**Advanced Analytics for Cardiovascular Disease Diagnosis**

An Application Development Report Submitted

In partial fulfillment of the requirements for the award of the degree of

**Bachelor of Technology in**

**Computer Science and Engineering**

By

N.B SRUTHI 22N31A05E6

NEELA SIVA KUSUMA 22N31A05F4

RAMAVATH JEEVAN 22N31A05K4

Under the esteemed guidance of

**V. Shilpa**

**Assistant Professor**



**Department of Computer Science and Engineering**

# Malla Reddy College of Engineering & Technology

(Autonomous Institution- UGC, Govt. of India)

(Affiliated to JNTUH, Hyderabad, Approved by AICTE, NBA &NAAC with ‘A’ Grade)

Maisammaguda, Kompally, Dhulapally, Secunderabad – 500100

Website: [www.mrcet.ac.in](http://www.mrcet.ac.in/)

2024-2025



# Malla Reddy College of Engineering & Technology

(Autonomous Institution- UGC, Govt. of India)

(Affiliated to JNTUH, Hyderabad, Approved by AICTE, NBA &NAAC with ‘A’ Grade) Maisammaguda, Kompally, Dhulapally, Secunderabad – 500100 website: [www.mrcet.ac.in](http://www.mrcet.ac.in/)

**CERTIFICATE**

This is to certify that this is the bonafide record of the Application entitled “**ADVANCED ANALYTICS FOR CARDIOVASCULAR DISEASE DIAGNOSIS**” submitted by N B SRUTHI (22N31A05E6), N SIVA KUSUMA (22N31A05F4) and RAMAVATH JEEVAN (22N31A05K4) of B.Tech in the partial fulfillment of the requirements for the degree of Bachelor of Technology in Computer Science and Engineering, Department of CSE during the year 2023-2024. The results embodied in this project report have not been submitted to any other university or institute for the award of any degree or diploma.

|  |  |
| --- | --- |
| **Internal Guide** | **Head of the Department** |
| V.Shilpa | Dr.S.Shanthi |
| Assistant professor | Professor |

**External Examiner**

# DECLARATION

We hereby declare that the Application “**ADVANCED ANALYTICS FOR CARDIOVASCULAR DISEASE DIAGNOSIS**” submitted to Malla Reddy College of Engineering and Technology (UGC Autonomous),affiliated to Jawaharlal Nehru Technological University Hyderabad (JNTUH) for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a result of original research carried-out in this thesis. It is further declared that the project report or any part there of has not been previously submitted to any University or Institute for the award of degree or diploma.

**N B Sruthi - 22N31A05E6**

**Neela Siva Kusuma - 22N31A05F4**

**Ramavath Jeevan - 22N31A05K4**

**ACKNOWLEDGEMENT**

We feel honored to place our warm salutation to our college Malla Reddy College of Engineering and Technology (UGC-Autonomous) for giving us opportunity to do this Application as part of our B. Tech Program. We are ever grateful to our Director **Dr. V.S.K Reddy** and Principal **Dr. S. Srinivasa Rao** who enabled us to have experience in engineering and gain profound technical knowledge.

We express our heartiest thanks to our HOD**, Dr. S. Shanthi** for encouraging us in every aspect of our course and helping us realize our full potential,

We would like to thank our Project Guide **Ms. V. Shilpa** for her regular guidance, suggestions, and constant encouragement. We are extremely grateful to our Project Coordinator **Ms. D. Sai Eswari** for her continuous monitoring and unflinching co-operation throughout project work.

We would also like to thank all the faculty members and supporting staff of the Department of CSE and all other departments who have been helpful directly or indirectly in making our project a success

We are extremely grateful to our parents for their blessings and prayers for the completion of our project that gave us strength to do our project.

**With regards and gratitude**

**N B Sruthi - 22N31A05E6**

**Neela Siva Kusuma - 22N31A05F4**

**R Jeevan- 22N31A05K4**

## ABSTRACT

Heart disease remains a significant health concern globally, necessitating effective predictive models to aid in early diagnosis and intervention. This study explores the application of machine learning techniques to predict the likelihood of heart disease based on patient data. The dataset used comprises various clinical parameters such as age, gender, Trest blood pressure, cholesterol levels, chest pain, and other medical attributes. The methodology involves preprocessing the application of supervised learning algorithms including Logistic Regression, Random Forest, and K-Nearest Neighbors. Performance metrics such as accuracy and precision. In conclusion, the Random Forest algorithm demonstrated superior performance in terms of accuracy and precision. The ensemble approach of Logistic regression is the most significant predictors of heart disease.

**Keywords:** Logistic Regression and K-Nearest Neighbors.

|  |  |
| --- | --- |
|  |  |
| |  |  |  | | --- | --- | --- | |  |  |  |   **TABLE OF CONTENTS**   |  |  |  | | --- | --- | --- | | **S.NO** | **TITLE** | **PG.NO** | | **1**  **2** | **INTRODUCTION**   * 1. PURPOSE AND OBJECTIVES   2. EXITING AND PROPOSED SYSTEM   3. SCOPE OF PROJECT   **LITERATURE SURVEY** | **01**  02  03  07  **10** | | **3** | **SYSTEM ANALYSIS**  3.1 HARDWARE AND SOFTWARE REQUIREMENTS | **13**  13 | | **4** | **SYSTEM DESIGN**  4.1 ARCHITECTURE  4.2 UML DIAGRAMS | 15  16 | | **5** | **METHODOLOGY**  5.1 TECHNOLOGIES USED  5.2 MODULES DESCRIPTION | **20**  20  21 | | **6** | **IMPLEMENTATION**  6.1 SAMPLE CODE  6.2 OUTPUT SCREENS | **26**  26  31 | | **7** | **CONCLUSION & FUTURE SCOPE** | **39** | | **8** | **BIBLIOGRAPHY** | **43** | |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE.NO** | **NAME** | **PG.NO** |
| **4.1.1** | **ARCHITECTUREDIAGRAM** | **14** |
| **4.2.1** | **CLASS DIAGRAM** | **15** |
| **4.2.2** | **USE CASE DIAGRAM** | **16** |
| **4.2.3** | **ACTIVITY DIAGRAM** | **17** |
| **4.2.4** | **SEQUENCE DIAGRAM** | **18** |

**CHAPTER 1**

**INTRODUCTION**

Heart disease is one of the leading causes of death globally, making early detection and prediction a critical task in healthcare. **Heart disease prediction** refers to the use of data-driven techniques to assess a person’s risk of developing heart-related conditions, such as coronary artery disease, heart attacks, or arrhythmias. Given the complexity and variety of risk factors involved including age, cholesterol levels, blood pressure, lifestyle habits, and genetic predispositions traditional methods of diagnosis can be time-consuming and prone to human error.

Recent advancements in artificial intelligence (AI) and machine learning (ML) have enabled the development of more efficient and accurate heart disease prediction systems. These systems analyze large amounts of patient data to identify patterns and correlations that may not be easily recognizable by human clinicians. By training models on historical health data, these prediction tools can assess a patient’s likelihood of developing heart disease, often with high accuracy.

Heart disease remains a leading cause of mortality worldwide, accounting for millions of deaths each year. It encompasses various conditions affecting the heart, including coronary artery disease, arrhythmias, and heart valve disorders. Early detection and intervention are crucial in managing heart disease and improving patient outcomes.With the rise of data analytics and machine learning, predicting heart disease has become more feasible and effective. These technologies leverage vast amounts of medical data—such as patient demographics, medical history, lifestyle choices, and laboratory results—to identify risk factors and predict the likelihood of developing heart conditions.

**1.1PURPOSE AND OBJECTIVES:**

**PURPOSE:**

The purpose of this project is to develop a comprehensive system that leverages advanced data analytics and machine learning to enhance the diagnosis of cardiovascular diseases. The system aims to provide accurate, efficient, and early detection of CVD, improving patient outcomes and aiding healthcare professionals in making informed decisions.

**OBJECTIVES:**

1. **Image Quality Assessment:**

○ Develop metrics to evaluate image quality, including sharpness, noise levels, and overall clarity.

○ Implement a standardized framework for assessing improvements in image quality.

1. **Enhancement Algorithms:**

○ Research and apply state-of-the-art image enhancement algorithms, such as deep learning-based techniques, to improve image resolution and clarity.

1. **Real-Time Processing:**

○ Create optimized algorithms that enable real-time image processing for applications requiring immediate feedback.

1. **User-Friendly Interface:**

○ Design an intuitive user interface that allows users to easily upload images, apply enhancement techniques, and view results.

1. **Performance Evaluation:**

○ Conduct rigorous testing to benchmark the performance of the developed techniques against existing solutions.

1. **Application Development:**

○ Explore potential applications for enhanced images in fields such as photography, medical imaging, satellite imagery, and security.

○ Collaborate with stakeholders to identify specific use cases and tailor solutions to meet their needs.

1. **Documentation and Training:**

○ Provide comprehensive documentation on the techniques and tools developed, including usage guidelines and best practices.

**1.2 EXISTING SYSTEM AND PROPOSED SYSTEM:**

**EXISTING SYSTEM:**

**1. Logistic Regression:**

Logistic regression is a popular algorithm for binary classification tasks, particularly well-suited for predicting whether a patient is at risk of heart disease or not. It models the probability of a binary outcome (e.g., heart disease presence) using a logistic function, which outputs a value between 0 and 1. The algorithm uses features such as age, cholesterol levels, and blood pressure to calculate the probability of heart disease. While logistic regression is simple and interpretable, it assumes a linear relationship between the input features and the log-odds of the outcome. This makes it effective for datasets where the relationship between features and heart disease risk is approximately linear, but it may struggle with more complex relationships.

**2. Decision Trees:**

Decision trees are a straightforward and interpretable machine learning algorithm used for both classification and regression tasks. In heart disease prediction, the algorithm splits the data based on specific features (e.g., cholesterol or age) at each node, creating a tree-like structure of decision rules. The decision-making process is intuitive and easy for doctors to understand, as each path from the root to a leaf node represents a series of questions and outcomes that lead to a heart disease prediction. However, decision trees can suffer from overfitting, where the model becomes too tailored to the training data, reducing its generalizability to unseen cases. Pruning techniques or limiting the depth of the tree can help reduce overfitting.

**3. Random Forest:**

Random Forest is an ensemble learning algorithm that creates a "forest" of decision trees to improve prediction accuracy. Each tree is trained on a random subset of the data and uses a random subset of features at each node, which helps reduce overfitting compared to a single decision tree. In heart disease prediction, Random Forest is highly effective because it handles large datasets with many features and can model complex, non-linear relationships between input variables (e.g., age, cholesterol, smoking habits) and heart disease risk. By averaging the predictions of many trees, Random Forest provides a robust and accurate prediction model, but it can be less interpretable compared to a single decision tree.

**4. Support Vector Machines (SVM):**

Support Vector Machines (SVM) is a powerful classification algorithm used in heart disease prediction to find the optimal hyperplane that best separates patients into "at risk" and "not at risk" groups. SVM is particularly useful in high-dimensional spaces, where it can find complex boundaries between classes. By using kernel functions, SVM can handle non-linear relationships in the data, making it suitable for heart disease prediction where the relationships between risk factors may be complex. However, SVM can be computationally expensive and may require careful parameter tuning to achieve optimal performance, particularly with large datasets.

**5. Neural Networks:**

Neural networks are a class of deep learning algorithms that mimic the structure of the human brain, consisting of layers of interconnected neurons that process input features. In heart disease prediction, neural networks can model highly complex and non-linear relationships between medical features (such as ECG results, cholesterol, and blood pressure) and the risk of heart disease. A deep neural network can automatically learn feature interactions, making it especially powerful when large amounts of data are available. However, neural networks require significant computational resources for training and are prone to overfitting if not properly regularized. Additionally, they are often considered "black-box" models due to their lack of interpretability, making it difficult to understand how specific predictions are made.

**PROPOSED SYSTEM:**

**Key Components of the Proposed System:**

1. **Data Collection Module**: This module serves as the backbone of the system, gathering a wide array of patient information. It collects data from clinical visits, lab reports, and patient questionnaires regarding lifestyle choices like smoking, diet, and physical activity levels. Ensuring the data is sourced from multiple reliable channels increases robustness and accuracy. The collected data undergoes rigorous cleaning and preprocessing to remove inconsistencies and fill in missing values, ensuring high-quality data is utilized for analysis.
2. **Feature Selection and Engineering**: After data collection, the system employs feature selection techniques to pinpoint the most significant factors affecting heart disease risk. This may involve statistical methods like correlation analysis or advanced techniques like recursive feature elimination. Feature engineering is also applied to create new variables that capture hidden patterns, such as interaction terms between cholesterol levels and age. This ensures that the model has the best possible predictors, enhancing the system's effectiveness.
3. **Machine Learning Model**: The heart of the proposed system is its machine learning model, which can be tailored based on specific needs.
   * **K-Nearest Neighbors (KNN)**: This algorithm classifies patients based on the proximity of their data to that of their nearest neighbors in the feature space. KNN is particularly useful for heart disease prediction as it can handle non-linear relationships and is intuitive to understand. The model requires careful tuning of the parameter 'k,' which determines the number of neighbors considered for classification. The choice of distance metric (e.g., Euclidean, Manhattan) is also critical for the model's performance.
   * **Logistic Regression**: Suitable for simpler, linear relationships between risk factors and the likelihood of heart disease.

The model is trained on a diverse dataset of historical patient records, with performance metrics such as accuracy, precision, recall, and F1-score used to evaluate efficacy. The system continually updates to reflect new data and improve its predictive capabilities.

1. **Prediction Module**: Once trained, the system's prediction module takes in patient data and generates real-time risk assessments. It categorizes patients into risk levels—low, moderate, or high—based on output probabilities. This risk stratification allows healthcare providers to prioritize high-risk patients for immediate attention, guiding them toward necessary interventions or further diagnostic testing. The prediction module also provides insights into the specific factors contributing to each patient's risk, empowering doctors with actionable information.
2. **User Interface**: The system includes a user-friendly interface designed for seamless interaction by healthcare providers. This interface allows clinicians to input patient data effortlessly and receive instant risk predictions. Additionally, it displays visualizations such as risk scores, trends over time, and comparative analyses with population data, facilitating informed discussions between doctors and patients. By presenting data in an intuitive format, the interface enhances user experience and encourages system adoption in clinical settings.
3. **Feedback and Model Improvement**: To ensure continuous improvement, a feedback mechanism is built in. As predictions are made, outcomes are monitored against actual patient results. This feedback loop allows the system to adjust and retrain the machine learning model, enhancing predictive accuracy over time. By integrating new data and user feedback, the system evolves to reflect the latest medical knowledge and practices.

**Benefits of proposed system**:

* **Early Detection**: The system's ability to analyze a comprehensive set of risk factors facilitates early identification of individuals at risk for heart disease, which is crucial for timely intervention.
* **Efficiency**: By automating the risk assessment process, the system significantly reduces the workload on healthcare professionals, allowing them to focus on patient care rather than manual data analysis.
* **Personalized Care**: With precise risk assessments, healthcare providers can tailor treatment plans to individual patients, optimizing care and improving health outcomes.
* **Scalability**: The system is designed to handle large volumes of patient data, making it suitable for deployment in various healthcare settings, including hospitals, outpatient clinics, and telemedicine platforms.

By harnessing the power of machine learning—including K-Nearest Neighbors—along with data analytics, the proposed heart disease prediction system aims to revolutionize cardiovascular care. It empowers healthcare professionals with timely insights and enhances patient management, ultimately contributing to improved public health outcomes through proactive prevention and personalized treatment strategies.

**1.2 SCOPE OF THE PROJECT:**

**Inclusions:**

* **Diverse Patient Populations**: Including patients from various demographics (age, gender, ethnicity) to ensure the model is representative and can generalize across different populations.
* **Comprehensive Health Data**: Utilizing a broad range of data, including electronic health records, lifestyle factors, genetic information, and socio-economic status, to enhance the model's predictive capabilities.
* **Longitudinal Data**: Incorporating data from long-term studies to track changes in health over time, allowing for more accurate predictions regarding heart disease progression and risk factors.
* **Pre-existing Conditions**: Including patients with known risk factors (e.g., hypertension, diabetes) as well as those without to create a robust understanding of how various conditions contribute to heart disease risk.
* **Access to Technology**: Patients who utilize wearable devices or telehealth services, providing real-time health data that can improve the accuracy of predictions.
* Development of noise reduction, sharpening, super-resolution, and color correction algorithms.
* Focus on digital photographs, medical images, satellite imagery, and scanned documents.
* Creation of a user-friendly application compatible with common image formats.
* Design of an intuitive interface for easy navigation and batch processing capabilities.
* Establishment of benchmarks and user feedback collection to refine algorithms.
* Comprehensive user manuals and tutorials for effective software use.

**Exclusions:**

* **Inconsistent Data Quality**: Excluding data with significant missing values, errors, or inconsistencies that could skew the model’s outcomes.
* **Short Follow-Up Duration**: Excluding studies or data from patients who have not been followed long enough to provide meaningful insights into heart disease progression.
* **Lack of Diagnostic Confirmation**: Excluding patients without a confirmed diagnosis of heart disease or relevant medical conditions to ensure the model focuses on accurately predicting outcomes.
* **Special Populations**: Excluding groups that may require separate analysis, such as pregnant women or individuals with rare diseases, to maintain the model's relevance and accuracy for the general population.
* **Ethical Concerns**: Excluding data from populations that may not have provided informed consent or where ethical considerations prevent the use of certain data sets.
* No real-time video or live stream enhancement.
* Focus solely on software solutions, no hardware integration.
* Initial scope does not include commercial deployment or sales.
* No specific adherence to industry regulations or standards.

**CHAPTER 2**

**LITERATURE SURVEY**

**Alotaibi, F. S. (2019) et al. [1]** Implementation of machine learning model to predict heart failure disease. International Journal of Advanced Computer Science and People with diabetes require lifelong access to healthcare services to delay the onset of complications. Their disease management processes generate great volumes of data across several domains, from clinical to administrative. Difficulties in accessing and processing these data hinder their secondary use in an institutional setting, even for highly desirable applications, such as the prediction of cardiovascular disease, the main driver of excess mortality in diabetes.

**Kim, J. K., & Kang, S. (2017) et al. [2]** Neural network-based coronary heart disease risk prediction using feature correlation analysis. Hindawi Journal of Healthcare Engineering. Machine learning is widely used to make the machine to learn and to predict when exposed to new data. Due to many advancements in machine learning, there are various methods that can be adopted to predict the heart disease of an individual. Heart Disease is one among the major diseases affecting the individual around the world. The risk factors which are not having the significant impact are identified and removed. The resultant significant factors are provided as input to the neural network. And neural network is trained for the risk factors that is obtained from logistic regression and used to test whether the person is having the heart disease or not

**Haq, A. U., Li, J.-P., Memon, M. H., Nazir, S., & Sun, R. (2018) et al. [3]** A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Hindawi Mobile Information System. The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time. In this paper different machine learning algorithms and deep learning are applied to compare the results and analysis of the UCI Machine Learning Heart Disease dataset. The dataset consists of 14 main attributes used for performing the analysis. Various promising results are achieved and are validated using accuracy and confusion matrix. Using deep learning approach, 94.2% accuracy was obtained.

**Jagtap, A., Malewadkar, P., Baswat, O., & Rambade, H. (2019) et al. [4]** Heart disease prediction using machine learning. International Journal of Research in Engineering, science and Management, 2 (2), 352-355.Healthcare occupies an indispensable part in human lives. The healthcare industry contains large amount of psychiatric data hence machine learning models were used to provide conclusion effectively in the heart disease prediction. The classification of healthy person and non-healthy person can be done reliably by using machine learning methods. DNN and ANN were used to analyse the efficiency of the model which accurately predicts the presence or absence of heart disease.

**Khan, S. N., Nawi, N. M., Shahzad, A., Ullah, A., & Mushtaq, M. F. (2019) et al. [5]** Comparative analysis for heart disease prediction. International Journal on Informatics Visualization, 1 (4-2), 227-231.Heart disease prediction is one among the foremost complicated tasks in medical field. This paper makes use of heart condition dataset available in UCI machine learning repository. The proposed work predicts the probabilities of heart condition and classifies patient's risk level by implementing different data processing techniques like Naive Bayes, Decision Tree, Logistic Regression and Random Forest. Thus, this paper presents a comparative study by analysing the performance of various machine learning algorithms. The trial result verifies that Random Forest algorithm has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented.

**Bo Jin, Chao Che (2018) et al. [6]** proposed a “Predicting the Risk of Heart Failure With EHR Sequential Data Modeling” model designed by applying neural network. This paper used the electronic health record (EHR) data from real-world datasets related to congestive heart disease to perform the experiment and predict the heart disease before itself. We tend to used one-hot encryption and word vectors to model the diagnosing events and foretold coronary failure events victimization the essential principles of an extended memory network model. By analyzing the results, we tend to reveal the importance of respecting the sequential nature of clinical records.

**Aakash Chauhan (2018) et al. [7]** presented “Heart Disease Prediction using Evolutionary Rule Learning”. This study eliminates the manual task that additionally helps in extracting the information (data) directly from the electronic records. To generate strong association rules, we have applied frequent pattern growth association mining on patient’s dataset. This will facilitate (help) in decreasing the amount of services and shown that overwhelming majority of the rules helps within the best prediction of coronary sickness.

**Ashir Javeed, Shijie Zhou (2017) et al. [8]** designed “An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection”. This paper uses random search algorithm (RSA) for factor selection and random forest model for diagnosing the cardiovascular disease. This model is principally optimized for using grid search algorithmic program. Two forms of experiments are used for cardiovascular disease prediction. In the first form, only random forest model is developed and within the second experiment the proposed Random Search Algorithm based random forest model is developed. This methodology is efficient and less complex than conventional random forest model. Comparing to conventional random forest it produces 3.3% higher accuracy. The proposed learning system can help the physicians to improve the quality of heart failure detection.

**Senthilkumar Mohan, Chandrasegar Thirumalai (2019) et al. [9]** was efficient technique using hybrid machine learning methodology. The hybrid approach is combination of random forest and linear method. The dataset and subsets of attributes were collected for prediction. The subset of some attributes were chosen from the pre-processed knowledge(data) set of cardiovascular disease. After prep-processing ,the hybrid techniques were applied and disgnosis the cardiovascular disease.

**K.Prasanna Lakshmi, Dr. C.R.K.Reddy (2015) et al. [10]** designed “Fast Rule-Based Heart Disease Prediction using Associative Classification Mining”. In the proposed Stream Associative Classification Heart Disease Prediction (SACHDP), we used associative classification mining over landmark window of data streams. This paper contains two phases: one is generating rules from associative classification mining and next one is pruning the rules using chi-square testing and arranging the rules in an order to form a classifier. Using these phase to predict the heart disease easily.

## CHAPTER 3

## SYSTEM ANALYSIS

System requirements are the functionality that is needed by a system in order to satisfy the customer's requirements. System requirements are abroad and a narrow subject that could be implemented to many items. The requirements document allows the project team to have a clear picture of what the software solution must do before selecting a vendor. Without an optimized set of future state requirements, the project team has no effective basis to choose the best system for your organization.

**2.1 SOFTWARE AND HARDWARE REQUIREMENTS:**

**SOFTWARE REQUIREMENTS:**

* **Operating System:** Windows 10+
* **Programming Language:** Python 3.8+
* **Frameworks:** TensorFlow 2.x / PyTorch, OpenCV
* **IDE:** Visual Studio Code, Jupyter Notebook
* **Libraries:** TensorFlow/PyTorch, OpenCV, Matplotlib, NumPy, Pandas
* **Additional Software:** Scikit-Learn, SQL databases, Apache Hadoop

**HARDWARE REQUIREMENTS:**

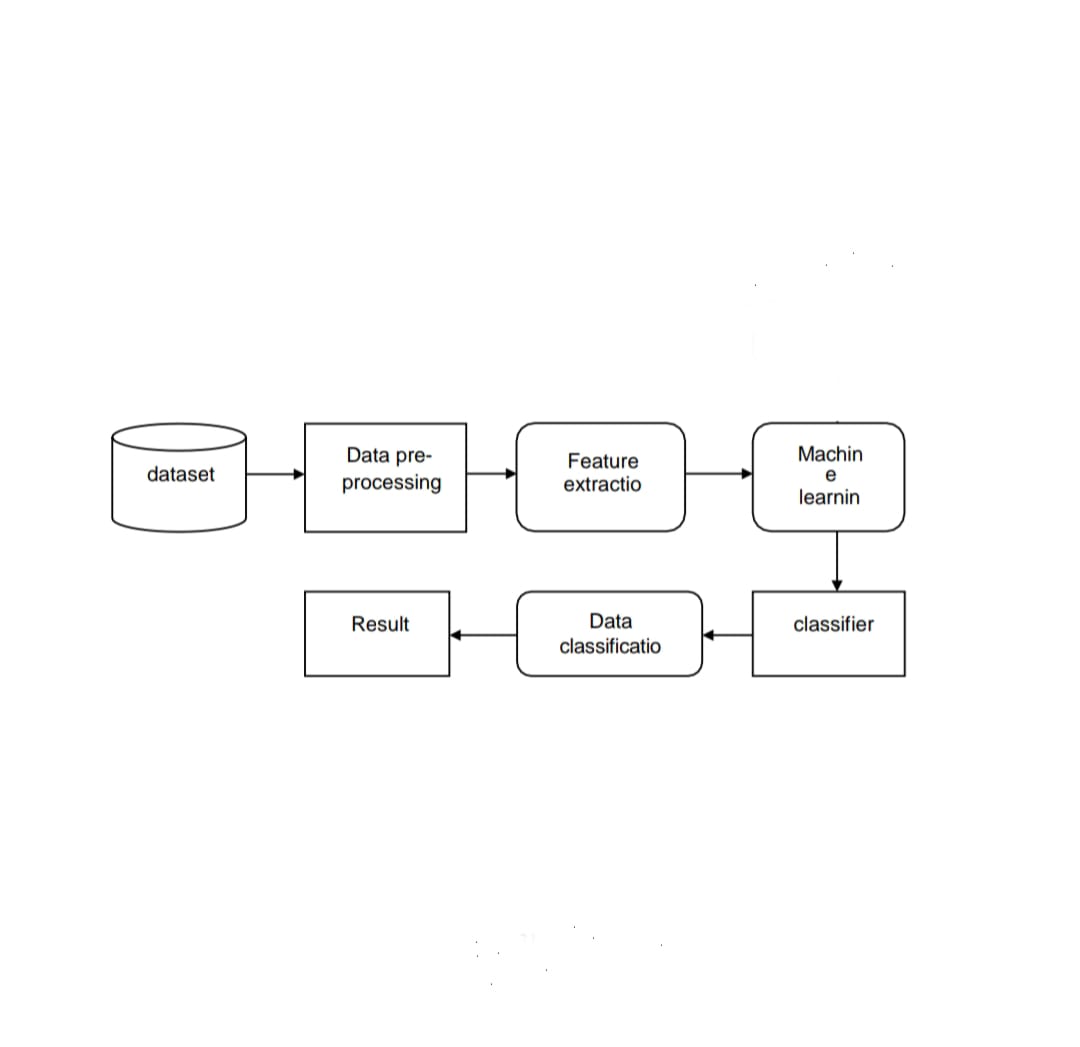
* **RAM:** 16 GB minimum (32 GB recommended)
* **Processor:** Intel Core i7+ / AMD Ryzen 7+
* **GPU:** NVIDIA GTX 1080 Ti+ (for deep learning)
* **Storage:** SSD, 256 GB free (1 TB recommended)
* **Display:** Full HD monitor
* **High-Performance Servers:** GPUs for machine learning model training
* **Cloud Storage Solutions:** Aws, Google Cloud

**CHAPTER 4**

**SYSTEM DESIGN**

**4.1** **ARCHITECTURE DIAGRAM:**

The architecture includes a Data Collection Layer that gathers patient data from various sources like electronic health records and lab results. The Data Preprocessing Layer cleans and normalizes the data, ensuring quality and consistency. The Machine Learning Model Layer houses algorithms such as KNN, Random Forest, and logistic regression for analyzing the data .Finally, the Feedback and Model Improvement Layer collects outcome data to enhance model accuracy over time.



**Fig. 4.1.1 Architecture image**

**4.2 UML DIAGRAMS:**

UML Diagrams are classified into different types such as :

1. USE CASE DIAGRAM
2. CLASS DIAGRAM
3. SEQUENCE DIAGRAM
4. ACTIVITY DIAGRAM

**CLASS DIAGRAM:**

A diagram of a dataset

Description automatically generatedThe class diagram defines the structure of the system. It includes classes like Patient, Doctor, and Prediction Model, showing their attributes and methods, as well as how they are related. For example, a doctor can have multiple patients, and the prediction model processes patient data to predict heart disease risk. The **class diagram** defines the system's structure by showing the key classes, their attributes, and the relationships between them. For example, the **Patient** class contains attributes such as age, gender, cholesterol, and blood Pressure, which represent the patient's medical data. The **Doctor** class has attributes like doctor ID and specialization, which identify the doctor using the system. The core class, **Prediction Model**, includes the algorithm used to predict heart disease risk based on patient data.

**Fig 4.2.1 Class diagram**

**USE CASE DIAGRAM:**

The **use case diagram** helps visualize how different users (or actors) interact with the heart disease prediction system. The primary actors include doctors, patients, and administrators. Doctors use the system to input patient data like medical history, test results, and symptoms. Once the data is entered, they can run a prediction model to determine the patient's likelihood of heart disease. The patients may have limited interaction, mainly to access prediction results or their medical records. Administrators manage the system by handling user accounts and ensuring data privacy and system functionality. This diagram clearly shows which users perform specific actions within the system, making it easy to understand user roles and functionality.

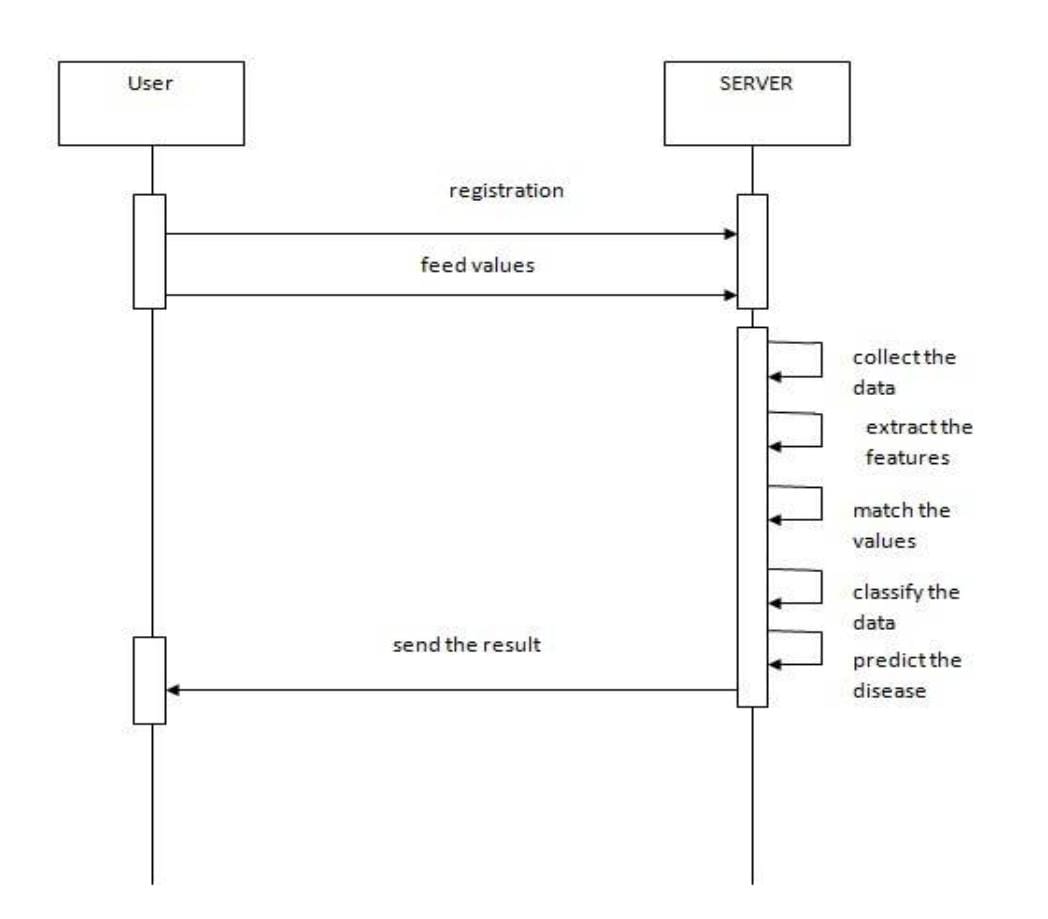
A diagram of a data flow

Description automatically generated

**Fig 4.2.2 Use case diagram**

**SEQUENCE DIAGRAM:**

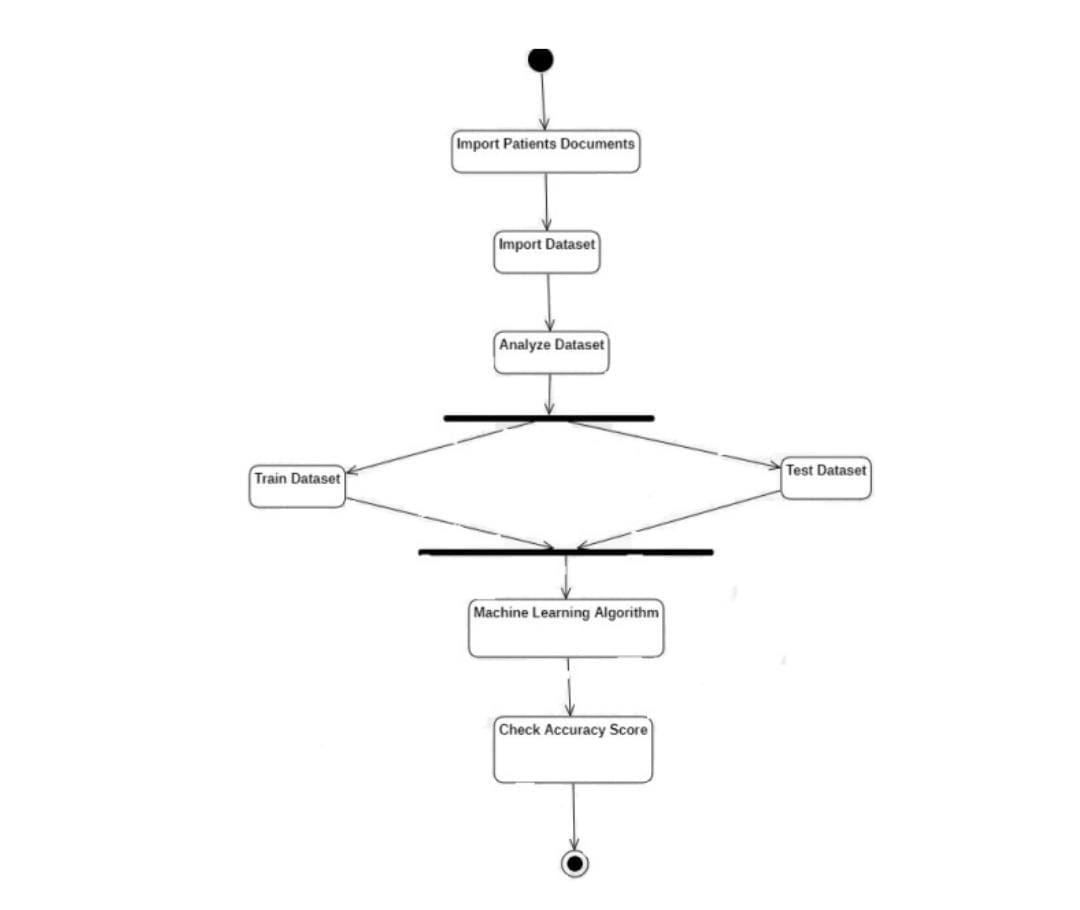
The sequence diagram details the order of interactions between the doctor, the heart disease prediction system, and the prediction model. The doctor inputs data, the system processes it via the prediction model, and the results are sent back to the doctor. The **sequence diagram** focuses on the time-ordered interactions between the system components and users. It begins with the doctor requesting to input patient data. The heart disease prediction system processes this request, sending the data to the **Prediction Model** component, which applies the algorithm to analyze the data. After the prediction result is returned to the system and finally displayed to the doctor.



**Fig 4.2.3 Sequence diagram**

**ACTIVITY DIAGRAM:**

The Activity Diagram demonstrates the step-by-step flow of the heart disease prediction process, from data input to the final prediction result. The Activity Diagram clearly illustrates the flow of the process, showing how data moves through different phases and how decisions (e.g., whether to retrain the model) affect the overall process. The arrows between actions show the transition from one stage to the next, providing a clear visual representation of how the system works from input to output.



**Fig 4.2.4 Activity diagram**

**CHAPTER 5**

**METHODOLOGY**

**5.1 TECHNOLOGIES USED:**

1. **Programming Language**:

* **Python 3.8+**: The primary programming language used for implementing machine learning models and data analysis. Python offers a wide range of libraries essential for machine learning and data preprocessing.

1. **Frameworks and Libraries**:

* **TensorFlow 2.x / PyTorch**: Used for building machine learning models, particularly for neural networks and other deep learning models.
* **OpenCV**: An open-source library focused on real-time computer vision, mainly for image processing tasks.
* **Scikit-Learn**: A robust library for machine learning models such as Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN).
* **Matplotlib**: For visualizing data and results, particularly for plotting graphs and metrics.
* **NumPy**: For numerical computations, handling large multi-dimensional arrays, and matrices.
* **Pandas**: For data manipulation and analysis, especially for handling tabular data efficiently.

1. **IDE and Tools:**

* **Visual Studio Code / Jupyter Notebook**: Used for code development, testing, and visualization of machine learning models and data preprocessing.
* **SQL Databases**: For managing and storing patient data securely.
* **Apache Hadoop**: A framework used for processing large datasets in a distributed computing environment.

1. **Hardware Requirements**:

* **RAM**: Minimum 16 GB, recommended 32 GB, especially for handling large datasets.
* **Processor**: Intel Core i7+ or AMD Ryzen 7+ for faster processing.
* **GPU**: NVIDIA GTX 1080 Ti+ for deep learning and accelerated computations.
* **Storage**: SSD with at least 256 GB free, recommended 1 TB.
* **Cloud Services**: AWS or Google Cloud for scaling and high-performance computing resources.

**5.2 MODULE DESCRIPTION:**

1. **Data Collection Module**:

* **Functionality**: Gathers patient data from various clinical sources like lab reports and medical history.
* **Process**: The module collects structured and unstructured data, ensuring robustness and completeness of the dataset.
* **Significance**: This module is critical for ensuring accurate and comprehensive data is available for analysis.

1. **Data Preprocessing Module**:

* **Functionality**: Handles missing data, normalization, and feature scaling to prepare the data for analysis.
* **Process**: Involves cleaning data, removing outliers, normalizing features (e.g., blood pressure, cholesterol), and splitting the dataset into training and testing sets.

1. **Feature Selection and Engineering Module**:

* **Functionality**: Identifies the most important features influencing the prediction of cardiovascular disease.
* **Process**: Applies statistical methods such as correlation analysis and Recursive Feature Elimination (RFE) to filter out less relevant features.
* **Significance**: This step improves the model’s efficiency and accuracy by focusing on key predictive features.

1. **Machine Learning Model Module**:

* **Algorithms Implemented**:
  + **Logistic Regression**: For binary classification, predicting whether a patient is at risk of heart disease.
  + **Random Forest**: A robust ensemble method using multiple decision trees to improve prediction accuracy.
  + **K-Nearest Neighbors (KNN)**: For classifying patients based on proximity to others in the feature space.
* **Functionality**: Trains the model on historical data and evaluates performance using metrics like accuracy, precision, recall, and F1-score.

1. **Prediction Module**:

* **Functionality**: Provides real-time cardiovascular disease risk assessment based on input data.
* **Output**: Generates a risk score (low, moderate, or high) for each patient.
* **Significance**: Enables healthcare providers to make informed decisions regarding patient care and prioritization.

1. **User Interface Module**:

* **Functionality**: A user-friendly interface that allows healthcare professionals to input patient data and receive predictions.
* **Features**: Displays risk scores, graphical insights, and trends over time for better patient risk management.
* **Significance**: Simplifies the interaction between healthcare providers and the prediction system, enhancing usability and efficiency.

1. **Feedback and Model Improvement Module**:

* **Functionality**: Continuously monitors the model’s performance and retrains it with new data to improve prediction accuracy over time.
* **Process**: Collects feedback on predictions and compares them with actual patient outcomes, updating the model as necessary.
* **Significance**: Ensures that the system remains accurate and up-to-date with the latest medical data and advancements.

1. **Learning Objectives:**

* **Understand the Basics of Machine Learning**: Grasp core ML concepts, including supervised and unsupervised learning, and familiarize yourself with key algorithms commonly used in healthcare.
* **Data Collection and Preprocessing**: Learn techniques for collecting and preparing diverse datasets, including electronic health records, demographic information, and real-time data from wearable devices.
* **Feature Engineering**: Gain skills in selecting and transforming relevant features to improve model accuracy, including the use of domain knowledge to identify important risk factors for heart disease.
* **Model Development and Evaluation**: Explore various ML algorithms such as logistic regression, decision trees, random forests, and neural networks. Understand how to evaluate model performance using metrics like accuracy, precision, recall, and ROC-AUC.
* **Interpretability of Models**: Learn techniques for interpreting ML models to ensure transparency and facilitate communication with healthcare providers and patients.
* **Ethical and Regulatory Considerations**: Discuss the ethical implications of using ML in healthcare, including data privacy, algorithmic bias, and regulatory frameworks that govern the deployment of predictive models.
* **Case Studies and Practical Applications**: Analyze real-world case studies showcasing successful implementations of ML in predicting heart disease, and participate in hands-on projects to apply learned techniques.
* **Future Directions**: Explore emerging trends in ML for heart disease prediction, including the integration of big data, advancements in wearable technology, and the potential for personalized medicine.

1. **Resources:**

* Recommended readings on machine learning and cardiovascular health
* Software tools and libraries for ML (e.g., Scikit-learn, TensorFlow)
* Online forums for discussion and collaboration

**CHAPTER 6**

**IMPLEMENTATION**

**6.1 SAMPLE CODE**

**Importing Libraries:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

import os

print(os.listdir())

import warnings

warnings.filterwarnings('ignore')

**Importing Dataset:**

dataset = pd.read\_csv(r"C:\Users\kusum\OneDrive\Desktop\heart.csv")

type(dataset)

dataset**.**shape

dataset**.**head(5)

dataset**.**sample(5)

dataset.describe()

dataset.info()

for i in range(len(info)):

print(dataset.columns[i]+":\t\t\t"+info[i])

dataset["target"].describe()

dataset["target"].unique()

print(dataset.corr()["target"].abs().sort\_values(ascending=False))

**Exploratory Data Analysis (EDA):**

y = dataset["target"]

sns.countplot(y)

target\_temp = dataset.target.value\_counts()

print(target\_temp)

print("Percentage of patience without heart problems: "**+**str(round(target\_temp[0]**\***100**/**303,2)))

print("Percentage of patience with heart problems: "**+**str(round(target\_temp[1]**\***100**/**303,2)))

**Analysing the 'Sex' feature**

dataset["sex"].unique()

sns.barplot(x=dataset["sex"],y=y)

**Analysing the 'Chest Pain Type' feature**

dataset["cp"].unique()

sns.barplot(x=dataset["cp"],y=y)

**Analysing the FBS feature**

dataset["fbs"].describe()

dataset["fbs"].unique()

sns.barplot(x=dataset["fbs"],y=y)

**Analysing the restecg feature**

dataset["restecg"].unique()

sns.barplot(x=dataset["restecg"],y=y)

**Analysing the 'exang' feature**

dataset["exang"].unique()

sns.barplot(x=dataset["exang"],y=y)

**Analysing the Slope feature**

dataset["slope"].unique()

sns.barplot(x=dataset["slope"],y=y)

**Analysing the 'ca' feature**

dataset["ca"].unique()

sns.countplot(dataset["ca"])

sns.barplot(x=dataset["ca"],y=y)

dataset["thal"].unique()

sns.barplot(x=dataset["thal"],y=y)

sns.distplot(dataset["thal"])

**Train Test split**

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

predictors **=** dataset**.**drop("target",axis**=**1)

target **=** dataset["target"]

X\_train,X\_test,Y\_train,Y\_test **=** train\_test\_split(predictors,target,test\_size**=**0.20,random\_state**=**0)

X\_train**.**shape

X\_test**.**shape

Y\_train**.**shape

Y\_test**.**shape

**Model Fitting**

**from** sklearn.metrics **import** accuracy\_score

**Logistic Regression**

from sklearn.linear\_model import LogisticRegression

lr = LogisticRegression()

lr.fit(X\_train,Y\_train)

Y\_pred\_lr = lr.predict(X\_test)

Y\_pred\_lr.shape

score\_lr = round(accuracy\_score(Y\_pred\_lr,Y\_test)\*100,2)

print("The accuracy score achieved using Logistic Regression is: "+str(score\_lr)+" %")

**K Nearest Neighbors**

from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n\_neighbors=7)

knn.fit(X\_train,Y\_train)

Y\_pred\_knn=knn.predict(X\_test)

Y\_pred\_knn.shape

score\_knn = round(accuracy\_score(Y\_pred\_knn,Y\_test)\*100,2)

print("The accuracy score achieved using KNN is: "+str(score\_knn)+" %")

**Output final score**

scores = [score\_lr,score\_knn]

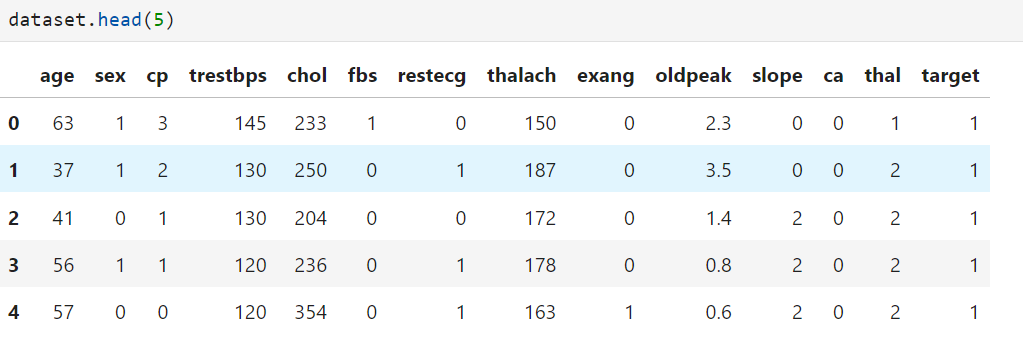
algorithms = ["Logistic Regression","K-Nearest Neighbors"]

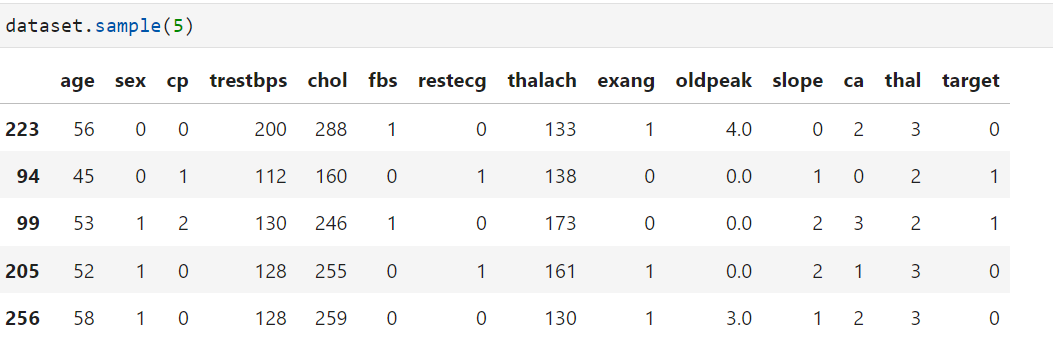
for i in range(len(algorithms)):

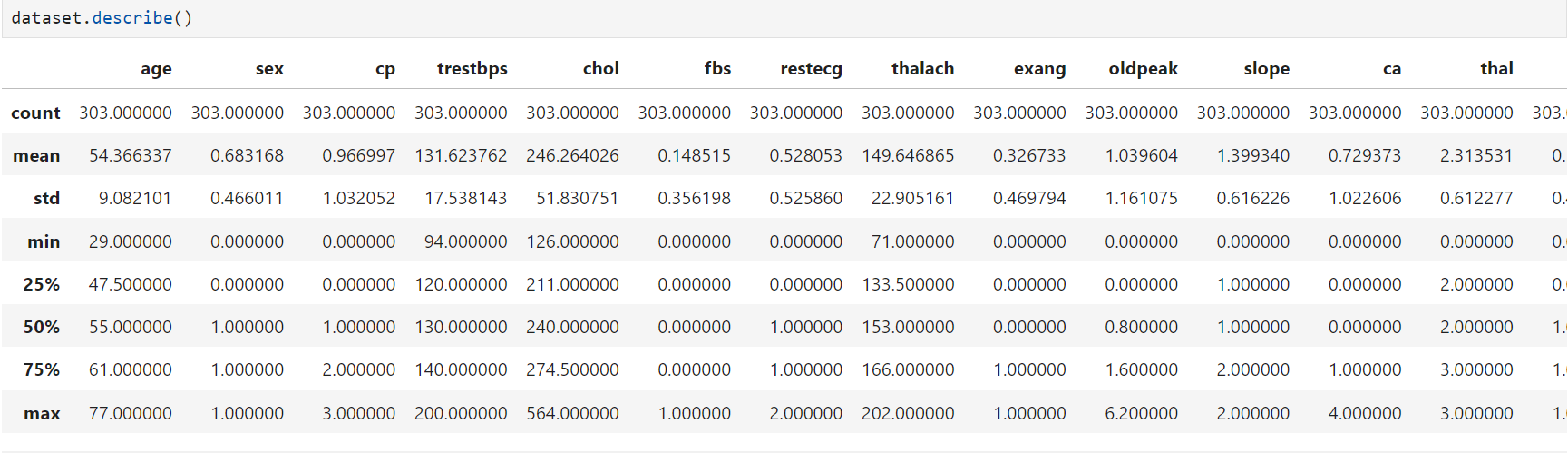
print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")

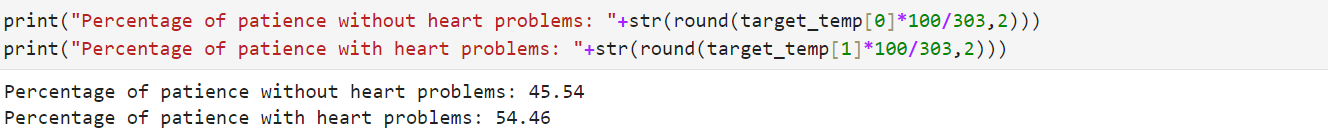
sns.barplot(x=algorithms, y=scores)

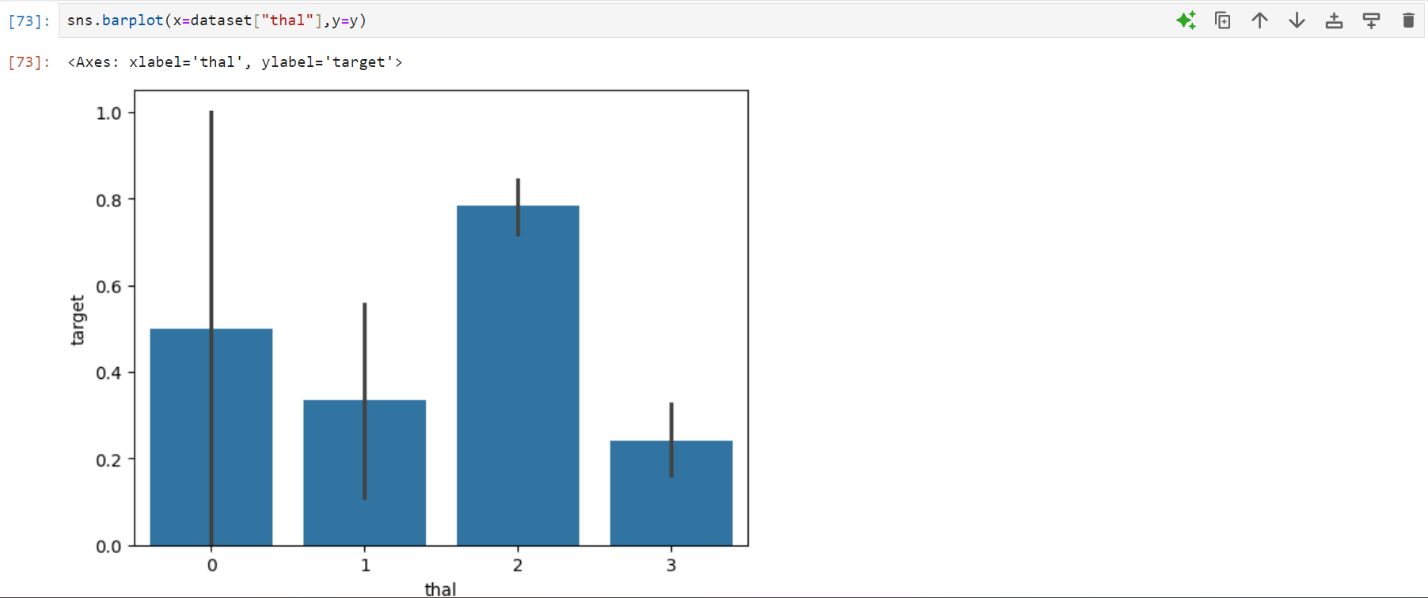
**6.2 OUTPUT SCREENS**

****

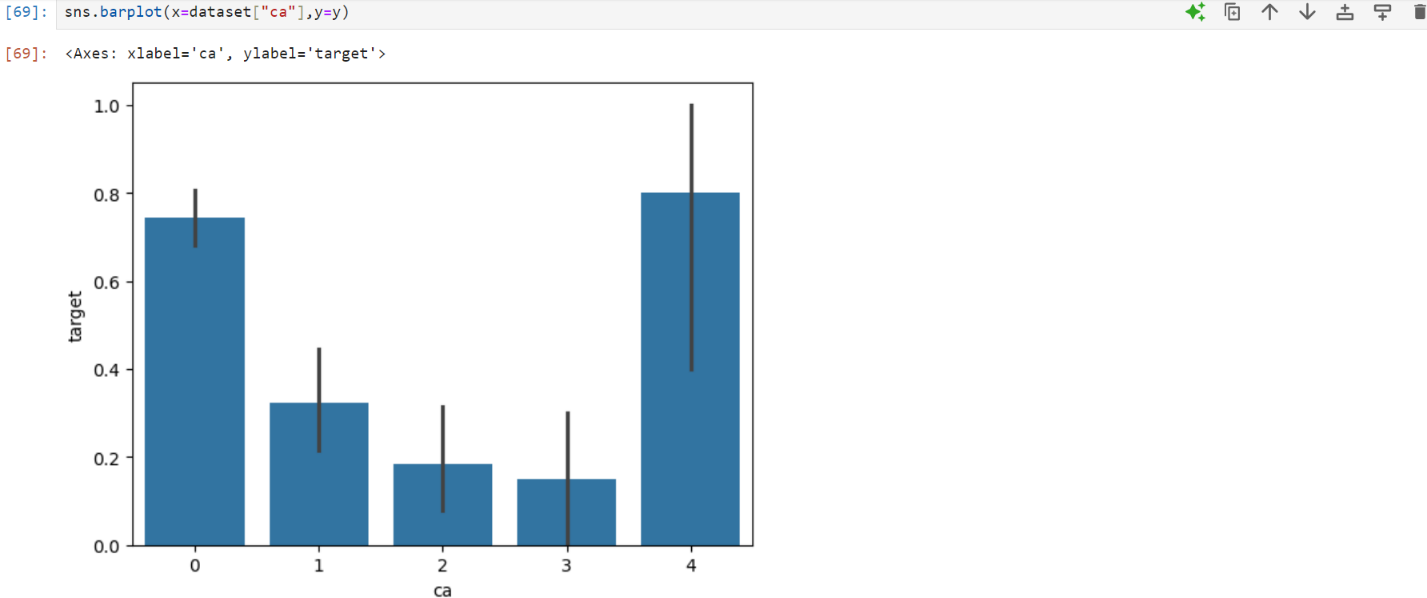
****

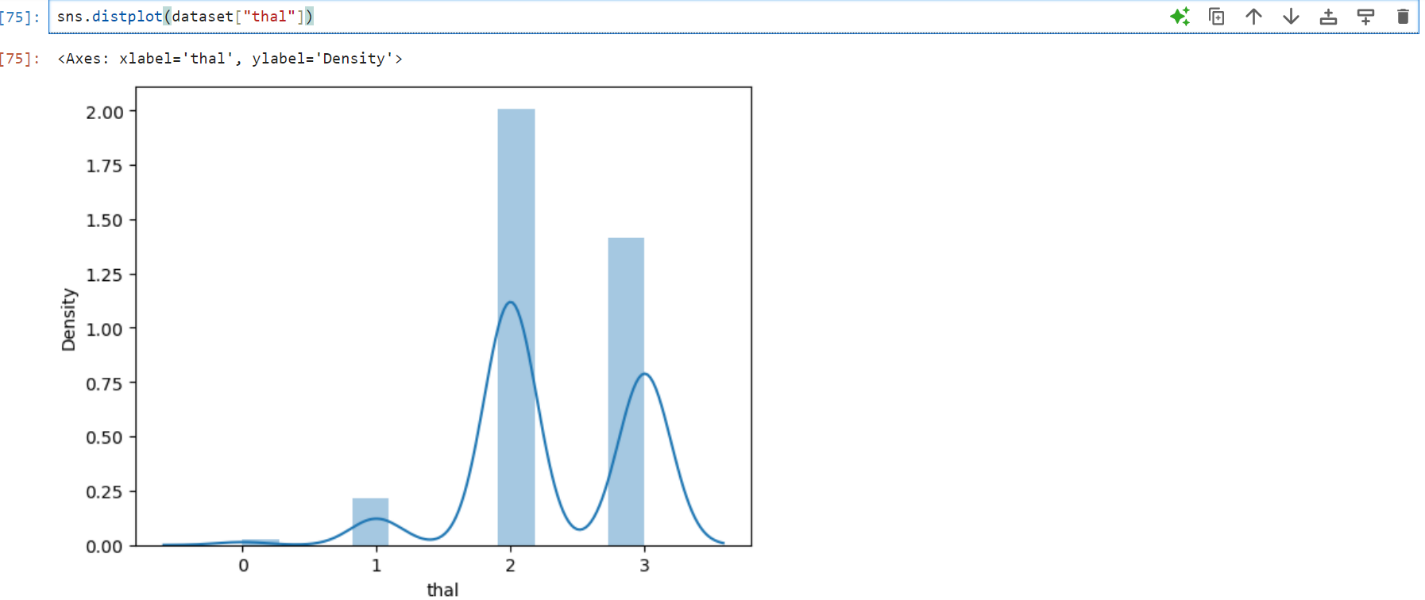
****

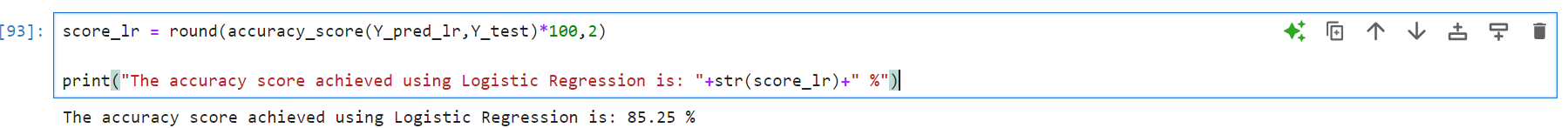
****

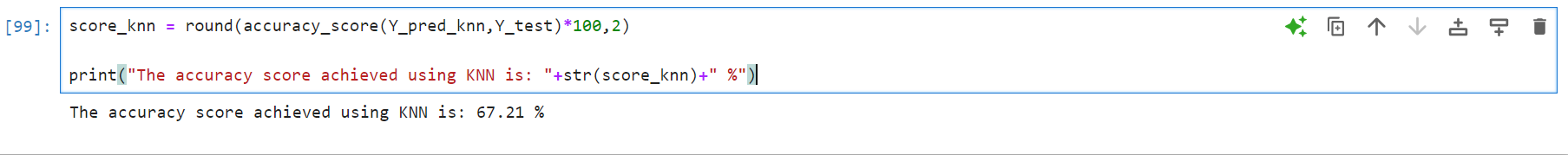
****

****

****

****

****

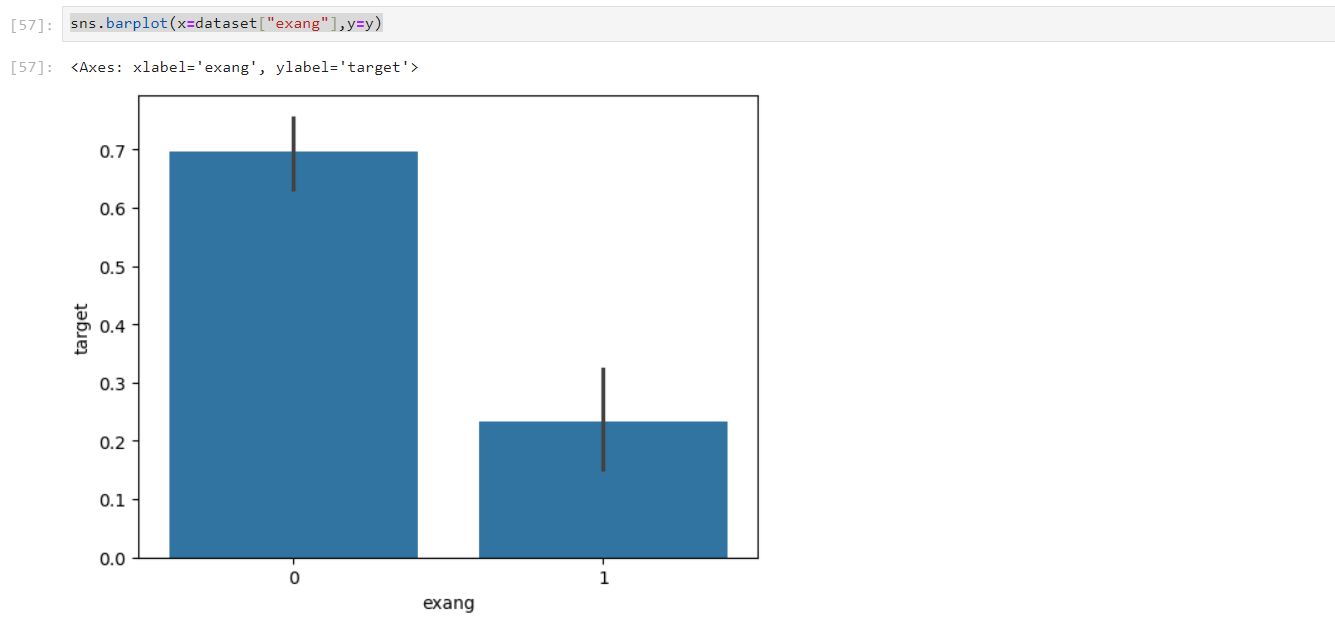
****

**A screenshot of a graph

Description automatically generated**

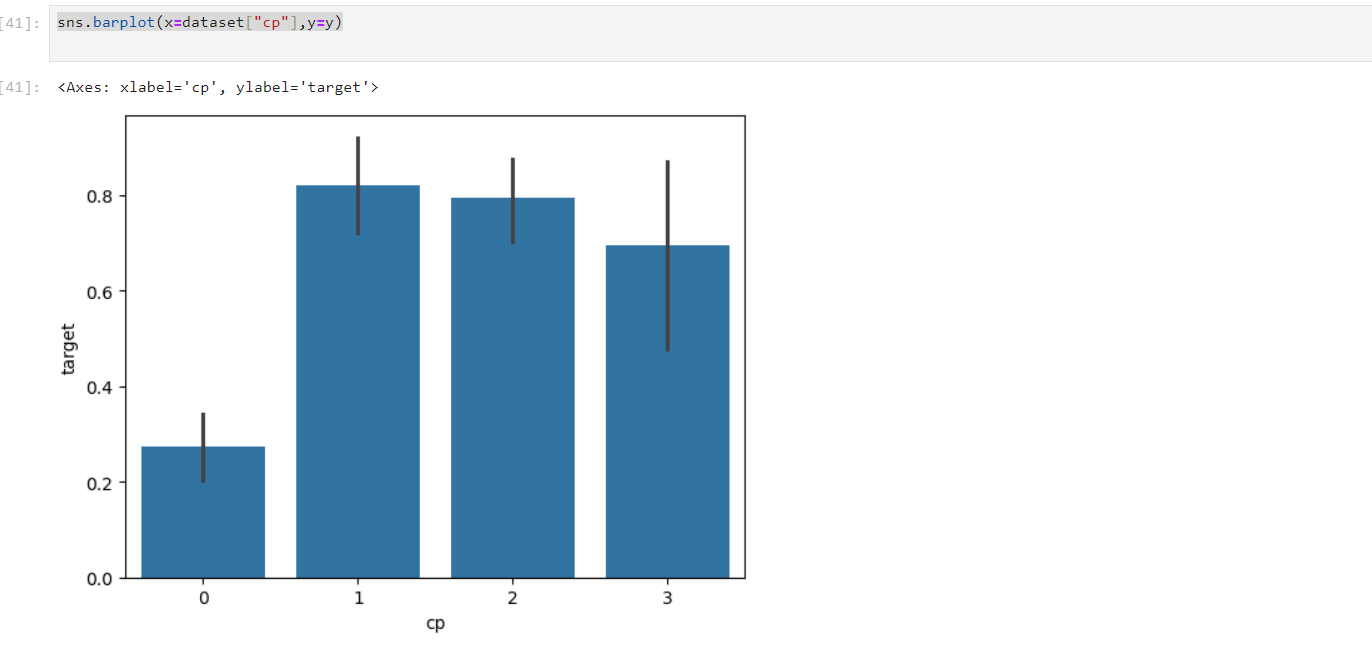
**A screenshot of a graph

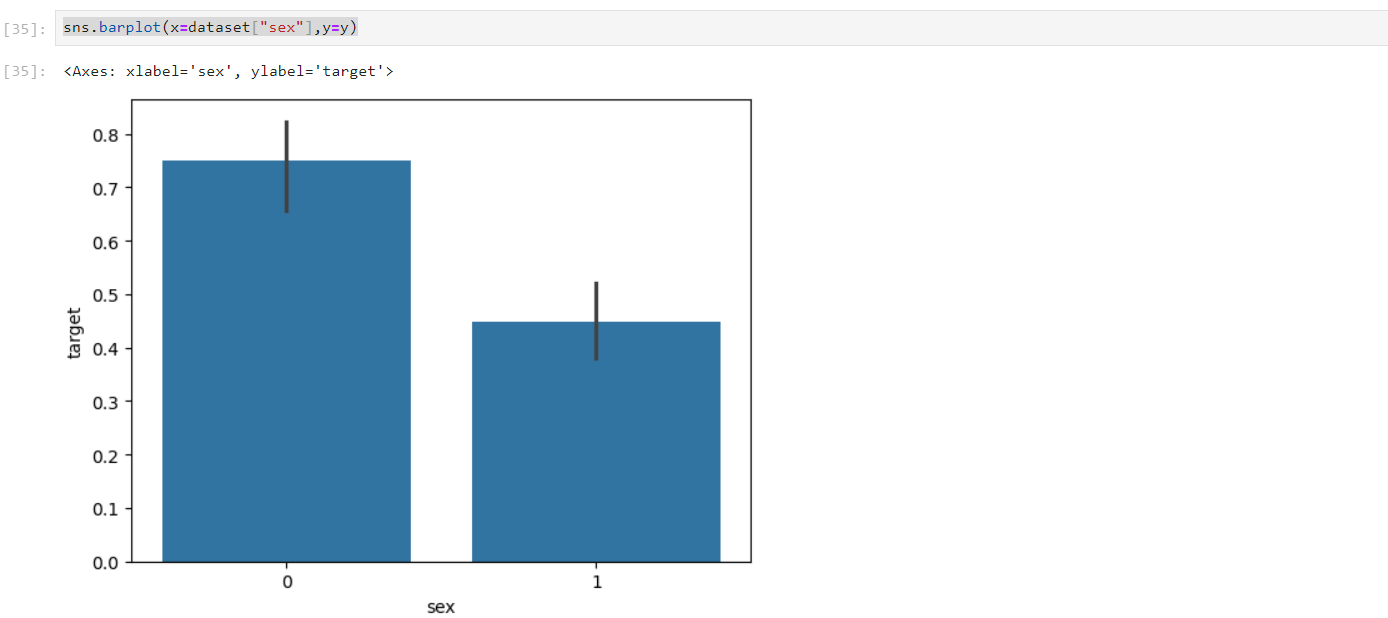
Description automatically generated**

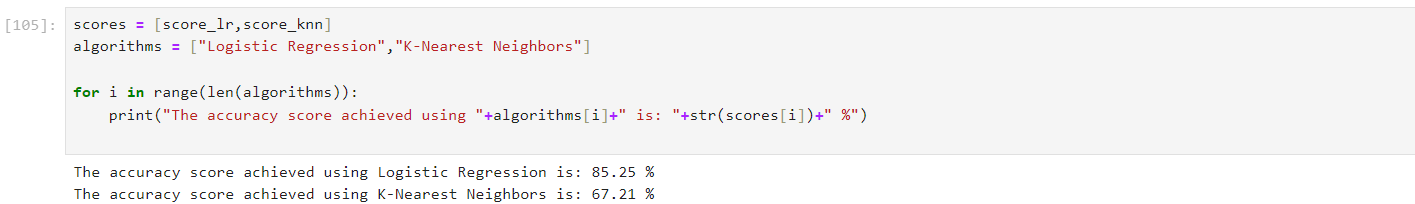
****

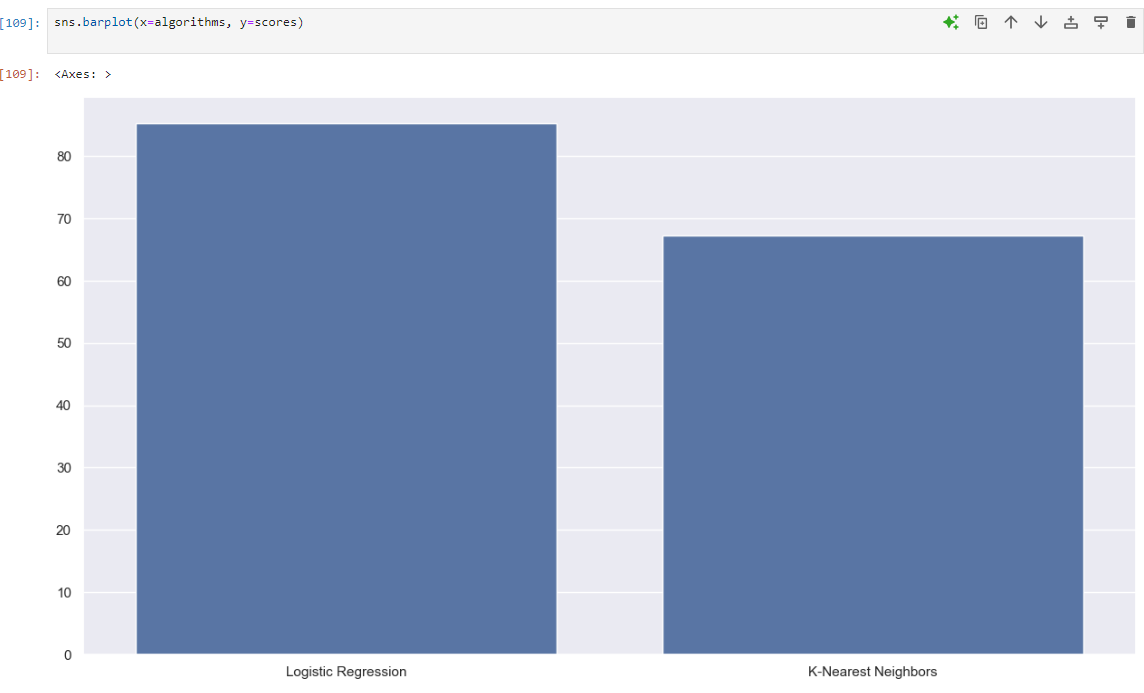
**A screenshot of a graph

Description automatically generated**









**CHAPTER 7**

**CONCLUSION AND FUTURE SCOPE**

**CONCLUSION**

Heart disease prediction using machine learning (ML) has emerged as a transformative approach in modern healthcare, leveraging data-driven techniques to identify patterns that predict the likelihood of heart-related conditions. Traditional diagnostic methods, while effective, often rely heavily on human expertise and can be time-consuming. Machine learning, on the other hand, utilizes vast datasets to automate the prediction process, potentially providing more efficient and accurate results. As heart disease remains a leading cause of death worldwide, incorporating ML tools in prediction models can offer significant advancements in early detection and prevention strategies.

A wide range of ML algorithms, including logistic regression, random forest and KNN algorithms have been applied to predict heart disease with promising results. Each algorithm processes data differently, capturing relationships and patterns in patient datasets that might be overlooked by conventional methods. The strength of machine learning lies in its ability to handle large and complex datasets, including various features like patient demographics, lifestyle habits, medical history, and lab results, providing a comprehensive risk assessment.

One of the key advantages of ML-based heart disease prediction is the ability to continuously learn from new data. With more medical institutions digitizing patient records and collecting health-related data through wearables, the influx of data allows models to evolve, improving accuracy and reducing prediction errors over time. This dynamic learning aspect can lead to personalized predictions for individual patients, tailoring recommendations for lifestyle changes or medical interventions based on evolving health data.

Despite the impressive capabilities of machine learning, the quality and reliability of predictions are highly dependent on the data used to train the models. Poor-quality data, such as incomplete or biased datasets, can lead to inaccurate predictions and misdiagnoses. Therefore, a critical aspect of machine learning in heart disease prediction involves proper data preprocessing, feature selection, and validation. Researchers must ensure the data used is diverse, balanced, and representative of the population to avoid bias in predictions.

In conclusion, machine learning holds immense potential to revolutionize heart disease prediction by providing more precise, personalized, and early detection mechanisms. However, successful implementation requires overcoming challenges related to data quality, interpretability, ethical concerns, and healthcare integration. As research and technology continue to evolve, ML-driven predictions may become a cornerstone of preventive cardiology, offering new hope in reducing the global burden of heart disease. Finally after implementing logistic regression, random forest and KNN algorithms, Random Forest algorithm gave 95% accuracy.

**FUTURE SCOPE**

The future scope of heart disease prediction using machine learning (ML) is vast, with several exciting advancements on the horizon. One of the key areas of development is the integration of more complex and diverse data sources, such as genetic information, real-time data from wearable devices, and environmental factors. As the availability of health data grows, ML models will become more sophisticated, offering highly personalized risk assessments and recommendations. Additionally, the use of federated learning will allow global institutions to collaboratively train ML models while maintaining data privacy, enhancing the generalizability and accuracy of predictions across different populations.

The dataset that is used in our thesis is very small and old. Moreover, no new dataset regarding heart disease has been introduced so far. There is a need of new dataset and we can collect that from various hospitals of India. We can also evaluate the efficiency of each individual classifier and also such classifiers in combination, by employing the bagging, boosting and stacking techniques.

Additionally, the integration of wearable technology is set to revolutionize heart disease prediction. As these devices become more prevalent, they will provide real-time data on heart rate, activity levels, and other vital signs. This continuous flow of information allows for dynamic models that adjust predictions based on a user’s current health status, ultimately promoting proactive health management.

Big data analytics will play a crucial role as well. By merging extensive datasets, such as electronic health records and genomic data, machine learning models can be trained more effectively. This will enhance our understanding of various risk factors associated with heart disease, leading to improved predictive capabilities. Furthermore, there is a growing emphasis on developing interpretable AI. Models that are not only accurate but also understandable will help healthcare professionals better communicate risks to patients and personalize treatment plans.

Telemedicine and remote monitoring are also expected to benefit significantly from ML applications. By analyzing data collected through telehealth platforms, ongoing assessments of heart health can be made without the need for frequent in-person visits, improving patient convenience and care continuity. Additionally, ML can assist in identifying at-risk populations, enabling targeted preventive strategies and interventions that could help reduce the overall incidence of heart disease.

Moreover, the integration of ML-driven tools into clinical decision support systems will enhance healthcare providers’ ability to make informed decisions regarding diagnostics and treatment options based on individual patient data. There is also an important focus on addressing ethical considerations, such as data privacy and algorithmic bias, to ensure equitable access to these predictive tools.

Moreover, collaboration between technologists, healthcare providers, and policymakers will be crucial for maximizing the impact of ML in heart disease prediction. Multi-disciplinary efforts can facilitate the design of comprehensive health solutions that leverage machine learning while ensuring they are clinically validated and user-friendly. Initiatives that promote data sharing among institutions and encourage the development of standardized protocols will further enhance the efficacy of predictive models. Ultimately, fostering a collaborative environment will be essential for realizing the full potential of machine learning in transforming heart disease prevention and management into a more proactive and personalized approach.

Top of Form

Bottom of Form

Finally, machine learning can be adapted to global health applications, predicting heart disease across diverse populations while accounting for varying risk factors and healthcare access. This adaptability will contribute to broader public health initiatives aimed at reducing heart disease incidence worldwide. In summary, the integration of machine learning into heart disease prediction holds significant potential for improving patient outcomes and advancing public health.

**CHAPTER 8**

**BIBLIOGRAPHY**

1. Alotaibi, F. S. (2019) et al. [1] “Implementation of machine learning model to predict heart failure disease” <https://en.wikipedia.org/wiki/Logistic_regression>
2. Kim, J. K., & Kang, S. (2017) et al. [2] “Random forest-based coronary heart disease risk prediction using feature correlation analysis” <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>
3. Haq, A. U., Li, J.-P., Memon, M. H., Nazir, S., & Sun, R. (2018) et al. [3] “A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms” <https://webfocusinfocenter.informationbuilders.com/wfappent/TLs/TL_rstat/source/DecisionTree>
4. Jagtap, A., Malewadkar, P., Baswat, O., & Rambade, H. (2019) et al. [4] “Heart disease prediction using machine learning” <https://www.investopedia.com/terms/n/neuralnetwork.asp>
5. Khan, S. N., Nawi, N. M., Shahzad, A., Ullah, A., & Mushtaq, M. F. (2019) et al. [5] “Comparative analysis for heart disease prediction” <https://www.ncbi.nlm.nih.gov/pubmed/19615487>
6. Bo Jin, Chao Che (2018) et al. [6] proposed a “Predicting the Risk of Heart Failure With EHR Sequential Data Modeling” <https://www.medicalnewstoday.com/articles/325861>
7. Aakash Chauhan (2018) et al. [7] presented “Heart Disease Prediction using Evolutionary Rule Learning” <http://cooleysanemia.org/updates/Cardiac.pdf>
8. Ashir Javeed, Shijie Zhou (2017) et al. [8] designed “An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection” <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3275732/>
9. Senthilkumar Mohan, Chandrasegar Thirumalai (2019) et al. [9] was efficient technique using hybrid machine learning methodology <https://www.who.int/features/qa/27/en/>
10. K.Prasanna Lakshmi, Dr. C.R.K.Reddy (2015) et al. [10] designed “Fast Rule-Based Heart Disease Prediction using Associative Classification Mining”.<https://www.heart.org/en/healthtopics/heart-attack/angina-chest-pain>.