using the logistic function. Mathematically, the logistic regression model can be represented as:

$$P(y = 1|\mathbf{x}; \mathbf{w}) = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

Where:

- (P(y = 1 | x; w) represents the probability that the input x belongs to class 1 (e.g., high risk of CAD).

- x is the input vector representing the health parameters.

- w is the weight vector.

- b is the bias term.

- e is the base of the natural logarithm.

Logistic regression can be extended to handle multiple classes using techniques like one-vs-all or softmax regression.

Description:

Logistic regression is a fundamental machine learning algorithm used for binary classification tasks.

In the context of the "AI Based Automated System for Coronary Artery Disease Prediction," logistic regression can be utilized to predict the likelihood of an individual belonging to a particular CAD risk category (e.g., no risk, low risk, high risk, severe risk) based on their health parameters.

By training the logistic regression model on a dataset containing labeled examples of CAD risk levels and corresponding health parameters, the system can learn to make accurate predictions for new, unseen data.

The logistic regression model's coefficients (weights) and bias are optimized during the training process to minimize the error between predicted probabilities and actual CAD risk labels.

Decision trees are a popular machine learning algorithm used for classification tasks, including CAD prediction.

A decision tree recursively splits the input space (represented by the health parameters in this case) into regions, with each split based on a chosen feature and a threshold value. The splits are determined based on criteria such as Gini impurity or entropy, which measure the homogeneity of the target variable (CAD risk level).

Mathematically, a decision tree can be represented as a tree structure consisting of decision nodes, which represent feature-value pairs for splitting, and leaf nodes, which represent the predicted CAD risk level.

The decision-making process follows a path from the root node to leaf nodes, where each node makes a decision based on the value of a specific feature.

In K-nearest neighbors (KNN) is a simple yet effective algorithm used for classification tasks. In KNN, the CAD risk level of a given input (set of health

parameters) is determined by the majority class among its K nearest neighbors in the training dataset.

The distance metric used to measure similarity between data points (e.g., Euclidean distance) determines the neighbors. Mathematically, KNN can be represented as:

[y = majority vote(y_1, y_2, ..., y_k)]

Where:

- (y) represents the predicted CAD risk level for the input.

- (y_1, y_2, ..., y_k) represent the CAD risk levels of the K nearest neighbors.

Support Vector Machine (SVM) is a powerful machine learning algorithm used for classification tasks, including CAD prediction.

In SVM, the algorithm finds the optimal hyperplane that best separates the data points belonging to different CAD risk categories.

The hyperplane is chosen to maximize the margin between the data points of different classes. Mathematically, SVM can be represented as:

$$\text{minimize} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i$$

Subject to:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \forall i$$

Where:

- w is the weight vector.

- b is the bias term.

- xi_i are slack variables.

- C is a regularization parameter.

- y_i are the CAD risk labels (-1 or 1).

- x_i are the input vectors representing health parameters.

Description:

Decision trees, K-nearest neighbors (KNN), and Support Vector Machine (SVM) are three common machine learning algorithms used for classification tasks, including CAD prediction.

In the context of the "AI Based Automated System for Coronary Artery Disease Prediction," these algorithms are utilized to analyze the input health parameters and predict the CAD risk level for an individual.

Decision trees split the input space into regions based on feature values, KNN determines the CAD risk level based on the majority class among nearest neighbors, and SVM finds the optimal hyperplane to separate different CAD risk categories.

By training these models on a dataset containing labelled examples of CAD risk levels and corresponding health parameters, the system can learn to make accurate predictions for new, unseen data.

ER DIAGRAM

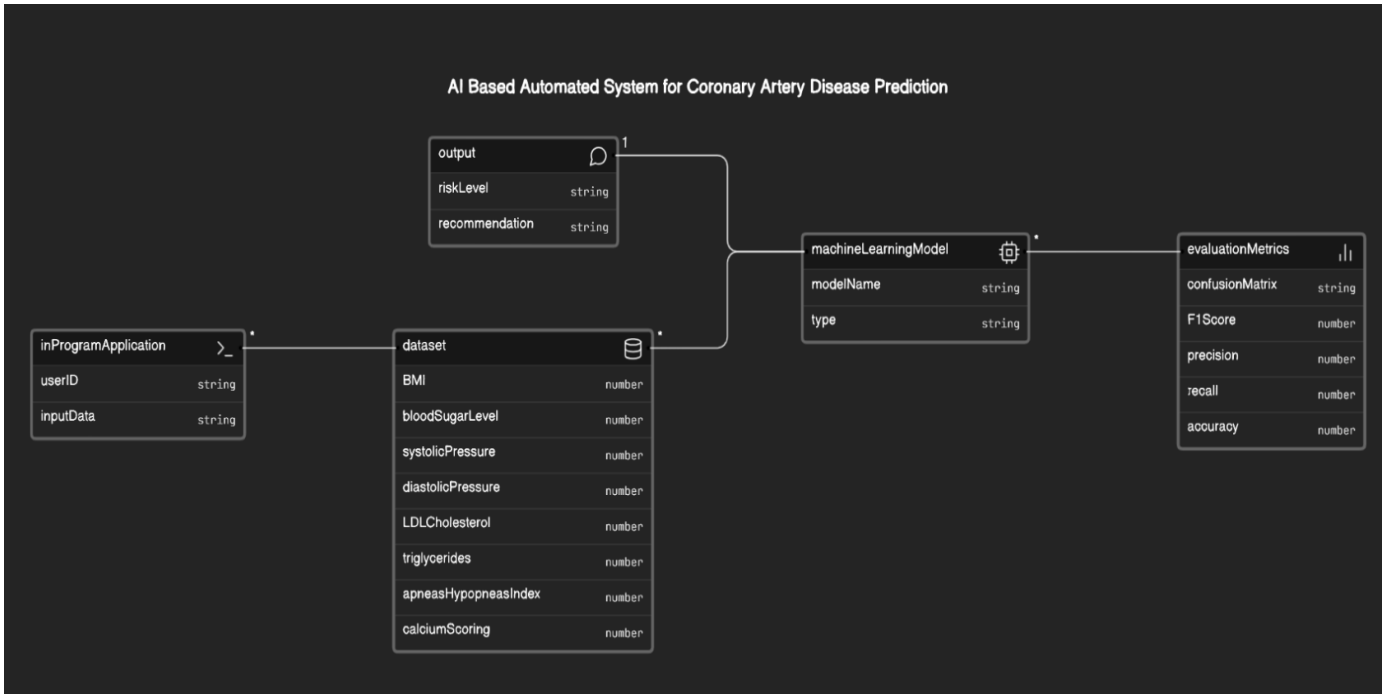

Fig 3.1 ER Diagram

DATA FLOW DIAGRAM (DFD)

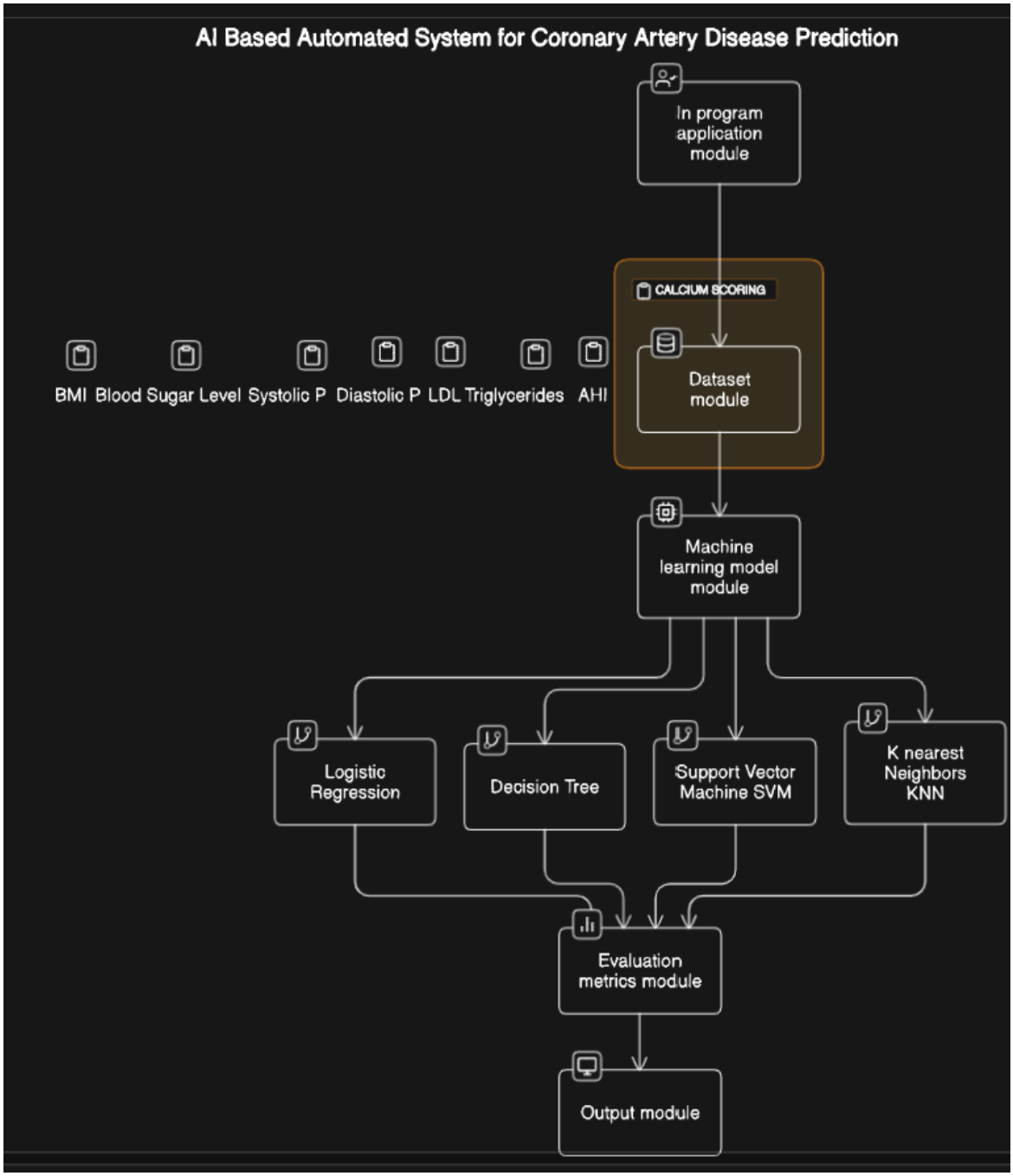

Fig 3.2 Data flow diagram

SYSTEM TRANSITION DIAGRAM (STD)

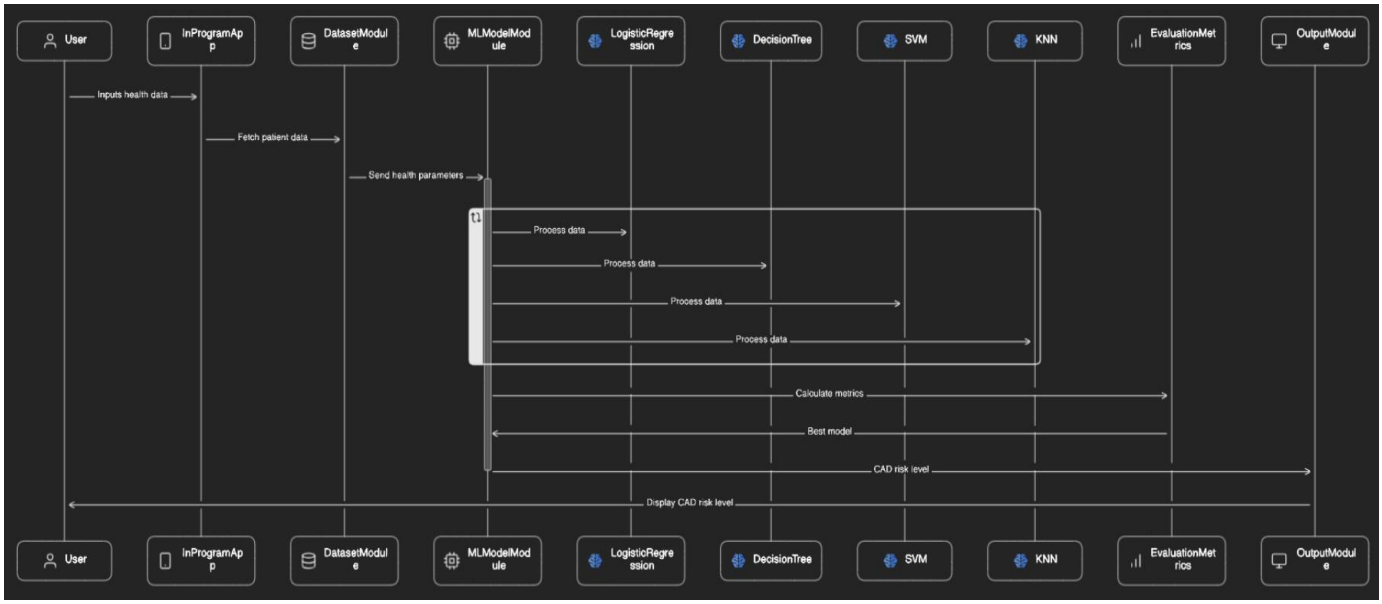

Fig 3.3 System transition diagram

3.1.4. ENGINEERING STANDARD REQUIREMENTS

- Economic:

  The system should be economically feasible, considering factors such as development costs, maintenance expenses, and potential cost savings for healthcare providers and individuals.

  By providing early detection and proactive management of coronary artery disease (CAD), the system can potentially reduce healthcare costs associated with treating advanced stages of the disease and its complications.

  Additionally, the economic impact of deploying the system in rural and remote areas, where access to healthcare services may be limited, should be assessed to ensure affordability and sustainability.

- Environmental:

  The development and deployment of the system should adhere to environmental standards to minimize its ecological footprint.

  This includes considerations such as energy efficiency in computing resources used for model training and inference, as well as responsible disposal of electronic waste generated during hardware upgrades or replacements.

  Sustainable practices should be prioritized throughout the system's lifecycle to minimize environmental impact.

- Societal Need:

  The system addresses a significant societal need by providing predictive insights into CAD risk levels, particularly for individuals in rural and remote areas who may have limited access to healthcare services.

  By empowering individuals to monitor their cardiovascular health and seek timely medical intervention when necessary, the system contributes to improving public health outcomes and reducing disparities in healthcare access.

- Political:

  The development and implementation of the system should align with relevant political regulations and policies governing healthcare technology and data privacy.

  Compliance with regulatory frameworks such as HIPAA and GDPR ensures that patient health information is handled securely and ethically, maintaining individuals' rights to privacy and confidentiality.

- Ethical:

  Ethical considerations are paramount in the development of the system, particularly regarding data privacy, algorithmic bias, and transparency in decision-making.

  The system should prioritize the ethical collection, storage, and use of health data, ensuring informed consent and respecting individuals' autonomy and rights. Transparency in the CAD risk prediction process, including explanations for model decisions and potential limitations, fosters trust and accountability among users and stakeholders.

- Health and Safety:

  The system should prioritize user safety and well-being, ensuring that CAD risk predictions are accurate and reliable to facilitate timely medical intervention when necessary.

  Additionally, the system should not pose any health risks to users, such as through the dissemination of misleading or inaccurate health information.

  Robust validation and testing procedures should be implemented to verify the system's performance and safety.

- Sustainability:

  The system should be designed with long-term sustainability in mind, considering factors such as scalability, adaptability to evolving healthcare needs, and ongoing support and maintenance.

  Sustainable practices, such as modular design and open data standards, enable the system to be easily integrated with existing healthcare infrastructure and adapted to future advancements in CAD detection and management.

- Inspectability:

  The system should be designed to facilitate inspectability and accountability, allowing for transparency and scrutiny of its inner workings by regulatory authorities, healthcare professionals, and users.

  Documentation of data sources, model architectures, training procedures, and performance evaluation metrics enables thorough auditing and validation of the system's functionality and adherence to engineering standards and best practices.

3.1.5. SYSTEM REQUIREMENTS

3.1.5.1. HARDWARE REQUIREMENTS

- Computing Devices: The system should be compatible with various computing devices such as desktop computers, laptops.
- Processor: The hardware should have a processor with sufficient computational power to handle data processing and machine learning model training and inference efficiently.
- Memory (RAM): Adequate RAM is required to store and manipulate datasets and perform computations during model training and prediction.
- Storage: Sufficient storage space is necessary to store datasets, machine learning models, and application files.
- Network Connectivity: Internet connectivity may be required for accessing online resources, downloading updates, and communicating with external servers if applicable.

3.1.5.2. SOFWARE REQUIREMENTS

- Operating System: The system should support major operating systems including Windows.
- Libraries: Required libraries include scikit-learn for machine learning algorithms, pandas for data manipulation, matplotlib. Pyplot for maps ,numpy ,seaborn .
- Development Tools: Integrated development environments (IDEs) such as Google colab has been used for software development, debugging, and testing purposes.

3.2. SYSTEM DESIGN

3.2.1. SYSTEM ARCHITECTURE

The system architecture for the "AI Based Automated System for Coronary Artery Disease Prediction" consists of several key components designed to achieve the project's objectives efficiently and effectively.

At the core of the architecture are the machine learning models, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-nearest Neighbors (KNN), which are utilized to predict CAD risk levels based on input health parameters.

These models are implemented using the Python programming language, leveraging libraries such as scikit-learn, numpy, pandas, seaborn, matplotlib.pyplot for model training and inference.

The system also incorporates an in-program and providing users with an intuitive interface to input their health data and receive prompt CAD risk predictions.

Data preprocessing techniques are employed to handle missing values, normalize features, and encode categorical variables, ensuring data quality and consistency.

Evaluation metrics such as confusion matrix, F1-score, precision, recall, and accuracy are utilized to assess the performance of the machine learning models.

The dataset used for training and testing encompasses essential health parameters, including weight, height, body mass index, blood sugar level, blood pressure, cholesterol levels, sleep patterns, apnea occurrences, hypopnea occurrences, apnea hypopnea index (AHI), calcium scoring, and patient status.

Notably, only eight features are utilized within the specified machine learning algorithms to streamline computational complexity and enhance model interpretability. Overall, the system architecture enables seamless integration of data processing, model training, prediction, and user interaction, culminating in an intuitive application that empowers users to proactively manage their cardiovascular health.

3.2.2. DETAILED DESIGN

In the detailed design of the "AI Based Automated System for Coronary Artery Disease Prediction," the system can be broken down into several modules, each responsible for specific functionalities. Here's a description of the modules along with UML diagrams:

Data Preprocessing Module:

Description: This module is responsible for preprocessing the input data before feeding it into the machine learning models. It handles tasks such as handling missing values, scaling or normalizing features, and encoding categorical variables.

UML Diagram:

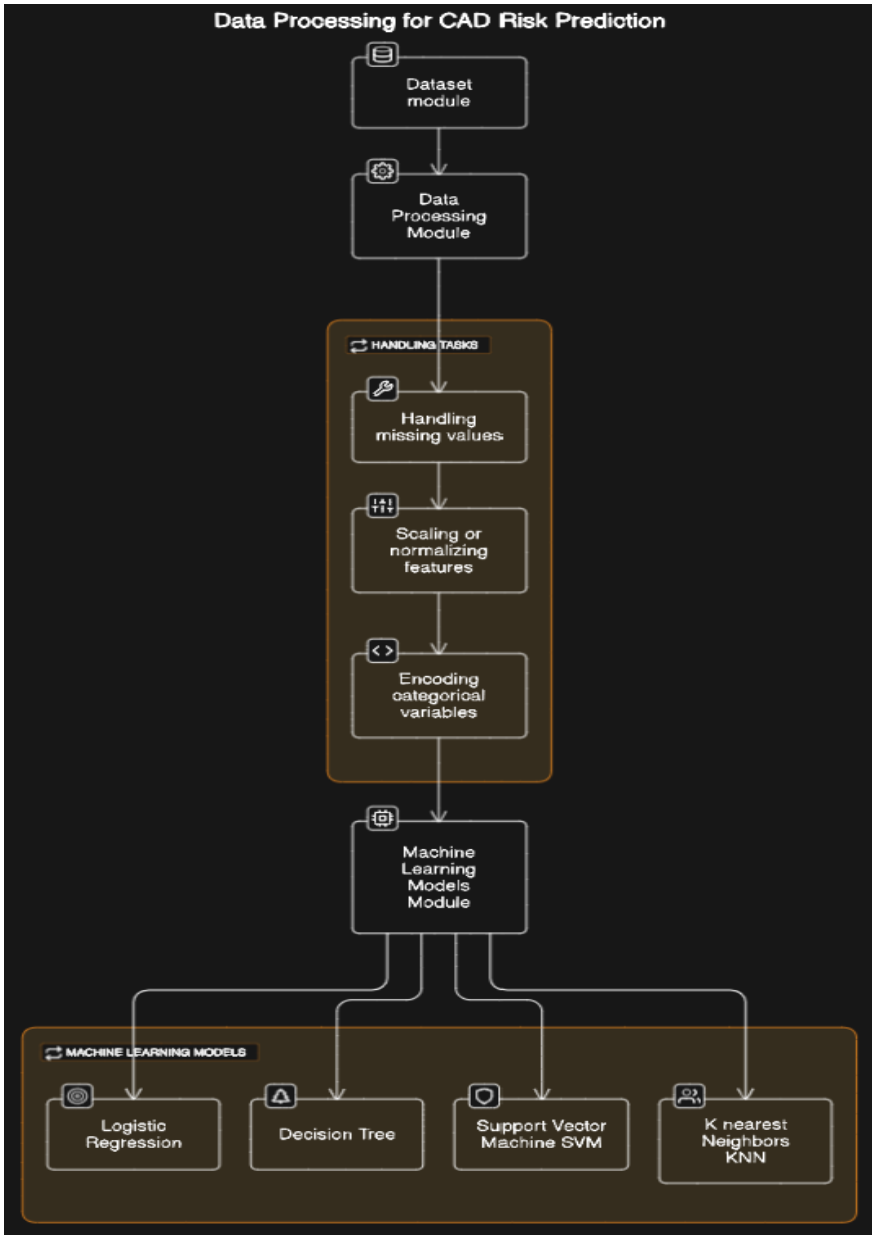

Fig 3.4 Data Preprocessing

Machine Learning Model Module:

Description: This module contains the implementation of various machine learning models such as Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-nearest Neighbors (KNN). Each model is trained on the preprocessed data to predict the CAD risk levels.

UML Diagram:

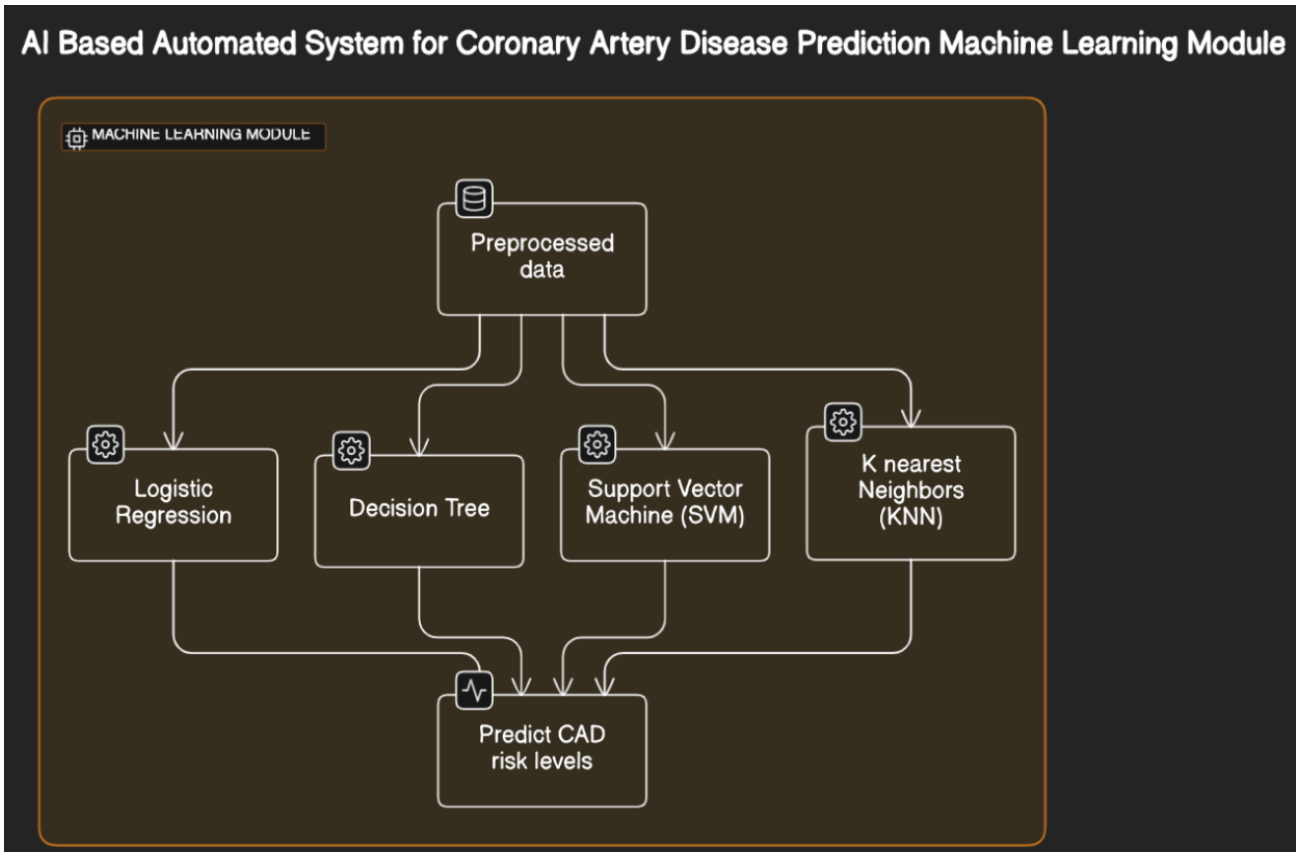

Fig 3.5 Machine Learning Model

Evaluation Module:

Description: This module evaluates the performance of the machine learning models using evaluation metrics such as confusion matrix, F1-score, precision, recall, and accuracy. It provides insights into how well the models are predicting CAD risk levels.

UML Diagram:

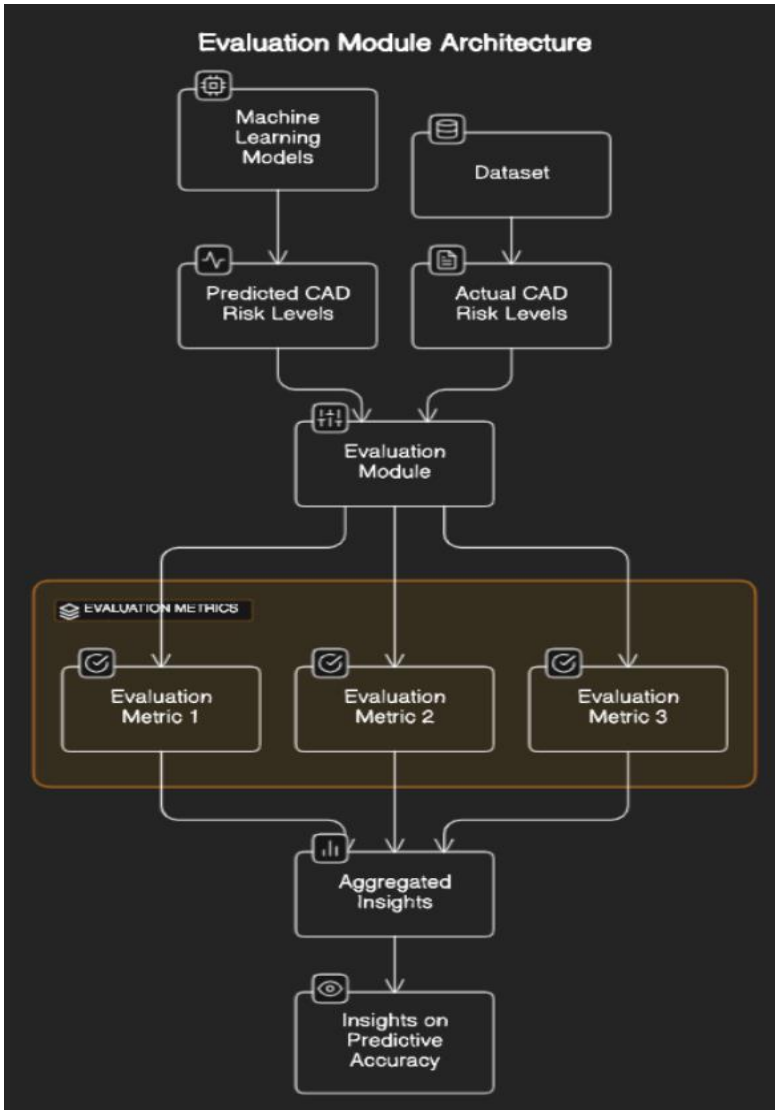

Fig 3.6 Evaluation Module

In-Program Application (command-line interface) Module:

Description: This module is, where users can input their health data and receive CAD risk predictions. It provides a user-friendly interaction with the system, (command-line interface).

UML Diagram:

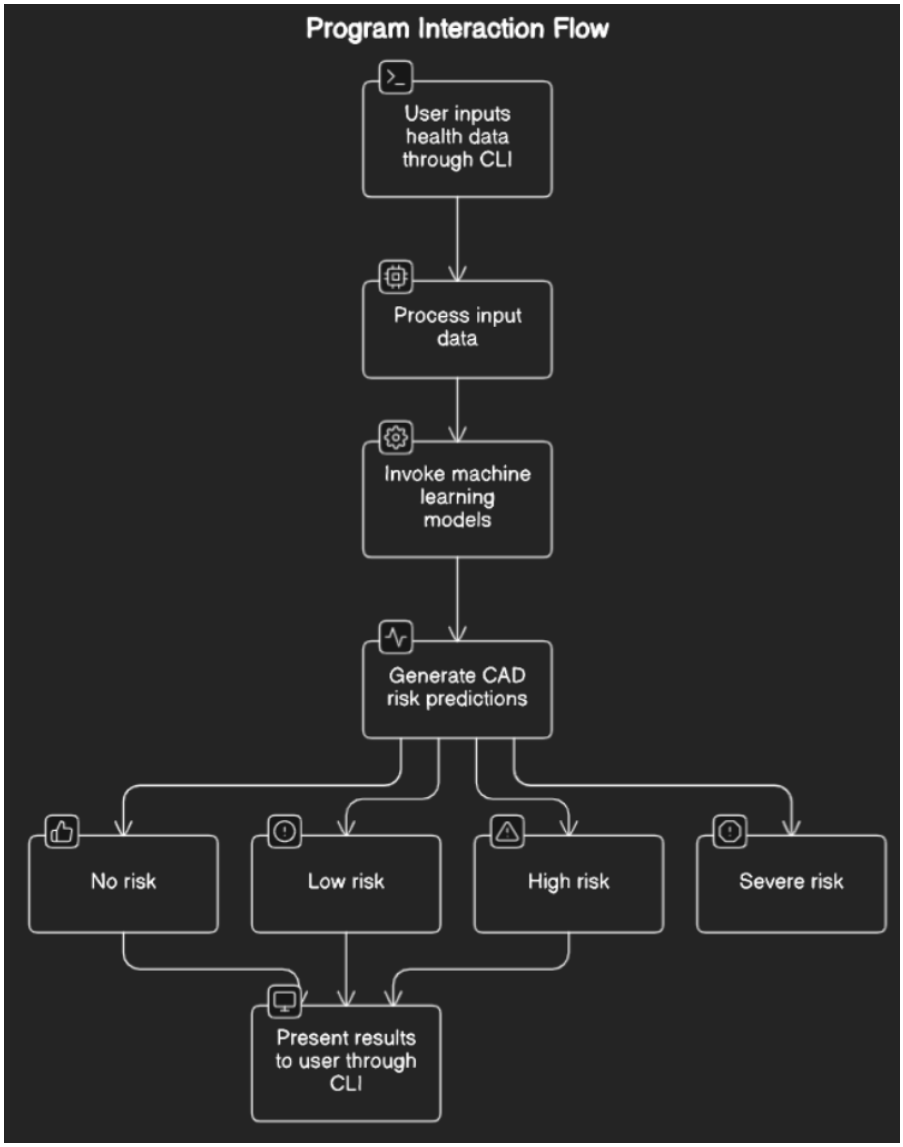

Fig 3.7 command In-line interface

Integration Module:

Description: This module integrates all other modules and orchestrates the flow of data and control throughout the system. It ensures seamless communication between different components of the system.

UML Diagram:

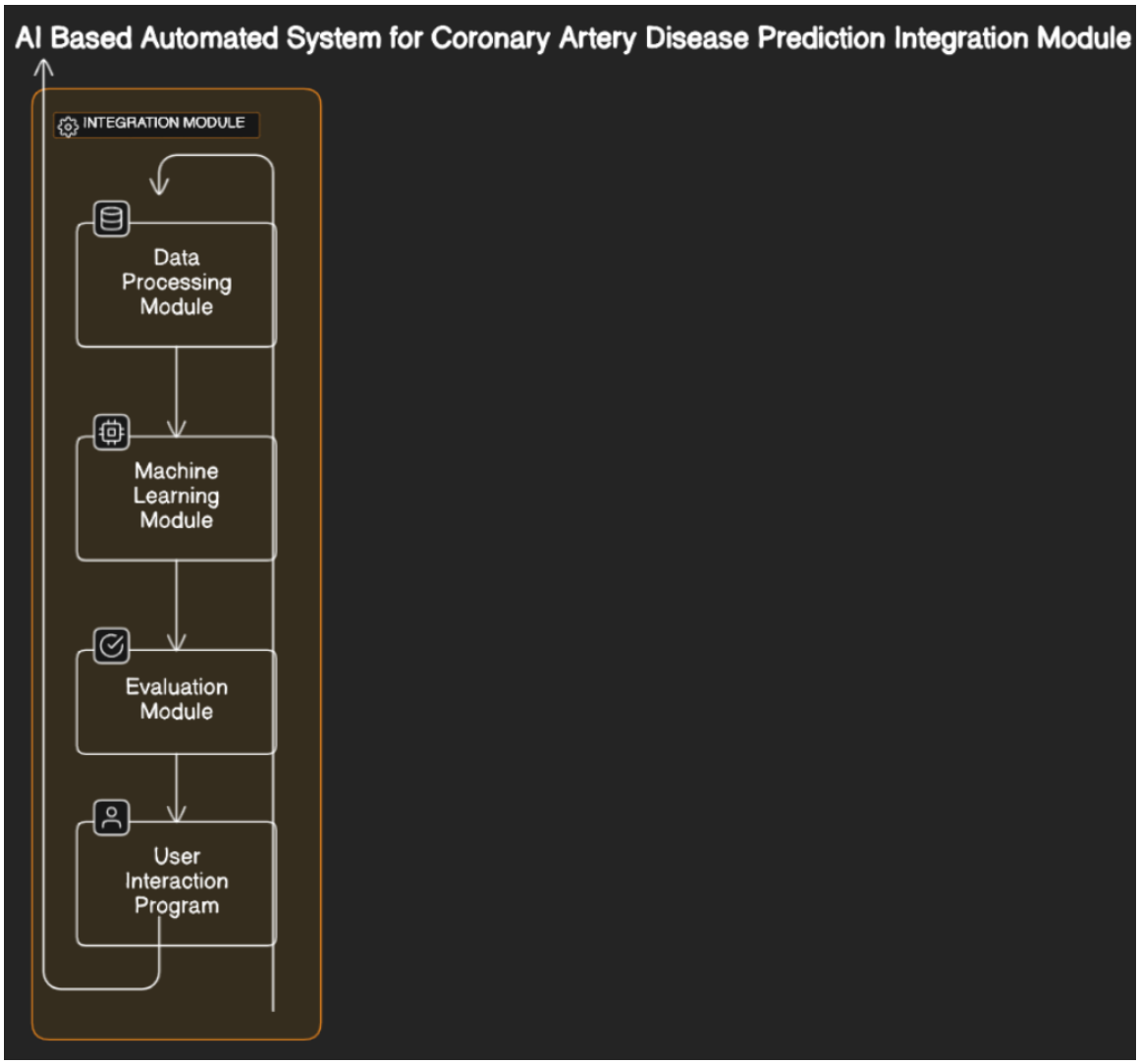

Fig 3.8 Integration Module

These modules work together to form the complete system for CAD risk prediction. Each module encapsulates specific functionalities, making the system modular, scalable, and maintainable.

CHAPTER -4

IMPLEMENTATION AND TESTING

4.1. METHODOLOGIES

- **Data Collection and Preprocessing:** -

  We assembled a dataset with (430 entries, 16 columns) comprising various health parameters relevant to CAD, including -

  - weight,
  - Height,
  - BMI (body mass index),
  - Blood sugar level,
  - Blood pressure (systolic, diastolic),
  - Cholesterol levels (total cholesterol, low density lipoprotein, high density lipoprotein, Triglycerides)
  - Apnea- hypopnea index (AHI) – (apnea, hypopneas, sleep hours) ,and
  - Calcium scoring

- **Model Training: -**

  We employed four distinct machine learning algorithms for CAD prediction: logistic regression, decision tree, support vector machine (SVM), and K-nearest neighbor (KNN).

  Only 8 features have been taken out of 16 and The dataset was split into training and testing sets, with an 80-20 ratio, to train and evaluate the performance of each model.

- ***Logistic Regression***

  It is a statistical model used for binary classification tasks. But Logistic regression for multiple targets, also known as multinomial logistic regression , extends logistic regression to handle more than two classes.

  It estimates the probability of each class using a logistic function, where each class has its own set of weights.

  The model is trained by optimizing the parameters (weights) through iterative method, which measures the difference between predicted and actual class probabilities across all classes.

- ***Decision tree***

  It recursively split the dataset into smaller subsets based on the most significant attribute at each node, aiming to maximize the homogeneity of the resulting subsets with respect to the target variable.

  The model is trained by selecting the best splitting criterion (e.g., Gini impurity, information gain) for each node until a stopping criterion is met (e.g., maximum depth reached).

  Note- Homogeneity: A perfectly homogeneous clustering is one where each cluster has data-points belonging to the same class label. Homogeneity describes the closeness of the clustering algorithm to this perfection.

- ***Support vector machine (SVM)-***

  SVM is a supervised learning algorithm used for classification and regression tasks. It constructs a hyperplane in a high-dimensional feature space, aiming to maximize the margin between different classes.

  The model is trained by finding the optimal hyperplane that separates the classes while minimizing classification errors and maximizing the margin, typically using optimization techniques such as quadratic programming or gradient descent.

- ***K-nearest neighbor-***

  KNN is a non-parametric algorithm used for classification and regression tasks. It classifies a new instance by a majority vote of its k nearest neighbors in the feature space.

  The model is trained by storing all available instances and their corresponding class labels or values, without explicit training, and classifying new instances based on the majority class or mean value of their nearest neighbors.

- **Model Evaluation: -**

  The accuracy of each model was assessed using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score.

- **Deployment**

  This deployment phase ensures accessibility and usability of the system for the target audience.

CHAPTER -5

RESULTS AND DISCUSSION

5.1. RESULT ANALYSIS

In this project, logistic regression, decision tree, SVM, KNN has been used but, with a highest accuracy rate, decision tree is the most accurate data analysis tool for this project. This is because it can catch complex, non-linear connections between the target variable and input variables, which other methods may overlook.

Decision tree does feature selection automatically by identifying the most informative attributes for categorization. Decision trees can concentrate on the most pertinent features for CAD prediction because to this feature selection technique, which increases accuracy.

Users can comprehend the underlying decision-making process by using decision trees, which provide transparency and interpretability. Healthcare practitioners can learn more about the factors determining CAD risk thanks to this transparency, which also helps to build trust in the model.

Due to its ability to divide the feature space using decision boundaries determined from data splits, decision trees exhibit robustness against outliers and noise in the data. The robustness of the model improves its overall predictive accuracy and allows it to generalize to new inputs.

5.2. EVALUATION MATRICS

LOGISTIC REGRESSION

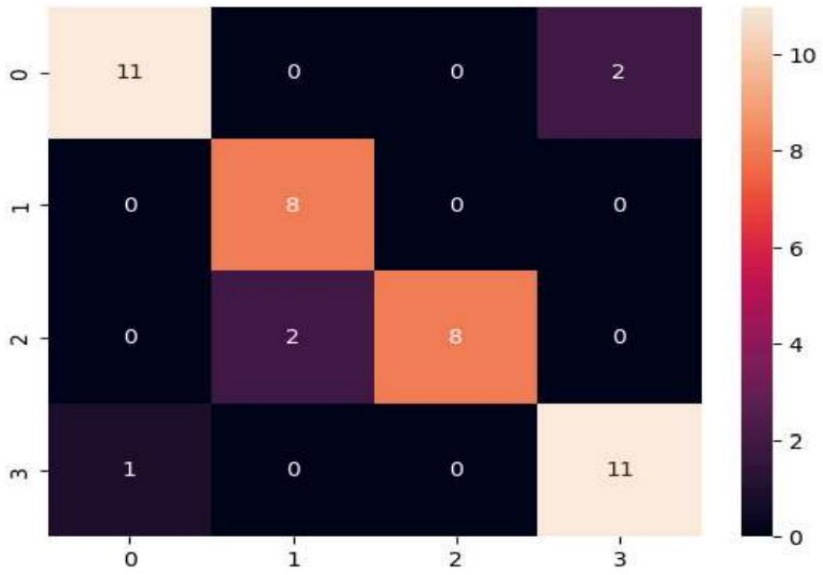

Fig 5.1 confusion matrix

	precision	recall	f1-score	support
0	0.92	0.85	0.88	13
1	0.80	1.00	0.89	8
2	1.00	0.80	0.89	10
3	0.85	0.92	0.88	12
accuracy			0.88	43
macro avg	0.89	0.89	0.88	43
weighted avg	0.89	0.88	0.88	43

Fig 5.2 LOGISTIC REGRESSION

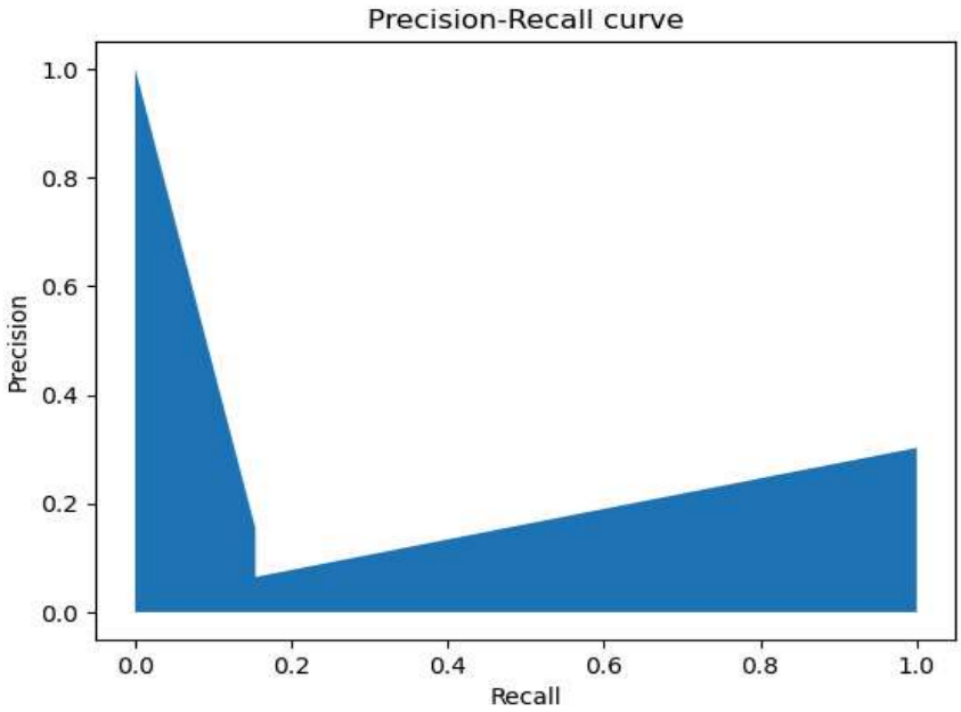

Fig 5.3 Precision-Recall Curve

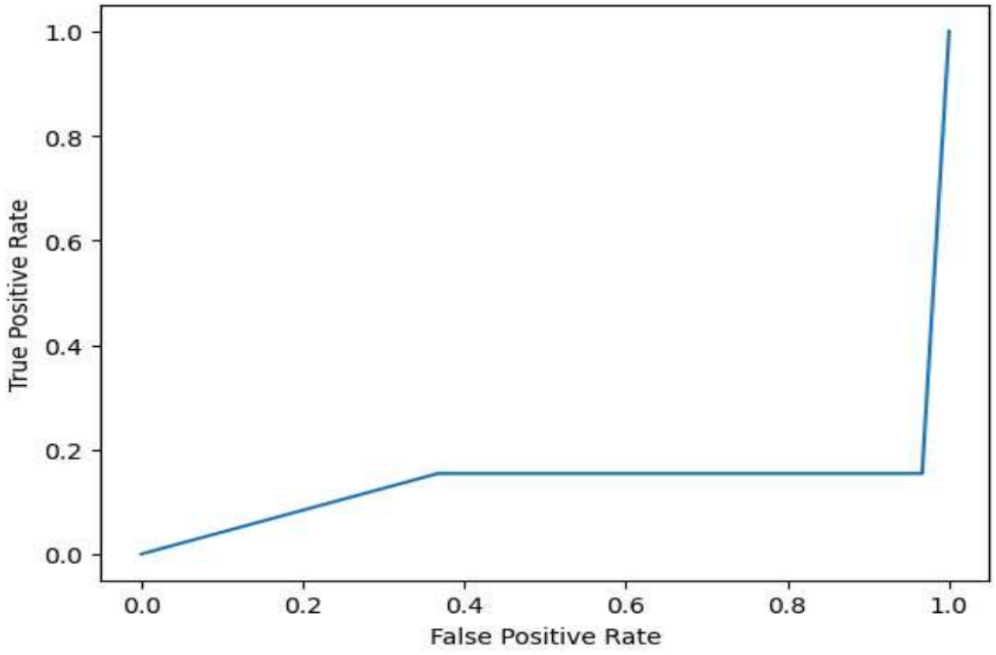

Fig 5.4 ROC curve

DECISION TREE

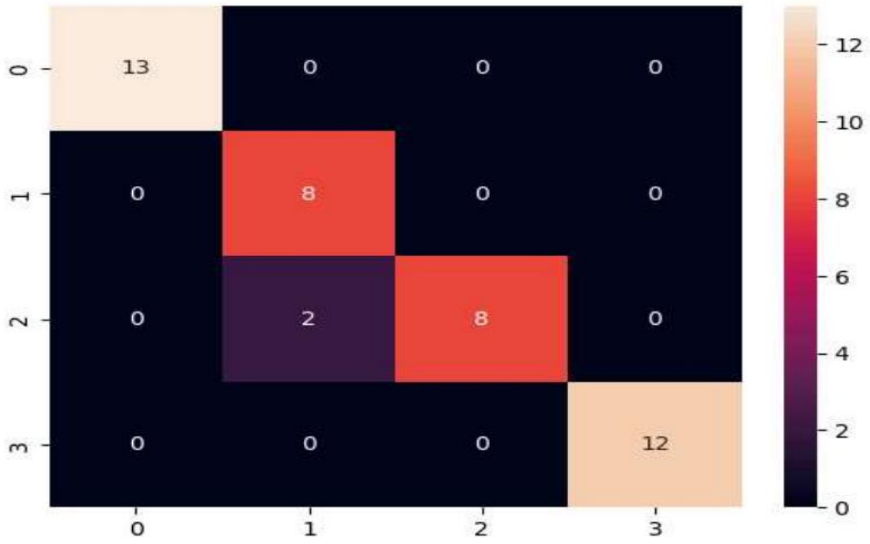

Fig 5.5 confusion matrix

	precision	recall	f1-score	support
0	1.00	1.00	1.00	13
1	0.80	1.00	0.89	8
2	1.00	0.80	0.89	10
3	1.00	1.00	1.00	12
accuracy			0.95	43
macro avg	0.95	0.95	0.94	43
weighted avg	0.96	0.95	0.95	43

Fig 5.6 DECISION TREE

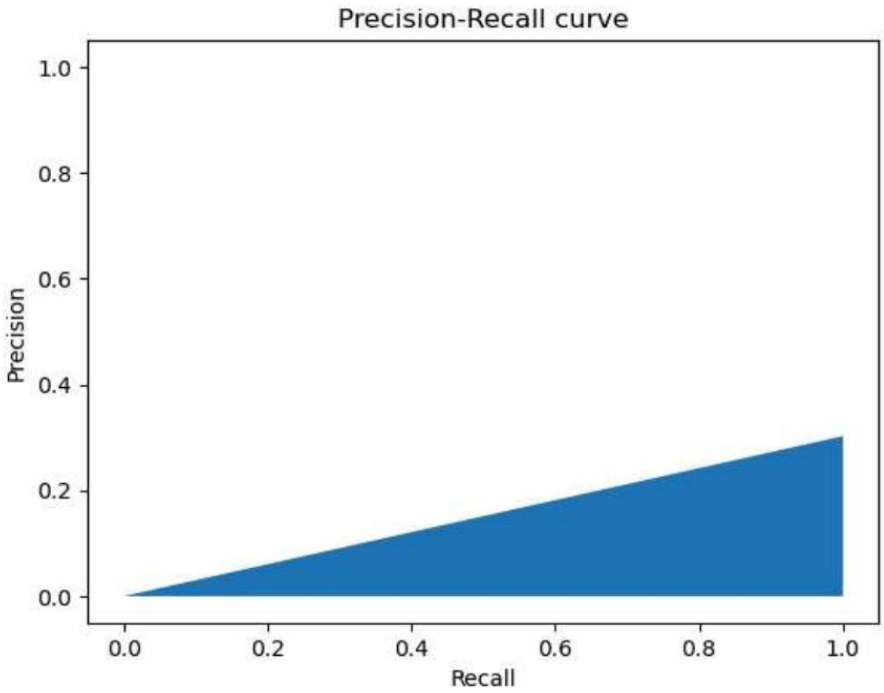

Fig 5.7 Precision-Recall Curve

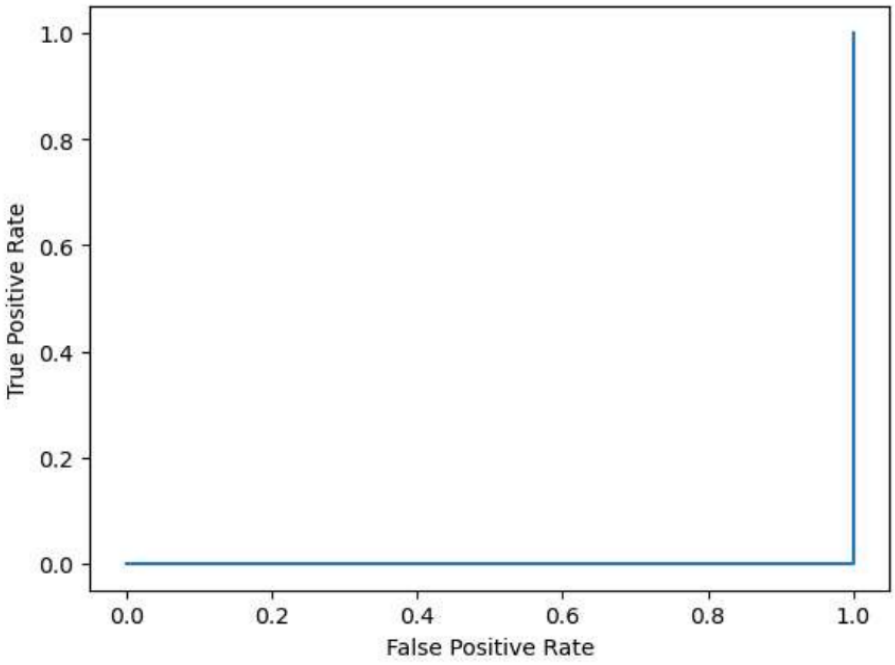

Fig 5.8 ROC curve

SUPPORT VECTOR MACHINE (SVM)

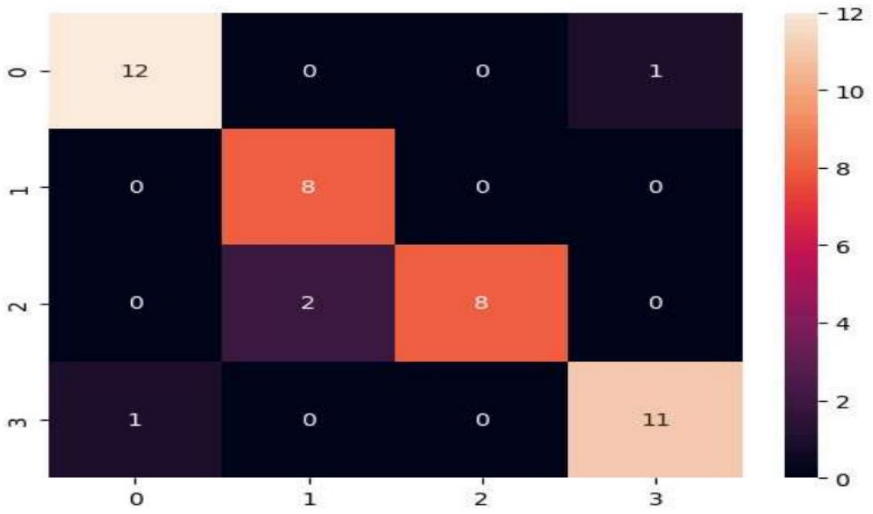

Fig 5.9 confusion matrix

	precision	recall	f1-score	support
0	0.92	0.92	0.92	13
1	0.80	1.00	0.89	8
2	1.00	0.80	0.89	10
3	0.92	0.92	0.92	12
accuracy			0.91	43
macro avg	0.91	0.91	0.90	43
weighted avg	0.92	0.91	0.91	43

Fig 5.10 SVM

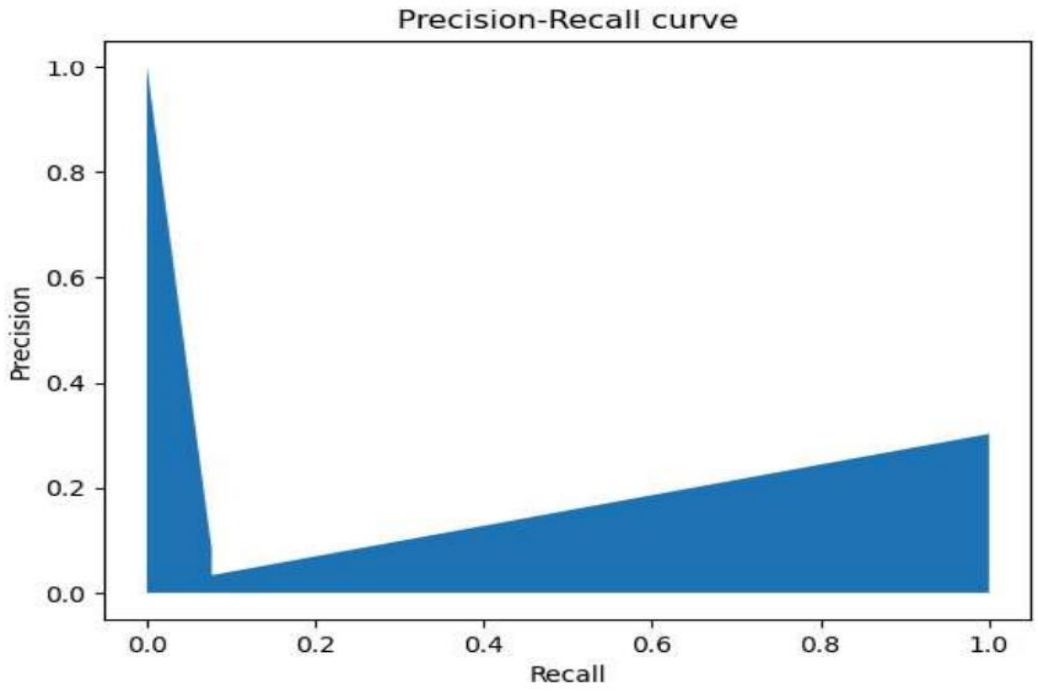

Fig 5.11 Precision-Recall curve

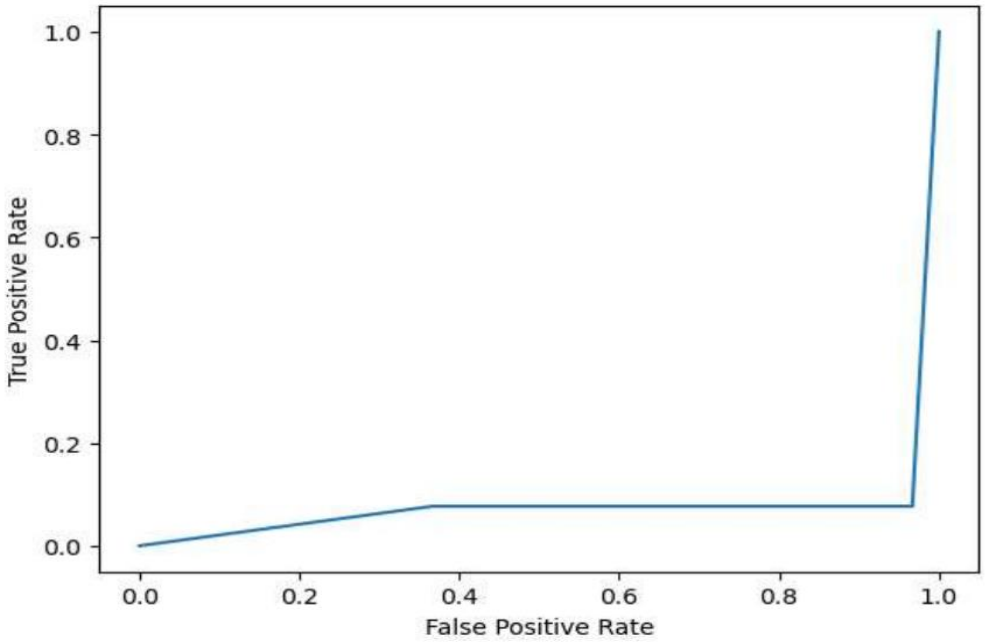

Fig 5.12 ROC curve

K-NEAREST NEIGHBOR (KNN)

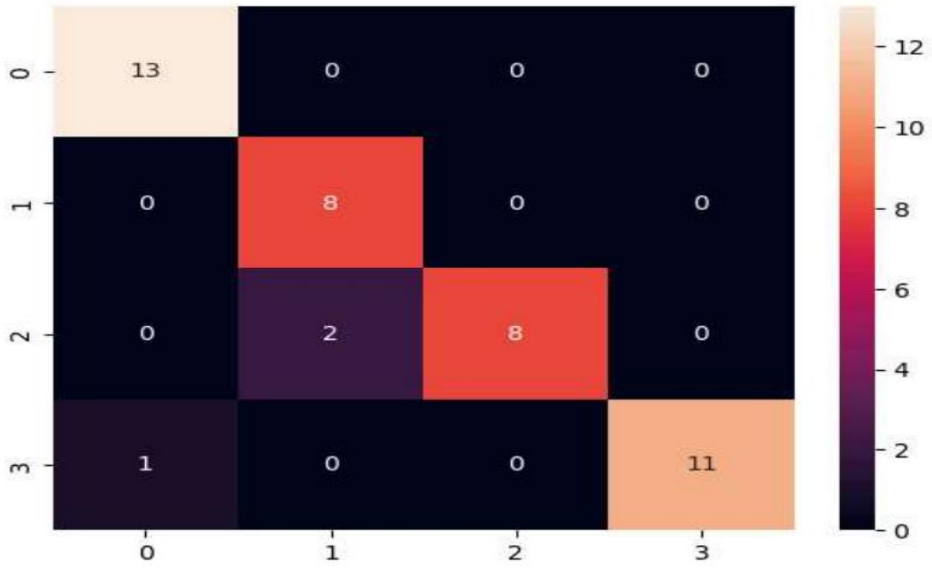

Fig 5.13 confusion matrix

	precision	recall	f1-score	support
0	0.93	1.00	0.96	13
1	0.80	1.00	0.89	8
2	1.00	0.80	0.89	10
3	1.00	0.92	0.96	12
accuracy			0.93	43
macro avg	0.93	0.93	0.92	43
weighted avg	0.94	0.93	0.93	43

Fig 5.14 KNN

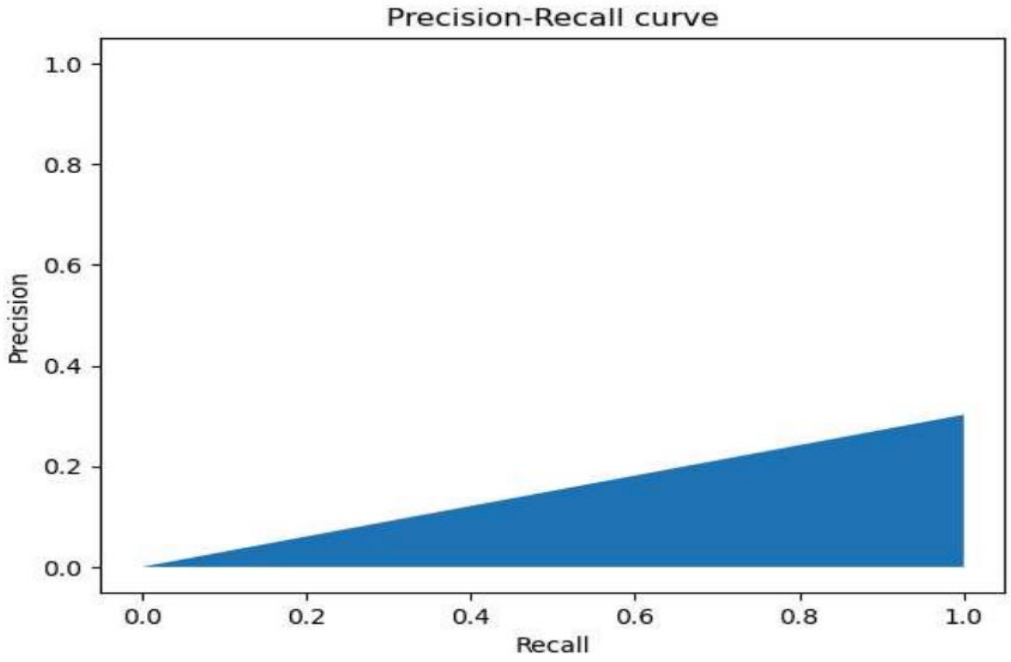

Fig 5.15 Precision-Recall curve

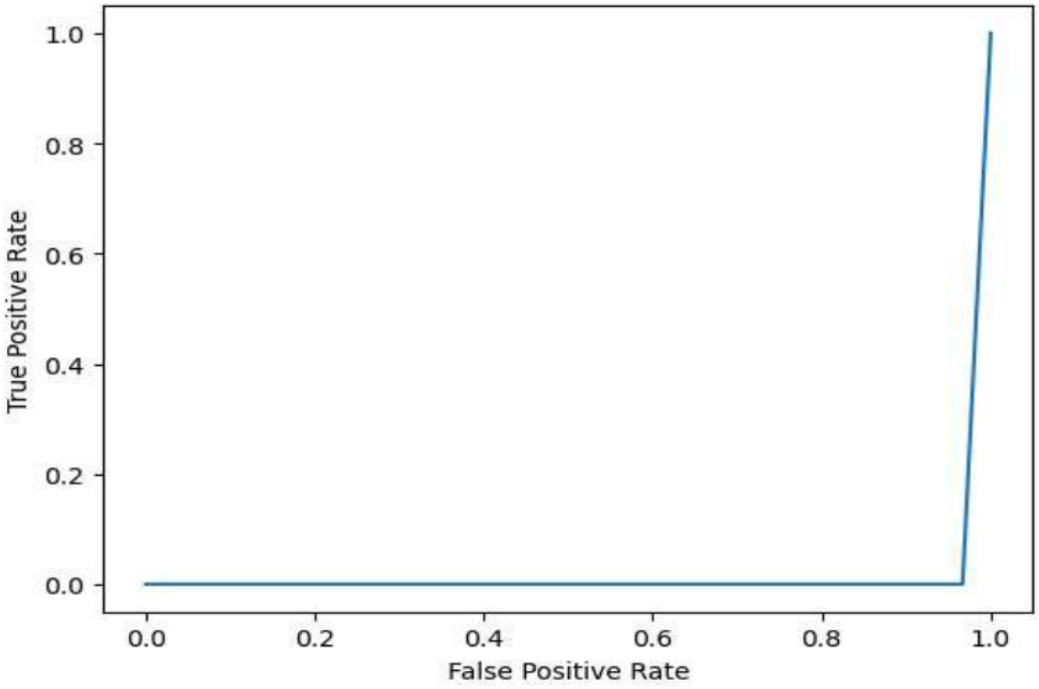

Fig 5.16 confusion matrix

CHAPTER -6

CONCLUSION AND SCOPE FOR FUTURE WORK

6.1. CONCLUSION

In conclusion, the "AI Based Automated System for Coronary Artery Disease Prediction" project addresses a critical need by providing predictive insights into CAD risk levels, particularly for individuals in rural and remote areas who may lack access to timely healthcare services.

By leveraging machine learning models such as Logistic Regression, Decision Tree, SVM, and KNN, implemented through an intuitive in-program application, the system empowers users to assess their CAD risk levels proactively.

Through the incorporation of evaluation metrics like confusion matrix, F1-score, precision, recall, and accuracy, the system ensures robust performance and reliability in predicting CAD risk categories ranging from "no risk" to "severe risk."

Utilizing a dataset encompassing essential health parameters and employing information extraction techniques, the system demonstrates its capability to process fresh data alongside existing datasets effectively. With only eight features utilized for training and testing within the specified machine learning algorithms, the system remains efficient and accessible.

Ultimately, the culmination of this project represents a significant step towards proactive health management, contributing to the overall well-being of individuals by enabling early detection and intervention in cases of coronary artery disease.

6.2. FUTURE WORK

There are multiple opportunities to improve the "AI Based Automated System for Coronary Artery Disease Prediction" project in the future. First, to increase prediction accuracy and robustness by experimenting with ensemble techniques or adding more machine learning algorithms.

Furthermore, broadening the dataset to incorporate a wider range of comprehensive health characteristics could enhance the system's forecasting skills. Furthermore, the system's reliability would be strengthened by performing longitudinal studies to validate its predictions over time and in various demographic groups.

Additionally, wearable technology and Internet of Things sensors could be used to integrate real-time data collecting capabilities and provide continuous monitoring of people's health state, allowing for early diagnosis and intervention in cases of CAD risk. Improving the models' interpretability and offering justifications for forecasts may help boost user confidence and system acceptability.

Last but not least, performing extensive deployment studies in rural and isolated areas and working with healthcare experts to integrate the system into clinical practice will prove its efficacy in real-world circumstances and support proactive CAD management on a larger scale.

APPENDIX

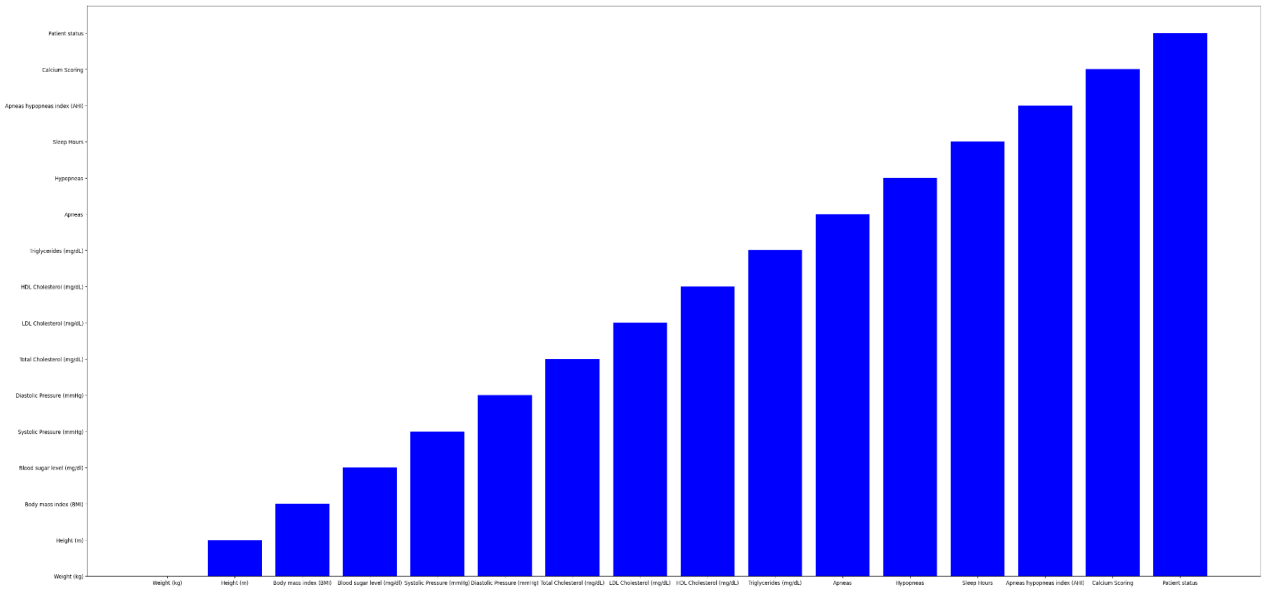

Fig 6.1

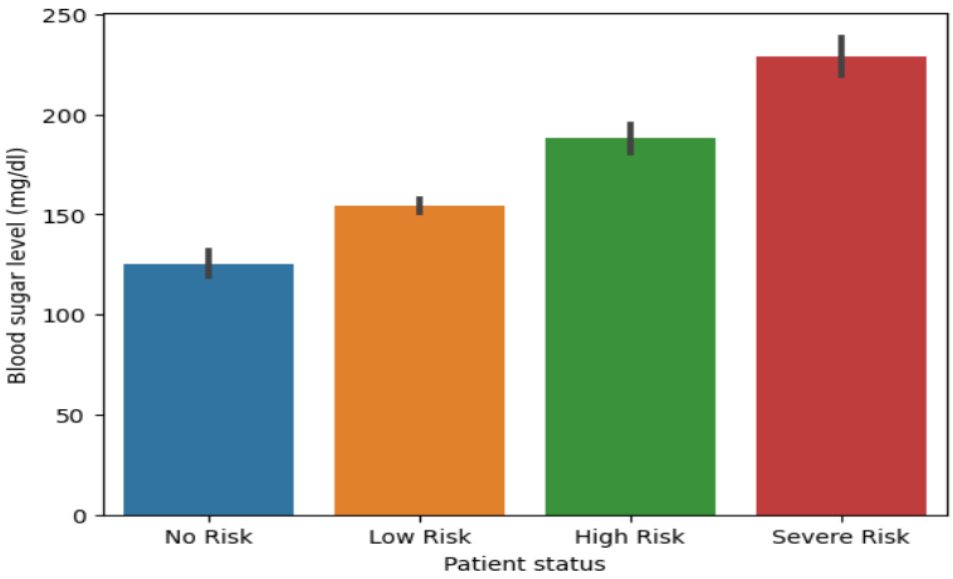

Fig 6.2 Patient status (through blood sugar level)

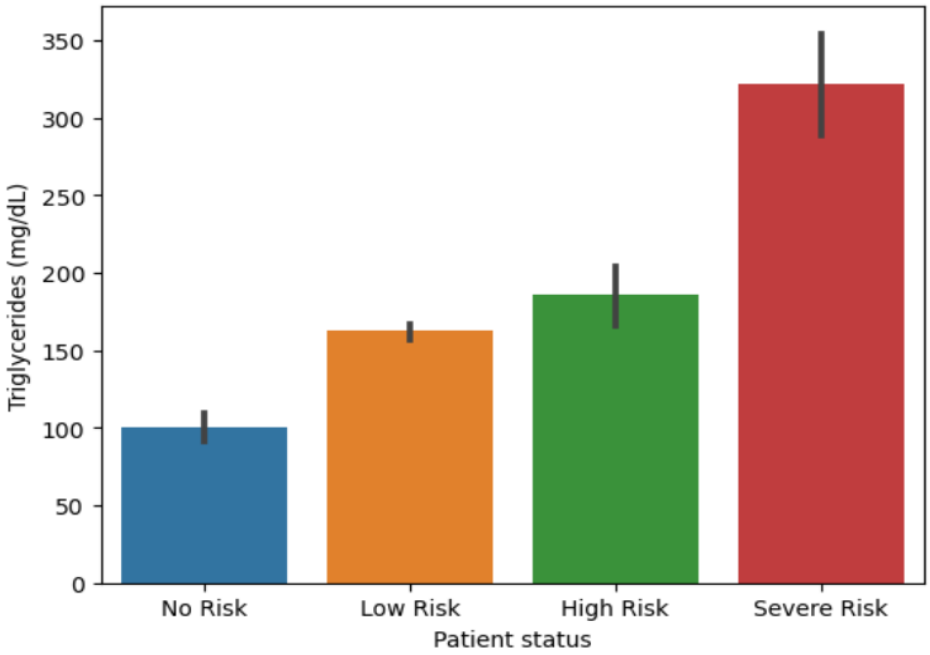

Fig 6.3 Patient status (through triglycerides)

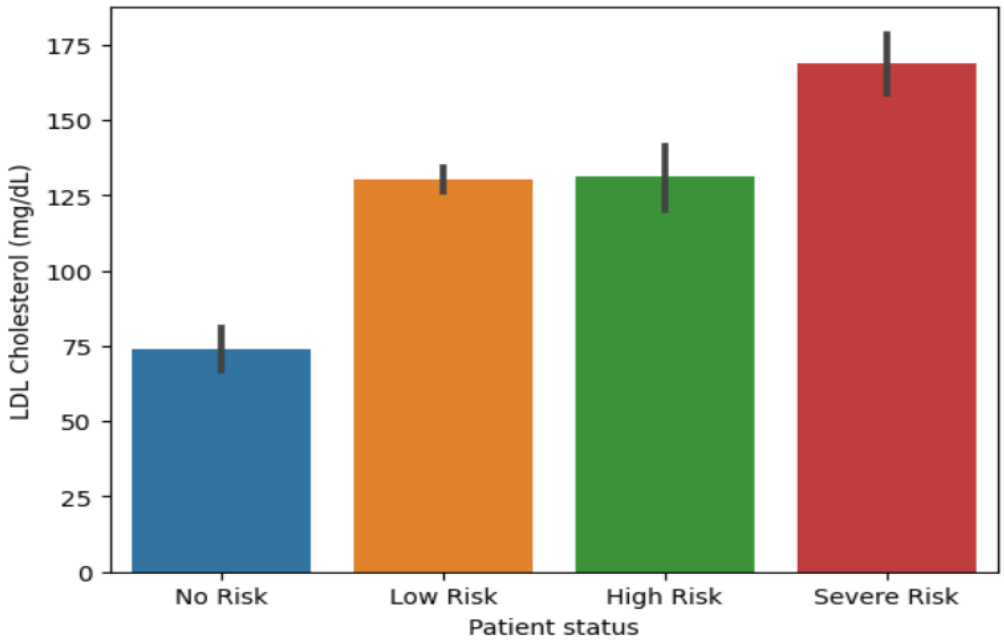

Fig 6.4 Patient status (through LDL)_

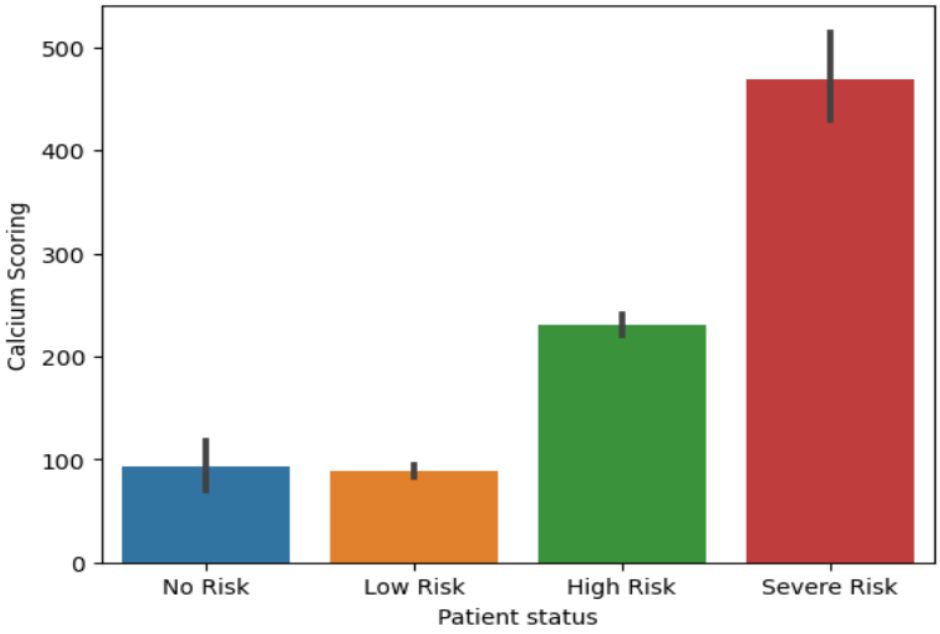

Fig 6.5 Patient status (through calcium scoring)

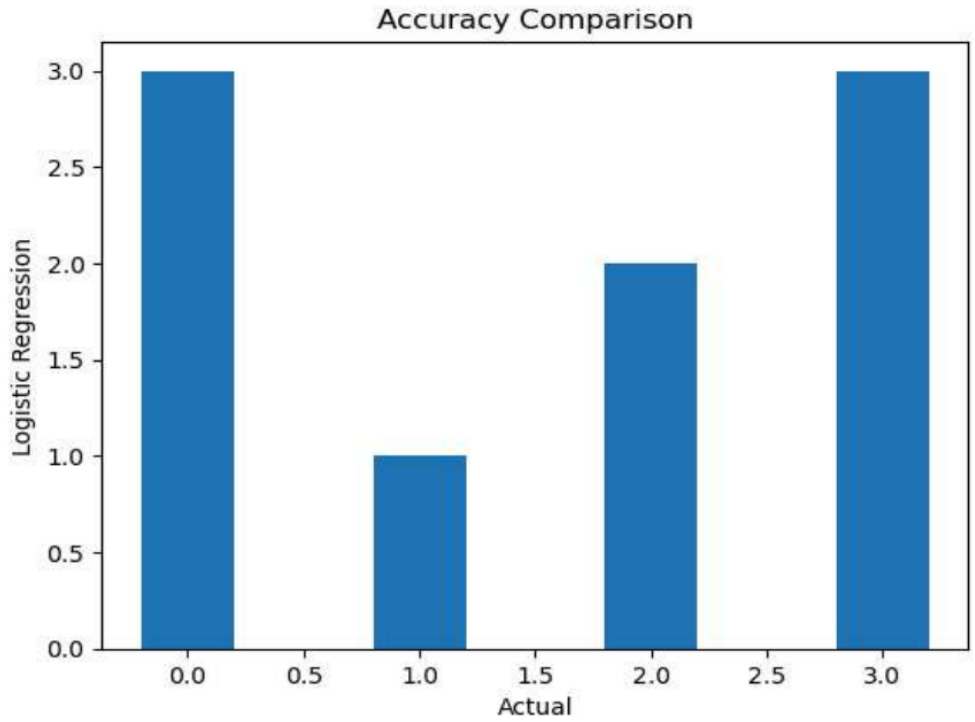

Fig 6.6 Logistic regression Accuracy

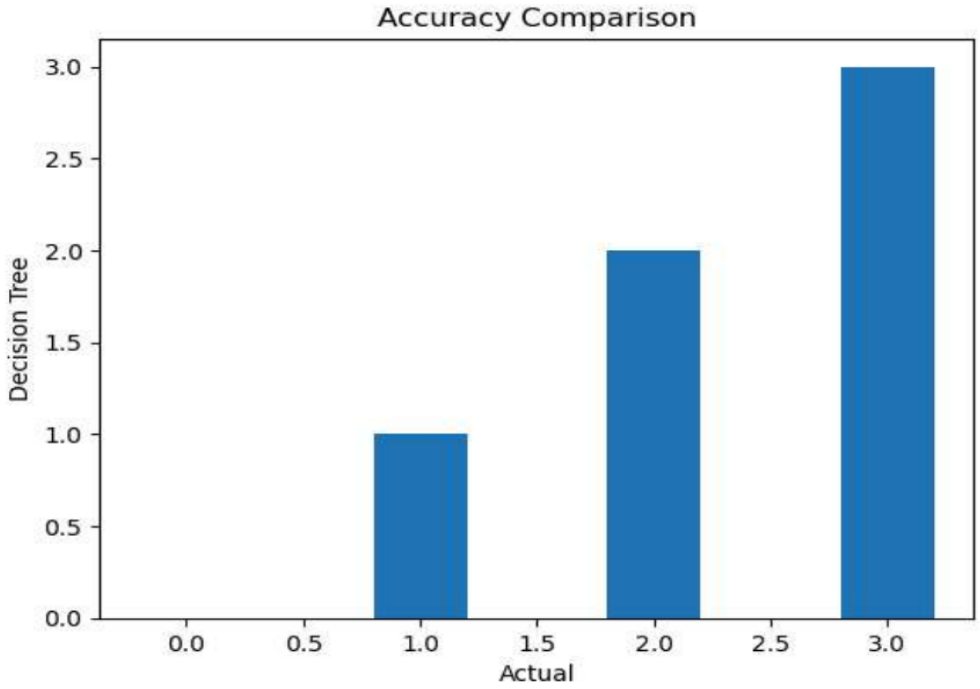

Fig 6.7 Decision tree Accuracy

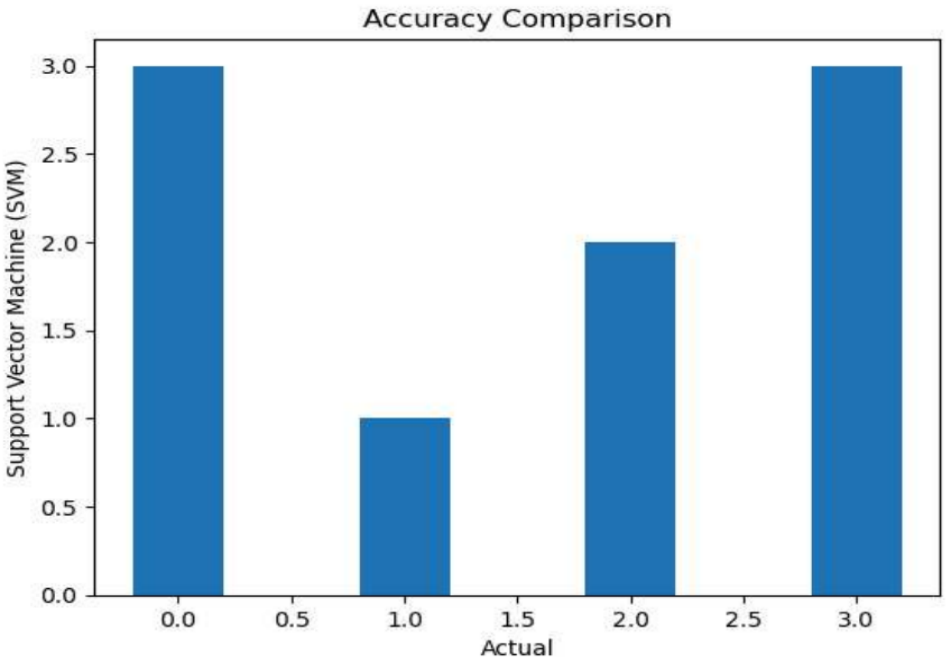

Fig 6.8 SVM Accuracy

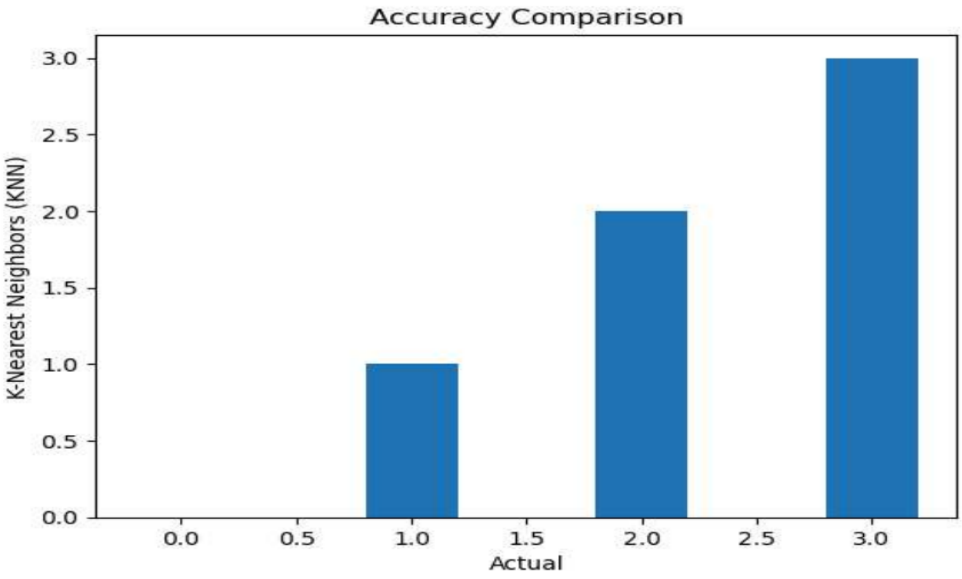

Fig 6.9 KNN Accuracy

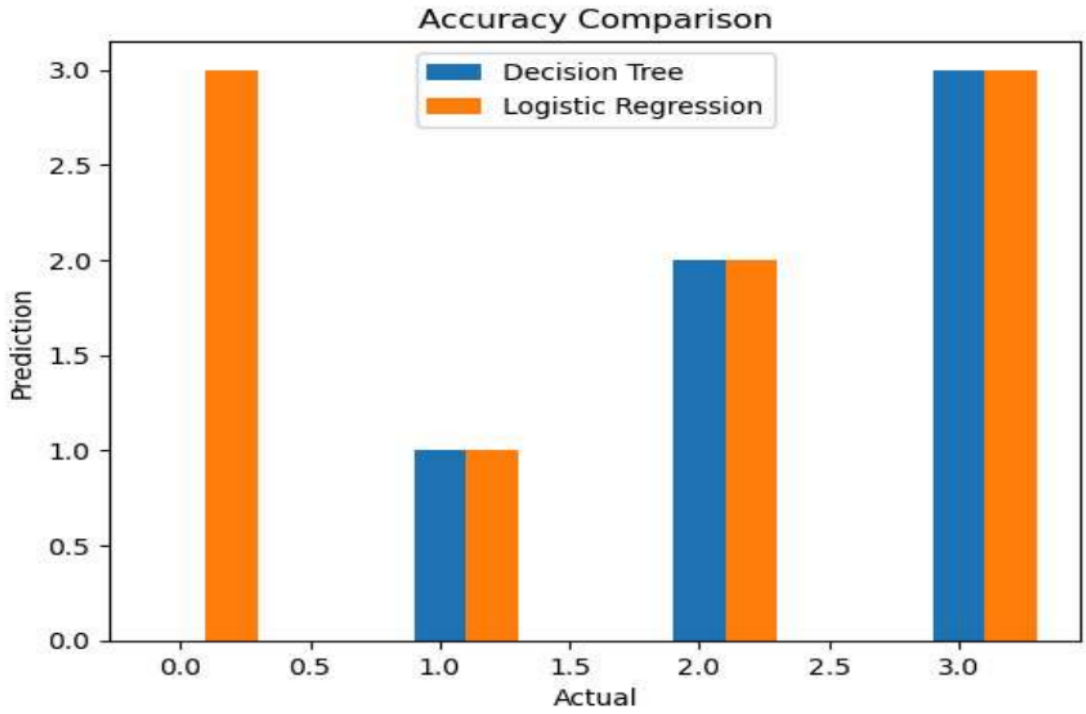

Fig 6.10 (DT, LR) Accuracy comparison

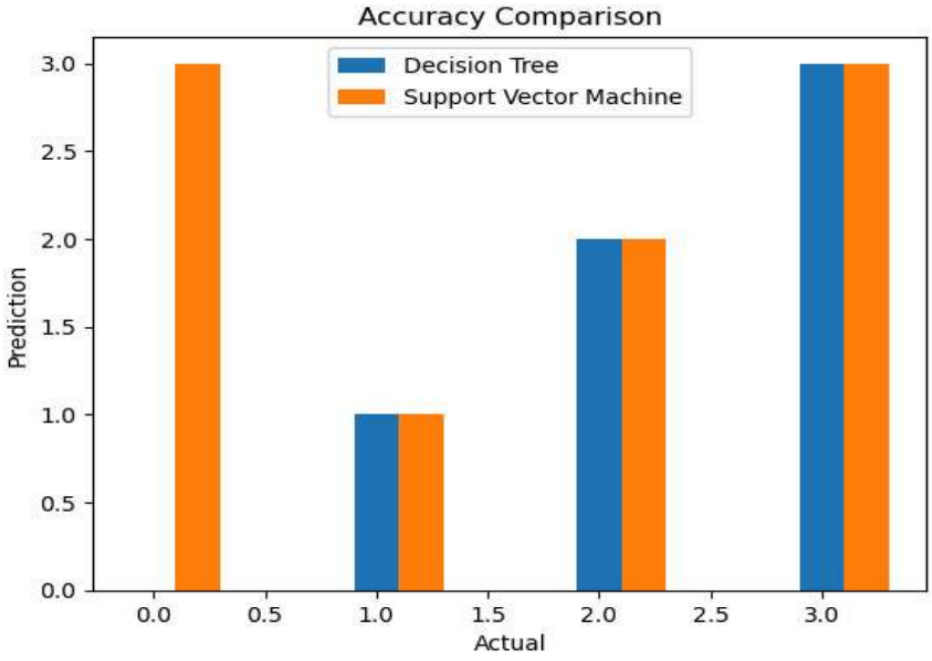

Fig 6.11 (DT, SVM) Accuracy comparison

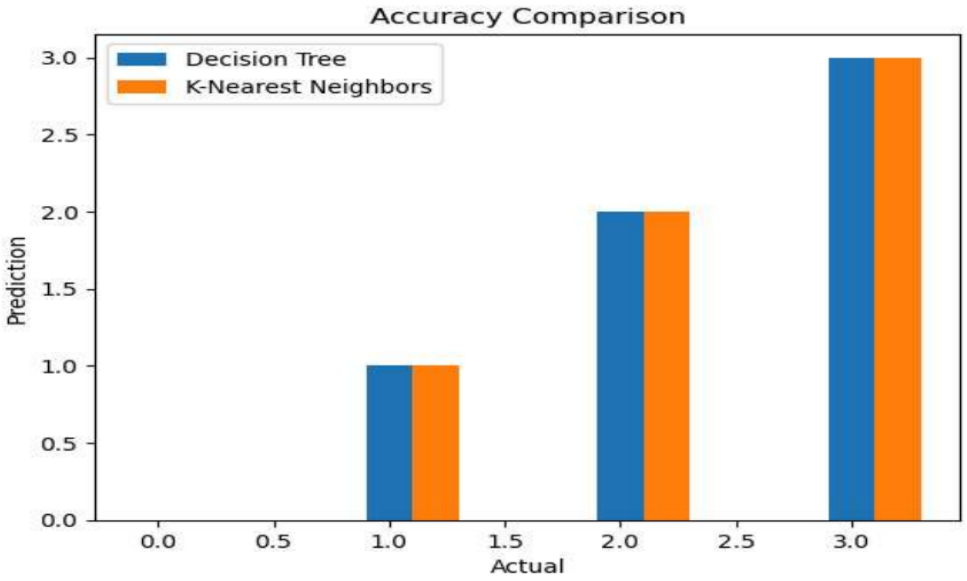

Fig 6.12 (DT, KNN) Accuracy comparison

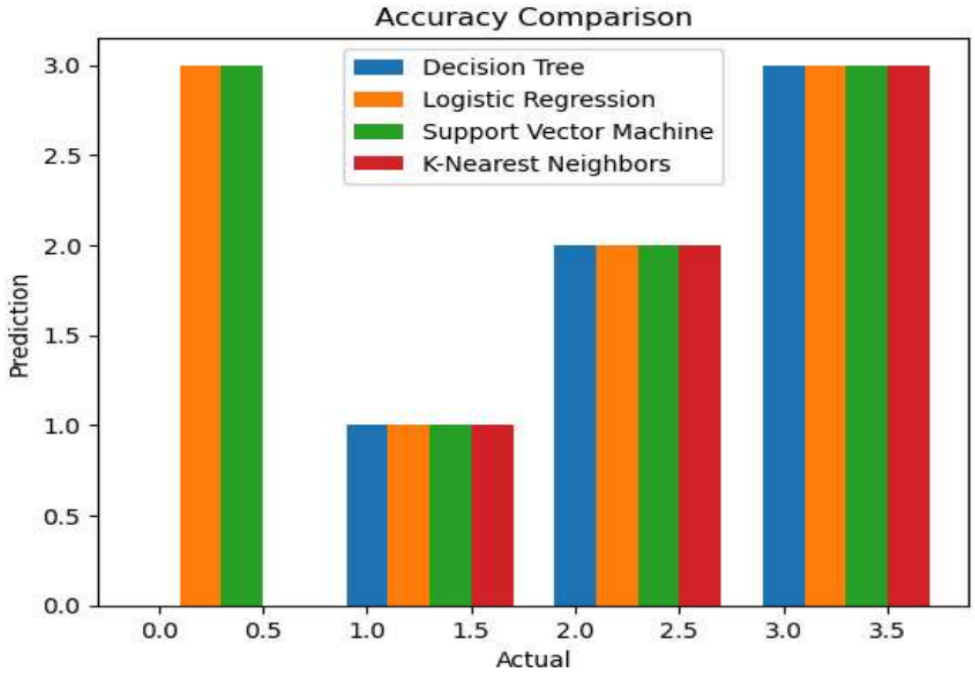

Fig 6.13 (decision tree, LR, SVM, KNN) Accuracy comparison

ANNEXURE-I -SAMPLE CODE

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy as sc
import sklearn as sk
import seaborn as sns
import pickle
import warnings
#reading the dataset
data = pd.read_csv('heart_attack_dataset.csv')
warnings.filterwarnings("ignore")
data.head(4)
data
data.describe()
#data visualization
df = pd.DataFrame(data)
fig = plt.figure()
ax = fig.add_axes([0,0,5,3])
m = list(df.iloc[:0])
n = list(df.iloc[:1])
ax.bar(m,n,color='b')
plt.show()
sns.barplot(x="Patient status",y="Blood sugar level (mg/dl)",data=data)
plt.show()
```

```
sns.barplot(x="Patient status",y="Triglycerides (mg/dL)",data=data)

plt.show()

sns.barplot(x="Patient status",y="LDL Cholesterol (mg/dL)",data=data)

plt.show()

sns.barplot(x="Patient status",y="Calcium Scoring",data=data)

plt.show()

#describing features and target

data.columns

#features

x = data[['Body mass index (BMI)', 'Blood sugar level (mg/dl)', 'Systolic Pressure
(mmHg)', 'Diastolic Pressure (mmHg)', 'LDL Cholesterol (mg/dL)', 'Triglycerides
(mg/dL)', 'Apneas hypopneas index (AHI)', 'Calcium Scoring']]

#target

y = data['Patient status']

x.shape

y.shape

data["Patient status"].value_counts()

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

y = le.fit_transform(y)

#spiliting dataset the rest

from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, train_size = 0.9)

x_train.shape

x_test.shape

y_train.shape

y_test.shape
```

```
def binarize_dataframe(y, classes):
  # to convert multi class dataframe to binary dataframe
  # Define the class you want to consider as the positive class (class 0)
    positive_class = 0
# Convert multiclass target variable y into a binary variable
    y_binary = np.where(y == positive_class, 1, 0)
    return y_binary
#logistic regression
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression(max_iter = 10000)
lr.fit(x_train, y_train)
y_pred = lr.predict(x_test)
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report,
precision_recall_curve, roc_auc_score, roc_curve
accuracy_score(y_test,y_pred)
confusion_matrix = confusion_matrix(y_test,y_pred)
confusion_matrix
sns.heatmap(confusion_matrix, annot=True)
plt.savefig('h1.png')

#Higher values is represented by darker colors,
#while lower values is represented by lighter colors.
print(classification_report(y_test,y_pred))
precision, recall, thresholds =
precision_recall_curve(binarize_dataframe(y_test,[0,1,2,3]), y_pred)
plt.fill_between(recall, precision)
plt.ylabel("Precision")
```

```
plt.xlabel("Recall")
plt.title("Precision-Recall curve");
def plot_roc_curve(true_y, y_prob):
    fpr, tpr, thresholds = roc_curve(true_y, y_prob)
    plt.plot(fpr, tpr)
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
plot_roc_curve(binarize_dataframe(y_test,[0,1,2,3]), y_pred)
#decision tree
from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier()
dtc.fit(x_train, y_train)
y_dt_pred = dtc.predict(x_test)
accuracy_score(y_test,y_dt_pred)
from sklearn.metrics import confusion_matrix
confusion_matrix = confusion_matrix(y_test, y_dt_pred)
confusion_matrix
sns.heatmap(confusion_matrix, annot=True)
plt.savefig('h1.png')
#Higher values is represented by darker colors,
#while lower values is represented by lighter colors.
print(classification_report(y_test,y_dt_pred))
precision, recall, thresholds =
precision_recall_curve(binarize_dataframe(y_test,[0,1,2,3]), y_dt_pred)
plt.fill_between(recall, precision)
plt.ylabel("Precision")
```

```
plt.xlabel("Recall")
plt.title("Precision-Recall curve");
plot_roc_curve(binarize_dataframe(y_test,[0,1,2,3]), y_dt_pred)
from sklearn import svm
sv = svm.SVC(kernel='linear', C=1.0)
sv.fit(x_train, y_train)
y_sv_pred = sv.predict(x_test)
accuracy_score(y_test,y_sv_pred)
from sklearn.metrics import confusion_matrix
confusion_matrix = confusion_matrix(y_test,y_sv_pred)
confusion matrix
sns.heatmap(confusion_matrix, annot=True)
plt.savefig('h1.png')
print(classification_report(y_test,y_sv_pred))
precision, recall, thresholds =
precision_recall_curve(binarize_dataframe(y_test,[0,1,2,3]), y_sv_pred)
plt.fill_between(recall, precision)
plt.ylabel("Precision")
plt.xlabel("Recall")
plt.title("Precision-Recall curve");
plot_roc_curve(binarize_dataframe(y_test,[0,1,2,3]), y_sv_pred)
#knn
from sklearn.neighbors import KNeighborsClassifier
knc = KNeighborsClassifier()
knc.fit(x_train, y_train)
y_kn_pred = knc.predict(x_test)
```

REFERENCES

[1] Alaa Khaleel Faieq, Maad M. Mijwil Department of Computer Techniques Engineering, Baghdad College of Economic Sciences University, Baghdad, Iraq, "Prediction of heart diseases utilising support vector machine and artificial neural network," Indonesian Journal of Electrical Engineering and Computer Science Vol. 26, No. 1, April 2022, pp. 374~380 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v26.i1.pp374-380.

[2] Girish S. Bhavekar & Agam Das Goswami ,A hybrid model for heart disease prediction using recurrent neural network and long short term memory, International Journal of Information Technology ,Published: 21 February 2022, Volume 14, pages 1781–1789, (2022).

[3] M. Akhil Jabbar; B. L Deekshatulu; Priti Chandra, Heart disease prediction using lazy associative classification, 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), DOI: 10.1109/iMac4s.2013.6526381, 10 June 2013.

[4] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz**,** Gaurav Dhiman, Sagar Pande, and Parneet singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," Computational Intelligence and Neuroscience/ 2021.

[5] Manoj Diwakar, Amrendra Tripathi, Kapil Joshi, Minakshi Memoria, Prabhishek Singh, Neeraj kumar ,"Latest trends on heart disease prediction using machine learning and image fusion," https://doi.org/10.1016/j.matpr.2020.09.078.

[6] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath, "Heart disease prediction using machine learning algorithms," doi:10.1088/1757-899X/1022/1/012072, IOP Conf. Series: Materials Science and Engineering.

[7] Melad Mizher Rahmah, Aymen Dawood Salman, "Heart Disease Classification– Based on the Best Machine Learning Model", September 2022 Iraqi Journal of Science, DOI:10.24996/ijs.2022.63.9.28

[8] Senthilkumar Mohan; Chandrasegar Thirumalai; Gautam Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”, Published by IEEE, 19 June 2019.

[9] Galla Siva, Sai Bindhika, Munaga Meghana, Manchuri Sathvika Reddy, “Heart Disease Prediction Using Machine Learning Techniques”, International Research Journal of Engineering and Technology (IRJET), October 2020.

[10] Miss. Kalyani S. Ubale, Dr. P. N. Kalavadekar, “EFFECTIVE HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES: A SURVEY” ,ISO 3297:2007 Certified, Volume 6 ,Issue 3 ,March 2021

[11] Marimuthu Muthuvel, Coimbatore Institute of Technology, M Abinaya, K S Hariesh, K Madhankumar, “A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach” , International Journal of Computer Applications 181(18):975-8887, DOI:10.5120/ijca2018917863, September 2018.

[12] S. P. Rajamhoana, C. Akalya Devi, K. Umamaheswari, R. Kiruba, K. Karunya, R. Deepika. Analysis of Neural Networks Based Heart Disease Prediction System. In Adam Bujnowski, Mariusz Kaczmarek, Jacek Ruminski, editors, 11th International Conference on Human System Interaction, HSI 2018, Gdansk, Poland, July 4-6, 2018. pages 233-239, IEEE, 2018. [doi]

[13] Kumar G Dinesh; K Arumugaraj; Kumar D Santhosh; V Mareeswari, “Prediction of Cardiovascular Disease Using Machine Learning Algorithms”, 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), DOI: 10.1109/ICCTCT42380.2018, 1-3 March 2018.

[14] Li Yang, Haibin Wu, Xiaoqing Jin, Pinpin Zheng, Shiyun Hu, Xiaoling Xu, Wei Yu & Jing Yan, “Study of cardiovascular disease prediction model based on random forest in eastern China”, Article number: 5245 (2020), Published: 23 March 2020.

[15] Michael Inouye, Gad Abraham, Christopher P Nelson, Angela M Wood, Michael J Sweeting, Frank Dudbridge, Florence Y Lai, Stephen Kaptoge, Marta Brozynska, Tingting Wang, Shu Ye, Thomas R Webb, Martin K Rutter, Ioanna Tzoulaki, Riyaz S Patel, Ruth J F Loos, Bernard Keavney, Harry Hemingway, John Thompson, Hugh Watkins, Panos Deloukas, Emanuele Di Angelantonio, Adam S Butterworth, John Danesh, Nilesh J Samani," Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention", PMID: 30309464 PMCID: PMC6176870 DOI: 10.1016/j.jacc.2018.07.079, 2018 Oct 16

[16] H. Benjamin Fredrick David and S. Antony Belcy, "heart disease prediction using data mining techniques", ictact journal on soft computing, october 2018, volume: 09, issue: 01, issn: 2229-6956 (online) ,doi: 10.21917/ijsc.2