Hearth Disease Prediction using Machine Learning

Taksh Rastogi

Business Analytics and Big Data

Thapar University

***Abstract-* In living creatures, the heart serves an important role. Diagnosis and forecasting heart illnesses necessitate greater precision, perfection, and accuracy because even a minor error can result in exhaustion or death. There are numerous death instances related to the heart, and the number is growing exponentially day by day. However, it is not feasible to monitor patients every day in all cases with good accuracy and precision. Doctors' consultation is not readily available to the patients because of their reduced count, limited time, and expertise. To address the issue, a predictive system for disease awareness is required. Machine learning is a branch of Artificial Intelligence (AI) that offers prestigious assistance in anticipating any event using natural events as training datasets. Using the same application, we calculated the accuracy of machine learning methods for predicting cardiac disease in this research. We used various heart attributes of patients and used a variety of machine learning algorithms like Logistic regression, Naïve Bayes algorithm, decision tree, and KNN classifier. The dataset used was available publicly on the Kaggle website. This project aims to realize the early prognosis of cardiovascular diseases, which can help make decisions on lifestyle changes in high-risk patients and reduce the complications, which can be a significant milestone in the field of medicine.**

1. **INTRODUCTION**
2. Every year, millions of people develop some form of heart disease, and heart disease is the leading cause of death in both men and women around the world. The World Health Organization (WHO) studied the twelve million fatalities caused by heart disease each year around the world. Heart disease kills one person nearly every 34 seconds around the world. Medical diagnosis is an important yet difficult process that must be completed quickly and precisely. Appropriate computer-based information and decision assistance should be helped to lower the cost of achieving clinical testing. The use of software tools for detecting patterns and consistency in sets of data is known as machine learning and data mining. In addition, since the emergence of data mining in the last two decades, there has been a surge in Allowing computers to generate and classify distinct qualities or classes directly presents a significant opportunity. Learning about the risk factors for heart disease aids medical professionals in identifying people who are at high risk for heart disease. Age, blood pressure, total cholesterol, diabetes, hypertension, family history of heart disease, obesity and lack of physical activity, fasting blood sugar, and other risk factors for heart disease have been identified through statistical analysis.*B.* ***About Problem***

The heart is one of the fastest and most crucial organs in the human body, and it requires special attention. Because most diseases are linked to nature, it is necessary to predict heart diseases, which necessitates a comparative study in this field. Today, most patients die because their conditions are detected at an advanced stage due to instrument inaccuracy, so there is a need to learn about more efficient algorithms for disease prediction. Having a huge chunk of data is a bane for this problem as we can use various machine learning algorithms to analyze the information for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

1. ***Motivation***

The primary motivation behind this research-based project is to explore the feature selection methods, data preparation, and processing behind the training models in machine learning. We need to have a mechanism in place to recognize the symptoms of a heart stroke early and so avoid it, given the rapid rise in heart stroke rates among children and adolescents. Because it is impractical for the average person to regularly undergo pricey tests like the ECG, there must be a convenient and reliable system for forecasting the risk of heart disease. As a result, we propose developing an app that can predict the vulnerability of a cardiac illness based on fundamental symptoms such as age, sex, pulse rate, and so on.

1. ***Objectives***

The prime objective of developing this project is to develop a machine learning model to predict the future possibility of heart disease by implementing algorithms like Logistic regression, Naïve Bayes, Decision Tree, and KNN. The project also aims to find the accuracy and precision of models created by the above algorithms.

1. **RELATED WORK**

Different methodologies and machine learning techniques Have been used to present a variety of models for product recommendation.

Kumar et al. [1] studied various machine learning and data mining methods using the UCI machine learning dataset, which contains 303 samples with 14 input features, and discovered that SVM is the best. Other algorithms studied include naive Bayes, knn, and decision tree.

Gavhane et al. [2] studied the multi-layer perceptron model for human heart disease prediction and the algorithm's accuracy using CAD technology. If more people use the prediction system to anticipate their ailments, the public's knowledge of those diseases will also rise.

Some researchers are working on one or two illness prediction systems. In their project, Krishnan et al. [3] demonstrated that decision trees are more accurate than the naive Bayes classification algorithm.

Machine learning algorithms are used to predict many diseases, and many researchers, such as Kohali et al., have worked on this. [4] conducted research on heart disease prediction using logistic regression, diabetes prediction using support vector machines, and breast cancer prediction using logistic regression, finding that logistic regression has an accuracy of 87.2 percent and support vector machines have an accuracy of 87.2 percent. Of 85.71%, the Adaboot classifier gives the accuracy up to 98.56 percent, which is good from a prediction point of view.

Table 1 Summary of related work

|  |  |  |  |
| --- | --- | --- | --- |
| References | Technique Used | Dataset Used | Performance measure used |
| Kumar et al. [ [1] | Svm,  naivy bayes, knn,  decision tree, | computer simulation on dataset available in UCI Machine Learning Repository | ACC |
| Gavhane et al. [2] | Naive Bayes, Decision Tree, KNN, Neural Network | Cleveland dataset from UCI library | Precision, recall, ACC, ROC curve, ROC value, |
| Krishnan et al. [3] | naivy bayes,  Decision tree,  WEKA (Waikato Environment for Knowledge Analysis) | taken from UCI Machine learning repository. | ACC |
| Kohali et al. [4] | logistic regression, Decision Trees, Random Forest, Support Vector Machine (SVM) and Adaptive Boosting, | UCI machine learning library | ACC |

**REFERENCES**

[1] M. Nikhil Kumar, K. V. S. Koushik, and K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools," IJSRCSEIT, 2019.

[2] S Geetha and Krishnan J Santhana, "Prediction of Heart Disease Using Machine Learning Algorithms," ICIICT, 2019.

[3] Aditi Gavhane, Gouthami Kokkula, Isha Panday, and Kailash Devadkar, "Prediction of Heart Disease Using Machine Learning," Proceedings of the 2nd International Conference on Electronics, Communication, and Aerospace Technology (ICECA) (ICECA), 2018.

[4] Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Disease Prediction," 4th International Conference on Computing, Communication, and Automation (ICCCA), 2018.

[5]A. H. M. S. U. Marjia Sultana, "Analysis of Data Mining Techniques for Heart Disease Prediction," 2018.

[6] M. I. K. ,. A. I. ,. S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms".

[7] M. A. K. S. H. K. M. a. V. P. M Marimuthu, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach".

[8] K. Bhanot, "towarddatascience.com," 13 Feb 2019. [Online]. Available:

<https://towardsdatascience.com/predicting-presence-of> heart diseases-using-machinelearning-36f00f3edb2c. [Accessed 2 March 2020].

[9] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

1. **DATASETS**

The dataset was available publicly on the Kaggle website. It provided patient information, which includes over 1026Records and 13 attributes. The attributes include: Age (age in years), sex, CP (chest pain type), TRESTBPS (resting blood pressure (in mm Hg on admission to the hospital)), CHOL (serum cholesterol in mg/dl), FPS (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false), RESEARCH (resting electrocardiographic results), THALACH (maximum heart rate achieved), EXANG (exercise-induced angina (1 = yes; 0 = no)), OLD PEAK (ST depression induced by exercise relative to rest), SLOPE (the slope of the peak exercise ST segment), THAL. The data set is in CSV (Comma Separated Value) format which is further prepared to data frame as supported by pandas library in python.

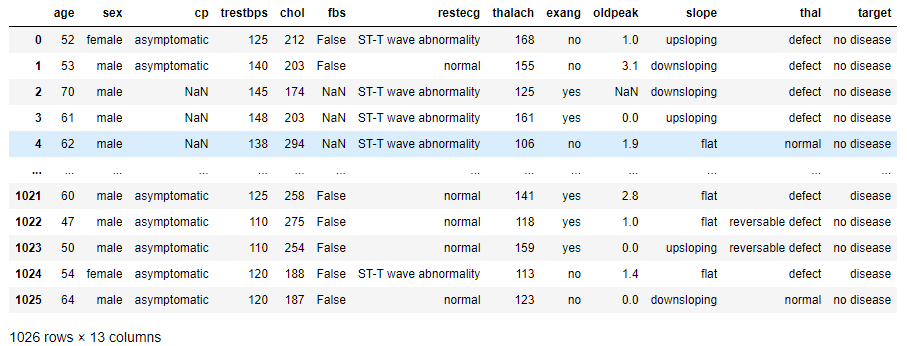
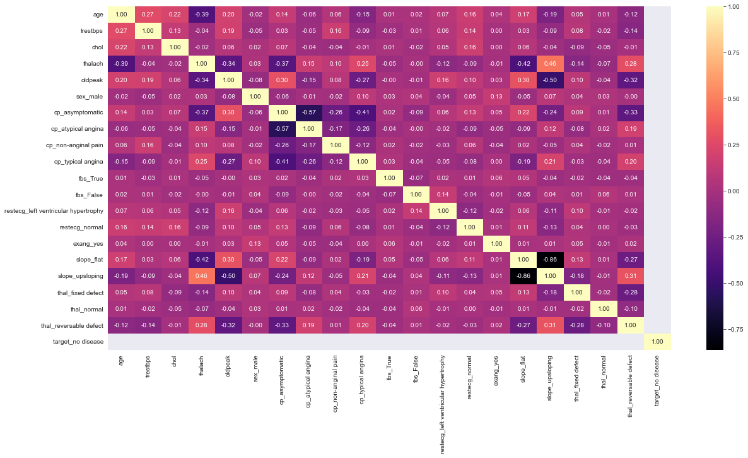
Fig 1: Snapshot of original dataset of patients

Fig 2: Correlation heatmap

A correlation heatmap is a valuable tool to represent how two features are related to each other graphically. We can see a positive correlation between chest pain (cp) & target (our predictor). This makes sense since the more significant amount of chest pain results in a greater chance of having heart disease. In addition, we see a negative correlation between exercise-induced angina (exam) & our predictor. This makes sense because when you exercise, your heart requires more blood, but narrowed arteries slow down blood flow.

1. **PROCEDURE, &METHODS(ALGORITHMS)**

1) Import all the modules for feature selection,

normalization, data splitting, ML models, accuracy

score, confusion matrix, classification report and

some other required modules.

2) Load the heart disease dataset.

3) Divide the datasets as feature and class.

4) Check the significant features for the prediction of class.

5) Normalize the features to scale in one range with

Standard scaling.

6) Split the dataset into two training and testing sets in

80:10 respectively.

7) Build a various machine learning models using

bagging techniques

8) Print accuracy and classification reports of different models as comparing the actual and predicted class

9) Plot confusion matrices of different models comparing the actual and predicted class.

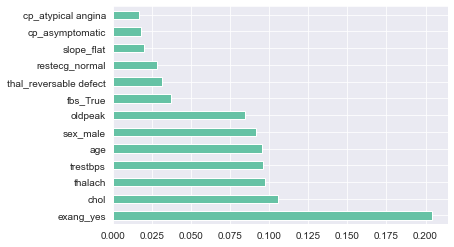
Significant Features

Fig 3: graph of feature importance

Feature importance improves the model’s performance. It is

essential to know the effect of a particular feature on the model’s performance. Some are significantly less important that can reduce model performance. A different tree classifier algorithm evaluates significant features with 250 estimation trees of the forest. This extra tree classifier worked based on a standard deviation of data. And indices of this algorithm sort features in decreasing order ranking.

Standard Scale or Z- Normalization

Features may differ in scale or units, and it is difficult for a classifier or regression to give optimal results. We need to scale it in one specific range to overcome this difficulty. For standard scaling, first, subtract its value from its mean to bring its mean around zero, then divide it by its normal deviation value to obtain its standard deviation near to one. The function/ Class used for this step is StandardScaler



**Machine Learning (ML) Approach**

This is a binary classification problem (has-disease or no-disease cases). Scikit learn offers a wide range of classification algorithms and is often the starting point in most/traditional machine learning challenges, so we start by exploring few of the classification alorithms from the sklearn libarary such as Logistic Regression, Nearest Neighbors, Decision Tree and Naive Bayes.

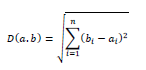
**Logistics Regression:**

Logistic regression is a supervised statistical machine learning model that classifies the data by considering outcome variables on extreme end class and tries to makes a logarithmic line that separates between them. It is a predictive analysis algorithm based on the concept of probability. It measures the relationship between response and predictors by estimating probabilities using underlying sigmoid function. Sigmoid functions act as a cost function used to limit the hypothesis of logistic regression between 0 and 1. Logistic regression gives 69.42 % accuracy in our approach.



1. ***K-Nearest Neighbor (KNN):***

KNN algorithm assumes the similarity between the new and available data and put the new data into the category that is most similar to the available categories. The classification is done on the basis of majority class neighbours. Here we have chosen k=5 and it has given us maximum accuracy of 67.4%



**Naïve Bayes Classifier**

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

This approach gave us the accuracy of 47.57%

**[lowest among all]**

1. ***Decision Tree:***

Decision tree algorithms uses the hierarchical tree

approach, there every node represents a feature, and branch

represents a decision and leaves represents an outcome (class). In this approach we got the accuracy of 64.56% taking information gain as evaluation measure and 68.44% taking Gini Index as the evaluating measure. Evidently Gini Index is giving us better results.



1. **EXPERIMENTAL RESULTS ANALYSIS**

There are several metrics that can be used to gauge the performance of a given classification algorithm.

1. **Confusion matrix**: A confusion matrix (aka an error matrix) is a specific table layout that allows visualization of the performance of a supervised learning algorithm. Each row of the matrix represents the instances in an *actual* class while each column represents the instances in a *predicted* class

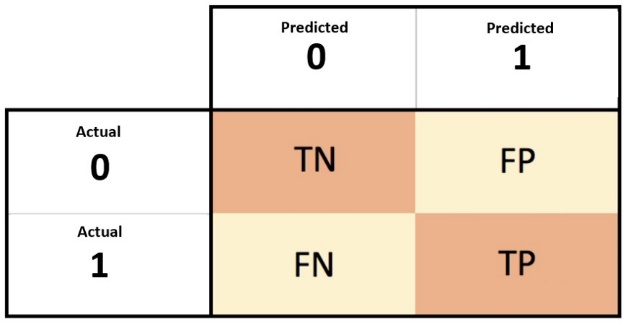
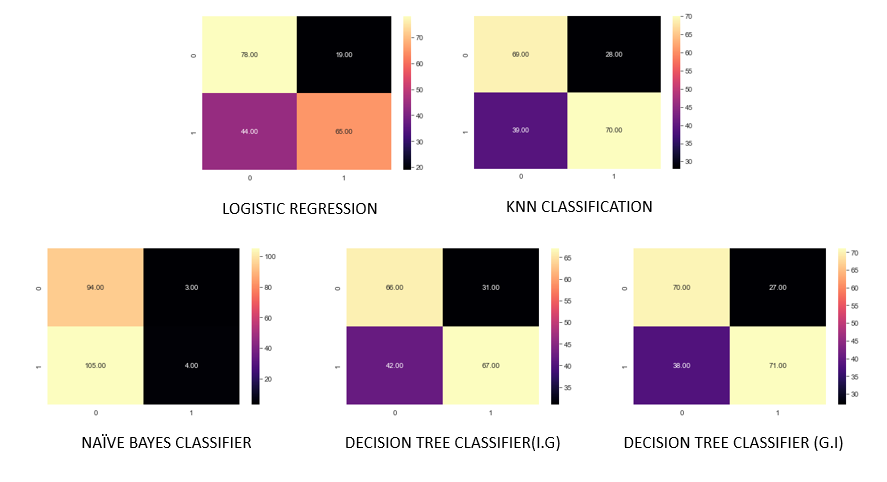


Fig:4 Confusion matrix for a binary classification

 Fig:5 Confusion matrices for all the classification algorithms used

1. **Accuracy**: Measures how many of the cases are correctly identified/predicted by the model, i.e correct prediction divided by the total sample size.



1. **Recall**: Measures the rate of *true positives*, i.e how many of the actual positive cases are *identified/predicted* as positive by the model. Defined as a ratio of correctly positive outcomes divide to the total number of positive outcomes



1. **Precision**: Measures how many of the positive predicted cases are actually positive



1. **F1-Score**: Combines the precision and recall of the model and it is defined as the harmonic mean of the model’s precision and recall.



Table 2 Classification Report (No disease)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CLASSIFIER | ACCURACY | PRECISION | RECALL | F-1 SCORE |
| Logistic Regression | 69.41% | 63.93 | 80.41 | 71.2 |
| KNN Classifier | 67.47% | 64.00 | 71.00 | 67.00 |
| Naïve Bayes Classifier | 47.57% | 47.00 | 97.00 | 64.00 |
| Decision Tree Classifier  (I.G) | 64.56% | 62.00 | 67.00 | 65.00 |
| Decision Tree Classifier  (G.I) | 67.96% | 64.00 | 71.00 | 68.00 |

The classification report gives the information about precision, recall, F1-score. Classification Report describe about the models behavior and its reports and thus helps in problems identification.

1. **CONCLUSION**

In the start of our project, we aimed to create a model which could best predict if someone will face heart disease problem in their future. For that we explored our dataset, replace null values with appropriate values, handled categorical values, performed standardization and normalized the dataset. We also tried to identify correlation between features and also with the target variable.

Later on, we practiced how to set-up binary classifiers using various base models. Out of the 4 machine learning algorithms used for model creation, logistic regression returned the highest accuracy.

Moreover, Out of the 13 features we examined, the top 3 significant features that helped us classify between a positive & negative Diagnosis were chest pain type (cp), maximum heart rate achieved (thalach) and ST depression induced by exercise relative to rest (oldpeak). These features are also among better correlated features from our EDA.