SISTER NIVEDITA UNIVERSITY

DG Block (Newtown), Action Area I, 1/2, Newtown, New Town, West Bengal 700156

HEART DISEASE PREDICTION USING MACHINE LEARNING

A project submitted in partial fulfilment of the requirements for the degree of
Bachelor of Technology

PRESENTED BY
SHUVRA CHAKRABARTY
REG. NO. – 180020328417
ENROLLMENT NO.- 1811100002032

UNDER THE SUPERVISION OF
DR. DEBBROTA PAUL CHOWDHURY
ASSISTANT PROFESSOR, DEPARTMENT OF CSE, SNU

CERTIFICATE

TO WHOM IT MAY CONCERN

This is to certify that the project "HEART DISEASE PREDICTION USING MACHINE LEARNING" is a work done by SHUVRA CHAKRABARTY student of 7th Semester in B. Tech (CSBS), department of Computer Science and Engineering, Sister Nivedita University, in partial fulfilment of requirement for the degree of "Bachelor of Technology in Computer Science and Business System".

……

CHAIRMAN,

BOARD OF EXAMINER

SISTER NIVEDITA UNIVERSITY

……

DR. DEBBROTA PAUL CHOWDHURY

(PROJECT GUIDE)

DEPARTMENT OF COMPUTER SC. AND ENGINEERING

SISTER NIVEDITA UNIVERSITY

……..

SIGNATURE OF EXTERNAL EXAMINER

ACKNOWLEDGEMENTS

First and foremost, I thankful to the Sister Nivedita University, Computer Science and Business System Engineering Department and dr. Debbrota Paul Chowdhury, Assistant Professor, Computer Science and Engineering Department, for his continued guidance and support for My project work.

- **BACKGROUND**

In comparison to the brain, which is the most important organ in the human body, the heart is the next most important organ. It circulates blood and supplies all of the body's organs. Heart disease ailments are the most common of all diseases. Medical experts undertake various reviews on heart diseases and compile information on heart patients, their symptoms, and disease development. Many harmful habits, such as excessive cholesterol, obesity, increased triglyceride levels, hypertension, and so on, raise the risk of heart disease. People in this fast-paced world want to live a very luxurious life, so they work like machines in order to earn a lot of money and live a comfortable life. As a result, they forget to look after themselves, and their food habits and lifestyles change as a result. They are more tense, have high blood pressure and sugar levels at a young age, and they don't give themselves enough rest and eat whatever they get.

The number of persons who have heart disease is increasing. According to the World Health Organization, an estimated 17 million people die each year from cardiovascular disease, particularly heart attacks and strokes. Heart disease is also stated as one of the greatest killers in Africa, India. Marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining, and mobile computing are just a few of the applications where machine learning has been applied. Of late, machine learning has been applied successfully in healthcare fraud and detecting abuse cases.

- **OBJECTIVES**

Objective of this project is to predict heart disease using machine learning. Also developed a web application for users using this machine learning model to predict heart disease.

- **USED METHODS**

In this project I used some machine learning algorithms like, - KNN, SVM, Naive Bayes, Decision Tree, Random Forest and also use Django for web application.

- **FINAL RESULT IN SHORT**

From all applied algorithms the K-nearest neighbors algorithms showing the best results in this model. So, I take this algorithm in my model and use it to develop the application.

- **KEYWORDS**

Machine Learning, Django, KNN, SVM, Random Forest, Naïve Bayes, Decision Tree

CONTENTS

1. INTRODUCTORY CONCEPTS

1.1. INTRODUCTION

Heart is an imperative organ of the human body. It pumps blood to every portion of our life structures. On the off chance that it comes up short to operate accurately, at that point the brain and different other organs will halt working, and inside few minutes, the individual will pass on. Alter in way of life, work related stress and awful nourishment propensities contribute to the increment in rate of several heart related maladies. Cardiovascular disease is increasing daily in this modern world. According to the World Health Organization, an estimated 17 million people die each year from cardiovascular disease, particularly heart attacks and strokes. The European Open Wellbeing Union detailed that heart assaults, strokes and other circulatory diseases account for 41% of all passing. In India, more than 2 lake open heart surgeries are done per year. The patients affected by the heart attack is growing in India is 20% to 30 % every year. The development of the sensor network in the human monitoring system is more applicable from recent years. A few diverse side effects are associated with heart infection, which makes it troublesome to analyze it faster and superior. Working on heart disease patients' databases can be compared to real-life application. Specialists' information to relegate the weight to each trait. More weight is allotted to the trait having tall effect on infection prediction. Therefore, it appears sensible to undertake utilizing the information undertake utilizing the information and encounter of a few pros collected in databases towards helping the Conclusion prepare. It moreover gives healthcare experts an additional source of knowledge for making decisions. Effective and efficient automated heart disease prediction systems can be beneficial for heart disease prediction. Our work attempts to present the detailed study about the different machine learning techniques which can be deployed in these automated systems. This automation will also reduce the number of tests to be taken by a patient.

In countries such as India and the United States, heart disease was the leading cause of death. Machine Learning techniques like Classification algorithms such as Decision Tree, K-Nearest Neighbors algorithm, Naive Bayes, Support Vector Machines, are used to explore the different kinds of heart - based problems. In this project we use these techniques and bud a Machin learning base web application for healthcare workers.

1.2. APPROACH METHODOLOGY

My Project aims to foresee the odds of having heart disease as probable cause of computerized prediction of heart disease and also using this machine learning model develop a web application that is helpful in the medical field for clinicians and patients. To accomplish the aim, we have discussed the use of various machine learning algorithms on the data set and dataset analysis, how I developed this web application all are mentioned in this research paper. This research also shows which characteristics contribute more than others to the prediction of higher precision. This could save money on different patient trials because all of the features may not have a significant role in predicting the outcome.

These objectives are set for this heart prediction system.

- This prediction system shouldn't assume any prior knowledge about the patient records it is comparing.
- The chosen system must be scalable to run against large database with thousands of data.

Those chosen approach is implemented using Anaconda which is a distribution of the Python programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.). Here I use Jupiter notebook from anaconda which is a an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources. To build a web application in this case I use Django which is a Python-based free and open-source web framework.

The following steps are performed to build this project

- I collect the dataset from Kaggle
- Check this dataset using pandas in Jupiter Notebook
- Train this model using machine learning algorithms
- Take highest accuracy model and implement it in web application

2. PROPOSEDMETHODOLOGY

2.1. DATA SOURCE

For this project I have collected the dataset from Kaggle. In this dataset there are 300 examples of data with 14 various attributes like age, Cholesterol, type of chest pain, electrocardiogram result, etc. In this research, we have used four algorithms to get reasons for heart disease and create a model with the maximum possible accuracy.

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0 = female 1 = male
Cp	Discrete	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain 4 =a symptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1-true 0=False
Exang Continuous Maximum heart rate achieved	Discrete	Exercise induced angina: 1 = Yes 0 = No
Thalach	Continuous	Maximum heart rate achieved
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment: 1 = up sloping 2 = flat 3 = down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6 = fixed defect 7= reversible defect

Table 1

This are all heart disease attributes.

2.2. UNDERSTANDING THE DATA

There are a total of 13 characteristics in the dataset, as well as one target variable. There are no missing values, thus we don't have to worry about null values. However, there is one value that is repeated. I removed this value in order to better comprehend the dataset. Following that, I used the describe() function. The method indicated that each variable's range is distinct. Because the maximum age value in this dataset is 77, while the maximum chol value is 564, so feature scaling must be performed on the dataset.

2.3. DATA PRE-PROCESSING

The methods utilized, as well as the quality of the dataset and preprocessing processes, have an impact on the prediction model's performance and accuracy. The actions taken before applying machine learning algorithms to a dataset are referred to as preprocessing. The preprocessing stage is critical because it prepares the dataset and converts it into an algorithm-friendly format. Data cleaning, data transformation, missing values imputation, data normalization, feature selection, and other stages are included in the data preprocessing steps, depending on the nature of the dataset. Many projects contain real-life data that contains high quantities of missing and noisy data. These data are pre-processed to eliminate these flaws and make confident forecasts You can see in this dataset there are no missing value so I no need to done any operation to fix this. So, the data are then classified and split into training data set and test data set which is run on various algorithms to achieve accuracy score results. To work with this data set I split this dataset into 80% training data and 20% testing data.

2.4. PROPOSED MODEL AND MATHEMATICAL BACKGROUND

2.4.1. CLASSIFICATION TECHNIQUE FOR HEART DISEASE PREDICTION

Some Classification Techniques are used in this project. The process of recognizing, comprehending, and organizing objects and thoughts into predetermined categories is referred to as classification. Now Classification algorithm is a Supervised Learning technique that uses training data to identify the category of fresh observations. Now a program learns from a dataset or observations and then classifies additional observations into a number of classes or groups using this approach. Such as Yes or No, 0 or 1, Spam or Not Spam, and so forth. Targets/labels or categories are all terms that can be used to describe classes. The classification technique uses data that has been labelled. That is to say, it has input and output that are cross-ponding.

$$Y = f(x), \text{ where } y = \text{categorical output}$$

Here discrete output is mapped to input variable x

There are various classification techniques used in this project for predicting heart disease. In this section we discuss about them

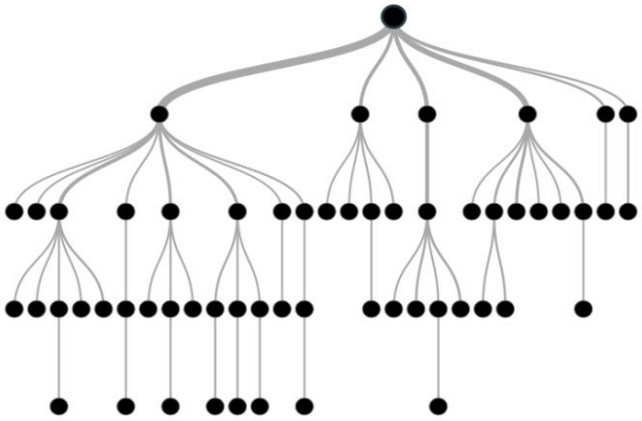
2.4.1.1. DECISION TREE

A decision tree is a classification algorithm that can be used to classify both categorical and numerical data. A decision tree is a type of structure that resembles a tree. It is a basic and extensively used tool for dealing with medical data. It's simple to use and analyses data on a tree-shaped graph. Three nodes form the basis of the decision tree model's analysis. Because it analyses the dataset in a tree-like graph, this technique is more accurate than other algorithms. However, the data could be overclassified, and only one attribute is examined for decision-making

at a time. The entropy of each and every attribute is initially calculated by the Decision Tree algorithm. Then the dataset is split with the help of the variables or predictors with maximum information gain or minimum entropy. The remaining attributes are processed in a recursive manner using these two processes.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$



An accuracy of 68.85 % has been achieved by the decision tree.

2.4.1.2. K-NEAREST NEIGHBOR

The K-nearest neighbor algorithm is a method for supervised classification. It categorizes objects based on their proximity to another object. It's an example of case-based learning. The Euclidean distance is used to calculate the distance between an attribute and its neighbor. It makes use of a set of named points to determine how to mark another point. The data are clustered based on their similarity, and K-NN can be used to fill in the missing values in the data. After the missing values have been filled in, the data set is subjected to a variety of prediction approaches. Using various combinations of these methods, it is feasible to improve accuracy. The K-NN technique is straightforward to implement without the need for a model or other assumptions. The knowledge is extracts based on the samples Euclidean distance function $d(x_i,x_j)$ and the majority of k-nearest neighbors.

$$d\left(x_{i,x_i,}\right) = \sqrt{\left(x_{i,1} - x_{j,1}\right)^2 + \cdots + \left(x_{i,m} - x_{j,m}\right)^2}$$

In This project I use KNN algorithm to predict the accuracy where K=13 then I got about 86.89% accuracy.

2.4.1.3. NAIVE BAYES

The Bayes' Theorem is used to create a collection of classification algorithms known as Naive Bayes classifiers. It is not a single algorithm, but rather a group of algorithms that all follow the same premise. It assumes predictor independence, which means that the traits or features should not be associated with one another or related in any manner. Even though there is a dependency, each of these characteristics or attributes contributes to the probability on its own. The model is trained through the Gaussian function with prior probability $P(X_f)$=priority∈

(0:1).

$$P(X_f) = priority \in (0:1)$$

$$P(X_{f1}, X_{f2}, \ldots, X_{f_n}|c) = \prod_{i=1}^{n} P(X_{fi}|c)$$

$$P(X_f|c_i) = \frac{P(c_i|X_f)\,P(X_f)}{P(c_i)} \quad c \in \{benign,\ malignant\}$$

In This project using Naive Bayes algorithm I get approx. 86% accuracy.

2.4.1.4. SUPPORT VECTOR MACHINE

Support Vector Machines, or SVM, are one of the most widely used and discussed algorithms. They were extremely popular at the time they were developed and refined in the 1990s, and they have remained popular ever since. SVM is one of the best choices for high-performance algorithms with a little tuning, and it provides one of the most robust prediction methods. The training data points are represented as points in the feature space by an SVM model, which is mapped in such a way that points belonging to different classes are separated by as wide a margin as possible. The test data points are then mapped into the same area and categorized according to where they fall on the margin.

Let the training samples having dataset Data=$\{y_i, x_i\}$; i=1,2,...,n where $x_i \in R^n$ represent the i^{th} vector and $y_i \in R^n$ represent the target item. The linear SVM finds the optimal hyperplane of the form f(x)=wTx+b where w is a dimensional coefficient vector and b is an offset. This is accomplished by resolving the following optimization problem:

$$Min_{w,b,\xi_i}\ \frac{1}{2}w^2 + C\sum_{i=1}^{n}\xi_i$$

$$s.t.\ y_i\left(w^{\mathrm{T}}x_{\mathrm{i}} + b\right) \geq 1 - \xi_i, \xi_i \geq 0, \quad \forall_i \in \{1, 2, \ldots, m\}$$

In This project using Support Vector Machines I get approx. 85.24% accuracy.

2.4.1.5. RANDOM FOREST

A supervised learning algorithm is the random forest. This method can be used for both regression and classification tasks, but it excels at classification. It creates a "forest" out of a collection of decision trees, which are commonly trained using the "bagging" method. The bagging method's general premise is that combining many learning models improves the final outcome. This method is based on the assumption that a larger number of trees will eventually reach the same conclusion. The more trees in the forest, the more accurate it is and the problem of overfitting is avoided.

In This project using Random Forest I get approx. 80.30% accuracy using 1000 tree.

3. EXPREMENTAL RESULT AND DISCUSSION

3.1. DATABASE

The Dataset For this project I have used the heart dataset from the KEGGLE Repository. The dataset consists of 302 individual clinical reports in which 164 did not have any disease. In this dataset there is a total of 96 female patients in which 24 people are healthy and other 72 have heart disease, also there are 206 male patients in which 114 are healthy and other 92 have heart disease. Data in this dataset are like

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

Table 2

3.2. EXPERIMENT

In this project I use the data set from Kaggle and using five machine learning algorithm I want to predict heart disease. During this process I can find any null value in this data set for that reason I scaled the data set and train it. Using those algorithms, I got different accuracy.

3.3. COMPARISON OF METHODOLOGIES

Comparison Of Different Algorithm Results

Algorithm	Result
DECISION TREE	68.85%
K-NEAREST NEIGHBORS	86.89%
NAIVE BAYES	86%
SUPPORT VECTOR MACHINE	85.24%
RANDOM FOREST	80.30%

Table 3

3.4. RESULT

The purpose of this work is to predict whether or not a patient will develop heart disease. This research was done on supervised machine learning classification techniques using decision tree and K-nearest neighbor, SVM, Random Forest, Naive Bayes on Kaggle repository. Various experiments using different classifier algorithms were conducted through the Jupiter notebook tool. Dataset was classified and split into a training set and a test set. Pre-processing of the data is done and supervised classification techniques to get accuracy score. The accuracy score results of different classification techniques were noted using Python Programming for training and test data sets.

4. APPLY MODEL IN WEB APPLICATION

From all algorithm the best result I got from KNN (86.89% accuracy). So, I take this algorithm tarin this model and convert this model int .sav file using python library joblib (Joblib is a set of tools to provide lightweight pipelining in Python). Then I use this .sav file and implement it in my Django web application where I take data from users and then predict those data output using this model then show output (Negative, Positive) to user.

To developed this web application, I use Django (Python base web framework), Html, CSS, JavaScript. I choose Django because the programming language it support is also Python.

Live Web Application Demo Link – Heart Disease Prediction

5. CONCLUSION AND FUTUREWORK

5.1. CONCLUSION

The goal of this study was to use machine learning to compare algorithms using different performance indicators. All of the data was pre-processed before being included in the test prediction. In some cases, each algorithm performed better than the others. The main goal is to establish several data mining approaches that can be used to accurately forecast cardiac disease. Our goal is to provide efficient and reliable prediction with a smaller number of features and tests. Only 14 critical features are considered in this study. I used five classification techniques: K-nearest neighbor, decision tree, SVM, Nave Bayes, and Random Forest, with percentage accuracy results of 86.86 percent, 68.85 percent, 85.25 percent, 86 percent, and 80.30 percent. The information was pre-processed before being utilized in the model. K-nearest neighbor algorithms showing the best results in this model. The best result is got from KNN 86.89% where k=13.

5.2. FUTURE WORK

To improve the scalability and accuracy of this prediction system, a number of changes could be investigated. I plan to enhance this project by including other approaches such as the XGBoost Classifier, Gradient Boost, and Ensemble Learning in the future. Given the study's limitations, more complicated and combined models are needed to improve early heart disease prediction accuracy. I'd like to work with a larger dataset on the same issue in the future, as well as optimize and add new features to the web application.

6. REFERENCES

- N. Khateeb and M. Usman, “Efficient heart disease prediction system using k-nearest neighbor classification technique,” in Proceedings of the International Conference on Big Data and Internet of Thing (BDIOT), New York, NY, USA: ACM, 2017, pp. 21–26. Chaurasia V, Pal S. Data mining approach to detect heart diseases. Int J Adv Compute Sci Inf Technol (IJACSIT). 2014; 2:56–66.
- Mr. Chala Beyene, Prof. Pooja Kamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques”, International Journal of Pure and Applied Mathematics, 2018.
- Animesh Hazra, Arkomita Mukherjee, Amit Gupta, Asmita Mukherjee, “Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review”, Research Gate Publications, July 2017, pp.2137-2159.
- T.Mythili, Dev Mukherji, Nikita Padaila and Abhiram Naidu, “A Heart Disease Prediction Model using SVMDecision Trees- Logistic Regression (SDL)”, International Journal of Computer Applications, vol. 68, 16 April 2013.
- Jayami Patel, Prof. Tejal Upadhay, Dr. Samir Patel, “Heart disease Prediction using Machine Learning and Data mining Technique”, March 2017. K. Polaraju, D. Durga Prasad, “Prediction of Heart Disease using Multiple Linear Regression Model”, International Journal of Engineering Development and Research Development, ISSN:2321-9939, 2017.
- “The Atlas of Heart Disease and Stroke”, [online]. http://www.who.int/cardiovascular_diseases/res ources/atlas/en/
- Ramadoss and Shah B et al. “A. Responding to the threat of chronic diseases in India”. Lancet. 2005; 366:1744–1749. doi: 10.1016/S0140-6736(05)67343-6.
- C. S. Dangare and S. S. Apte, “Improved study of heart disease prediction system using data mining classification techniques,” International Journal of Computer Applications, vol. 47, no. 10, pp. 44–48, 2012.
- Y. E. Shao, C.-D. Hou, and C.-C. Chiu, “Hybrid intelligent modelling, schemes for heart disease classification,” Applied Soft Computing, Vol. 14, pp. 47–52,2014
- World Health Organization, *Cardiovascular Diseases*, WHO, Geneva, Switzerland, 2020, https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.
- American Heart Association, *Classes of Heart Failure*, American Heart Association, Chicago, IL, USA,2020, https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure.
- Soni, J., Ansari, U., & Sharma, D. (2011). Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. Heart Disease, 3(6), 2385–2392.