A MINI PROJECT REPORT

On

ENSEMBLE MODEL FOR EXPLORATORY DATA ANALYSIS AND PREDICTION OF CARDIOMYOPATHY

Submitted to

OSMANIA UNIVERSITY

In partial fulfillment of the requirements for the award of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING (AI & ML)

BY

SHAIK ALTAAF	**245521748119**
SADDANAPU RAHUL	**245521748115**
ALUGALA SAI TEJA	**245521748065**

Under the esteemed guidance of

Dr.R.MADHAVI

Assistant Professor, Dept. of CSE (AI & ML)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (AI & ML)

KESHAV MEMORIAL ENGINEERING COLLEGE

(Approved by AICTE, New Delhi & Affiliated to Osmania University, Hyderabad)

D.No. 10 TC-111, Kachavanisingaram (V), Ghatkesar (M), Medchal-Malkajgiri, Telangana – 500088

(2023-2024)

KESHAV MEMORIAL ENGINEERING COLLEGE

Department of Computer Science and Engineering (AI & ML)

CERTIFICATE

This is to certify that the project report entitled **ENSEMBLE MODEL FOR EXPLORATORY DATA ANALYSIS AND PREDICTION OF CARDIOMYOPATHY** that is being submitted by **SHAIK ALTAAF (245521748119), SADDANAPU RAHUL (245521748115), ALUGALA SAI TEJA (245521748115),** under the guidance of **Dr.R.Madhavi** with fulfillment for the award of the Degree of **Bachelor of Engineering in Computer Science and Engineering (AI & ML)** to the **Osmania University** is a record of bonafide work carried out by his under my guidance and supervision. The results embodied in this project report have not been submitted to any other University or Institute for the award of any graduation degree.

Dr.R.MADHAVI
Assistant Professor,
Internal Guide,
CSE (AI & ML) Dept.

Dr. B. Devender
Assosiate Professor,
Head of the Department,
CSE (AI & ML) Dept.

EXTERNAL EXAMINER

Submitted for Viva Voce Examination held on _____

Vision & Mission of KMEC

Vision of KMEC:

To be a leader in producing industry-ready and globally competent engineers to make India a world Leaders in software products and services.

Mission of KMEC:

1. To provide a conducive learning environment that includes problem solving skills, professional and ethical standards, lifelong learning through multimodal platforms and prepare students to become successful professionals.
2. To forge industry-institute partnerships to expose students to the technology trends, work culture and ethics in the industry.
3. To provide quality training to the students in the state-of-art software technologies and tools.
4. To encourage research-based projects/activities in emerging areas of technology.
5. To nurture entrepreneurial spirit among the students and faculty.
6. To induce the spirit of patriotism among the students that enables them to understand India's challenges and strive to develop effective solutions.

Vision & Mission of CSE (AI & ML)

Vision of the CSE (AI & ML):

To be a global center of excellence in Artificial Intelligence and Machine Learning, producing socially responsible graduates, excelling in education, research, and innovation for transformative societal impact.

Mission of the CSE (AI & ML):

1. To cultivate global expertism in Artificial Intelligence and Machine Learning for lifelong impact.
2. To lead in ethical innovation, training experts in cutting-edge Artificial Intelligence technologies.
3. To provide top-tier education in Artificial intelligence, fostering innovation and ethics.
4. To establish Centers of Excellence in Artificial Intelligence and Machine Learning, emphasizing Research and Development collaboration, professional development, and community engagement.
5. To create an open growth environment, producing Industry-ready graduates, and partnering globally in technical education and research.
6. To be a global Center of Excellence for Artificial intelligence, promoting industry collaboration and instilling self-learning, team work, and professional ethics.

PROGRAM OUTCOMES (POs)

1. **Engineering Knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis:** Identify formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences
3. **Design/development of solutions:** Design solutions for complex engineering problem and design system component or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural societal, and environmental considerations.
4. **Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
5. **Modern tool usage:** Create select, and, apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society:** Apply reasoning informed by the contextual knowledge to societal, health, safety. legal und cultural issues and the consequent responsibilities relevant to professional engineering practice.
7. **Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams and in multidisciplinary settings.
10. **Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation make effective presentations, and give and receive clear instructions.
11. **Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. **Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

PEO-1: Graduates can apply foundational computer science knowledge adeptly to solve real-world challenges in professional roles or advanced academic pursuits.

PEO-2: Graduates can utilize a comprehensive understanding of computer science and related engineering disciplines for success in diverse careers through effective collaboration and trade-off navigation.

PEO-3: Graduates can adapt quickly to evolving technological landscapes, applying ethical considerations and actively engaging in continuous learning and professional development.

PEO-4: Graduates can demonstrate exceptional communication, collaboration, and professionalism, leveraging specialized knowledge to contribute significantly to specific disciplines and foster societal progress.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1: Apply Artificial Intelligence and Machine Learning knowledge to design automation solutions for real-world challenges in software development.

PSO2: Demonstrate expertise in algorithmic design and contribute to the development of optimized solutions in Artificial Intelligence, Machine Learning and Emerging technologies.

PSO3: Utilize Artificial Intelligence and Machine Learning principles to design intelligent subsystems, address real-world business problems, and adapt to the dynamic Artificial Intelligence landscape with ethical values.

PROJECT OUTCOMES

P1: To create a model to detect cardiomyopathy using Random Forest algorithm.

P2: To predict the genetic basis of cardiomyopathy using Decision Tree.

P3: Apply project management skills i.e. scheduling work, procuring parts and documenting expenditures and working within the confines of a deadline.

P4: Work with team mates, sharing due and fair credits and collectively apply effort for making project successful.

P5: Communicate technical information by means of written and oral reports.

1 – LOW

2 - MEDIUM

3 - HIGH

PROJECT OUTCOMES MAPPING PROGRAM OUTCOMES:

PO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
P1	2	2	2	2	2	3	2	2	2	3	3	3
P2	2	2	2	2	2	2	2	2	2	2	3	3
P3	2	3	2	3	3	3	3	3	2	2	3	2
P4	2	3	3	3	3	2	3	2	3	3	3	2
P5	3	2	2	2	2	2	2	3	3	3	3	3

PROJECT OUTCOMES MAPPING PROGRAM SPECIFIC OUTCOMES:

PSO	PO1	PO2	PO3
P1	2	2	3
P2	2	2	3
P3	3	3	3
P4	3	3	3
P5	3	3	3

PROJECT OUTCOMES MAPPING PROGRAM EDUCATIONAL OUTCOMES:

PEO	PEO1	PEO2	PEO3	PEO4
P1	3	2	3	2
P2	3	2	3	3
P3	3	3	3	3
P4	3	3	3	3
P5	3	3	3	3

DECLARATION

This is to certify that the mini project titled **ENSEMBLE MODEL FOR EXPLORATORY DATA ANALYSIS AND PREDICTION OF CARDIOMYOPATHY** is a bonafide work done by us in fulfillment of the requirements for the award of the degree **Bachelor of Engineering** in Department of Computer Science and Engineering (AI & ML), and submitted to the **Department of CSE (AI & ML)**, **Keshav Memorial Engineering College, Hyderabad.**

We also declare that this project is a result of our own effort and has not been copied or intimated from any source. Citations from any websites are mentioned in the bibliography. This work was not submitted earlier at any other university for the award of any degree.

SHAIK ALTAAF (245521748119)
SADDANAPU RAHUL (245521748115)
ALUGALA SAI TEJA (245521748115)

ACKNOWLEDGEMENT

This is to place on our record my appreciation and deep gratitude to the persons without whose support this project would never been this successful.

We are grateful to **Mr. Neil Gogte,** Founder Director, for facilitating all the amenities required for carrying out this project.

It is with immense please that we would like to express our indebted gratitude to the respected **Prof. P.V.N Prasad, Principal, Keshav Memorial Engineering College,** for providing a great support and for giving us the opportunity of doing the project.

We express our sincere gratitude to **Mrs. Deepa Ganu,** Director Academics, for providing an excellent environment in the college.

We would like to take this opportunity to specially thank to **Dr. Birru Devender, Professor & HoD, Department of CSE (AI & ML), Keshav Memorial Engineering College,** for inspiring us all the way and for arranging all the facilities and resources needed for our project.

We would like to take this opportunity to thank our internal guide **Dr.R.Madhavi, Assistant Professor, Department of CSE (AI & ML), Keshav Memorial Engineering College,** who has guided us a lot and encouraged us in every step of the project work. Her moral support and guidance throughout the project helped us to a greater extent.

We would like to take this opportunity to specially thank our Project Coordinator**, Mr.P.Naresh Kumar, Asst Professor, Department of CSE (AI & ML), Keshav Memorial Engineering College,** who guided us in successful completion of our project.

Finally, we express our sincere gratitude to all the members of the faculty of Department of CSE (AI & ML), our friends and our families who contributed their valuable advice and helped us to complete the project successfully.

SHAIK ALTAAF (245521748119)
SADDANAPU RAHUL (245521748115)
ALUGALA SAI TEJA (245521748115)

CONTENTS

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

One American dies every 36 seconds due to CVD. More than.665 million people die due to heart disease which 1 in every 4 deaths. Cardiovascular disease costs a lot to the US healthcare system. In the years 2014 and 2015, it cost about $219 billion per year in terms of healthcare services, medicine, and lost productivity due to death. Early diagnosis can also help to prevent heart failure which can lead to the death of a person. Angiography is considered as the most precise and accurate method for the prediction of cardiac artery disease (CAD), but it is very costly which makes it less accessible to low-income families.

It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Heart disease is a collection of diseases impacting the heart and veins of human beings. Cardiac disease symptoms vary depending on the specific type of cardiac disease. Detecting and diagnosing the cardiovascular disease is an on-going job that can be achieved with enough experience and knowledge by a qualified professional

There are many factors including age, diabetes, smoking, overweight, junk foods diet and so on. Several factors/parameters have been identified that cause heart disease or increase cardiac disease. Most hospitals have management software for monitoring their clinical and/or patient data. It is popular now and Such systems produce enormous amounts of patient information. These data are seldom used for clinical decision-making support. This factor led to research on the processing of medical data due to the lack of experts and the number of cases incorrectly diagnosed, a rapid and efficient automated detection system was required. The main purpose is to classify the key features of the medical data using the classifier model and use the models for the early prediction of cardiac disease. According to WHO, Heart Diseases are a leading cause of death worldwide. It is quite difficult to identify the cardiovascular disease (CVD) because of some contributory factors which contribute to CVD like high blood pressure, cholesterol level, diabetics, abnormal pulse rate, and many other factors.

1. 2 **RESEARCH MOTIVATION**

Cardiomyopathy is a complex and multifaceted medical condition characterized by structural and functional abnormalities in the heart muscle. Cardiomyopathy is a significant public health concern worldwide. It can lead to heart failure, arrhythmias, and other life-threatening complications. Understanding the underlying causes, mechanisms, and potential treatments for cardiomyopathy is crucial to improve the overall health and well-being of affected individuals and reduce the burden on healthcare systems.

Cardiomyopathy can result from various causes, including genetic mutations, infections, autoimmune diseases, metabolic disorders, and exposure to toxins or drugs. Researchers are motivated to study these different etiologies to develop targeted therapies and interventions based on the specific underlying factors. Genetic factors play a significant role in the development of certain types of cardiomyopathy, such as hypertrophic cardiomyopathy and dilated cardiomyopathy. Identifying the genetic mutations responsible for these conditions can lead to better risk assessment, early diagnosis, and potentially gene-based therapies. Advances in medical imaging, genetics, and molecular biology have improved our ability to diagnose and monitor cardiomyopathy. Researchers are motivated to explore the potential of these technologies to enhance early detection and personalized treatment strategies.

Improving the quality of life for individuals living with cardiomyopathy is a primary goal. Research in this area seeks to address not only the physical aspects of the disease but also its psychosocial and emotional impacts on patients and their families. Researchers are motivated to identify effective prevention strategies. Cardiomyopathy affects people of all ages and backgrounds globally. The advocacy efforts of patient and caregiver organizations play a significant role in motivating research into cardiomyopathy.

1.3 PROBLEM STATEMENT

Cardiomyopathy, a heterogeneous group of heart muscle diseases, presents a significant challenge in contemporary healthcare due to its diverse etiology, complex pathophysiology, and variable clinical outcomes. Despite advances in diagnostic methods and therapeutic interventions, there is an urgent need for further research to address the following critical issues:

- Etiological Heterogeneity: Cardiomyopathy encompasses a wide range of etiological factors, including genetic mutations, viral infections, toxic exposures, and autoimmune reactions. Understanding the specific mechanisms underlying different subtypes of cardiomyopathy is essential for precise diagnosis and targeted treatment strategies.
- Early Detection and Risk Stratification: Early detection of cardiomyopathy remains a challenge, often leading to delayed diagnosis and intervention. Developing reliable biomarkers and risk assessment tools that can identify individuals at high risk of developing cardiomyopathy is crucial for timely intervention and improved patient outcomes.
- Therapeutic Advancements: While conventional heart failure therapies are beneficial for many cardiomyopathy patients, there is a need for novel, disease-specific treatment options. Research into innovative pharmacological, genetic, regenerative, or device based therapies is essential to enhance treatment efficacy and patient quality of life.
- Genetic Insights: Genetic mutations play a significant role in certain cardiomyopathy subtypes. Investigating the genetic basis of cardiomyopathy and developing gene specific therapies hold promise for precision medicine approaches.

1. 4 APPLICATIONS

Some applications related to cardiomyopathy research and treatment are as follows:

➢ Early Detection and Diagnosis:

- Development of improved diagnostic tools, including genetic testing, biomarker identification, and advanced imaging techniques, to enable early detection of cardiomyopathy subtypes.
- Application of artificial intelligence (AI) and machine learning algorithms to analyse cardiac imaging data, aiding in the early identification of structural and functional abnormalities associated with cardiomyopathy.

➢ Genetic Screening and Counselling:

- Genetic screening programs to identify individuals with familial cardiomyopathy, enabling early intervention and family-based counselling.
- Genetic counselling services to provide information and support to individuals and families at risk of hereditary cardiomyopathy.

➢ Therapeutic Interventions:

- Development of novel drug therapies targeting specific molecular pathways involved in cardiomyopathy, with the aim of slowing disease progression and improving heart function.
- Investigation of gene therapies and genome editing techniques to correct genetic mutations responsible for certain cardiomyopathy subtypes. o Research into regenerative medicine approaches, such as stem cell therapy and tissue engineering, for the repair and regeneration of damaged heart tissue.

➢ Personalized Medicine:

- Tailoring treatment plans based on the individual patient's genetic profile, disease subtype, and response to therapy, leading to more effective and personalized care.
- Pharmacogenomic studies to identify genetic factors that influence an individual's response to medications commonly used in cardiomyopathy treatment.

CHAPTER 2

LITERATURE SURVEY

LITERATURE SURVEY

Rani et. al proposed a hybrid decision support system that can assist in the early detection of heart disease based on the clinical parameters of the patient. Authors have used multivariate imputation by chained equations algorithm to handle the missing values. A hybridized feature selection algorithm combining the Genetic Algorithm (GA) and recursive feature elimination has been used for the selection of suitable features from the available dataset. Further for preprocessing of data, SMOTE (Synthetic Minority Oversampling Technique) and standard scalar methods have been used. In the last step of the development of the proposed hybrid system, authors have used support vector machine, naive bayes, logistic regression, random forest, and Adaboost classifiers. It has been found that the system has given the most accurate results with random forest classifier. It was tested on the Cleveland heart disease dataset available at UCI (University of California, Irvine) machine learning repository. It has achieved an accuracy of 86.6%, which is superior to some of the existing heart disease prediction systems found in the literature.

Kavitha et. al proposed a novel machine learning approach to predict heart disease. The proposed study used the Cleveland heart disease dataset, and data mining techniques such as regression and classification are used. Machine learning techniques Random Forest and Decision Tree are applied. The novel technique of the machine learning model is designed. In implementation, 3 machine learning algorithms are used, they are 1. Random Forest, 2. Decision Tree and 3. Hybrid model (Hybrid of random forest and decision tree). Experimental results show an accuracy level of 88.7% through the heart disease prediction model with the hybrid model. The interface is designed to get the user's input parameter to predict the heart disease, for which they used a hybrid model of Decision Tree and Random Forest.

Mohan et. al proposed 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 prediction model is introduced with different combinations of features and several known classification techniques. They produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

Shah et. al presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm. It uses the existing dataset from the Cleveland database of UCI repository of heart disease patients. The dataset comprises 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of different algorithms. This research paper aims to envision the probability of developing heart disease in the patients. The results portray that the highest accuracy score is achieved with K-nearest neighbor.

Aniruddha Dutta et. al proposed a multi-stage model to predict CHD using a highly imbalanced clinical data containing both qualitative and quantitative attributes. For such clinical data, imbalance is an imminent challenge that exists due to the limited availability of data. Such data imbalance adversely affects the performance of any state-of-the-art clinical classification model. As a remedy to the imbalance problem, one cannot efficiently apply conventional techniques, such as data augmentation.

Ishaq et. al analyzes the heart failure survivors from the dataset of 299 patients admitted in hospital. The aim is to find significant features and effective data mining techniques that can boost the accuracy of cardiovascular patient's survivor prediction. To predict patient's survival, this study employs nine classification models: Decision Tree (DT), Adaptive boosting classifier (AdaBoost), Logistic Regression (LR), Stochastic Gradient classifier (SGD), Random Forest (RF), Gradient Boosting classifier (GBM), Extra Tree Classifier (ETC), Gaussian Naive Bayes classifier (G-NB) and Support Vector Machine (SVM). The imbalance class problem is handled by Synthetic Minority Oversampling Technique (SMOTE). Furthermore, machine learning models are trained on the highest ranked features selected by RF. The results are compared with those provided by machine learning algorithms using full set of features.

CHAPTER 3
EXISTING SYSTEM

3.1.1 RANDOM FOREST ALGORITHM

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

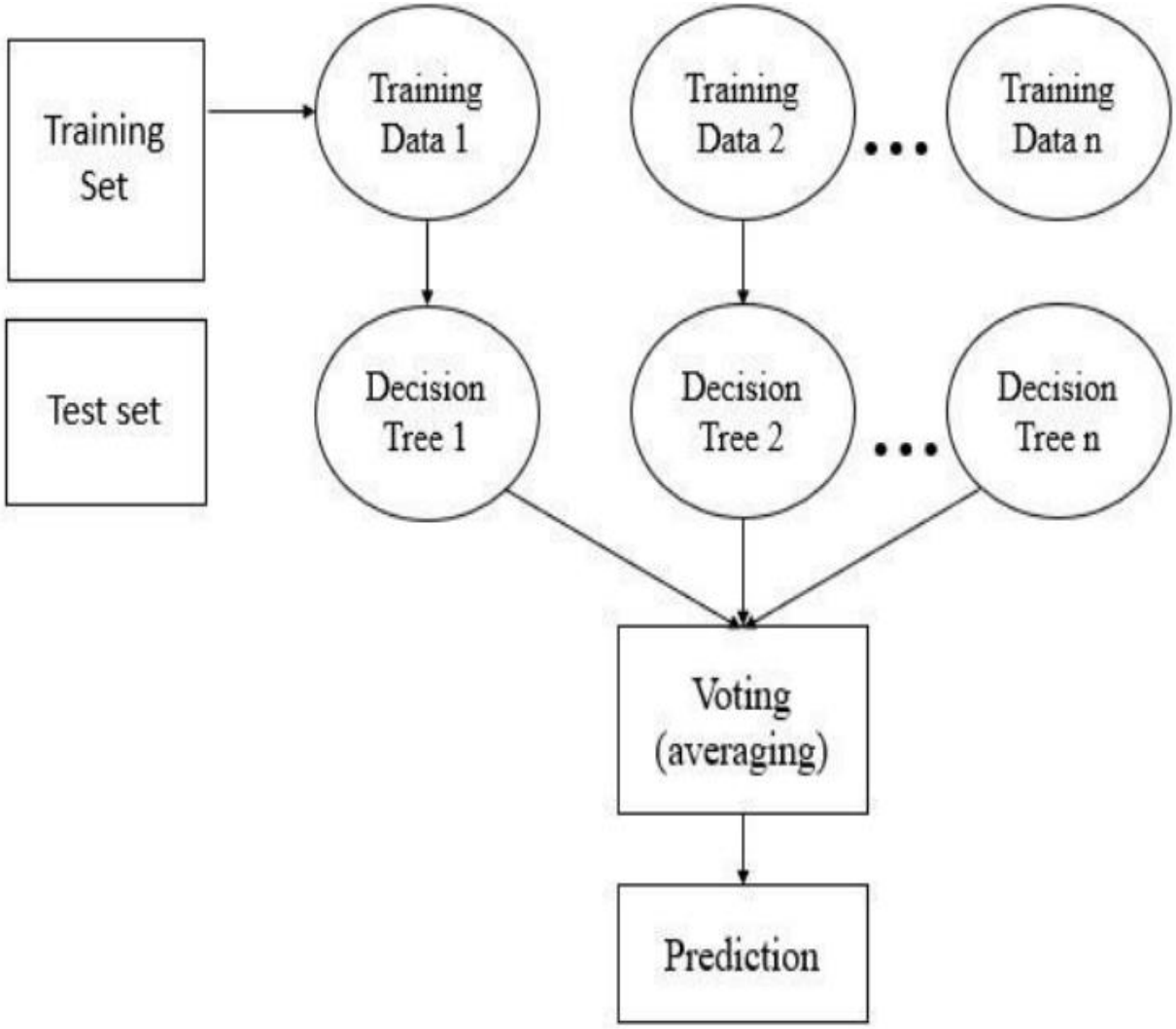

Fig. 3.1.1: Random Forest algorithm.

3.1.2 IMPORTANT FEATURES OF RANDOM FOREST

- **Diversity-** Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- **Immune to the curse of dimensionality**- Since each tree does not consider all the features, the feature space is reduced.

3.1.3 ASSUMPTIONS FOR RANDOM FOREST

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

3.2 DRAWBACKS OF EXISTING SYSTEM

Random Forest Classification (RFC) is a machine learning algorithm used for classification tasks. The drawbacks of RFC is as follows:
Lack of Interpretability: Random Forests are often considered "black box" models. It can be challenging to interpret the underlying decision-making process because they involve an ensemble of many decision trees. Understanding which features are the most important and how they influence the predictions can be less straightforward compared to some other algorithms.

Resource Intensive: Training a Random Forest can be computationally expensive, especially when dealing with a large number of trees and features. This can make it less suitable for realtime or online learning applications and may require substantial computing resources.

CHAPTER 4

PROPOSED SYSTEM

4.1 OVERVIEW

Addressing critical issues through multidisciplinary research efforts will not only advance our understanding of cardiomyopathy but also lead to the development of more effective diagnostic and therapeutic strategies, ultimately improving the lives of individuals affected by this challenging cardiac condition. shows the block diagram of proposed system. The detailed operation is illustrated as follows:

Step 1: Dataset Acquisition: Start by obtaining a cardiomyopathy dataset from a reliable source. Ensure that the dataset contains relevant features (independent variables) and a target variable (dependent variable) indicating the presence or absence of cardiomyopathy.

Step 2: Exploratory Data Analysis (EDA): EDA is the major step in understanding dataset:

- Data Summary: Compute summary statistics (mean, median, standard deviation, etc.) for numerical features and frequency counts for categorical features.
- Data Visualization: Create visualizations (histograms, box plots, scatter plots, etc.) to explore data distributions, identify outliers, and visualize relationships between features.
- Missing Data: Check for missing values and decide how to handle them (imputation or removal).

Step 4: Existing Decision Tree Classifier (DTC): Train a Decision Tree Classifier on the preprocessed training data. Tune hyperparameters (e.g., tree depth, criterion) using techniques like cross-validation and grid search. Evaluate the DTC's performance on the testing dataset using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

Step 5: Existing Random Forest Classifier (RFC): Train a Random Forest Classifier on the preprocessed training data. Optimize hyperparameters (e.g., number of trees, maximum depth, number of features to consider) through cross-validation. Assess the RFC's performance on the test dataset using the same evaluation metrics as in the DTC step.

Step 6: Proposed k-Nearest Neighbors (KNN) Classifier: Implement a KNN classifier using the preprocessed training data. Experiment with different values of k (number of neighbors) to find the optimal value through cross-validation. Evaluate the KNN classifier's performance on the test dataset, employing the same evaluation metrics used for DTC and RFC.

Step 7: Performance Evaluation: Compare the performance of the DTC, RFC, and KNN classifiers using metrics such as accuracy, precision, recall, F1-score

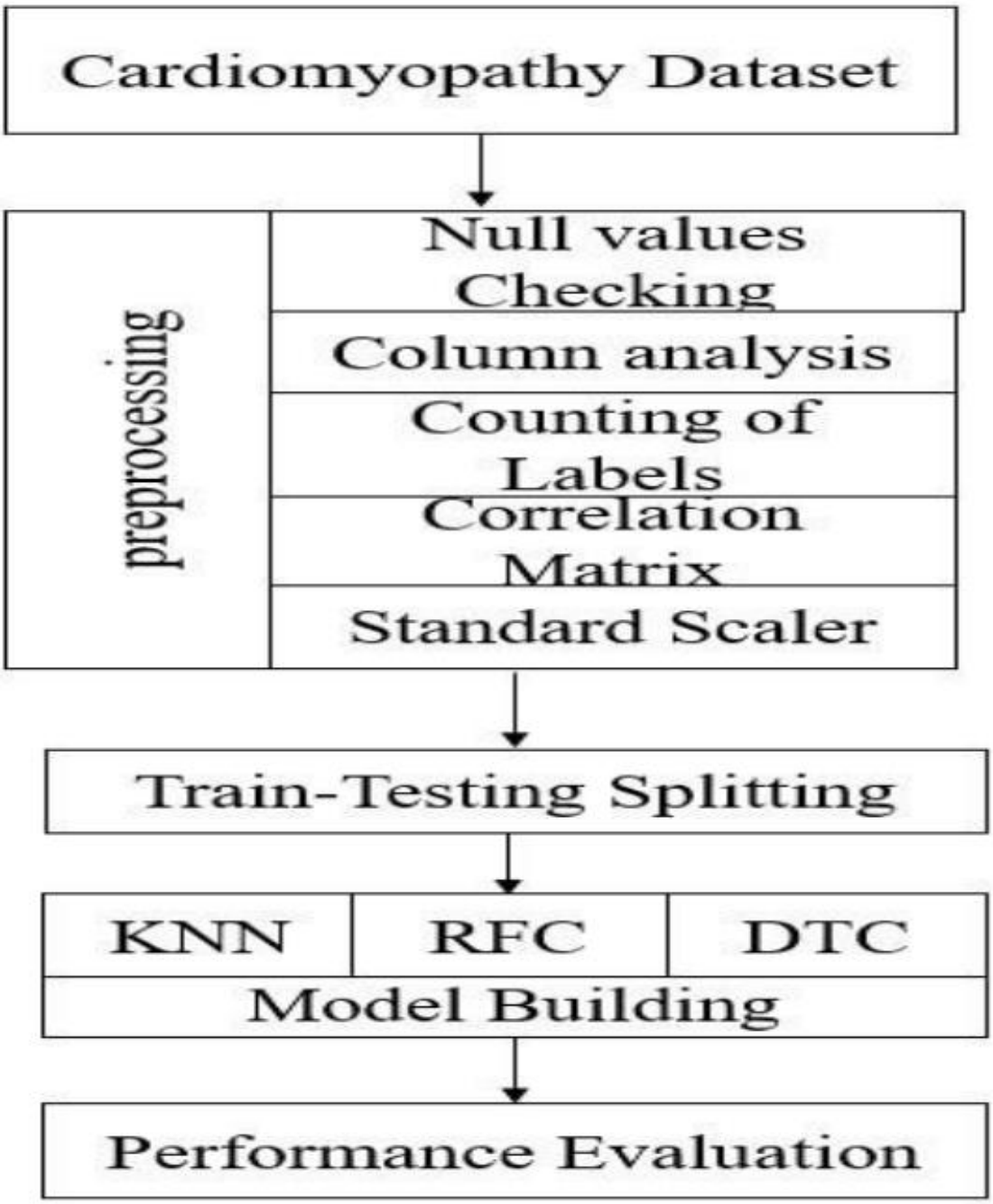

Fig. 4.1: Block diagram of proposed system

4.2 K-NEAREST NEIGHBORS (KNN)

K-Nearest Neighbors (KNN) is a simple yet powerful supervised machine learning algorithm used for classification and regression tasks. KNN is a distance-based classification algorithm. It assigns a new data point to the majority class of its k-nearest neighbors. The choice of 'k' (the number of neighbors) is a crucial hyperparameter that impacts the model's performance. KNN is an instance-based learning method, meaning it doesn't build a model during training. Instead, it memorizes the entire training dataset and uses it for predictions.

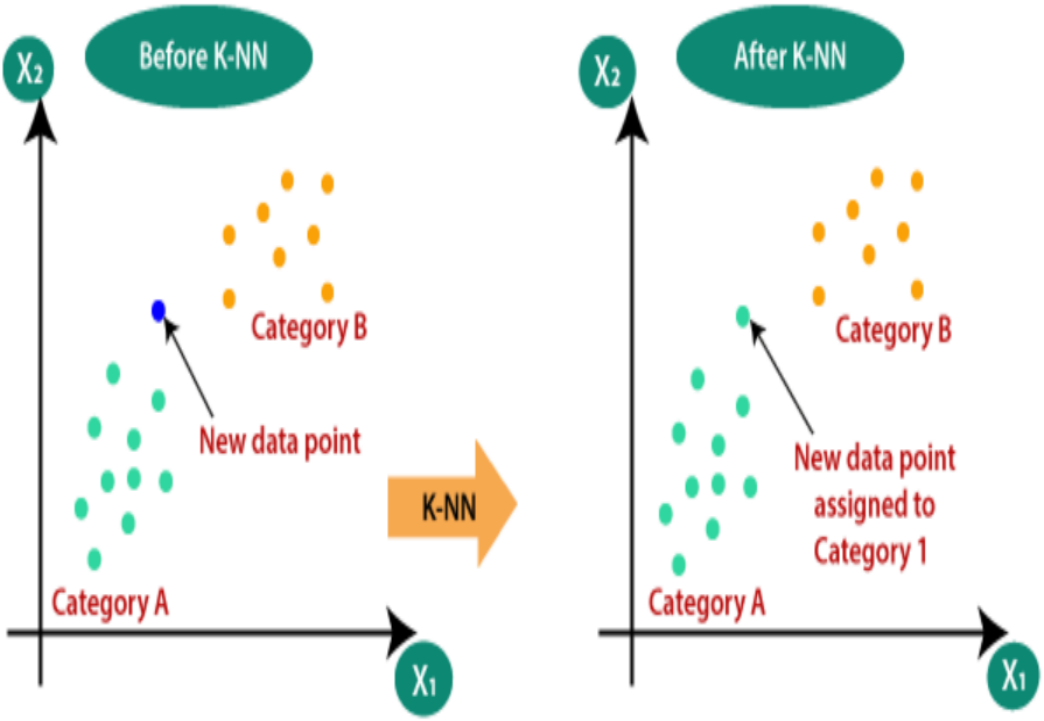

Figure 4.3 KNN initialization

CHAPTER 5
UML DAIGRAMS

5.1 CLASS DIAGRAM

The class diagram classifies the actors defined in the use case diagram into a set of interrelated classes. The relationship or association between the classes can be either an "is-a" or "has-a" relationship. Each class in the class diagram may be capable of providing certain functionalities.

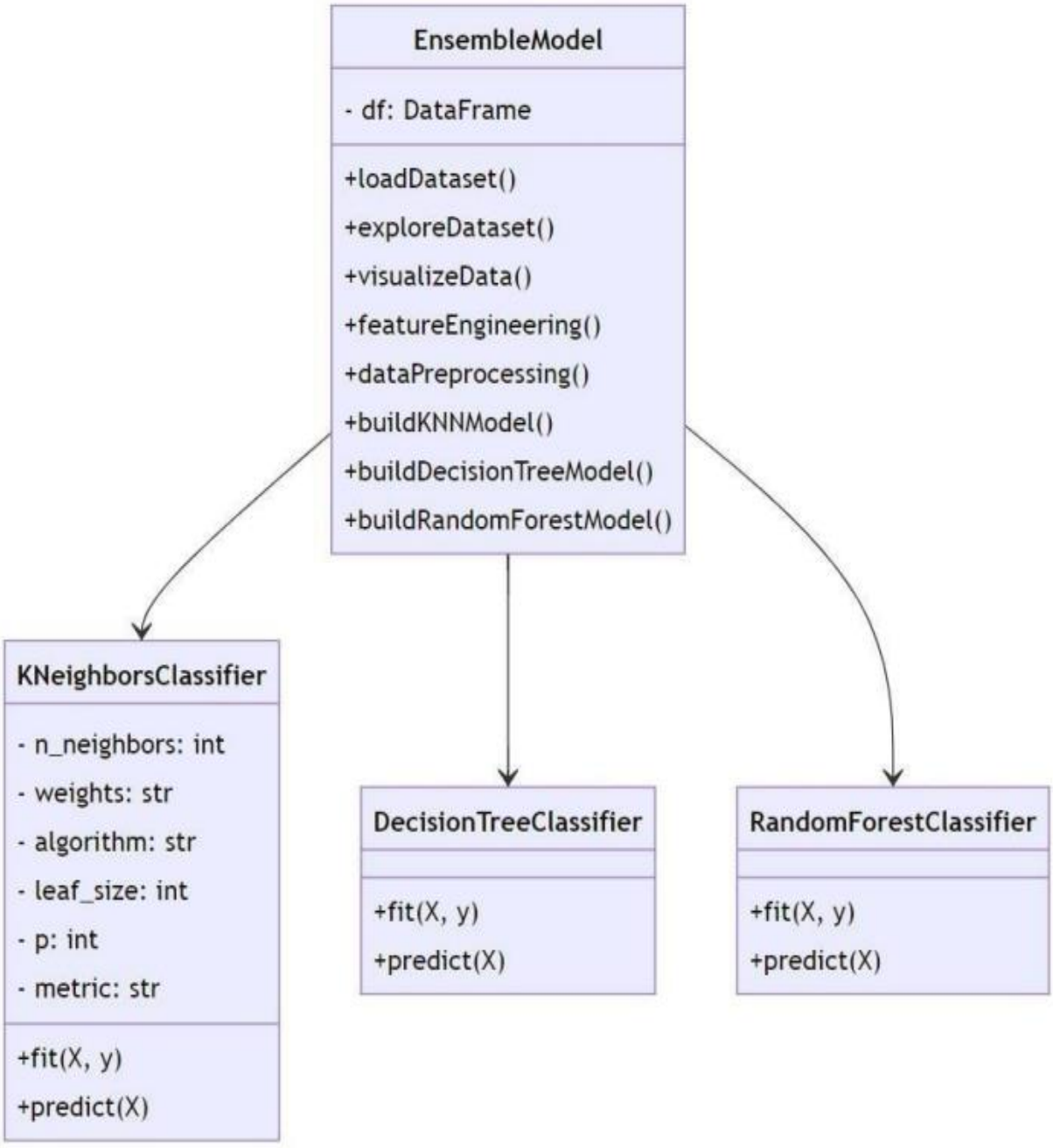

Figure 5.1. Class Diagram

5.2 USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

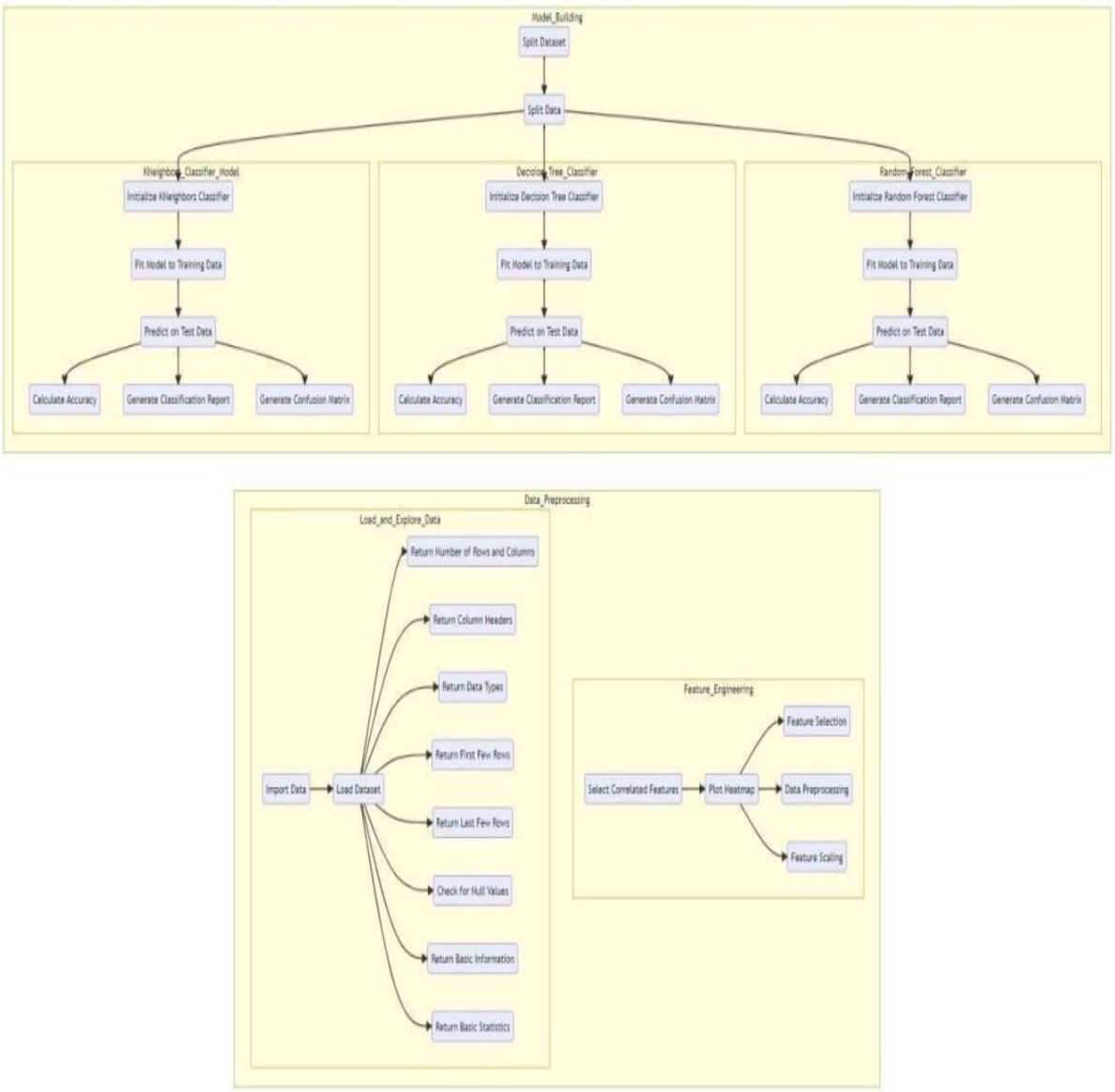

Figure 5.2 Use case Diagram

5.3 COMPONENT DIAGRAM

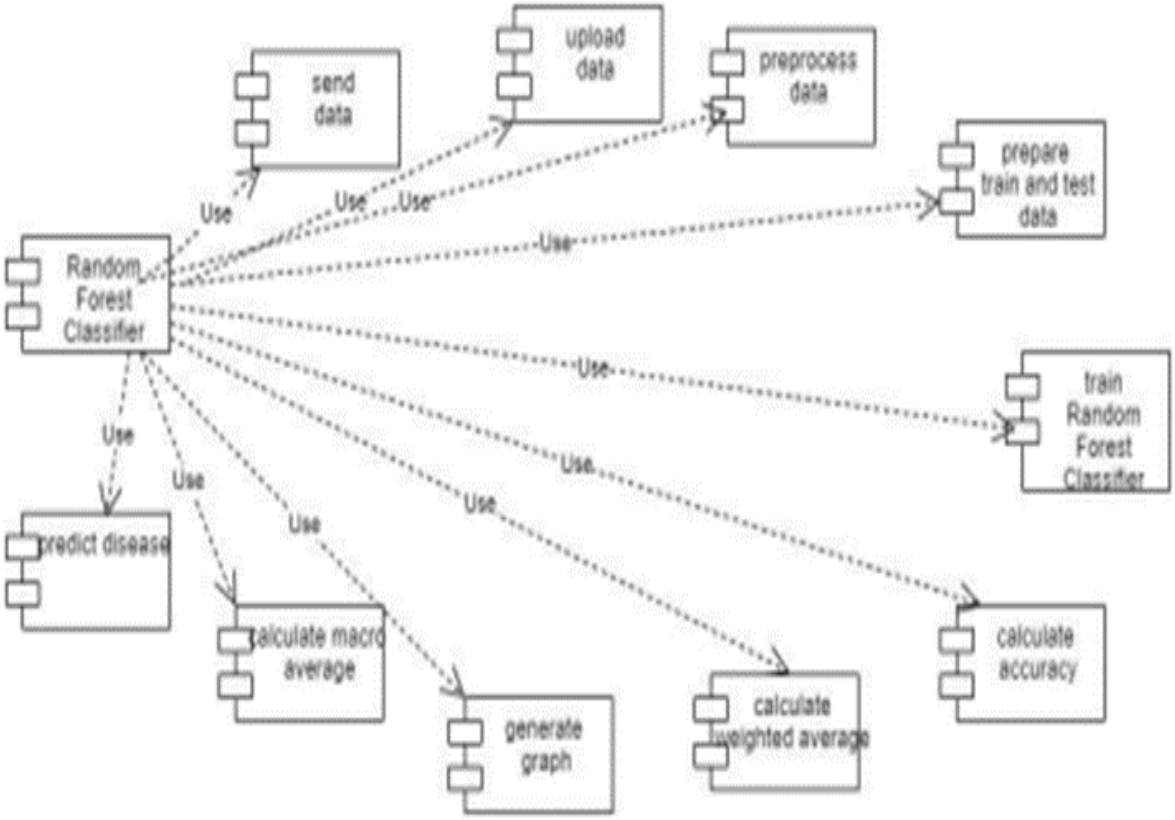

Figure 5.3. Component Diagram

5.4 SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows, as parallel vertical lines ("lifelines"), different processes or objects that live simultaneously, and as horizontal arrows, the messages exchanged between them, in the order in which they occur.

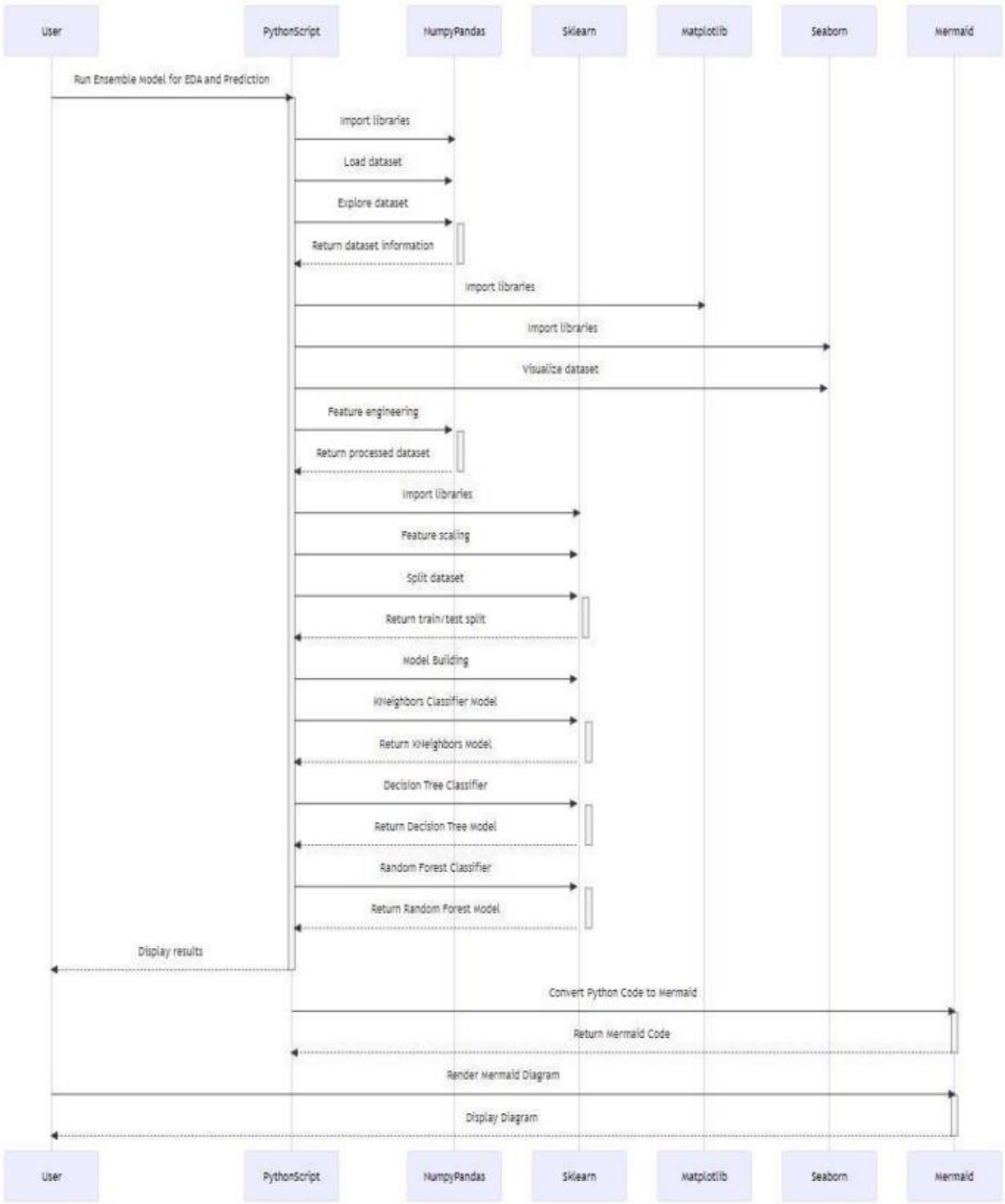

Figure 5.4. Sequence Diagram

5.5 DEPLOYMENT DIAGRAM

The deployment diagram visualizes the physical hardware on which the software will be deployed.

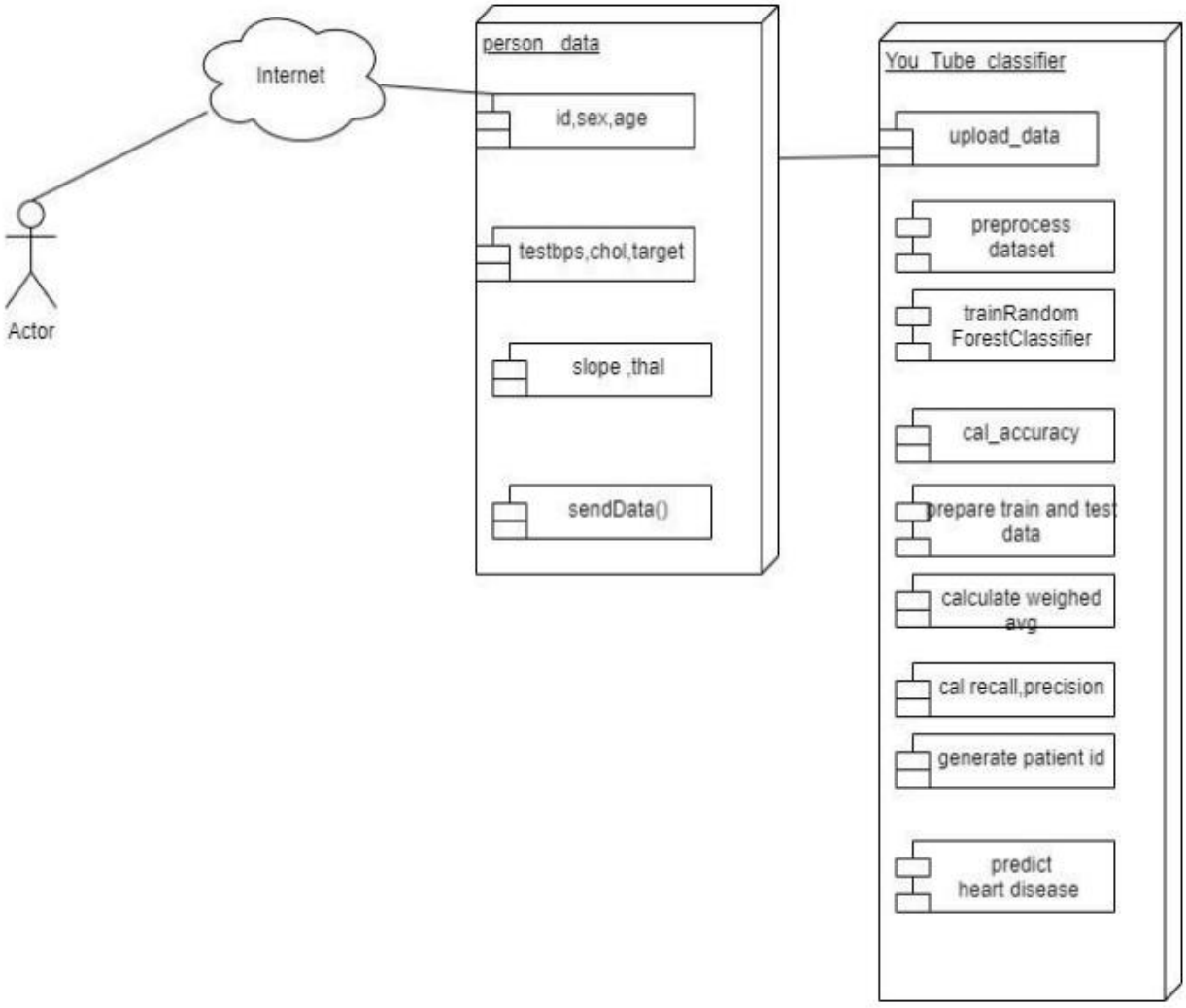

Figure 5.5. Deployment Diagram

CHAPTER 6

SYSTEM REQUIREMENTS SPECIFICATIONS

SOFTWARE REQUIREMENTS

The functional requirements or the overall description documents include the product perspective and features, operating system and operating environment, graphics requirements, design constraints and user documentation

- Python IDLE 3.7 version (or)
- Anaconda 3.7 (or)
- Jupiter (or)
- Google collab

HARDWARE REQUIREMENTS

Minimum hardware requirements are very dependent on the particular software being developed by a given Enthought Python / Canopy / VS Code user. Applications that need to store large arrays/objects in memory will require more RAM, whereas applications that need to perform numerous calculations or tasks more quickly will require a faster processor.

- Operating system : Windows, Linux
- Processor : minimum intel i3
- Ram : minimum 4 GB
- Hard disk : minimum 250GB

CHAPTER 7

RESULTS AND DISCUSSION

7. 1 implementation

The implementation focused on python script for performing cardiomyopathy classification on a dataset using three different machine learning algorithms: K-Nearest Neighbors (KNN), Decision Tree Classifier,
and Random Forest Classifier. Below is a detailed explanation of the code:

Importing Libraries and Loading Data:

- The ode begins by importing necessary libraries such as NumPy, Pandas, and scikitlearn for data manipulation and machine learning.
- It loads a dataset from a CSV file named 'heart.csv' into a Pandas DataFrame called 'df'.

Exploratory Data Analysis (EDA):

- Several exploratory data analysis steps are performed, including:
- Checking the shape of the dataset (number of rows and columns).
- Displaying column names.
- Checking data types of columns.
- Displaying the first and last few rows of the dataset.
- Checking for missing values in the dataset.
- Providing basic information and statistics about the dataset.

Data Visualization:

Data visualization is performed using Matplotlib and Seaborn libraries. Key visualizations include:

- Histograms of the dataset's numerical features.
- A count plot to visualize the balance of the target variable.

Train-Test Split:

- The dataset is split into independent features (X) and the target variable (y).
- A train-test split is performed, allocating 20% of the data to the test set.

K-Nearest Neighbors (KNN):

- KNN classification is applied using scikit-learn's KNeighborsClassifier.
- The code calculates accuracy, prints a classification report, and visualizes the confusion matrix for KNN.

Decision Tree Classifier:

- Decision tree classification isapplied using scikit-learn's DecisionTreeClassifier.
- Accuracy, a classification report, and a confusion matrix are generated for the Decision Tree Classifier.

Random Forest Classifier:

- Random forest classification is applied using scikit-learn's RandomForestClassifier.
- Accuracy, a classification report, and a confusion matrix are generated for the Random Forest Classifier.

7.2 DATASET DESCRIPTION

The columns of dataset is described as follows:

- **age:** This column represents the age of the patient. Age is an important factor in assessing the risk of heart disease, as the likelihood of developing heart-related issues often increases with 62 age. sex: This column indicates the gender of the patient. It's typically encoded as 1 for male and 0 for female. Gender can play a role in the prevalence and presentation of heart disease.
- **sex:** This column indicates the gender of the patient. It's typically encoded as 1 for male and 0 for female. Gender can play a role in the prevalence and presentation of heart disease.
- **cp:** This column stands for "chest pain" and represents the type of chest pain experienced by the patient. Chest pain is a common symptom of heart disease and can be categorized into different types based on its characteristics.
- **trestbps:** Short for "resting blood pressure," this column records the patient's blood pressure

while at rest. Blood pressure is an essential measure of cardiovascular health and can influence the risk of heart disease.

- **chol:** This column represents the serum cholesterol levels of the patient. Cholesterol is a type of fat that can accumulate in arteries, potentially leading to heart problems.
- **fbs:** Stands for "fasting blood sugar." It indicates the patient's blood sugar level after fasting. Elevated fasting blood sugar can be indicative of diabetes, which is a risk factor for heart disease.
- **restecg:** This column records the results of a resting electrocardiogram (ECG or EKG), which measures the electrical activity of the heart. Abnormal ECG results can suggest heart conditions.
- **thalach:** Represents the maximum heart rate achieved by the patient during exercise. The maximum heart rate is an indicator of cardiovascular fitness and can be related to heart disease risk.
- **exang:** Short for "exercise induced angina." Angina is chest pain caused by reduced blood flow to the heart. This column indicates whether angina was induced by exercise (1 for yes, 0 for no).
- **oldpeak**: This column represents the ST depression induced by exercise relative to rest. ST depression can be an ECG finding associated with reduced blood flow to the heart during exercise. slope: Refers to the slope of the peak exercise ST segment on the ECG. It provides information about the rate of change in ST segment values during exercise. • ca: Represents the number of major blood vessels colored by fluoroscopy. The presence of more blocked or narrowed vessels can indicate more advanced heart disease.
- **thal:** Stands for "thalassemia," a genetic blood disorder. It can have different types, and its presence might provide additional information about the patient's health status.
- **target:** This column indicates whether the patient has heart disease. It's often encoded as 1 for the presence of heart disease and 0 for its absence. This is the target variable used for predictive modeling.

7.3 RESULTS

Figure 7.1 provides a representation of the dataset itself, displaying a portion of the data rows. It shows specific instances or examples from the dataset, each with its own values for the different columns. This sample dataset is useful for understanding the type of data you're working with and gaining an initial sense of the values and structure of the dataset.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 7.1. Sample dataset.

FIGURE 7.2. COLUMN WISE DISTRIBUTION:In this figure, each column of the dataset is visualized individually. It showcases how the values are distributed within each column. This visualization can help you understand the range of values, potential outliers, and the overall pattern of the data within each feature or attribute.

FIGURE 7.3. COUNT OF EACH CLASS: This figure displays the count of each class or category in the target variable (such as "presence" or "absence" of heart disease). It provides insights into the balance of classes and helps you understand if one class dominates the dataset. This is crucial for building accurate predictive models.

Figure 10.4. Correlation Matrix: The correlation matrix shows the correlation coefficients between pairs of numerical features in the dataset. Each cell in the matrix represents the correlation between two variables. This figure helps identify potential relationships between features; positive values indicate a positive correlation, negative values indicate a negative correlation, and values closer to zero indicate weaker correlations

FIGURE 7.5. CORRELATION MATRIX FEATURES:

Similar to the previous figure, this figure focuses on the correlation matrix, but specifically for the features (columns) related to heart disease. It provides a more detailed view of how these features correlate with each other. Strong correlations between certain features might suggest important interactions or redundancies in the data

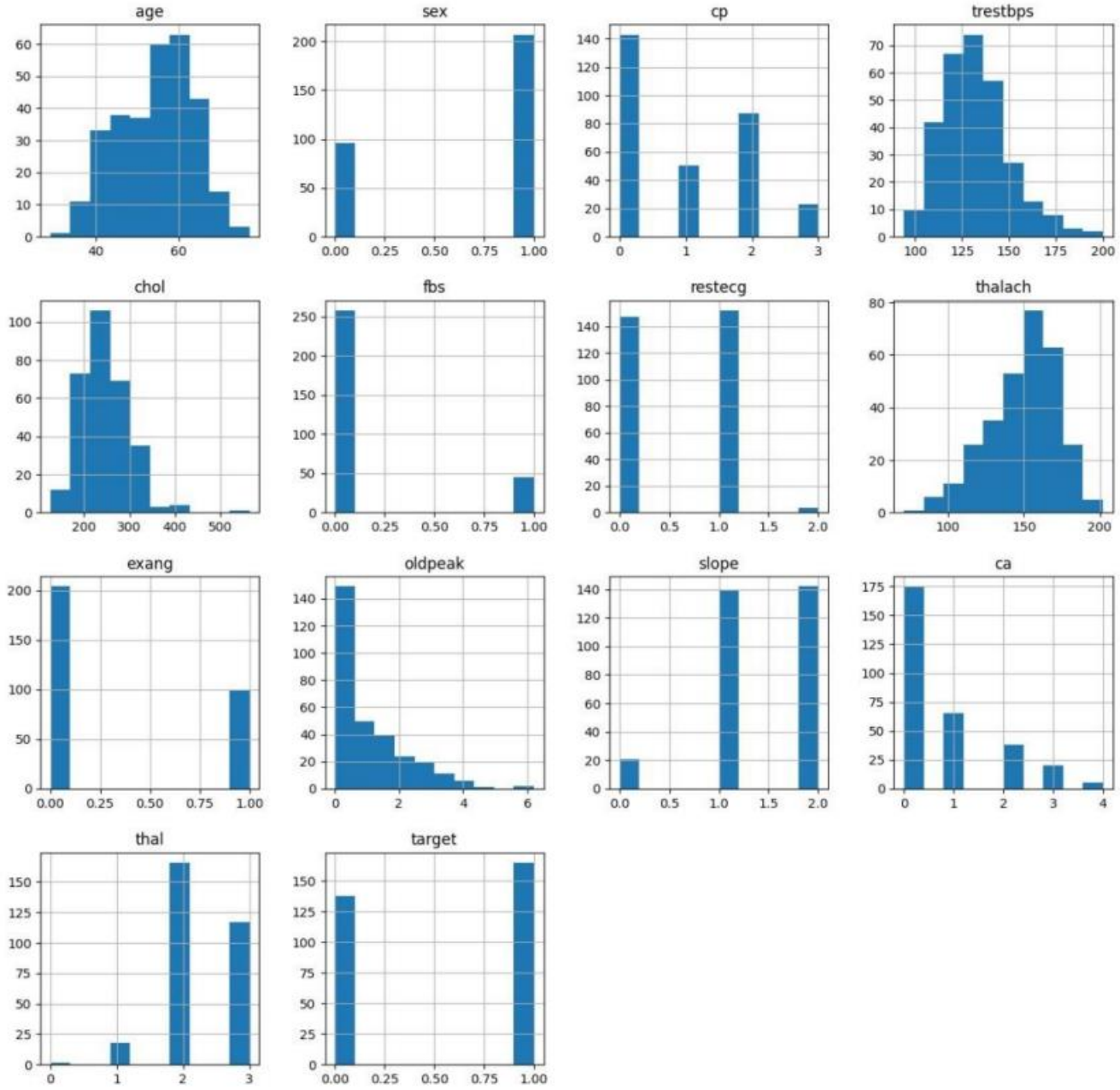

Figure 7.2. Column wise distribution

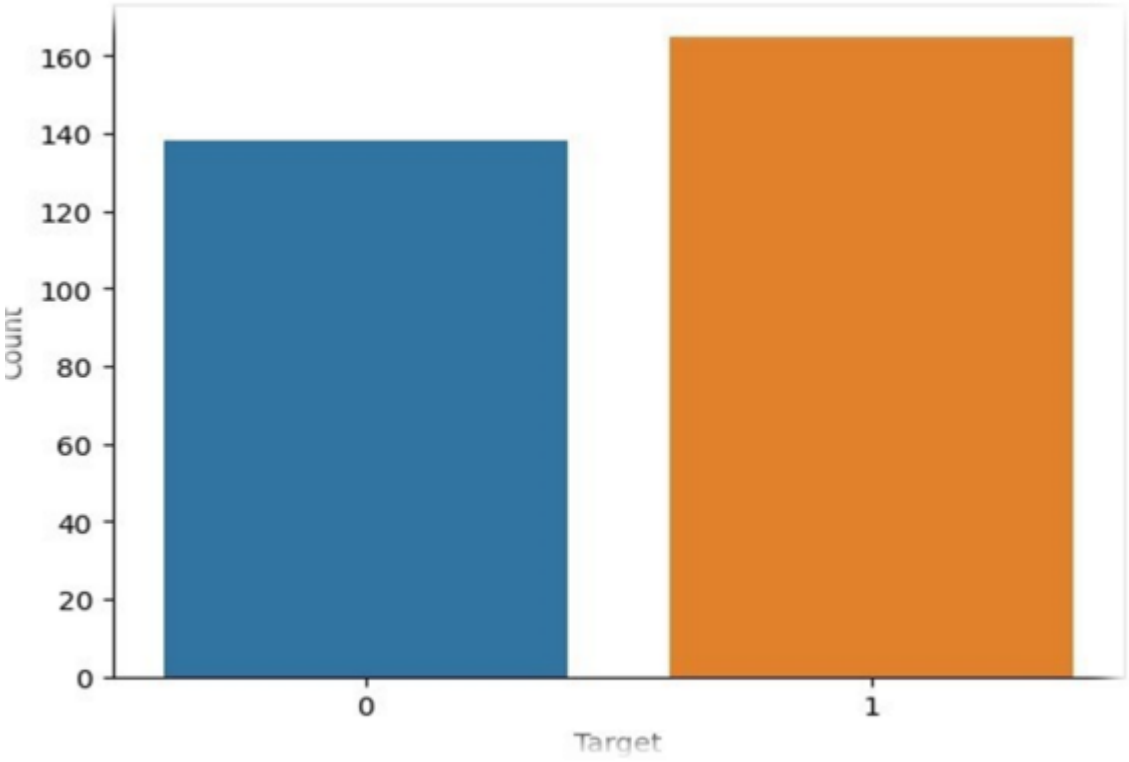

Figure 7.3. Count of each class

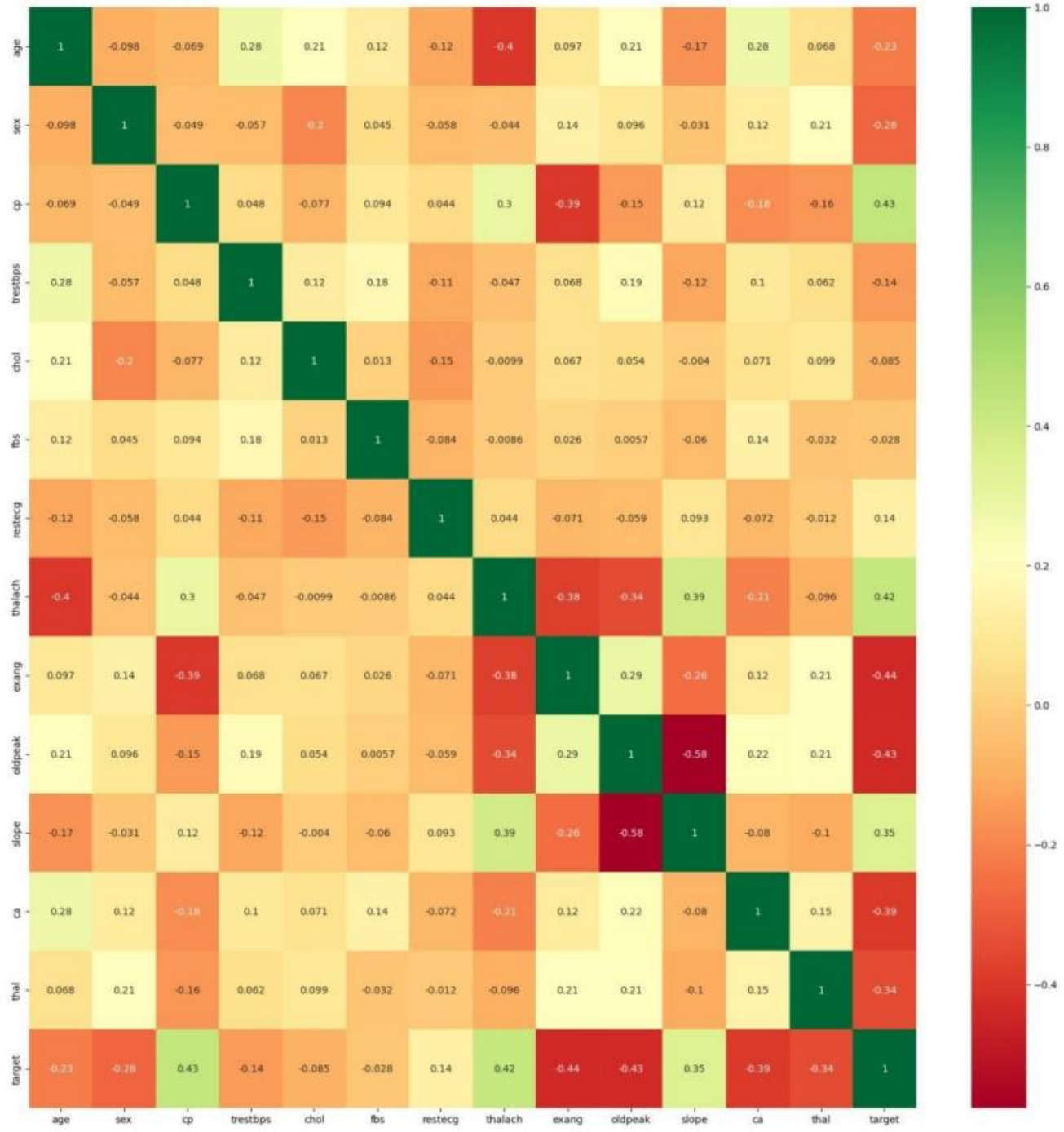

Figure 7.4. Correlation Matrix

age	trestbps	chol	thalach	oldpeak	target	sex_0	sex_1	cp_0	cp_1	...	slope_2	ca_0	ca_1	ca_2	ca_3	ca_4	thal_0	thal_1	thal_2	thal_3
0.952197	0.763956	-0.256334	0.015443	1.087338	1	0	1	0	0	...	0	1	0	0	0	0	0	1	0	0
-1.915313	-0.092738	0.072199	1.633471	2.122573	1	0	1	0	0	...	0	1	0	0	0	0	0	0	1	0
-1.474158	-0.092738	-0.816773	0.977514	0.310912	1	1	0	0	1	...	1	1	0	0	0	0	0	0	1	0
0.180175	-0.663867	-0.198357	1.239897	-0.206705	1	0	1	0	1	...	1	1	0	0	0	0	0	0	1	0
0.290464	-0.663867	2.082050	0.583939	-0.379244	1	1	0	1	0	...	1	1	0	0	0	0	0	0	1	0

Figure 7.5. Correlation matrix features.

The Random Forest Classifier (RFC) is assessed. For class 0, precision is 0.83, recall is 0.86, and the F1-score is 0.85. Similarly, for class 1, precision is 0.87, recall is 0.84, and the F1- score is 0.86. The macro average F1-score, precision, and recall is 0.85, while the weighted average aligns with these metrics, suggesting consistent performance. The overall accuracy is also 0.85.

Metric	Precision	Recall	F1-Score	Support
Class 0	0.83	0.86	0.85	29
Class 1	0.87	0.84	0.86	32
Macro Avg	0.85	0.85	0.85	61
Weighted Avg	0.85	0.85	0.85	61
Accuracy	0.85	0.85	0.85	

Table 7.3. RFC Classification report.

CHAPTER 8

CONCLUSION AND FUTURE SCOPE

8.1 CONCLUSION

In summary, the multifaceted analysis conducted in this study, which included in-depth Exploratory Data Analysis (EDA), meticulous data preprocessing, and a thorough evaluation of the Decision Tree Classifier (DTC), Random Forest Classifier (RFC), and the novel k- Nearest Neighbors (KNN) Classifier, was instrumental in unraveling the intricacies of the cardiomyopathy dataset. This comprehensive approach provided invaluable insights into the dataset's inherent characteristics, the nuanced performance of each classifier, and their respective strengths and limitations. The outcomes of this study not only advanced our understanding of how these machine learning models can be applied to cardiomyopathy classification but also illuminated potential avenues for further research and refinement. These findings have the potential to shape the future of cardiac health diagnostics by guiding the development of more accurate and effective tools for the early detection and management of cardiomyopathy, ultimately improving patient outcomes and healthcare practices in this critical domain.

8.2 FUTURE SCOPE

The future scope in the field of cardiomyopathy research and diagnostics is promising and encompasses several areas of development and innovation. Continued advancements in medical imaging technologies, such as magnetic resonance imaging (MRI) and echocardiography, will provide more detailed and accurate information about heart function, helping in early detection and personalized treatment planning. Further exploration of the genetic underpinnings of cardiomyopathy will lead to more precise diagnosis and targeted therapies based on an individual's genetic profile.

Advances in stem cell research and tissue engineering may offer regenerative treatments for repairing damaged heart tissue, potentially reversing the effects of cardiomyopathy. The 71 development of patient-specific treatment strategies, considering genetic, lifestyle, and environmental factors, will improve therapeutic outcomes and reduce adverse effects. As technology and genetic testing continue to advance, ethical and legal frameworks surrounding patient data privacy, genetic counseling, and informed consent will need to evolve accordingly.

The future of cardiomyopathy research and diagnostics holds great potential for improving patient care, early detection, and treatment outcomes. Continued collaboration among researchers, healthcare professionals, and technology innovators will play a crucial role in driving these advancements and ensuring that they benefit patients worldwid

CHAPTER 9

REFERENCES

1. Bakar, Wan Aezwani Wan Abu, et al. "A Review: Heart Disease Prediction in Machine Learning & Deep Learning." 2023 19th IEEE International Colloquium on Signal Processing & Its Applications (CSPA). IEEE, 2023.

2. Dileep, P., et al. "An automatic heart disease prediction using cluster-based bidirectional LSTM (C-BiLSTM) algorithm." Neural Computing and Applications 35.10 (2023): 7253-7266.

3. Mishra, Nilamadhab, et al. "Visual Analysis of Cardiac Arrest Prediction Using Machine Learning Algorithms: A Health Education Awareness Initiative." Handbook of Research on Instructional Technologies in Health Education and Allied Disciplines. IGI Global, 2023. 331-363.

4. Guo, Saidi, et al. "Survival prediction of heart failure patients using motion-based analysis method." Computer Methods and Programs in Biomedicine 236 (2023): 107547.

5. Nandy, Sudarshan, et al. "An intelligent heart disease prediction system based on swarm-artificial neural network." Neural Computing and Applications 35.20 (2023): 14723-14737.

6. Pant, Aman, et al. "Heart disease prediction using image segmentation Through the CNN model." 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2023.

7. Nandy, Sudarshan, et al. "An intelligent heart disease prediction system based on swarm-artificial neural network." Neural Computing and Applications 35.20 (2023): 14723-14737. 73

8. Saranya, G., and A. Pravin. "A novel feature selection approach with integrated feature sensitivity and feature correlation for improved prediction of heart disease." Journal of Ambient Intelligence and Humanized Computing 14.9 (2023): 12005- 12019.

9. Chandrasekhar, Nadikatla, and Samineni Peddakrishna. "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization." Processes 11.4 (2023): 121

10. Sherly, S. Irin, and G. Mathivanan. "An efficient honey badger based Faster region CNN for chronc heart Failure prediction." Biomedical Signal Processing and Control 79 .

11. Rani, P., Kumar, R., Ahmed, N.M.O.S. et al. A decision support system for heart disease prediction based upon machine learning. J Reliable Intell Environ 7, 263–275 (2021).

12. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

13. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 8154281554, 2019, doi: 10.1109/ACCESS.2019.

CHAPTER 10

APPENDIX

Annexure-1: Sample Coding

```
# Importing essential libraries import numpy as np
 import pandas as pd from sklearn.model_selection import
 train_test_split
# Loading the dataset
df=pd.read_csv('heart.csv')
 # Returns number of rows and columns of the dataset df.shape
 # Returns an object with all of the column headers df.columns
# Returns different datatypes for each columns (float, int, string, bool, etc.) df.dtypes
# Returns the first x number of rows when head(x). Without a number it returns 5 df.head()
 # Returns the last x number of rows when tail(x). Without a number it returns 5 df.tail()

# Returns true for a column having null values, else false df.isnull().any()
# Returns basic information on all columns df.info()
# Returns basic statistics on numeric columns df.describe().T
# Importing essential libraries
import matplotlib.pyplot as plt %matplotlib
inline import seaborn as sns
 # Plotting histogram for the entire dataset
fig = plt.figure(figsize = (15,15)) ax = fig.gca() g = df.hist(ax=ax)
# Visualization to check if the dataset is balanced or not g = sns.countplot(x='target', data=df)
plt.xlabel('Target') plt.ylabel('Count')
# Selecting correlated features using
Heatmap
# Get correlation of all the features of the dataset corr_matrix = df.corr() top_corr_features =
corr_matrix.index

# Plotting the heatmap plt.figure(figsize=(20,20)) sns.heatmap(data=df[top_corr_features].corr(),
annot=True, cmap='RdYlGn') dataset = pd.get_dummies(df, columns=['sex', 'cp', 'fbs', 'restecg',
'exang', 'slope', 'ca', 'thal']) dataset.columns from sklearn.preprocessing import StandardScaler 58
standScaler = StandardScaler() columns_to_scale = ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']
dataset[columns_to_scale] = standScaler.fit_transform(dataset[columns_to_scale]) dataset.head()
# Splitting the dataset into dependent and independent features
```

```
X = dataset.drop('target', axis=1) y = dataset['target']
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.20,random_state=42)

# Importing essential libraries from sklearn.neighbors
import KNeighborsClassifier from sklearn.model_selection
import cross_val_score
KNN=KNeighborsClassifier(n_neighbors=5,weights='uniform',algorithm='auto',leaf_size=30
,p=2,metric='minkowski') KNN.fit(X_train,y_train) y_pred=KNN.predict(X_test) y_pred from
sklearn.metrics import accuracy_score,classification_report,confusion_matrix
ac=accuracy_score(y_test,y_pred)*100
 print("KNN_Accuracy:",ac) print("Classification Report:") print(classification_report(y_test,
y_pred)) cm=confusion_matrix(y_test, y_pred) print("Confusion Matrix:") plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues") plt.xlabel("Predicted Label")
 plt.ylabel("True Label") plt.title("KNeighborsClassifier_Confusion Matrix") plt.show()
# Importing essential libraries from sklearn.tree import DecisionTreeClassifier
dt=DecisionTreeClassifier() dt.fit(X_train,y_train) y_pred1=dt.predict(X_test) y_pred1
ac1=accuracy_score(y_test,y_pred1)*100 print("DecisionTreeClassifier Acuracy:",ac1)
print("Classification Report:") print(classification_report(y_test, y_pred1))
cm=confusion_matrix(y_test, y_pred1) print("Confusion Matrix:") plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues") plt.xlabel("Predicted Label")
 plt.ylabel("True Label") plt.title("DecisionTreeClassifier_Confusion Matrix") plt.show()
 # Importing essential libraries from sklearn.ensemble import RandomForestClassifier
rf=RandomForestClassifier() rf.fit(X_train,y_train) y_pred2=dt.predict(X_test) y_pred2 from
 sklearn.metrics import accuracy_score accuracy_score(y_test,y_pred2) print("Classification
Report:")
 print(classification_report(y_test, y_pred2)) cm=confusion_matrix(y_test, y_pred2)
print("Confusion 59 Matrix:") plt.figure(figsize=(8, 6)) sns.heatmap(cm,
 annot=True, fmt="d", cmap="Blues") plt.xlabel("Predicted Label") plt.ylabel("True Label")
plt.title("RandomForestClassifier_Confusion Matrix")
plt.show()
```

ANNEXURE-2: LIST OF FIGURES

ANNEXURE-3: LIST OF OUTPUT SCREENS

ANNEXURE-5: BASE PAPER