A PROJECT ON

ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR PREDICTION OF CARDIAC DISEASES

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

Ayush Gupta	**2014605**
Ananya Singh	**2014561**
Gautam Raj	**2014661**
Pragya Kumar	**2014766**

Under the Guidance of

Mr. Vivek Tomar

Assistant Professor

Project Team ID: MP22CSE14

Department of Computer Science and Engineering

Graphic Era (Deemed to be University)

Dehradun, Uttarakhand

MAY-2023

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the Project Report entitled **"ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR PREDICTION OF CARDIAC DISEASES"** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering and submitted in the Department of Computer Science and Engineering of the Graphic Era (Deemed to be University), Dehradun is an authentic record of my own work carried out during a period from **August-2022 to May-2023** under the supervision of **Mr. Vivek Tomar**, **Assistant Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University).

The matter presented in this dissertation has not been submitted by me/us for the award of any other degree of this or any other Institute/University.

Ayush Gupta	2014605	Ayush
Ananya Singh	2014561	CAShaurhan
Gautam Raj	2014661	Gautam Raj
Pragya Kumar	2014761	Pkumar

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

Vtomar 27/05/2023
Supervisor

pvr 27/5/23
Head of the Department

External Viva

Name of the Examiners: **Signature with Date**

1.

2.

Abstract

Early detection and diagnosis are crucial because cardiovascular diseases are the leading cause of deaths in the world. Timely diagnosis and intervention can significantly improve patient outcomes and reduce healthcare costs. In this study, we used various machine learning techniques to predict the risk of cardiovascular diseases based on two datasets obtained from the UCI library, including Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbour, and Support Vector Machines. We assessed each algorithm's performance based on its accuracy, precision, recall, and F1 Score. Our findings show that KNN performed better than the other algorithms, with accuracy, precision, recall, and F1-score values of 90.16%, 84.37%, 88.23%, and 81.81%, respectively. This study shows how machine learning algorithms can be used to predict cardiac diseases and offers guidance for further study in this area.

Keywords: Cardiovascular Diseases, Artificial Intelligence, Risk Assessment, Treatment Strategies, Performance Evaluation, Supervised Learning

Acknowledgement

Any achievement, be it scholastic or otherwise does not depend solely on the individual effort but on the guidance, encouragement and co-operation of intellectuals, elders and friends. A number of personalities in their own capacity have helped me in carrying out this project work.

Our sincere thanks to project guide **Mr. Vivek Tomar, Assistant Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), for his valuable guidance and support throughout the course of project work and for being a constant source of inspiration.

We extend our thanks to **Prof. (Dr.) Guru Prasad M.S.**, Project coordinator, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), for his valuable suggestions throughout all the phases of the Project Work.

We are extremely grateful to **Prof. (Dr.) D. P. Singh**, HOD of the Computer Science and Engineering Department, Graphic Era (Deemed to be University), for his moral support and encouragement.

We thank the **management of Graphic Era (Deemed to be University)** for the support throughout the course of our Bachelor's Degree and for all the facilities they have provided.

Last, but certainly not least we thank all teaching and non-teaching staff of Graphic Era (Deemed to be University) for guiding us in the right path. Most importantly we wish to thank our parents for their support and encouragement.

Ayush Gupta 2014605

Ananya Singh 2014571

Gautam Raj 2014661

Pragya Kumar 2014766

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Project Introduction

This project is a study that utilizes the different Machine Learning (ML) techniques that can be used to detect cardiovascular conditions. Cardiovascular diseases (CVD) are conditions impacting the heart. They account for more than 30 percent of deaths globally. It's projected that by 2030, over 22 million people worldwide will have some form of heart-related problems. [1] The leading threat factors for heart complaints are high blood pressure & cholesterol, diabetes, smoking/alternate-hand bank exposure, rotundity, unhealthy diet, and physical inactivity. If left unchecked, these conditions may result in a heart attack or stroke. To treat these ails and help patients, early discovery of these conditions is pivotal.

Machine learning (ML) is a branch of Artificial Intelligence (AI) that is based on models that accept input data and through a combination of fine optimization and statistical analysis, predict the output. ML algorithms learn from historical data, identify patterns and relationships, and generalize that knowledge to make predictions or take actions on new, unseen data. [2] With the same objective in mind, this project aims towards working out the best classification algorithm that will predict the possibility of a heart disease in a patient with at most accuracy. The work in this project is validated by performing a comparative study and analysis over five classification algorithms namely - Logistic Regression, Random Forest, Decision Tree, Support Vector Machine and K-Nearest Neighbor, using two databases to ensure comprehensive results.

The main idea of this research is to give doctors a way to detect heart problems at an early stage. As a result, it will be easier to deliver applicable treatment to cases while avoiding severe conditions. In the area of healthcare, ML- a grounded clinical decision tree has been applied lately. [3] With the recent advances in machine learning like representing discriminative classifiers advantages for automatic cardiac complaint discovery. Studies have preliminarily shown that machine learning algorithms like SVM (Support Vector Machine), RF (Random Forest), LR (Logistic Regression), BPNN (Back Propagation Neural Network), and MLP (Multilayer Perceptron) [4] have been successfully employed previously for decision-making tools to predict heart diseases grounded on individual information.

1.2 Problem Statement

Day by day the cases of heart disease are increasing at a rapid rate and it's very important and concerning to predict any such diseases beforehand. Diagnosis of diseases by the traditional approach has forever been in practice and is still quite prevalent, but it faces many constraints

and challenges, hence the diagnosis should be performed precisely and efficiently In the vast domain of health-related issues that require immediate attention, is the risk of heart attack and other heart-related diseases. To overcome these restrictions and improve the state of healthcare infrastructure and availability, the project "Analysis of machine learning techniques for prediction of cardiac diseases" has been developed. This project mainly focuses on which patient is more likely to have heart disease based on various medical attributes. A quite helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of heart attacks in any individual.

1.3 Objectives

- **Analyze the Dataset:** The first step is to acquire the dataset and understand and classify the data of a patient's heart rate and blood pressure based on its analysis.
- **Implementing Existing Techniques:** After the analysis of the patient's medical data, try to predict the condition of a patient's heart using existing machine learning techniques.
- **Analyzing the Techniques:** Finally, analyzing the Machine Learning techniques that predict the heart condition to be normal/abnormal based on the readings from various IoMT devices like smartwatches, pulse oximeters, pacemakers, etc.

Chapter 2

Literature Survey/ Background

In recent years, cardiovascular diseases have emerged as an important public health problem causing significant morbidity and mortality worldwide. Scientists and doctors want to use machine learning algorithms to create accurate and efficient models of heart disease. In this section, we provide an overview of the current literature in this area, focusing on research using machine learning techniques to predict cardiovascular disease.

Early research by Gandhi et al. [5] utilized information mining strategies to foresee heart problems by making use of different algorithms like Decision Tree, Neural Networks as well as Naïve Baye's. The methods used to uncover hidden patterns and identify solutions to healthcare organizations. Building on this work, Beunza et al [12] explored the application of Support Vector Machines (SVM) for heart disease prediction. Their study demonstrated the effectiveness of SVM in accurately classifying patients into different heart disease categories. dtgvtv5yh

More recent exploration by Mohammed et al. [6] proposed a method that involves the integration of Flask Web Framework with the RF algorithm to predict heart diseases at different stages. They highlighted the potential of machine learning techniques in this domain and specifically discussed the Random Forest algorithm as a suitable choice due to its ability to handle complex datasets and provide robust predictions.

Bhavani et al [7] employed various ML models like Logistic Regression, KNN, Random Forest, and others and compared the accuracies obtained from each of the algorithms. The proposed system not only facilitates medical treatment but also lowers costs. The Random Forest algorithm had the highest accuracy of 97.54%.

Nagavelli U. et al [8] used different decision tree classification algorithms using XGBoost to increase the accuracy of heart disease detection. Four different machine learning (ML) model types are contrasted in terms of performance parameters such as recall, precision, accuracy, and f1-measure. The XGBoost algorithm achieved a 95% accuracy rate.

Gomathy C K [9] proposed a system where symptoms of the patients are given to the machine learning model to predict the disease. They used Naïve Bayes for disease detection using symptoms, and KNN for the classification, to extract features with the most impact value Logistic Regression is used, and to divide the dataset Decision Tree is used. The highest accuracy was shown by Random Forest i.e., 98.95%.

Kathiseran et al [10] used the data obtained from an IoMT device and applied various machine-learning algorithms to find out which one gives the best accuracy. Algorithms like Logistic Regression, Naïve Bayes, SVM, and others were used among which Logistic Regression gave the highest accuracy of 92.009%.

A prediction algorithm for cardiovascular illnesses in the patient was created by Mezzatesta et al [11] In two datasets—Italian datasets and American datasets—the author employed machine learning methods. Several techniques are employed, with Support Vector Machine producing the best results with an accuracy of 95.25 percent for Italian datasets and 92.1 percent for American datasets.

The classification or regression techniques for predicting clinical occurrences are contrasted by Beunza et al [12]. The strategy is examined using a database from the Framingham heart research. Using the Support Vector Machine, the author was able to achieve the greatest AUC value of 0.75 out of all the algorithms, including Decision Tree, Random Forest, Neural Network, and Logistic Regression.

Uddin, S. , Khan, A. , Hossain, M. et al. [13] In this study, 48 articles were used to predict 49 diseases, and 50 supervised machine learning algorithms were found to have higher accuracy. Support Vector Machine was used the most frequently, it was noticed (29 out of 49 diseases that were predicted). Naïve Bayes, which was used in 23 articles, comes next. RF was ranked second in terms of the number of times it was considered, but it had the highest percentage (53%) overall. It was noted that Support Vector Machine consistently demonstrated superior accuracy for three diseases (such as Parkinson's disease, diabetes, and heart disease). [9]

Jindal et al. [14] utilized a number of ML models including Logistic Regression and K-NN. The chosen approaches demonstrated more accuracy when compared to previously used classifiers like Naive Bayes and others. The suggested approach cuts expenses while simultaneously facilitating medical care. The K-NN algorithm achieved an accuracy rate of 88.52%. Based on clinical information about a patient's prior heart disease diagnosis, this heart disease detection system helps the patient. The formula for constructing the presented model includes KNN, Random Forest Classifier, and Logistic Regression. Our model has an accuracy rate of 87.5%.

Develops a disease predictor, which is capable of predicting more than one disease. It also lowers mortality rates. This model uses two algorithms – Logistic Regression and Support vector machine. Under logistic regression, a categorical dependent variable may be predicted using this technique using a collection of independent factors. Such a system delivers probabilistic values between 0 and 1. They used hyper parameters adjustment to acquire the best results for all individual datasets while testing numerous parameters. Support Vector Machine was used to discover the best decision boundary for categorizing n dimensional space into classes so that subsequent data points can be easily placed in the right category.

Despite the progress made in heart disease prediction using machine learning algorithms, there are still certain limitations and challenges. For instance, the interpretability of black-box models, similar to neural networks, remains a concern in the medical domain, where explain ability is pivotal for gaining trust and acceptance from healthcare professionals. Also, the availability of large- scale, different datasets with standardized features poses a challenge for developing robust prediction models.

In summary, previous studies have shown the eventuality of various machine learning algorithms, including logistic regression, decision trees, SVM, KNN, and deep learning techniques, in predicting heart disease. Still, there's a need for further exploration to address challenges related to including the need for larger and more diverse datasets for further validation.

Chapter 3

System Analysis

3.1 Existing System

By utilizing machine learning (ML) techniques, the current system for the prediction of cardiac disease has achieved considerable advancements in the precision and effectiveness of diagnosis. These systems use a variety of machine learning (ML) methods, including logistic regression, random forests, decision trees, and support vector machines, to analyze huge datasets and identify significant patterns relating to heart disease. In addition, the most important qualities for prediction are found using feature selection approaches. The existing system for heart disease prediction using machine learning is a valuable tool in the field of healthcare. By leveraging advanced algorithms and predictive models, it assists healthcare professionals in making informed decisions, enabling early detection and intervention in cases of heart disease. This contributes to improved patient outcomes and the effective management of cardiovascular health.

3.2 Proposed System

Upon reviewing the papers mentioned earlier, our proposed system aims to develop a heart disease prediction system using the input variables outlined in Table 2. Our study involves analyzing several classification algorithms including LR, RF, DT, SVM and KNN on 2 datasets from UCI Library (Hungarian and Cleveland dataset). We evaluated these algorithms based on their performance in terms of Accuracy, Precision, Recall, and f-measure scores, and identified the best algorithm for heart disease prediction.

The gathering of data and selection of the most important attributes is the first step in the system's working. The relevant data is then preprocessed into the format required. After that, the data is split into a training set and a testing set. The algorithms are implemented, and the training data is used to train the model. By testing the system using testing data, the correctness of the system is determined. The modules listed below are used to implement this system:

3.3 Dataset Description

For this proposed system, we have used Cleveland and Hungarian cardiac dataset from UCI research repository. There are 14 properties with 303 and 308 occurrences respectively in the dataset. 8 categorical and 6 numerical characteristics are available. [15-18]

Table 3.1 Dataset description

S. No.	Attribute Name	Description	Range of values
1	Age	Age of person in years	29-79
2	Sex	Gender of person [1-male 0-female]	0.1
3	Trestbps	Resting blood pressure in mmHg	94-200
4	Cp	Chest pain type [1-typical type 1 angina, 2-atypical type angina, 3-non-angina pain, 4-asymptomatic	1, 2, 3, 4
5	Fbs	Fasting blood sugar in mg/dl	0.1
6	Choi	Serum cholesterol in mg/dl	126-564
7	Thalac	Maximum heart rate achieved	71-202
8	Restecg	Resting electrocardiographic results	0, 1, 2
9	Old peak	ST depression induced by exercise relative to rest	1-3
10	Exang	Exercise induced angina	0.1
11	Ca	Number of major vessels coloured by fluoroscopy	0-3
12	Slope	Slope of the peak exercise ST segment	1, 2, 3
13	Thai	3-normal, 6-fixed defect, 7-reversible defect	3, 6, 7
14	Num	Class attribute	0 or 1

Each study has chosen patients between the ages of 29 and 79. The gender value of the males is 1 and the value for female gender is 0. Symbolic term is given to the four forms of discomfort heart disease. Trestbps attribute will be the measurement of blood pressure levels. Chol would be the quantity of cholesterol. FBS is actually worth your fast blood sugar level. This means whether the value of it is actually one consequently fasting blood sugar level is actually under 120 mg/dl, as well as zero is actually above. Restecg may be the residual electrocardiographic score; thalach is actually the optimum heart rate.

3.4 Relation of attributes with target features

Following are some of the essential attribute comparisons for both the datasets:

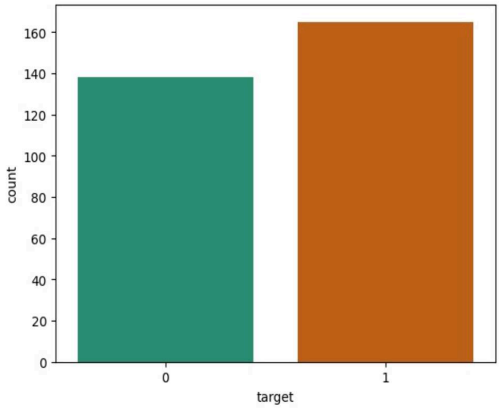

Fig. 3.1 Disease vs target for Cleveland dataset

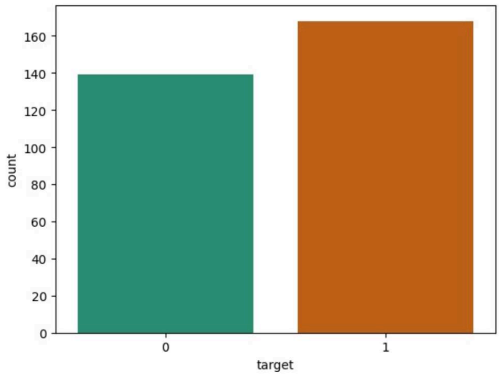

Fig. 3.2 Disease vs target for Hungarian dataset

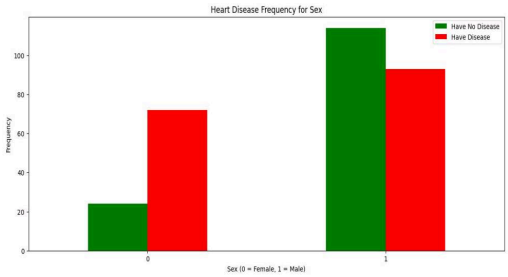

Fig. 3.3 Sex vs target for Cleveland dataset

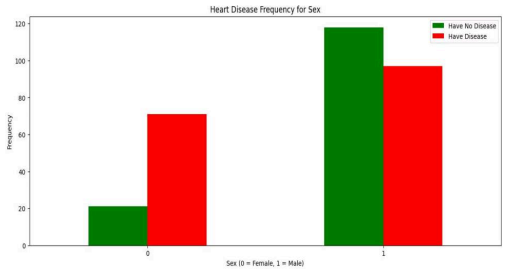

Fig. 3.4 Sex vs target for Hungarian dataset

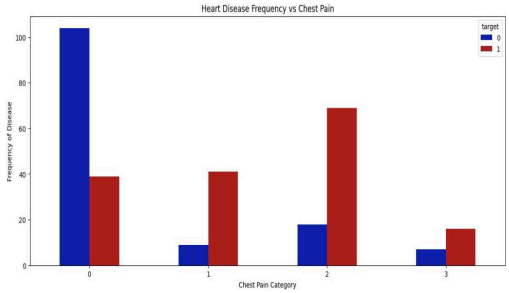

Fig. 3.5 Chest pain vs target for Cleveland dataset

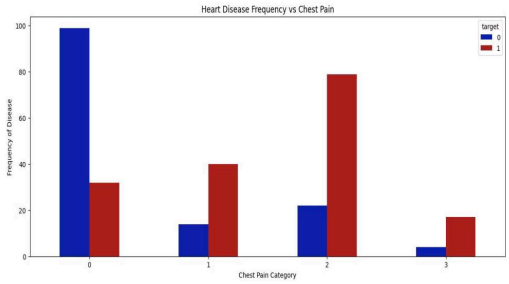

Fig. 3.6 Chest pain vs target for Hungarian dataset

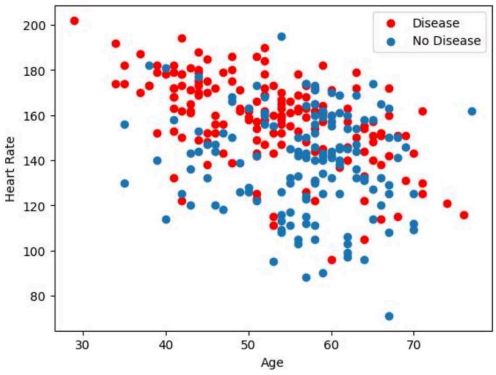

Fig. 3.7 Heart rate vs target for Cleveland dataset

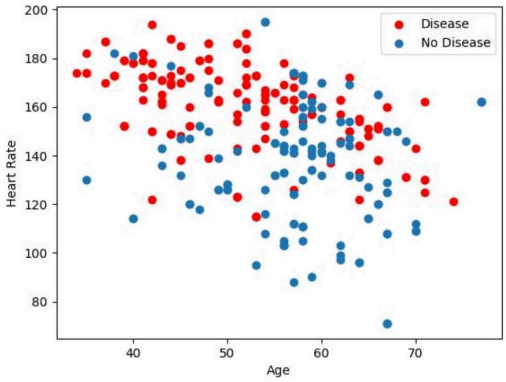

Fig. 3.8 Heart rate vs target for Hungarian dataset

3.5 Heatmap of features

Heatmap comparison of the features for both datasets have been done to predict how the data points can be distinguished from each other (linear or overlapping).

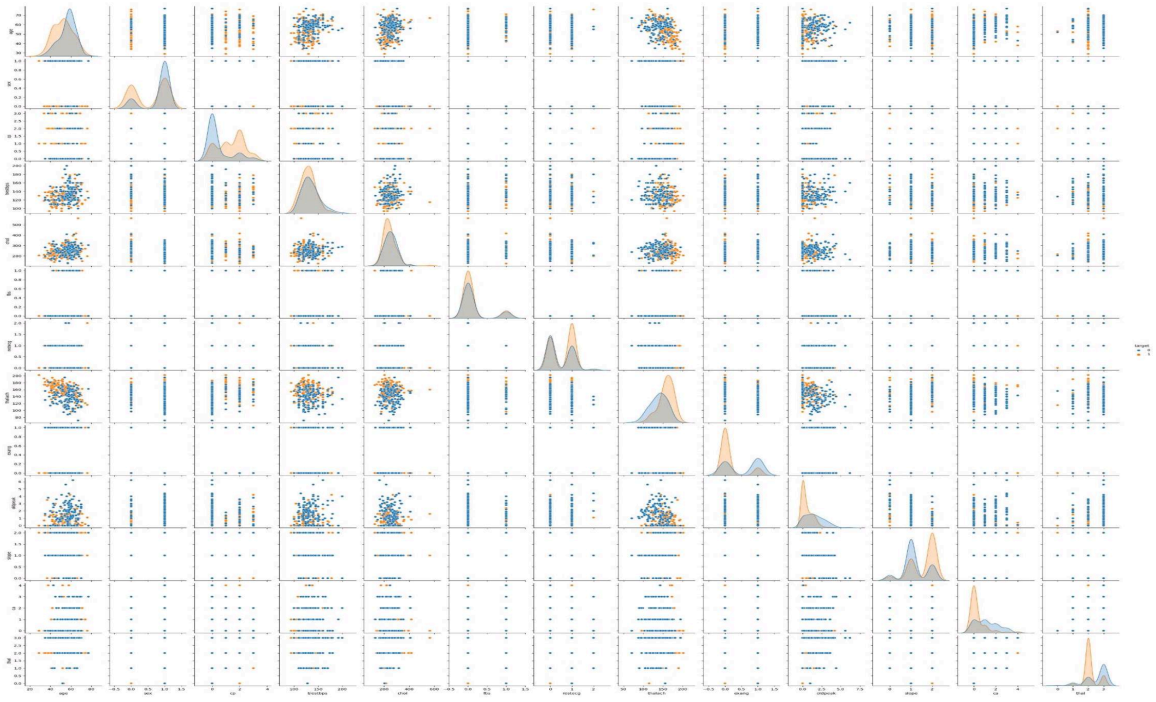

Fig. 3.9 Heatmap of features for Cleveland dataset

Fig. 3.10 Heatmap of features for Hungarian dataset

From the above heatmap comparisons we observe that there is only overlapping of data points, so linear regression algorithm is not suitable for prediction on the given datasets.
The suitable algorithms for the disease prediction are selected to be : Logistic Regression, SVM, KNN, Decision Tree, and Random Forest.

3.6 Machine Learning

ML is a subfield of artificial intelligence (AI) that focuses on developing algorithms and models that enable computers to learn and make predictions or decisions without being explicitly programmed. It involves the study of computational methods and statistical models that allow systems to automatically learn from data, improve performance over time, and adapt to new information. [19]
The learning process involves extracting meaningful features from the data and using them to build predictive or descriptive models. These models are then used to make predictions, classify new data, or uncover hidden patterns in the data.
There are different types of machine learning techniques, including supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.
Supervised Learning is defined use of labeled datasets to train algorithms in order to classify data or to predict outcomes accurately. This sort of learning helps organizations solve particular real-world problems, such as predicting heart disease problems using patient past data.

Whereas Unsupervised Learning models themselves find the hidden patterns and insights from given data. In simpler words, the models are trained using unlabeled datasets and are allowed to act on the data without supervision.
ML algorithms are particularly valuable because these algorithms are capable of helping medical practitioners in making sense of the massive amounts of healthcare data that is generated every day within electronic health records.
Using ML in healthcare like ML algorithms are capable of helping us find patterns and insights in medical data that would be impossible to find manually. The most common use cases for machine learning in healthcare among healthcare professionals are automating medical billing, clinical decision support and the development of clinical practice guidelines within health systems.
As of then, there are a wide range of potential uses for machine learning technologies in healthcare from improving patient data, medical research, diagnosis and treatment, to reducing costs and making patient safety more efficient.

3.7 Classification Techniques

The input dataset is split into train and test datasets for supervised algorithms. The supervised algorithms used in this study are the Decision tree, Random Forest, SVM, KNN, and logistic regression.

Logistic Regression
LR is a member of the supervised machine learning model family in the context of artificial intelligence. This type of statistical model, often known as a logit model, is extensively used in categorization and predictive analytics.
It determines the probability that an event, such as voting or not voting, will occur based on a collection of independent variables. [20]
As the outcome is a probability, the range of the dependent variable is 0 to 1.
The odds, or likelihood of success divided by the probability of failure, are transformed using the logit formula in logistic regression.
We use the sigmoid function as a cost function.
The sigmoid function transforms a real value prediction into a probabilistic value between '0' and '1'.
Sigmoid logistic function:

$P(x)=1/(1+e\wedge(-x))$ (1) -------- (1)

Where the mathematical constant e is Euler's number, and its value is roughly equal to 2.71828 as given in Eq, x is the input to the probability function (the algorithm's prediction value), and P(x) is the probability estimation function with a value between 0 and 1.

Advantages:
Rather than straight away starting with a complex model, LR is sometimes used as a benchmark model to measure performance, as it is relatively quick and easy to implement.

1. It is very fast at classifying unknown records.
2. Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.

Disadvantages:

1. Non-linear problems can't be solved with logistic regression since it has a linear decision surface.
2. Only important and relevant features should be used to build a model otherwise the probabilistic predictions made by the model may be incorrect and the model's predictive value may degrade.
3. The presence of data values that deviate from the expected range in the dataset may lead to incorrect results as this algorithm is sensitive to outliers.

Random Forest

Random Forest is a supervised learning system. A huge number of decision trees are created during the training phase of this ensemble learning technique for classification, regression, and other problems. The class that results corresponds to the mean prediction of the individual trees or the class mode**.** [21]

A supervised learning system is Random Forest. An ensemble learning approach for classification, regression, and other issues generates many decision trees during training. The class that emerges is equivalent to the class's mode or the individual trees' mean prediction.
Using data samples, this method builds decision trees. The number of trees in the Random Forest and the outcomes it can produce is directly correlated. Following the creation of the decision trees, it obtains the prediction from each tree before using voting to choose the optimal course of action. A more accurate result would be obtained with a greater number of trees. Accuracy will increase as trees get bigger.

Advantages:

1. RF is capable of both classification and regression.
2. Random Forest can measure the importance of each feature in the classification or regression task. This information can be useful for feature selection, as it helps identify the most influential variables.
3. RF is less prone to overfitting compared to individual decision trees. By combining predictions from multiple trees and averaging them, it reduces the risk of overfitting and provides more generalized results.

Disadvantages:

1. RF is a black-box model, meaning it doesn't provide readily interpretable explanations for its predictions. While it can estimate feature importance, understanding the underlying decision-making process of the ensemble is challenging.
2. RF is not good at extrapolating beyond the range of the training data. It may struggle to provide accurate predictions for data points that are significantly different from the training distribution.
3. Although Random Forest is faster than certain complex algorithms like gradient boosting, it can still be slower than simple models. Training time increases with the number of trees in the ensemble, affecting real-time or time-sensitive applications.

Decision Tree

A decision tree is a non-parametric technique that can handle large, complex data sets effectively without utilizing numerous parametric structures. Study data can be split into training and validation data sets if the sample size is big enough. In order to create the best final model, the proper tree size should be chosen using the training data set and validation data set. [22]

The decision tree algorithm, a common machine learning technique, provides insight into how to deal with target variable prediction based on provided input parameters. Depending on the state of the variables, decision trees can be seen in one of two ways. Regression can be used to solve the issue if the target variable is numerical; otherwise, it must be classified.

Advantages:

1. DT can provide a measure of feature importance. By evaluating how much each feature contributes to the decision-making process, it helps identify the most relevant features for the prediction task.
2. DT has relatively fast training times compared to more complex algorithms. They can handle large datasets efficiently, and once trained, making predictions for new instances is computationally inexpensive.
3. DT provides easily interpretable and understandable rules. The tree structure allows users to visualize and comprehend the decision-making process. The rules can be expressed in if-then-else statements, which can be helpful in explaining the model to non-technical stakeholders.

Disadvantages:

1. DT are prone to overfitting, especially when the tree becomes too deep and complex. They can capture noise and outliers in the training data, leading to poor generalization and decreased performance on unseen data.
2. DT can be sensitive to small changes in the training data. A slight modification in the data or the addition of new data points can result in a different tree structure and, consequently, different predictions.

3. DT are not well-suited for capturing linear relationships in the data. They often require a large number of splits to approximate linear patterns, leading to complex and less interpretable trees. Linear models may perform better for such scenarios.

Support Vector Machine (SVM):

Support Vector Machine (SVM) is a popular Supervised Learning algorithm that is primarily used for classification problems. This algorithm works by creating the best-line/decision-boundary that can segregate an n-dimensional space into classes so thatwe can easily put the new data points in the correct category in the future. In other words, the algorithm separates the data into various classes by finding the optimal hyperplane which maximizes the distance between these classes. SVM chooses the extreme points/vectors that help in creating the hyperplane. [23]

The best decision boundary is known as the hyperplane and the data points (vectors) which are the closest to the hyperplane (that affect its position) are called Support Vectors as these vectors support the hyperplane, and hence algorithm is termed as Support Vector Machine.

There are 2 types of SVM classifiers:

- **Linear SVM:** is used for linearly separable data i.e., if a dataset can be classified into two classes by using a single straight line. In this case, the data is termed as linearly separable data, and the classifier is used as Linear SVM classifier.
- **Non-linear SVM:** is used for data that cannot be separated linearly. This means that if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data. In this case, the classifier used is called Non-linear SVM classifier.

Advantages:

1. SVMs perform well even when the number of features is larger than the number of samples. They can handle complex relationships and capture intricate decision boundaries.
2. SVMs are less prone to overfitting compared to other machine learning algorithms, through the utilization of the use of a regularization parameter (C) that helps control the trade-off between maximizing the margin and minimizing the classification error.
3. SVMs work well with small to medium-sized datasets. They can handle situations where the number of samples is limited, making them suitable for applications with limited data availability.

Disadvantages:

1. SVMs can be computationally expensive, especially when dealing with large datasets. The training time and memory requirements increase significantly as the dataset size grows.

2. SVMs provide effective predictions but lack interpretability. It can be difficult to understand the rationale behind the classification decisions made by SVMs, as they rely on complex mathematical transformations.
3. SVMs are sensitive to noisy data, particularly when outliers are present. Noisy data can disrupt the margin and negatively impact the classifier's performance.

K-Nearest Neighbour (KNN)

K-Nearest Neighbour is a non-parametric supervised learning classifier that makes use of proximity in order to classify/predict how a given data point will be grouped by searching the dataset for correlations between the predictors and the values. It makes use of the Euclidean distance Formula to find out how close each data point in the trained dataset is to the newly observed data point. KNN is a typical illustration of a "lazy learner" because it memorizes the training data rather than learning a discriminative function leading to all the computation occurring during the timeframe of a classification or prediction being made. [24]

Advantages:

1. KNN is easy to understand and implement, making it a popular choice for beginners and as a baseline algorithm for comparison.
2. KNN does not require an explicit training phase. It stores the entire training dataset in memory, allowing for immediate predictions once new data is presented.
3. KNN can dynamically update its model with new training instances without retraining the entire model. This makes it useful in scenarios where new data becomes available incrementally.

Disadvantages:

1. The main drawback of KNN is its computational cost during the prediction phase. For each new instance, KNN requires calculating distances to all training samples, making it inefficient for large datasets. As the number of training instances grows, the prediction time increases significantly.
2. KNN stores the entire training dataset in memory, which can be a limitation for large datasets. As the number of training instances and features increases, the memory requirements also increase.
3. KNN calculates distances between data points, and the choice of distance metric can be influenced by the scale of features.

Chapter 4

Requirements and Methodology

This chapter gives the information about the hardware and software requirements for the working of the project.

4.1 Hardware Requirements

Processor : Any updated processor

Ram : Min 4 GB

Hard Disk : Min 100 GB

4.2 Software Requirements

Operating System: Windows family

Technology : Python 3.7 and above

IDE : Jupyter notebook

4.1 System Architecture

The system architecture gives an overview of the working of the project.

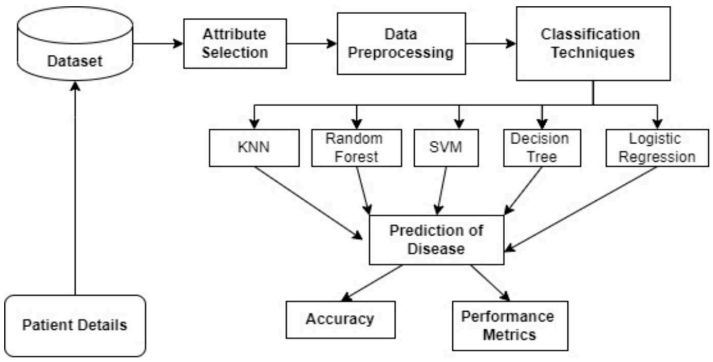

Figure 4.1 System Architecture

Collecting relevant data sets is the first step in using machine learning algorithms to predict heart disease. The data is preprocessed, then divided into training and testing sets. The machine learning algorithms KNN, SVM, LR, RF, and DT are trained using the training set.

Following training, the algorithms' performance is assessed on the testing set. Metrics like accuracy, precision, and recall are utilized to estimate how well the algorithms perform.

The main objective of this research paper is to summarize the recent research with comparative results that has been done on heart disease prediction and also make analytical conclusions. In this paper commonly used data mining and machine learning techniques and their complexities are summarized.

4.2 Data Flow Diagram

The data flow diagram maps out the flow of the data for the processes of the system.

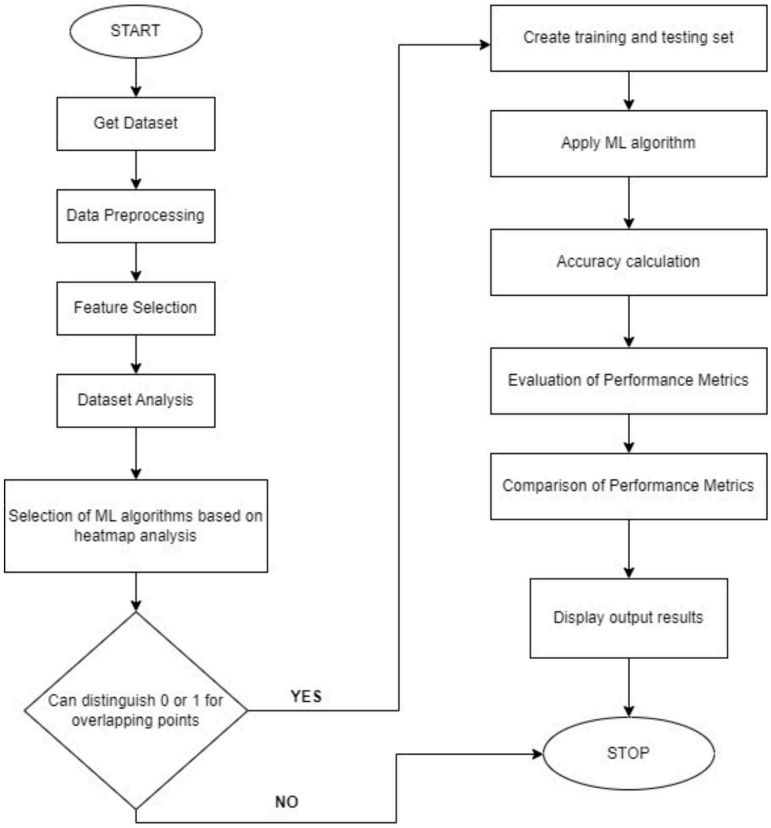

Figure 4.2 Data Flow Diagram

Explaination of workflow of the system:

- Our system starts with the collection of datasets from Kaggle.
- The data is preprocessed and analysed for missing or incorrect values.
- Features which will be used for the evaluation and prediction of disease using ML algorithms are selected.
- Dataset is analysed by the use of heatmap to check for correlation between attributes.
- Based on the heatmap analysis ML algorithms are selected which are suitable for classification.

- The training and testing sets are divided in 80% and 20% respectively.
- The next step is to apply the ML algorithms and check for accuracy.
- To support the result obtained above, performance metrics namely, confusion matrix, precision, recall and f1 score are evaluated.
- The performance metric values for particular algorithms are compared against each other, and are cross validated by calculating their standard deviations.
- The least deviated value gives the most precise algorithm, best suited for classification and prediction of diseases in the future.

Chapter 5

Pseudo Code

```
# Step 1: Read the Dataset
dataset = read_dataset() # Read the dataset from a file or source

# Step 2: Clean and Check for Null Values
cleaned_dataset = clean_dataset(dataset) # Clean the dataset and handle any null values

# Step 3: Split the Dataset into Training and Testing Sets
train_data, test_data = split_dataset(cleaned_dataset, train_ratio)  # Split the dataset into training and testing sets

# Step 4: Training and Evaluation of Classification Algorithms
algorithms = [KNN, SVM, LogisticRegression, DecisionTree, RandomForest]
results = []

for algorithm in algorithms:
   # Step 4.1: Train the Algorithm
    model = train_algorithm(algorithm, train_data)  # Train the algorithm using the training data

   # Step 4.2: Calculate Accuracy
    accuracy = calculate_accuracy(model, test_data) # Calculate the accuracy using the test data
```

```
    # Step 4.3: Calculate Confusion Matrix
        confusion_matrix = calculate_confusion_matrix(model, test_data)  # Calculate the confusion matrix

    # Step 4.4: Calculate Recall, Precision, and F1 Score
        recall = calculate_recall(confusion_matrix)  # Calculate the recall using the confusion matrix
        precision = calculate_precision(confusion_matrix)  # Calculate the precision using the confusion matrix
      f1_score = calculate_f1_score(recall, precision) # Calculate the F1 score using the recall and precision

    # Step 4.5: Store the Results
    result = {
        "Algorithm": algorithm,
        "Accuracy": accuracy,
        "Confusion Matrix": confusion_matrix,
        "Recall": recall,
        "Precision": precision,
        "F1 Score": f1_score
    }
    results.append(result)

# Step 5: Output the Results
display_results(results) # Display the results for each algorithm
```

Chapter 6

Results and Discussion

6.1 Evaluation of Data

The measures used to assess prediction models and demonstrate their effectiveness on both datasets are listed below. The following is a list of the factors used to predict performance:

1) Confusion Matrix: The algorithm's categorization measurements on data are expressed using the Confusion Matrix. [25]

    a) True Positive (TP): This value predicts "Yes", if the patient has the disease.

    b) True Negative (TN): This value predicts "No", if the patient does not have the disease.

    c) False Positive (FP): This value predicts "Yes", even if the patient does not have the disease.

    d) False Negative (FN): This value predicts "No", even if the patient has the disease.

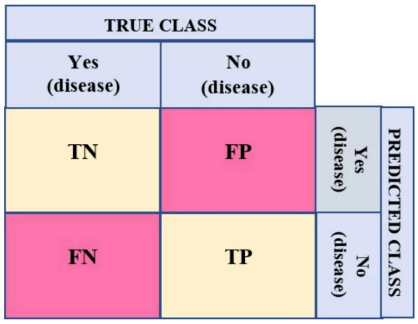

Figure 6.1 Confusion Matrix

2) Accuracy: The accuracy is calculated as the number of examples that were accurately identified divided by the total number of occurrences in the dataset.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

3) Precision: the average likelihood of retrieving relevant information.

$$\text{Precision} = \frac{TP}{TP + FP}$$

4) Recall: The recall predicts how well the model can identify Positive samples.

$$\text{Recall} = \frac{TP}{TP + FN}$$

5) F-Measure: The classification problem's precision and recall are computed, and the two scores are then combined to derive the F-Measure. [10]

$$\text{F1 Score} = \frac{(2*Precision*Recall)}{Precision + Recall}$$

After implementing various models of machine learning classification algorithms, we obtained the following values of confusion matrices.

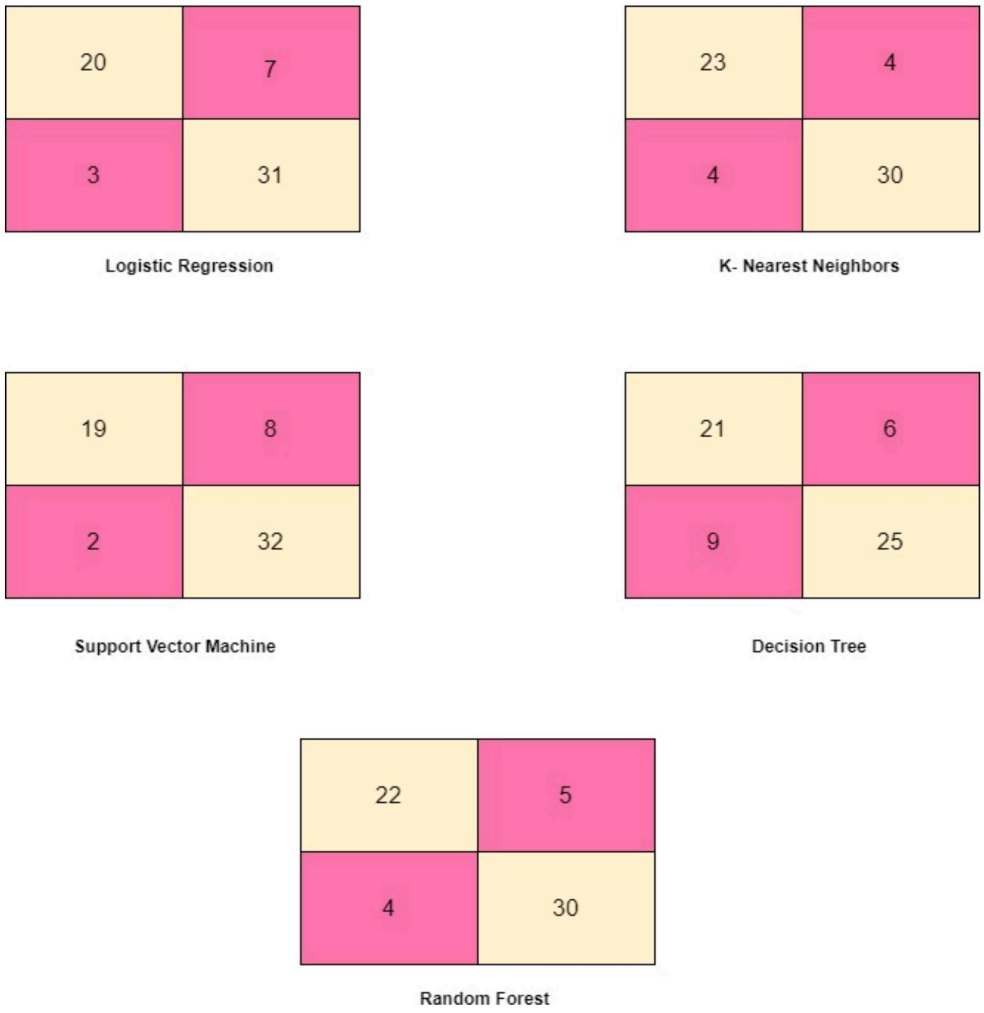

Fig 6.2 Confusion Matrices obtained using Cleveland Dataset

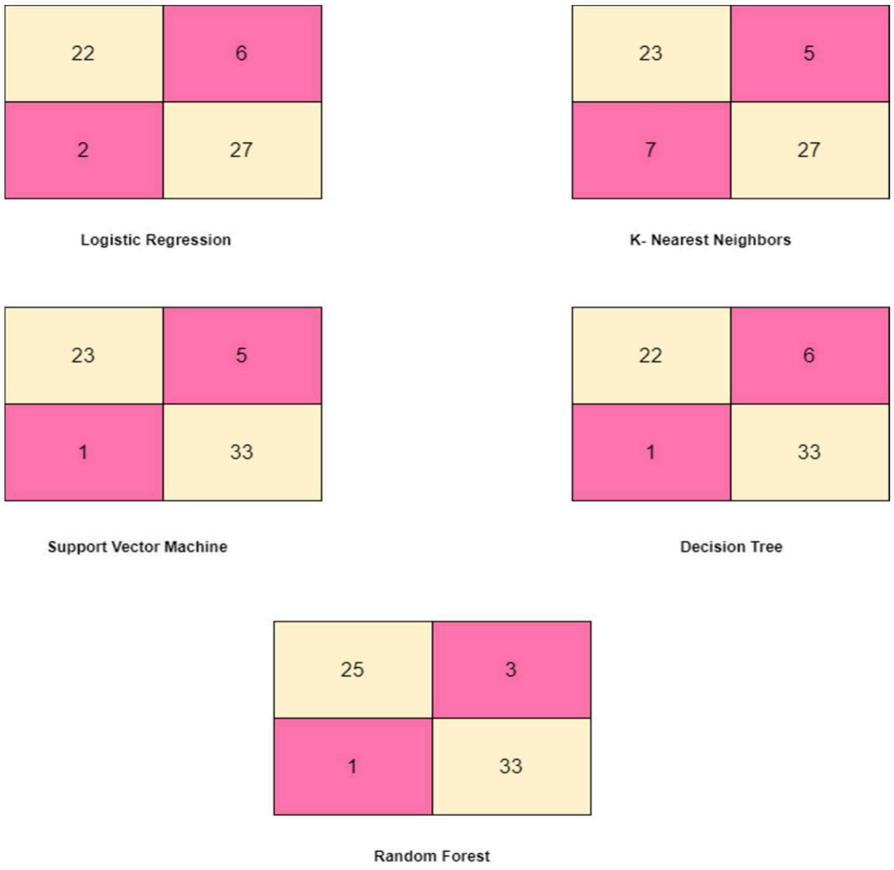

Fig 6.3 Confusion Matrices obtained using Hungarian Dataset

The performance of Machine Learning algorithms in terms of Accuracy is listed below in the table:

TABLE 6.1 Accuracy obtained using Cleveland Dataset

Sl.No.	Algorithm Applied	Accuracy Obtained
1	Logistic Regression	83.61%
2	Random Forest	85.25%
3	Decision Tree	72.13%
4	SVM	83.61%
5	KNN	90.16%

TABLE 4.2 Accuracy obtained using Hungarian Dataset

Sl.No.	Algorithm Applied	Accuracy Obtained
1	Logistic Regression	86.89%
2	Random Forest	90.16%
3	Decision Tree	88.52%
4	SVM	88.52%
5	KNN	93.44%

The performance metrics for both datasets as per the derived confusion matrix:

TABLE 6.3.Results of the Model Performances using Cleveland Dataset

Algorithms Used/ Factors Compared	Logistic Regression	Random Forest	Decision Tree	KNN	SVM
Precision	0.8157	0.8571	0.8064	0.8823	0.8
Recall	0.9117	0.8823	0.7352	0.8823	0.9411
F1 Score	0.8611	0.8695	0.7692	0.8823	0.8648

TABLE 6.4 Results of the Model Performances using Hungarian Dataset

Algorithms Used/ Factors Compared	Logistic Regression	Random Forest	Decision Tree	KNN	SVM
Precision	0.8421	0.9166	0.8461	0.8437	0.8684
Recall	0.9411	0.9705	0.9705	0.7941	0.9705
F1 Score	0.8888	0.9428	0.9041	0.8181	0.9166

However, the obtained values are not discrete enough to give a precise result for the predictions. Hence, we find the standard deviation of different performance metrics. The results obtained are as follows:

TABLE 6.5 Standard deviations of precision obtained using Cleveland and Hungarian Datasets

Algorithm Applied	Standard deviation (Cleveland)	Standard deviation (Hungarian)
Logistic Regression	0.0334	0.0292
Random Forest	0.0302	0.0237
Decision Tree	0.0341	0.0287
KNN	0.0242	0.0291
SVM	0.0351	0.0266

TABLE 6.6. Standard deviations of recall obtained using Cleveland and Hungarian Datasets

Algorithm Applied	Standard deviation (Cleveland)	Standard deviation (Hungarian)
Logistic Regression	0.0461	0.0798
Random Forest	0.0278	0.0814
Decision Tree	0.0685	0.0814
KNN	0.0278	0.0802
SVM	0.0819	0.0814

Based on the above-obtained result, we observe that KNN has the lowest value of standard deviation in both sets and thus is the most precise among the other algorithms.

Chapter 7

Conclusion and Future Work

Based on the obtained results of accuracy from both datasets, we observe that KNN has the highest accuracy among all the applied algorithms. To support this conclusion obtained results of recall, precision, and f1 score are evaluated and cross validated through standard deviation, where it is observed that KNN has the most precise and significant result in comparison to other algorithms and it can be inferred that KNN is the most suitable algorithm for the prediction of heart diseases.

Future studies could compare the performance of the k-modes clustering algorithm with other widely used clustering algorithms, like k-means or hierarchical clustering, to gain a more thorough understanding of its performance and address the limitations of this study. It would also be helpful to assess how missing data and outliers affect the model's accuracy and come up with solutions for these situations. More features can be added if they are required to increase algorithm implementation accuracy. To determine the model's generalizability to fresh, untested data, it would also be advantageous to assess the model's performance on a held-out test dataset. The goal of future research should be to determine the interpretability of the clusters created by the algorithm, and the robustness and generalizability of the results, which could help in understanding the outcomes and supporting decisions based on the findings of the research.

Details of Research Publication

The details of our research publication are as follows:

Ayush G., Ananya S., Gautam R., Pragya K., "Analysis of Machine Learning Techniques for Predicting Cardiac Diseases."
(Completed and yet to be communicated)

References

[1] A. B. C. Patil, ‘‘An IoT based health care and patient monitoring system to predict medical treatment using data mining techniques: Survey,’’ Int. J. Adv. Res. Comput. Commun. Eng., vol. 6, no. 3, pp. 24–26, Mar. 2017, doi: 10.17148/IJARCCE.2017.6306

[2] Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., ... & Tang, W. W. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. Scientific reports, 10(1), 16057.

[3] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms, *16*(2), 88.

[4] Divya, K., Sirohi, A., Pande, S., & Malik, R. (2021). An IoMT assisted heart disease diagnostic system using machine learning techniques. Cognitive Internet of Medical Things for Smart Healthcare: Services and Applications, 145-161.

[5] Gandhi, M., & Narayan Singh, S. (2015). Prediction in Heart Disease Using Techniques of Data Mining. In Proceedings of the 1st International Conference on Futuristic Trends in Computational Analysis and Knowledge Management (pp. 520-525). Springer.

[6] Mohammed, A. A., Basa, R., Kuchuru, A. K., Nandigama, S. P., & Gangolla, M. (2020). Random Forest Machine Learning technique to predict Heart disease. European Journal of Molecular & Clinical Medicine, 7(4), 2453-2459.

[7] Ramya, P., Bhavani, A., & Viswanadham, S (2022). Heart Diseases Detection by Machine Learning Classification Algorithms.

[8] Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine Learning Technology-Based Heart Disease Detection Models. Journal of Healthcare Engineering, 2022, 7351061. doi: 10.1155/2022/7351061

[9] Gomathy, C. K., & Naidu, M. A. R. (2021). The prediction of disease using machine learning. International Journal of Scientific Research in Engineering and Management (IJSREM) Volume, 5.

[10] Kathiresan, S. (2020). Analysis on cardiovascular disease classification using machine learning framework. *ICTACT Journal on Data Science and Machine Learning*, *2*(1), 153-6.

[11] Mezzatesta, G. (2019). A Machine Learning-Based Approach for Predicting the Outbreak of Cardiovascular Diseases in Patients on Dialysis. Computer Methods and Programs in Biomedicine, 177, 9-15. doi: 10.1016/j.cmpb.2019.03.019

[12] Beunza, J.-J., et al. (2019). Comparison of Machine Learning Algorithms for Clinical Event Prediction (Risk of Coronary Heart Disease). Journal of Biomedical Informatics, 97, 103257. doi: 10.1016/j.jbi.2019.103257

[13] Uddin, S., Khan, A., Hossain, M., et al. (2019). Comparing Different Supervised Machine Learning Algorithms for Disease Prediction. BMC Medical Informatics and Decision Making, 19, 281. doi: 10.1186/s12911-019-0953-9

[14] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[15] Andras J. M.D. Hungarian Institute of Cardiology, Budapest [Heart Disease Data Set]. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[16] William S. M.D. University Hospital, Zurich, Switzerland [Heart Disease Data Set]. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[17] Matthias P. M.D. University Hospital, Basel, Switzerland [Heart Disease Data Set]. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[18] Robert D., M.D., Ph.D. V.A. Medical Center Long Beach, and Cleveland Clinic Foundation [Heart Disease Data Set]. Retrieved-from_https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[19] IBM. (n.d.). AI and Machine Learning. Retrieved from https://www.ibm.com/design/ai/basics/ml/

[20] Ambrish, G., Ganesh, B., Ganesh, A., Srinivas, C., & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings, 3(1), 127-130.

[21] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2016). Prediction of heart disease using random forest and feature subset selection. In *Innovations in Bio-Inspired Computing and Applications: Proceedings of the 6th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2015) held in Kochi, India during December 16-18, 2015* (pp. 187-196). Springer International Publishing.

[22] Shouman, M., Turner, T., & Stocker, R. (2011). Using Decision Tree for Diagnosing Heart Disease Patients. *AusDM*, *11*, 23-30.

[23] Bhatia, S., Prakash, P., & Pillai, G. N. (2008, October). SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features. In *Proceedings of the world congress on engineering and computer science* (pp. 34-38).

[24] Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology*, *2*(3), 220-223.

[25] Sarveshvar, M. R., Gogoi, A., Chaubey, A. K., Rohit, S., & Mahesh, T. R. (2021, December). Performance of different machine learning techniques for the prediction of heart diseases. In *2021 international conference on forensics, analytics, big data, security (FABS)* (Vol. 1, pp. 1-4). IEEE.