IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A STUDY ON HEART ATTACK PREDICTION USING MACHINE LEARNING ALGORITHMS AND PROVIDE EARLY SUGGESTION TO REDUCE FATALITY

Monalisha Sahoo[1], *Purnalaxmi Panda*[2], *Smrutirekha Das*[3], *Priyansa Priyadarsni Pani*[4]. *Chandan Kumar Panda*[5]

[1,2,3,4,5] C V Raman Global University, Bhubaneswar, Odisha

1.1 ABSTRACT

To detect and predict heart attack, we use the dataset from the UCI repository by the use of machine learning classifiers. Enhancing accuracy, facilitating early detection, and allocating healthcare resources as efficiently as possible are the objectives. Our main goal is to identify most suitable classifier which we can use for diagnostic applications. To predict heart disease, we use several machine learning approaches are used and compare their accuracy to get best result. This study discovered that the RF approach obtained 78% accuracy coupled with 100% sensitivity and specificity utilizing a heart attack dataset that we gathered from Kaggle three-classification based on k-nearest neighbour (KNN), decision tree (DT), and random forests (RF) algorithms. To construct a model training and testing of data evaluated. Accuracy, precision, recall, and F1-score were used to evaluate the models. With an accuracy of 80.33%, the SVM model outperformed the others.

Keywords: Model fitting, Random Forest, Decision Tree, K nearest neighbours, Support vector machine, Naïve Bayes, Logistic Regression, EDA (Exploratory data analysis)

1.2 INTRODUCTION

The heart, a muscular organ at the core of the circulatory system, works in conjunction with the lungs and a network of blood vessels, including veins, arteries, and capillaries, to pump blood throughout the body. Cardiovascular diseases (CVDs), such as myocarditis, vascular disease, and coronary artery disease, are leading causes of death globally. According to the World Health Organization (WHO), approximately 17.8 million people die from cardiovascular illnesses annually. Conditions such as myocarditis, vascular disease, coronary artery disease, and others are classified as heart and blood vessel diseases or cardiovascular diseases (CVDs). Heart disease and stroke account for 80% of CVD-related deaths, with individuals under 70 representing nearly three-quarters of these fatalities. The main cause of cardiovascular disease includes age, smoking, family history, Alcohol, lack of physical exercise, high blood pressure, high lipid levels, weight gain. To reduce mortality rates effective monitoring and early checkup of CVD are necessary. For disease detection and diagnosis, we use two techniques Data Mining and Internet of things. Data mining techniques enable the extraction of hidden knowledge and identification of correlations within datasets, while IoT technology facilitates the detection of health details, sharing and collection through interconnected sensors and devices. An integration of IoT with machine learning methods can significantly improve the monitoring and management of cardiovascular health, enabling more timely and accurate diagnoses. The implementation of cloud-based systems for data analysis and decision-making further supports this goal, helping to track the history and progression of CVDs and ultimately contributing to better health outcomes. The best Solution we improve heart attack disease outcomes by monitoring a cardiac

patient with the help of IoT.The IoT is the leading technique that allows sensors and data collections to connect, share, and interact over public, private, or Internet Protocol networks. Sensors gather patient data, process it, and provide the necessary insights for clinical decision-making. These systems leverage machine learning methods to forecast CVD history and support clinical goals.

1.3 LECTURE REVIEW

To enhance classification performance, Jyoti Soni et al. (2011) used a Decision Tree (DT) algorithm in conjunction with a genetic algorithm. This approach was compared with two other algorithms, namely Naive Bayes (NB) and classification via cluster approaches. The proposed technique achieved an accuracy of 99.2%. In 2017, Hend Mansoor et al. examined to determine the risk factor of patients with cardiovascular disease through logistic regression and random forest classification algorithm. To predict and detect heart disease Prediction we find out the benefits of using machine learning based systems. With the help of artificial intelligence, we can detect and predict more accurately cardiovascular disease as per the research of American Heart association department, particularly in detecting cardiac diseases. To detect cardiac Abnormalities a new improvement comes under machine learning algorithm and classifier is the weighted associative classifier. Shiva Kazempour Dehkordi and Hedieh Sajedi achieved 73.17% accuracy by the use of data mining prediction model based on prescription. Though, this method performed low as compared to other classification algorithm. The research aims to develop a Decision Support System for heart attack detection using data mining techniques which provide high accuracy and performance among machine learning algorithms. The research will analyze cardiovascular parameters such as age, blood pressure, electrocardiogram results, sex, and blood sugar levels to show the risk of heart attack. The system will take input as medical parameters and provide us the probability of heart disease as output. This study will also involve the design and development of android application where effective machine learning technique used to predict heart attack. To support 100% accuracy, we use 10-fold cross validation by three algorithms. Ultimately, our main goal is to find out classifiers capable of accurately predicting heart disease to be clinically useful.

Summary:

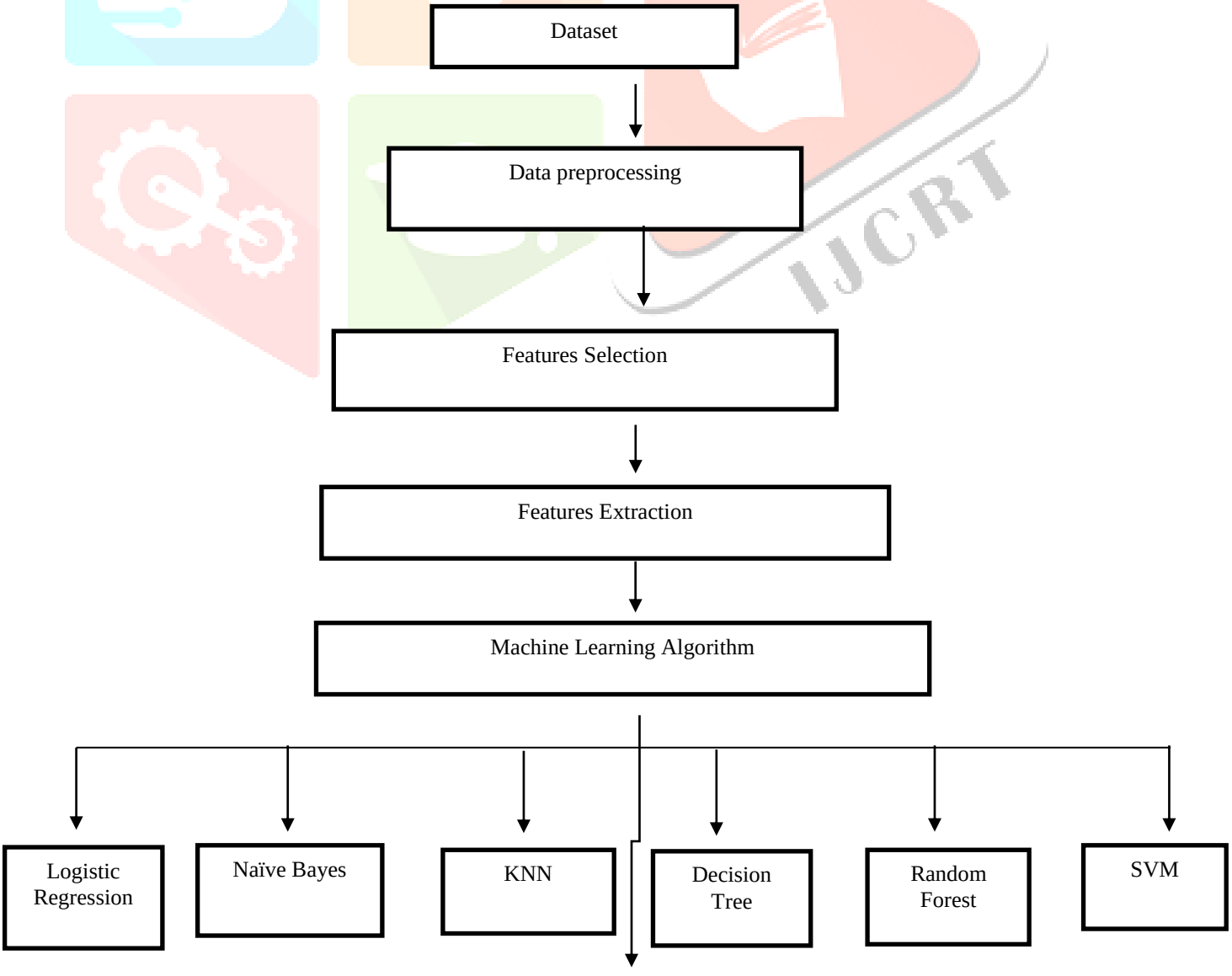
Index	Characteristics	Outline
1.	Age	Age in years
2.	Sex	Male=1; Female=0
3.	Chest pain	Chest pain type(4 values)
4.	Resting Blood Pressure	Resting blood pressure(in mm Hg on admission to hospital)
5.	Cholesterol	Serum cholesterol in mg/dl
6.	Fasting Blood sugar	Fasting blood sugar > 120 mg/dl (1=true;0=false)
7.	Resting electrocardiogram	Resting electrocardiographic results
8.	Maximum heart rate(thalassic)	Maximum heart rate achieved
9.	Exercise induced angina(exang)	Exercise induced angina(1=yes; 0=no)

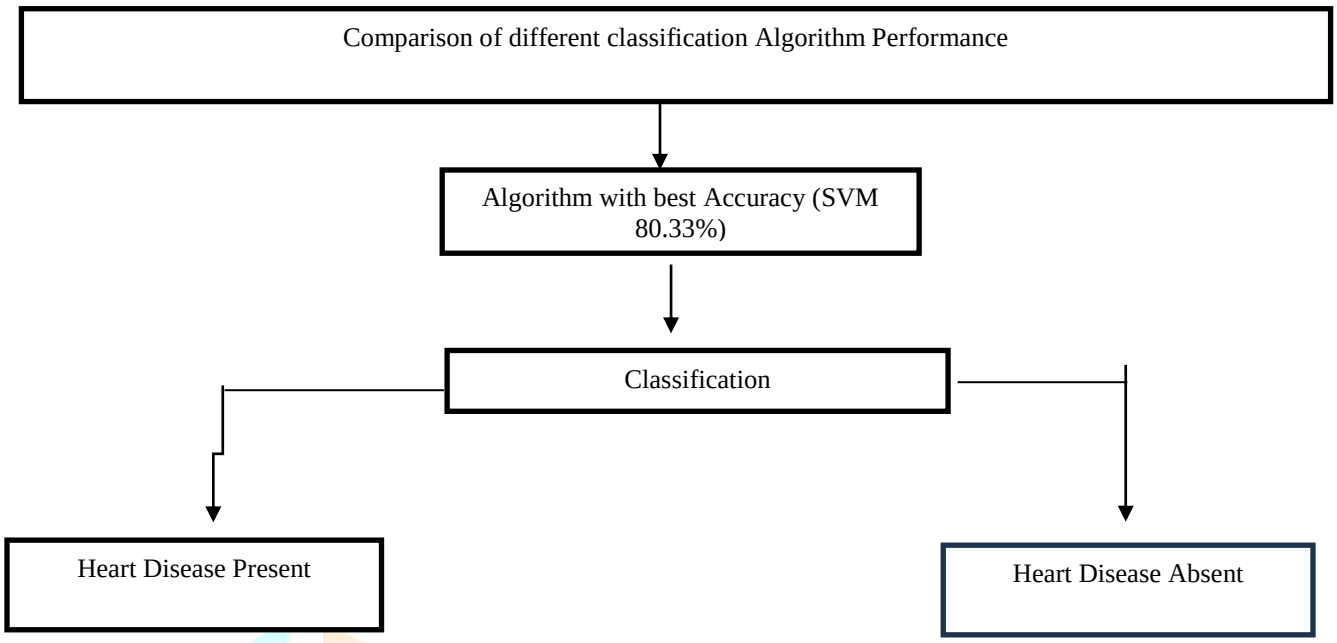
10.	Old peak	Segment-T depression induced by exercise relative to rest
11.	Slope	The slope of the peak exercise Segment- T segment
12.	Coronary Artery	Number of major vessels(0-3) colored by fluroscopy
13.	Thalassemia	1=normal;2=fixed defect;3=reversible defect

Then , we prepare a hypothetical model with different classification algorithms to show their implemented model gave better result than other classification algorithms.

1.4 IMPLEMENTATION

The study provides a comprehensive summarization of the heart disease prediction system, outlining its key components, techniques, and tools. To compare multiple machine learning algorithms and to train large datasets we will develop a system which is adaptive and user-oriented. Following the most reliable algorithm with the highest accuracy and performance, we will implement it in a smartphone-based application designed to detect and predict heart disease risk levels.





1.5 FINDINGS

This study has done using machine learning algorithm that is supervised machine learning which is based on where machine receives label data with outputs. This study utilized WEKA version 3.8.2. To extract insides from raw data in New Zealand a university named Waikato developed a software WEKA. WEKA supports various data mining algorithms, including:

-Classification
- Regression
- Feature selection
-Feature extraction
- Data preprocessing
- Clustering
- Visualization

In WEKA there are 20 machine learning algorithms develop to predict heart attack. In the following subsection a detailed description of the proposed supervised machine learning methods for disease prediction provided.

Logistic Regression (LR)

In supervised machine learning technique, a popular algorithm used for prediction known as Logistic Regression. With the help of independent variables, logistic regression used to predict categorical dependent variable. It extends the general regression model by assigning a dataset's target variable as either an occurrence or non-occurrence of a particular event. This algorithm predicts categorical dependent variable output. Where the result shows as categorical value such as 'yes' or 'No' and '0' or '1', and 'True' or 'false'. Though the outcome shows probabilistic value in between '0' or '1', not the exact value as 0 and 1. To understand relationship between variables where the model shows probabilistic and predictive outcome where logistic regression used. It provides the rationale for a prediction in probabilistic terms, offering clear insights into how certain input attributes influence the output variable. The accuracy score which we achieved using Logistic Regression is: 78.69 %.

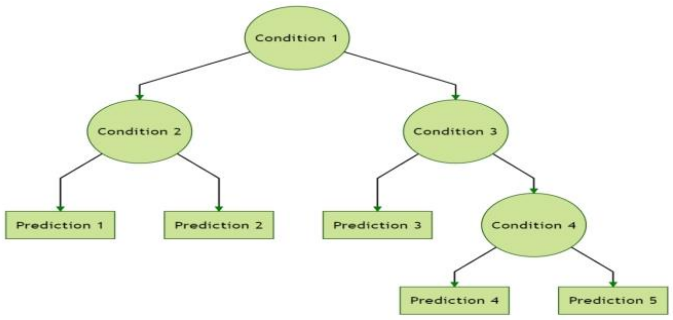


Fig.1: logistic regression

Naïve Bayes

Naive Bayes is a surprisingly powerful algorithm for predictive modelling. It is a statistical classifier which assumes no dependency between attributes attempting to maximize the posterior probability in determining the class. Theoretically, this classifier has the minimum error rate, but may not be the case always. There are three points to find naïve bayes classifier:

a) Convert dataset into frequency tables.

b) Generate likelihood table by finding the probabilities of given features.

C)Now use bayes theorem to calculate the posterior probability.

According to Bayesian theorem P (A| B) = P(A) * P(B|A) /P (B), where P(B/A) is likelihood probability, P(A) is prior/posterior probability, P(B) is marginal probability. Bayesian classifier calculates conditional probability of an instance belonging to each class, based on the above formula, and based on such conditional probability data, the instance is classified as the class with the highest conditional probability. The accuracy score achieved using Naive Bayes is: 78.69 %

KNN (K Nearest Neighbour)

KNN is one of the simplest supervised machine learning algorithms mostly used for classification. It classifies a data plot based on how its neighbour are classified stores available cases & classifies new cases based on similarity measure is a parameter that refers to the no. of nearest neighbour include in the majority visiting process. As it works with a single instance, this classifier is very effective and performs well in disease prediction especially in HD prediction. In this study, the value of neighbour 2 and leaf size 40 were the best fit parameter for the data set. The accuracy score achieved using KNN is: 63.93 %.

Fig.2: knn

Decision Tree (DT)

Decision Tree prediction technique in machine learning is easy to understand for learning of many protocols in medical diagnosis. From the given dataset we find the maximum depth for the classification and it produced the best result. Among the machine learning algorithm decision tree is the most common and oldest one. Decision Tree classifies the data items and design the decision logically where it evaluates and matches results for the classification into a structure as like a tree. Generally, a DT has multiple levels of nodes, the topmost level is known as root or parent node and others are child nodes. The accuracy score achieved using Decision Tree is: 77.05 %.

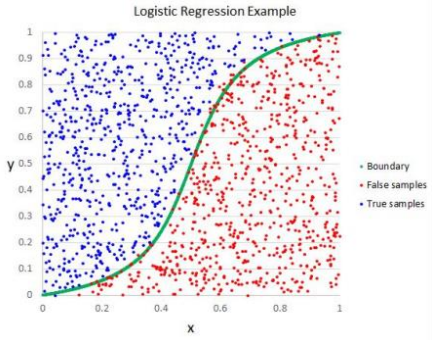


Fig.3: decision tree

Random Forest (RF)

Random forest is a classifier that contains a number of decision trees on various subsets of the given data set and taken the average to improve predictive accuracy. The accuracy score achieved using Decision Tree is: 78.69 %.

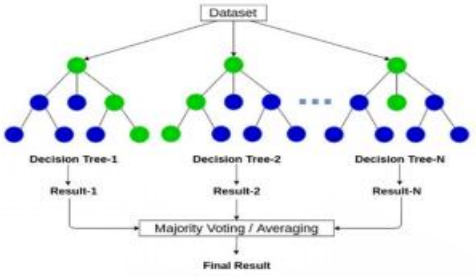


Fig.4: random forest

SVM(Support Vector Machine)

Data which are linear and non-linear, SVM algorithm used for classification. To transforming the training data into a higher , it applies a non-linear mapping methodism is a supervised learning method used for multiple purpose as classification, regression, and outlier detection. It establishes a decision boundary, for label prediction using one or more feature vectors between different classes as shown in the figure-

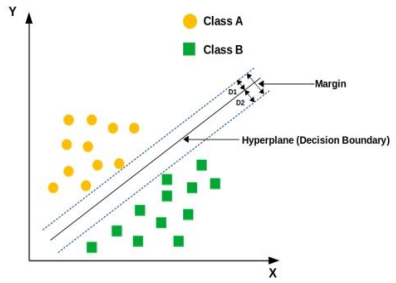


Fig.5: svm

1.6 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we are going to find the proposed technique which is the best machine learning algorithm and the operation of the cardiac monitoring system. We use Jupyter Notebook 7 to predict heart diseases from a dataset. It simplifies the visualization of different data relationship graphs in the data set and facilitates document creation, including live coding.

1. The first step of this research involves cleaning the data set using Python's Pandas and NumPy libraries (version 24.2.0). Next, the Standard Scalar method from Python's Scikit-learn module preprocesses the dataset34.

2. Then using a feature selection approach, the second step of the process calculates the importance of each feature, generating three sets of features (SF).

3. At last, we split the data set into training and testing sets. We use 75% of the data for training and the remaining 25% for testing.

Finally, we trained ten distinct Machine Learning algorithms using this 75% of test data. To predict heart disease35 , we selected the method with the best performance.

In this subsection, we evaluate and explain the proposed system's performance. Here, we presented various algorithms and their comparative performances using evaluation metrics such as accuracy, sensitivity, specificity, and F1-score. We evaluated these performance measures using true positive (TP), true negative (TN), false positive (FP), and false negative (FN) data. The next subsection we are going to focus on these measurements. After the evaluation, we would provide the algorithm that produced the best results.

The following measures are calculated for performance analysis:

Precision = true positive/true positive + false positive

Recall = true positive/true positive + false negative

Fscore = 2*precision*recall/precision + recall

Accuracy = true postive + true negative/n

Sensitivity = true positive/true positive + false negative

Specificity = true negative/true negative + false positive

Here,

n = Total number of instances.

correlation between the important features of SF-2 using SMOTE. The y-axis values include thalach, Chol, sex, age, slope, exang, old-peak, ca., cp., and thal. Between the two variables, Positive or negative correlation coefficients show a significant relationship whereas −1 and 1 indicate no association.

Linear link that exists between the variables only that can be detected via the use of correlation. The prediction for the patient is correlated with each of those variables at a level of at least 70% correlation.

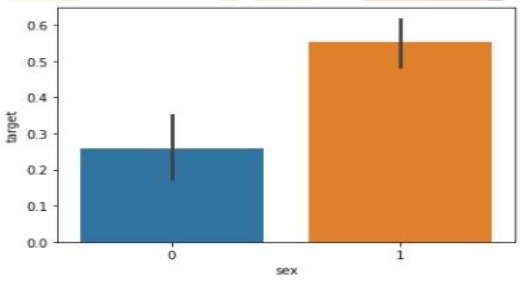
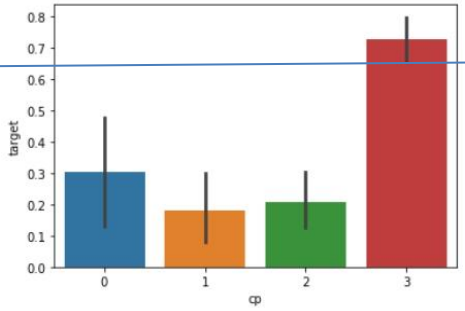


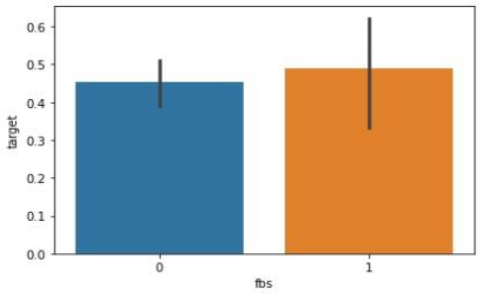
Fig.6: experimental result and analysis

1.7 RESULT ANALYSIS

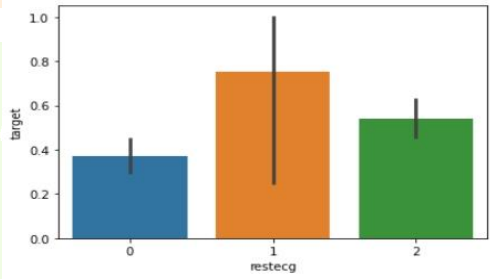


evaluating the 'sex' feature

We observe, from above graph that females has less heart problems than male.

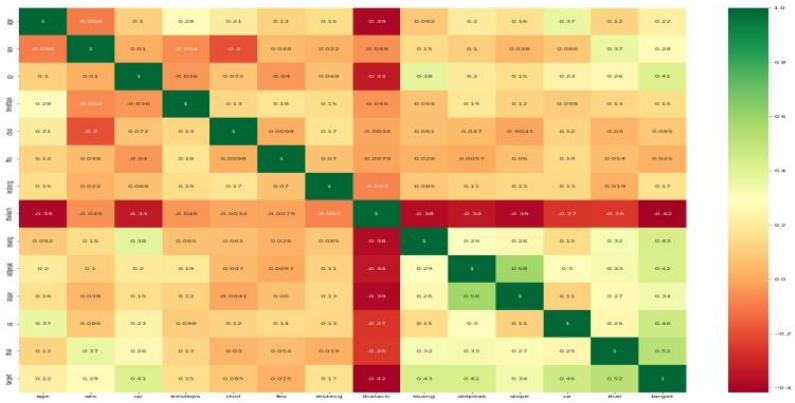


1.

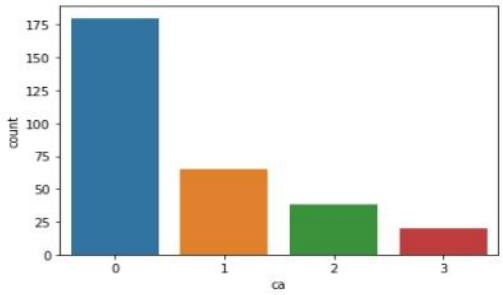


evaluating the 'chest pain type'

Here, we observe, that the ones with typical angina are much less likely to have heart problems, the chest pain of ‘1’.

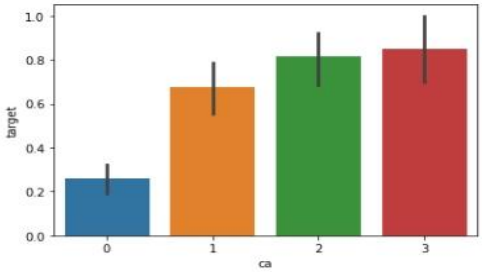


evaluating the fasting blood sugar feature Nothing extraordinary here



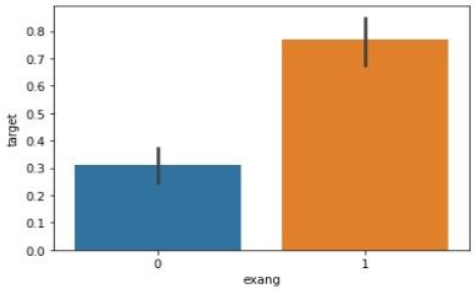
evaluating the rest ecg feature

We realize that people with restecg '0' are lesser heart disease than restecg '1' and



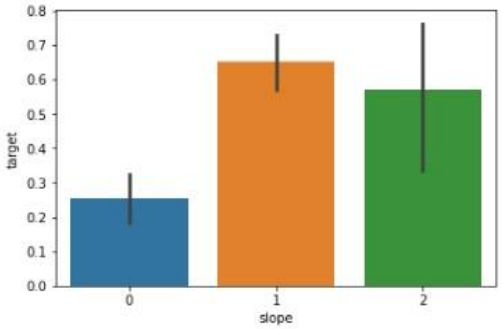
evaluating the 'exercise induced' feature

Exercise induced angina are much less likely to have heart problems. People with exang=0



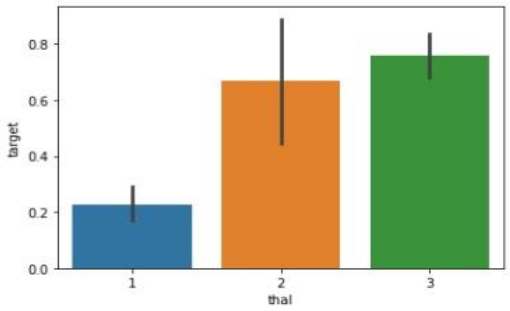
evaluating the slope feature

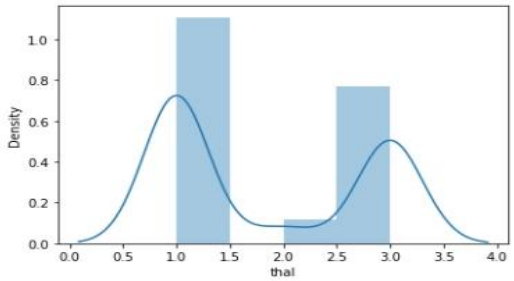
We notice, that from the graph, Slope '1' causes heart pain much more than Slope '0' and '2'.



evaluating the 'ca' feature

Larger number heart patients from above graph has ca=4.





evaluating the ‘thalassemia’ feature

1.8 CONCLUSION AND FUTURE WORK

Heart disease, which is a deathly condition that give rise to certainly severe like heart failure. We use data mining and machine learning techniques to predict heart disease due to its potential for accurate prediction .The objective of this study aims to predict by machine learning approaches. Specific health measurements are used in heart disease patients .

In heart disease prediction, dataset of the heart disease was used for evaluating that impact of the machine learning approaches. Three classification algorithms, which were tested are used. Future enhancements to this work would involve:

1.To apply more extensive data analysis.

2. To achieve the highest possible accuracy, to explore additional algorithms.

1.9 REFERENCES

[1] https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 [Accessed 02 June 2021].

[2] R.D. Canlas, Data Mining in Healthcare: Current Applications and Issues, School of Information Systems & Management, Carnegie Mellon University, Australia, 2009.

[3] Christoph Helma, Eva Gottmann, Stefan Kramer, Knowledge discovery and data mining in toxicology, Stat. Methods Med. Res. 9 (4) (2000) 329–358.

[4] I.-N. Lee, S.-C. Liao, M. Embrechts, Data mining techniques applied to medical information, Med. Inf. Internet Med. 25 (2) (2000) 81–102.

[5] L. Parthiban, R. Subramanian, Intelligent heart disease prediction system using CANFIS and genetic algorithm, Int. J. Biol., Biomed. Med. Sci. 3 (3) (2008).

[6] P.C. Austin, J.V. Tu, J.E. Ho, D. Levy, D.S. Lee, Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes, J. Clin. Epidemiol. 66 (4) (2013) 398–407.

[7] S.K. Dehkordi, H. Sajedi, Prediction of disease based on prescription using data mining methods, Health Technol. 9 (1) (2018) 37–44.

[8] M. Jan, A.A. Awan, M.S. Khalid, S. Nisar, Ensemble approach for developing a smart heart disease prediction system using classification algorithms, Res. Rep. Clin. Cardiol. 9 (2018) 33–45.

[9] J. Soni, U. Ansari, D. Sharma, S. Soni, Predictive data mining for medical diagnosis: an overview of heart disease prediction, Int. J. Comput. Appl. 17 (8) (2011) 43–48.

[10] H.M. Islam, Y. Elgendy, R. Segal, A.A. Bavry, J. Bian, Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: a machine learning approach, J. Heart & Lung (2017) 1–7.
[11] H.M. Le, T.D. Tran, L.A.N.G. Van Tran, Automatic heart disease prediction using feature selection and data mining technique, J. Comput. Sci. Cybern. 34 (1) (2018) 33–48.
[12] M. Tarawneh, O. Embarak, February. "Hybrid approach for heart disease prediction using data mining techniques, Acta Sci. Nutr. Health 3 (7) (2019) 147–151, 2019.
[13] S. Rehman, E. Rehman, M. Ikram, and Z. Jianglin, "cardiovascular disease (CVD): assessment, prediction and policy implications," BMC Public Health, vol. 21, no. 1, p. 1299, 2021, doi: 10.1186/s12889-021-11334-2.
[14] O. Atef, A. B. Nassif, M. A. Talib, and Q. Nassir, "Death/Recovery Prediction for Covid-19 Patients using Machine Learning," 2020.
[15] A. B. Nassif, I. Shahin, M. Bader, A. Hassan, and N. Werghi, "COVID-19 Detection Systems Using Deep-Learning Algorithms Based on Speech and Image Data," Mathematics, 2022.
[16] H. Hijazi, M. Abu Talib, A. Hasasneh, A. Bou Nassif, N. Ahmed, and Q. Nasir, "Wearable Devices, Smartphones, and Interpretable Artificial Intelligence in Combating COVID-19," Sensors, vol. 21, no. 24, 2021, doi: 10.3390/s21248424.