**Heart-Attack-prediction-using-different-ML-models**

**Project report in partial fufilllment of the requirement for the award of the degree of**

**Bachelor of Technology**

**In**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted By**

<<Student\_1 Name>> University Roll No. XXXXXXXXX

<<Student\_2 Name>> University Roll No. XXXXXXXXX

<<Student\_3 Name>> University Roll No. XXXXXXXXX

<<Student\_4 Name>> University Roll No. XXXXXXXXX

**Under the guidance of**

PROF. XXXXXXXXXXXX XXXXXX

Department of Computer Science & Engineering



UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160.

**CERTIFICATE**

This is to certify thatthe project titled **<<Project Title>>** submitted by **<<Student Name\_1>>(University Roll No. XXXXXXXXX),<<Student Name\_2>>(University Roll No. XXXXXXXXX)**,**<<Student Name\_3>>(University Roll No. XXXXXXXXX)** and **<<Student Name\_4>>(University Roll No. XXXXXXXXX)**, students of UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA, in partial fulfilment of requirement for the degree of Bachelor of Computer Science, is a bonafide work carried out by them under the supervision and guidance of Prof. XXXXXXXXXXXX during 7thSemester of academic session of 2020 - 2021. The content of this report has not been submitted to any other university or institute. I am glad to inform that the work is entirely original and its performance is found to be quite satisfactory.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. XXXXXXXXXXXX XXXXXX Prof. Sukalyan Goswami

Assistant Professor Head of the Department

Department of CSE Department of CSE

UEM, Kolkata UEM, Kolkata

**ACKNOWLEDGEMENT**

We would like to take this opportunity to thank everyone whose cooperation and encouragement throughout the ongoing course of this project remains invaluable to us.

We are sincerely grateful to our guide Prof. XXXXXXXXXXXX of the Department of Computer Science& Engineering, UEM, Kolkata, for his wisdom, guidance and inspiration that helped us to go through with this project and take it to where it stands now.

We would also like to express our sincere gratitude to Prof. Sukalyan Goswami, HOD, Computer Science& Engineering, UEM, Kolkata and all other departmental faculties for their ever-present assistant and encouragement.

Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

<<Student Name\_1>>

<<Student Name\_2>>

<<Student Name\_3>>

<<Student Name\_4>>

**TABLE OF CONTENTS**

**ABSTRACT........................................................................<<PAGE NO>>**

**CHAPTER – 1: INTRODUCTION...............................................<<PAGE NO>>**

**CHAPTER – 2: LITERATURE SURVEY**

**2.1 <<XXXX>> .....................................................................<<PAGE NO>>**

**2.2<<XXXX>> .....................................................................<<PAGE NO>>**

**CHAPTER – 3: PROBLEM STATEMENT**

**3.1<<XXXX>> .....................................................................<<PAGE NO>>**

**3.2<<XXXX>> .....................................................................<<PAGE NO>>**

**CHAPTER – 4: PROPOSED SOLUTION**

**4.1<<XXXX>> .....................................................................<<PAGE NO>>**

**4.2<<XXXX>> .....................................................................<<PAGE NO>>**

**CHAPTER – 5 : EXPERIMENTAL SETUP AND RESULT ANALYSIS**

**5.1 <<XXXX>> .....................................................................<<PAGE NO>>**

**5.2 <<XXXX>> .....................................................................<<PAGE NO>>**

**CHAPTER – 6 : CONCLUSION & FUTURE SCOPE**

**6.1<<XXXX>> .....................................................................<<PAGE NO>>**

**6.2<<XXXX>> .....................................................................<<PAGE NO>>**

**BIBLIOGRAPHY ..........................................................................<<PAGE NO>>**

**ABSTRACT**

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. For providing appropriate results and making effective decisions on data, some advanced data mining techniques are used. In this study, a Heart Disease Prediction System (HDPS) is developed using Logistic regression, **K nearest neighbor** and Decision Tree algorithms for predicting the risk level of heart disease. The system uses 14 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The HDPS predicts the likelihood of patients getting heart disease. It enables significant knowledge. E.g. Relationships between medical factors related to heart disease and patterns, to be established. We have employed K nearest neighbor with backpropagation as the training algorithm. The obtained results have illustrated that the designed diagnostic system can effectively predict the risk level of heart diseases.

**Keywords: Machine Learning, Logistic regression, K nearest neighbor, Decision Tree, Disease Diagnosis .**

**CHAPTER – 1**

**Introduction**

Machine Learning is used across many spheres around the world. The healthcare industry is no exception. Machine Learning can play an essential role in predicting presence/absence of Locomotor disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis.

In this article, I’ll discuss a project where I worked on predicting potential Heart Diseases in people using Machine Learning algorithms. The algorithms included Logistic Regression, K Neighbors Classifier and Decision Tree Classifier. The dataset has been taken from [Kaggle](https://www.kaggle.com/ronitf/heart-disease-uci) (https://www.kaggle.com/johnsmith88/heart-disease-dataset).

**CHAPTER – 2**

**LITERATURE REVIEW**

Different researchers have contributed for the development of this field. Predication of heart disease based on machine learning algorithm is always curious case for researchers recently there is a wave of papers and research material on this area. Our goal in this chapter is to bring out all state of art work by different authors and researchers. Marjia Sultana, Afrin Haider and Mohammad ShorifUddin have illustrated about how the datasets available for heart disease are generally a raw in nature which is highly redundant and inconsistent. There is a need of pre-processing of these data sets; in this phase high dimensional data set is reduced to low data set. They also show that extraction of crucial features from the data set because there is every kind of features. Selection of important features reduces work of training the algorithm and hence resulted in reduction in time complexity. Time is not only single parameter for comparison other parameters like accuracy also play vital role in proving effectiveness of algorithm similar. An approach proposed in have worked to improved the accuracy and found that performance of Bayes Net and SMO classifiers are much optimal than MLP, J48 and KStar. Performance is measured by running algorithms (Bayes Net and SMO) on data set collected from WEKA software and then compared using predictive accuracy, ROC curve, ROC value. It is an approach for early detection of heart disease by utilizing variety of feature. These kind of approach can also be utilize for other sphere of research. Here input output are depressive and non depressive categories in the hidden layer scaled conjugate gradient algorithm is used for training to achieve efficient result. authors have got efficiency up to 95% with help of trained neural network watching the success of neural network researches working in the domain of KNN have used this technique to classify and achieve more better result in case where the feature vector are multi dimensional and non linear these method defeated all other existing quantum contemporary techniques because it has capability to work under dataset of high dimensionality. After going through majority of state of art technique we have pointed out certain loop holes existed in them. Some of them are discussed below

• There is wide need for more robust algorithm which can minimised the noise in the dataset because medical dataset may consists of various types of redundancy and noise in them.

• Recently with advancement in deep learning there could be chance to enhance efficiency and accuracy for detection heart disease

• Dimensionality of medical dataset is very high these put ergs to find such algorithm which can compress and reduce higher dimensionality, resulting in gaining execution time.

**CHAPTER – 3**

**Problem Statement**

Heart disease has created a lot of serious concerned among researches; one of the major challenges in heart disease is correct detection and finding presence of it inside a human. Early techniques have not been so much efficient in finding it even medical professor are not so much efficient enough in predicating the heart disease. There are various medical instruments available in the market for predicting heart disease there are two major problems in them, the first one is that they are very much expensive and second one is that they are not efficiently able to calculate the chance of heart disease in human. According to latest survey conducted by WHO, the medical professional able to correctly predicted only 67% of heart disease so there is a vast scope of research in area of predicating heart disease in human. With advancement in computer science has brought vast opportunities in different areas, medical science is one of the fields where the instrument of computer science can be used. In application areas of computer science varies from metrology to ocean engineering. Medical science also used some of the major available tools in computer science; in last decade artificial intelligence has gained its moment because of advancement in computation power. Machine Learning is one such tool which is widely utilized in different domains because it doesn’t require different algorithm for different dataset. Reprogrammable capacities of machine learning bring a lot of strength and opens new doors of opportunities for area like medical science. In medical science heart disease is one of the major challenges; because a lot of parameters and technicality is involve for accurately predicating this disease. Machine learning could be a better choice for achieving high accuracy for predicating not only heart disease but also another diseases because this vary tool utilizes feature vector and its various data types under various condition for predicating the heart disease, algorithms such as Logistic Regression, Decision Tree, KNN are used to predicate risk of heart diseases each algorithm has its speciality such as Naive Bayes used probability for predicating heart disease, whereas decision tree is

used to provide classified report for the heart disease, whereas the Neural Network provides opportunities to minimize the error in predication of heart disease. All these techniques are using old patient record for getting predication about new patient. This predication system for heart disease helps doctors to predict heart disease in the early stage of disease resulting in saving millions of life. This survey paper is dedicated for wide scope survey in the field of machine learning technique in heart disease. Later part of this survey paper will discusses about various machine learning algorithm for heart disease and their relative comparison on the various parameter. It also shows future prospectus of machine learning algorithm in heart disease. This paper also does a deep analysis on utilization of deep learning in field of predicting heart disease.

**CHAPTER – 4**

**Proposed Solution**

**In this study, performances of classification techniques were compared in order to predict the presence of the patients getting a heart disease. A retrospective analysis was performed in 303 subjects. We compared the performance of Logistic Regression(LR), decision trees(DTs), and K Nearest Neighbor(KNN). The variables were medical profiles are age, Sex, Chest Pain Type, Blood Pressure, Cholesterol, Fasting Blood Sugar, Resting ECG, Maximum Heart Rate, Induced Angina, Ole Peak, Slope, Number Colored Vessels, Thal and Concept Class. We have created the model using logistic regression classifiers, K Nearest neighbor and decision trees that they are often used for classification problems. Performances of classification techniques were compared using lift chart and error rates. In the result, artificial neural networks have the greatest area between the model curve and the baseline curve. KNN exhibited accuracy of 88.2%, while the decision tree provided the prediction performance with an accuracy of 81.96%. And the logistic regression provided the prediction performance with an accuracy of 85.24%. Logistic Regression have the least of error rate and has the highest accuracy, therefore Logistic Regression is the best technique to classify in this data set.**

|  |  |
| --- | --- |
| **Algoritm Used** | **Accuracy** |
| **Logistic Regression** | **85.24%** |
| **KNN** | **88.2%** |
| **Decision Tree** | **81.96%** |

**Logistic Regression :**

It’s a type of statistical regression analysis method used for approximation and prediction of result of a definite dependent attributes. Dependent means it can take only some set of values for example binary values such as true or false, good or bad, on or off likewise. Logistic regression is mainly used for prediction besides that it can also be used in calculating the probability of success. Basically Logistic Regression involves fitting an equation of the form to the data:

Y = ß0 + ß1x1 + ß2x2 + … + ßnxn – eq. 1

The regression coefficients are usually estimated using maximum likelihood estimation. The maximum likelihood ratio helps to determine the statistical significance of independent variables on the dependent variables. The likelihood-ratio test assesses the contribution of individual predictors (independent variables). Then the probability (p) of each case is calculated using odds ratio,

P/(1-P) = eY – eq. 2

From this p–value is found out. This gives the probability or chance for the individual to have coronary heart disease.

**K- NEAREST NEIGHBOUR ALGORITHM (KNN) :**

KNN is slow supervised learning algorithm, it take more time to get trained classification like other algorithm is divided into two step training from data and testing it on new instance . The K Nearest Neighbour working principle is based on assignment of weight to the each data point which is called as neighbour. in K Nearest Neighbour distance is calculate for training dataset for each of the K Nearest data points now classification is done on basis of majority of votes there are three types of distances need to be measured in KNN Euclidian, Manhattan, Minkowski distance in which Euclidian will be consider most one the following formula is used to calculate their distance:

𝐸𝑢𝑐𝑙𝑒𝑑𝑖𝑎𝑛 𝐷𝑖𝑠𝑡𝑎𝑛𝑐𝑒 = 𝐷 𝑥, 𝑦

= (𝑥𝑖 −𝑦𝑖)2𝑘𝑖 = 1 (1)

K=number of cluster x ,

y=co-ordinate sample spaces

𝑀𝑎𝑛ℎ𝑎𝑡𝑡𝑎𝑛 𝑑𝑖𝑠𝑡𝑎𝑛𝑐𝑒 = (𝑥𝑖 − 𝑦 )𝑖=1 (2)

x&y are co-ordinates

Minkowski distances are generally Euclidian distance

𝑀𝑖𝑛 = ( 𝑖 − 𝑦𝑖 𝑝 )1 𝑝 (3)

Grouping of sample is based on super class in the KNN reduction of sample is the result of proper grouping which is used for further training. Selection of k value plays a pivotal role, if the k value is large then it precise and less noisy.

**DECISION TREE :**

Decision tree is a graphical representation of specific decision situation that used for predictive model, main component of decision tree involves root, nodes, and branching decision. There are few approaches for building tree such as ID3, CART, CYT, C5.0 and J48 has used the approaches to classify the dataset using J48, similarly have compared decision tree with classification output of other algorithm. Decision tree is used in those area of the medical science where numerous parameters involved in classification of data set.

Since decision tree is most compressive approach among all machine learning algorithm. These clearly reflect important features in the data set. In heart disease where number of parameter affect patient such as blood pressure, blood sugar, age, sex, genetic and other factor. By seeing decision tree, doctor can clearly identifies the most effecting feature among all the parameter. They can also generate the most affecting feature in the mass of population. Decision tree is based on entropy and Information gain clearly signifies the importance of dataset. Drawback of decision tree is that it suffers from two major problems over fitting and it is based on greedy method. over fitting happened due to decision tree spilt dataset aligned to axis it means it need a lot of nodes to spilt data, this problem is resolved by J48 explained in based on greedy method lead to less optimal tree, if dynamic approach is taken it lead to exponential number of tree which is not feasible.

**CAPTER – 5**

**Experimental Setup And Result Analysis**

**Classifiers Used for the Experiment:**

# **Import libraries**

I imported several libraries for the project:

1. **numpy**: To work with arrays
2. **pandas**: To work with csv files and dataframes
3. **matplotlib**: To create charts using pyplot, define parameters using rcParams and color them with cm.rainbow
4. **train\_test\_split**: To split the dataset into training and testing data
5. **StandardScaler**: To scale all the features, so that the Machine Learning model better adapts to the dataset

Next, I imported all the necessary Machine Learning algorithms.

**There are four phases that involve in the spiral model:**

**1) Planning phase**

Phase where the requirement are collected and risk is assessed. This phase where the title of the project has been discussed with project supervisor. From that discussion, Heart Prediction System has been proposed. The requirement and risk was assessed after doing study on existing system and do literature review about another existing research.

**2) Risk analysis Phase**

Phase where the risk and alternative solution are identified. A prototype are created at the end this phase. If there is any risk during this phase, there will be suggestion about alternate solution.

**3) Engineering phase**

At this phase, a software are created and testing are done at the end this phase.

**4) Evaluation phase**

At this phase, the user do evaluation toward the software. It will be done after the system are presented and the user do test whether the system meet with their expectation and requirement or not. If there is any error, user can tell the problem about system.

**Data Flow Diagram:**

**Dataset**

**Rule Generation**

**Result**

**Accuracy Calculation**

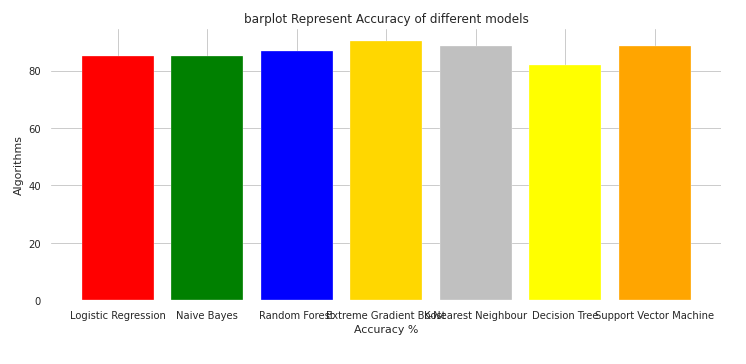
**Prediction**

**Pattern Matching**

**Preprocessing**

**Accuracy Comparison** :

Accuracy comparison between Logistic Regression, Naïve Bayes, Rando Forest Classifier, Extreme Gradient Boost, K-Nearest Neighbour, Decision Tree and Support Vector Machine.



**What is Flask ?**

**Flask** is a micro web Framework written in Python. It is classified as a microframework because it does not require particular tools or libraries.It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

## Scope and Limitation

## Scope.

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

## Limitations.

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

**Conclusion And Future Work**

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio-patients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health’s diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.