**K L HYDERABAD**

**COMPUTER SCIENCE AND ENGINEERING DEPARTMENT**

**Project-Based Lab Report**

**On**

**Heart Disease Prediction**

**SUBMITTED BY:**

I.D NUMBER NAME

2110030099 P Praneetha

2110030359 B Chihnita

2110030412 Yeshaswini

2110030432 Amar Goud

2110030444 Dhana Sree

**UNDER THE ESTEEMED GUIDANCE OF**

**Dr. Figlu Mohanty**

**COURSE CORDINATOR** 

**KONERU LAKSHMAIAH EDUCATION FOUNDATION**

(Deemed to be University)

Moinabad Road, Aziz Nagar, Hyderabad - 500075

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

****

**CERTIFICATE**

This is to certify that the project-based laboratory report entitled “Heart Disease Prediction” submitted by Student bearing Regd. No. 2110030399, 2110030359, 210030412, 2110030432, 2110030444 to the **Department of Computer Science and Engineering, KL University** in partial fulfillment of the requirements for the completion of a project in the “Artificial Intelligence for Data Structures”course in II B Tech IV Semester, is a bonafide record of the work carried out by him/her under my supervision during the academic year 2022-23.

PROJECT SUPERVISOR HEAD OF THE DEPARTMENT

Dr Figlu Mohanty Dr. Arpita Gupta

**ACKNOWLEDGEMENTS**

It is a great pleasure for me to express my gratitude to our honorable President **Sri. Koneru Satyanarayana**, for giving me the opportunity and platform with facilities in accomplishing the project-based laboratory report.

I express my sincere gratitude to our Principal **Dr. A Ramakrishna** for his administration of our academic growth.

I express sincere gratitude to our Coordinator **Dr Figlu Mohanty** and HOD-BES-1 **Dr. Arpita Gupta** for her leadership and constant motivation provided in the successful completion of our academic semester. I record it as my privilege to deeply thank you for providing us with the efficient faculty and facilities to make our ideas into reality.

I express my sincere thanks to our project supervisor **Dr. Figlu Mohanty**

for her novel association of ideas, encouragement, appreciation, and intellectual zeal which motivated us to venture into this project successfully.

Finally, it is pleased to acknowledge the indebtedness to all those who devoted themselves directly or indirectly to making this project report successful.

2110030099 P Praneetha

2110030359 B Chihnita

2110030412 Yesashwini

2110030432 Amar Goud

2110030444 Dhana Sree

**ABSTRACT**

Heart disease (HD) is one of the most common diseases nowadays, and an early diagnosis of such a disease is a crucial task for many healthcare providers to prevent their patients from such a disease and to save lives. In this paper, a comparative analysis of different classifiers was performed for the classification of the Heart Disease dataset to correctly classify and or predict HD cases with minimal attributes. The set contains 76 attributes including the class attribute, for 1025 patients collected from Cleveland, Hungary, Switzerland, and Long Beach, but in this paper, only a subset of 14 attributes are used, and each attribute has a given set value. The algorithms used K- Nearest Neighbor (K-NN), Naive Bayes, Decision tree J48, Rip, SVM, Ad boost, Stochastic Gradient Decent (SGD), and Decision Table (DT) classifiers to show the performance of the selected classifications algorithms to best classify, and or predict, the HD cases.

The healthcare 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 Naive Bayes and Decision Tree algorithms for predicting the risk level of heart disease. The system uses 15 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 the multilayer perceptron neural network with backpropagation as the training algorithm. The obtained results have illustrated that the designed diagnostic system can effectively predict the risk level of heart disease.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **TITLE** | **PAGE NO** |
| 1 | Introduction | 6 |
| 2 | Aim of the Project | 7-8 |
| 2.1 | Future Implementation | 11-13 |
| 3 | Software & Hardware Details | 15 |
| 4 | Flowcharts | 8-11 |
| 5 | Implementation | 14 |
| 6 | Outputs/ScreenShots | 15-16 |
| 7 | Conclusion | 16 |

**INTRODUCTION**

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining combines statistical analysis, machine learning, and database technology to extract hidden patterns and relationships from large The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure - a condition that causes around half of all deaths from stroke and heart disease. Heart disease, also known as cardiovascular disease (CVD), encloses several conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy, and cardiovascular disease are some categories of heart diseases. The term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and how blood is pumped and circulated through the body. Diagnosis is a complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on the doctor’s experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial

**PROBLEM STATEMENT**

Heart disease can be managed effectively with a combination of lifestyle changes, medicine, and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce the cost of surgical treatment and other expenses. The overall objective of my work will be to predict accurately with a few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors’ intuition and experience rather than on the knowledge-rich data hidden in the data set and databases. This practice leads to unwanted biases, errors, and excessive medical costs which affects the quality of service provided to patients. Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. According to (Wurz & Takala, 2006) ⁠the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. The successful application of data mining in highly visible fields like e-business, marketing, and retail has led to its application in other industries and sectors. Among these sectors just discovered is healthcare. The healthcare environment is still „information rich‟ but „knowledge poor‟. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in the data for African genres.

**PROPOSED ARCHITECTURE**

In this system, we are implementing an effective heart attack prediction system using the Naïve Bayes algorithm. We can give the input as in a CSV file or manual entry to the system. After taking input the algorithms apply on that input that is Naïve Bayes. After accessing the data set the operation is performed and an effective heart attack level is produced. The proposed system will add some more parameters significant to a heart attack with their weight, age, and priority levels by consulting expert doctors and medical experts. The heart attack prediction system is designed to help identify different risk levels of a heart attack like normal, low, or high and also give the prescription details related to the predicted result

**NAÏVE BAYES CLASSIFER**

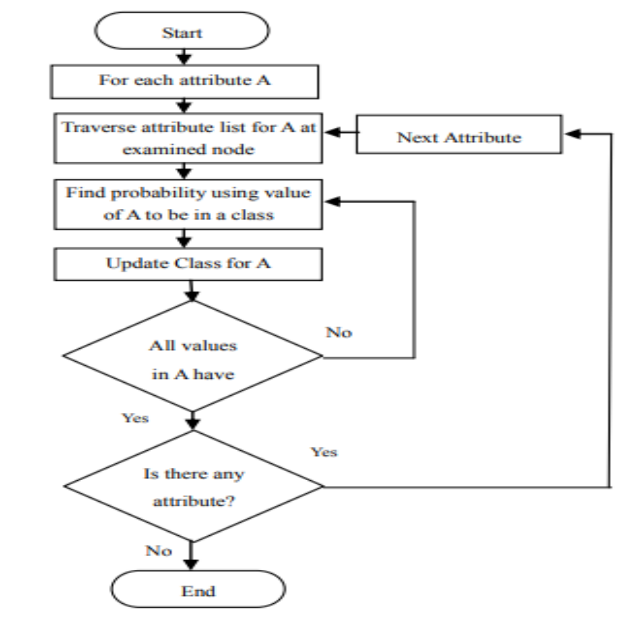
Naïve Bayes classifier is based on Bayes theorem. This classifier uses conditional independence in which the attribute value is independent of the values of other attributes. The Bayes theorem is as follows: Let X= {x1, x2, ......, Xn} be a set of n attributes. In Bayesian, X is considered as evidence and H is some hypothesis means, the data of X belongs to specific class C. We have to determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as P(H|X) = P(X|H) P(H) / P(X).4 Algorithm Used Accuracy Time Taken Naïve Bayes 52.33% 609 m Decision Tree 52% 719 m K-NN 45.67% 1000 m Using Bayesian classifiers, the system will discover the concealed knowledge associated with diseases from historical records of the patients having heart disease. Bayesian classifiers predict the class membership probabilities, in a way that the probability of a given sample belonging to a particular class statistically. A Bayesian classifier is based on Bayes’ theorem. We can use the Bayes theorem to determine the probability that a proposed diagnosis is correct, given the observation. A simple probabilistic, the naive Bayes classifier is used for classification based on which is based on Bayes’ theorem. According to the naïve Bayesian classifier, the occurrence or an occurrence of a particular feature of a class is considered independent of the presence or absence of any other feature. When the dimension of the inputs is high and a more efficient result is expected, the chief Naïve Bayes Classifier technique is applicable. The Naïve Bayes model identifies the physical characteristics and features of patients suffering from heart disease. For each input, it gives the possibility of the attribute of the expectable state. Naïve Bayes is a statistical classifier that assumes no dependency between attributes. This classifier algorithm uses conditional independence, which means it assumes that an attribute value of a given class is independent of the values of other attributes. The advantage of using Naïve Bayes is that one can work with the Naïve Bayes model without. using any Bayesian methods.

**DECISION TREE**

The decision tree approach is more powerful for classification problems. There are two steps in this technique building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48. From this J48 algorithm is used for this system. J48 algorithm uses the pruning method to build a tree. Pruning is a technique that reduces the size of the tree by removing overfitting data, which leads to poor accuracy in predictions. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides a balance of flexibility & accuracy.

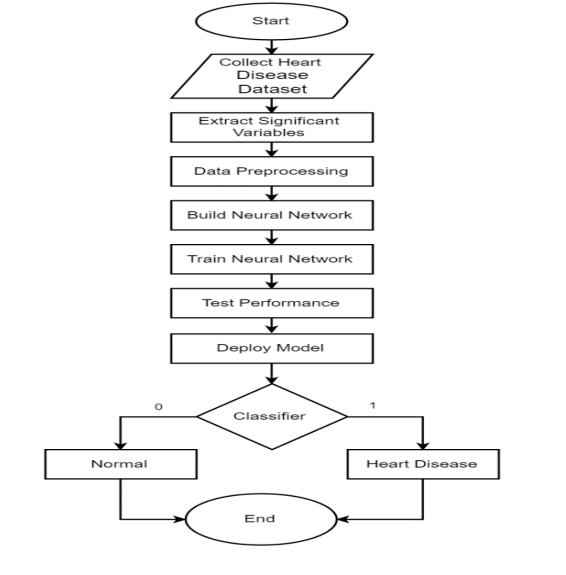
**FLOW CHART OF NAÏVE BAYES DECISION TREE ALGORITHM**

The classification tree creates a tree with branches, nodes, and leaves that lets us take an unknown data point and move down the tree, applying the attributes of the data point to the tree until a leaf is reached and the unknown output of the data point can be determined. To create a good classification tree model, we need to have an existing data set with known output from which we can build our model. We also divide our data set into two parts: a training set, which is used to create the model, and a test set, which is used to verify that the model is accurate and not overfitted

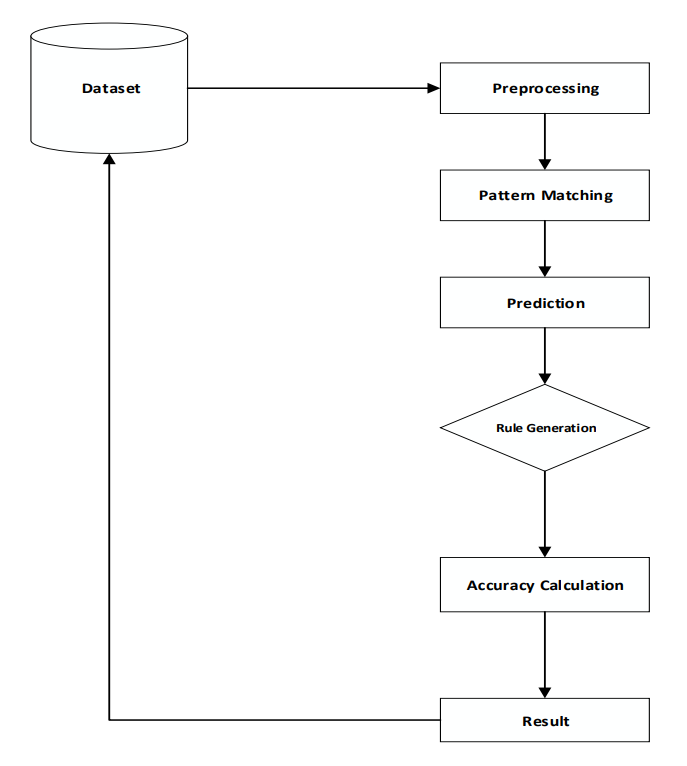
****

**PROJECT FLOW DIAGRAM**

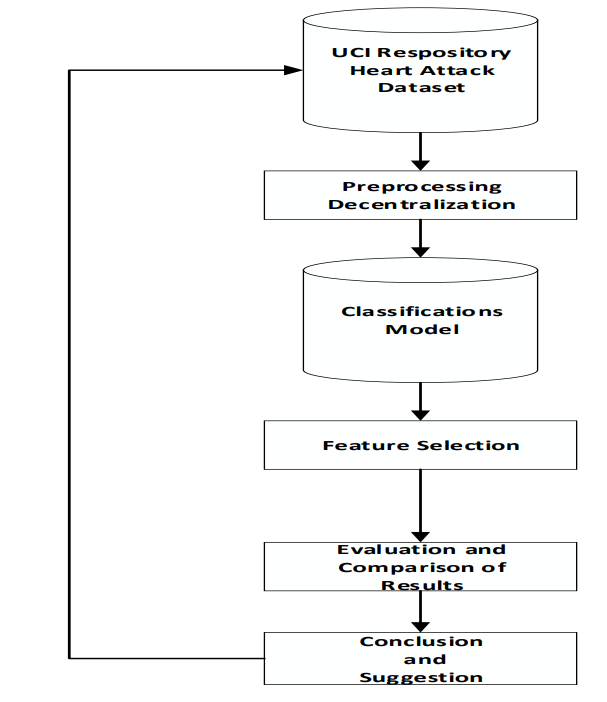
This will be the proposed flow chart that the system will look like

****

**DATAFLOW DIAGRAM**

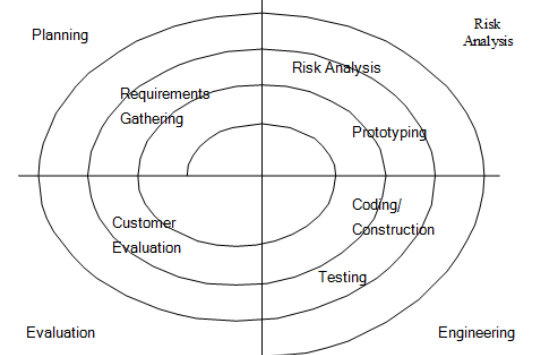
****

**PROPOSED MODEL**

****

**SYSTEM DEVELOPMENT METHODOLOGY**

The methodology of software development is the method of managing project development. There are many models of the methodology available such as the Waterfall model, Incremental model, RAD model, Agile model, Iterative model, and Spiral model. However, it still needs to be considered by the developer to decide which will be used in the project. The methodology model is useful to manage the project efficiently and able to help developers from getting any problems during the time of development. Also, it helps to achieve the objective and scope of the projects. To build the project, it needs to understand the stakeholder requirements. The methodology provides a framework for undertaking the proposed DM modeling. The methodology is a system comprising steps that transform raw data into recognized data patterns to extract knowledge for users.

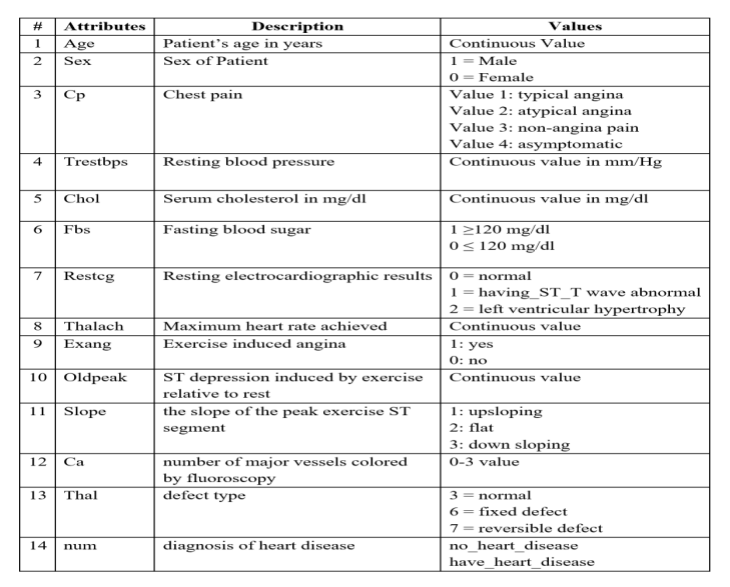
****

Four phases are involved in the spiral model: 1) Planning phase Phase where the requirement is collected and risk is assessed. This phase is where the title of the project has been discussed with the project supervisor. From that discussion, Heart Prediction System has been proposed. The requirement and risk were assessed after doing a study on the existing system and doing a literature review about other existing research. 2) Risk analysis Phase Phase where the risk and alternative solution are identified. A prototype is created at the end of this phase. If there is any risk during this phase, there will be suggestions about alternate solutions. 3) Engineering phase At this phase, the software is created and testing is done at the end of this phase. 4) Evaluation phase At this phase, the user evaluates the software. It will be done after the system is presented and the user tests whether the system meets their expectation and requirement or not. If there is any error, the user can tell the problem about the system.

**DATA COLLECTION AND PREPROCESSING**

The data set for this research was taken from the UCI data repository.14 Data accessed from the UCI Machine Learning Repository is freely available. In particular, the Cleveland and Hungarian databases have been used by many researchers and found to be suitable for developing a mining model, because of lesser missing values and outliers. The data is cleaned and preprocessed before it is submitted to the proposed algorithm for training and testing. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The overall objective of our work is to predict more accurately the presence of heart disease. In this paper, UCI repository datasets are used to get more accurate results. Two data mining classification techniques were applied namely Decision trees and Naive Bayes his database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is an integer value from 0 (no 15 presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (value 1,2,3,4) from absence (value 0). Attributes with categorical values were converted to numerical values since most machine learning algorithms require integer values. Additionally, dummy variables were created for variables with more than two categories. Dummy variables help Neural Networks learn the data more accurately.

**CLASSIFIERS USED FOR EXPERIMENTS**

****

**SYSTEM REQUIREMENTS**

* **SOFTWARE REQUIREMENTS:**

The major software requirements of the project are as follows:

* Operating System:- Any OS with clients to access the internet
* Network Wi-Fi Internet or cellular Network
* Visio Studio Create and design Data Flow and Context Diagram
* Github Versioning Control
* Google Chrome Medium to find references to do system testing, display, and run the shiny app.

* **HARDWARE REQUIREMENTS:**

The hardware requirements that map to the software are as follows:

* Processor: Intel or high
* RAM: 1024 MB
* Space on disk: minimum 100MB
* For running the application:
* Device: Any device that can access the internet
* Minimum space to execute: 20 MB

The effectiveness of the proposal is evaluated by conducting experiments with a cluster formed by 3 nodes with identical settings, configured with an Intel CORE™ i7-4770 processor (3.40GHZ, 4 Cores, 8GB RAM, running Ubuntu 18.04 LTS with 64-bit Linux 4.31.0 kernel)

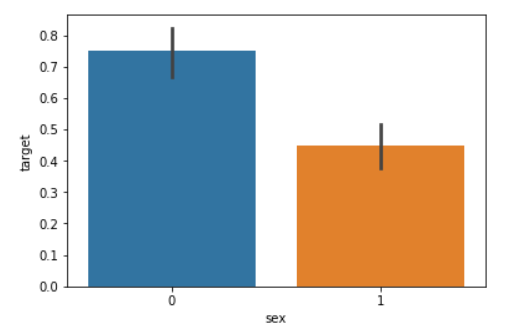
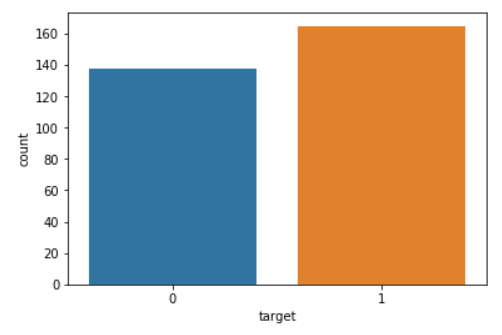
* **TOOLS**
* Operating System: Windows 7 or any Linux Debian Distro.
* Language: R and Shiny
* Tools: RStudio IDE, Microsoft Excel (Optional).
* Technologies used: R, Unix, Shiny.

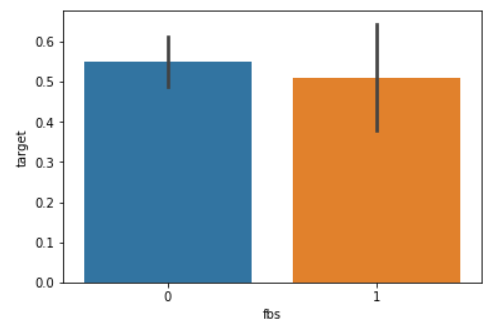
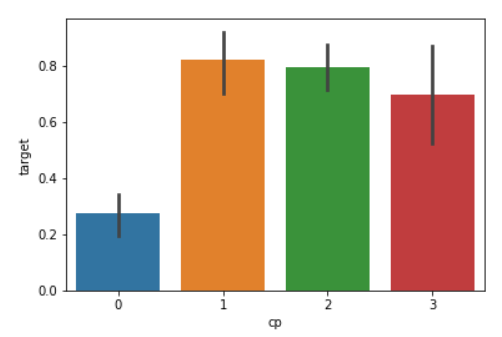
**IMPLEMENTATION**

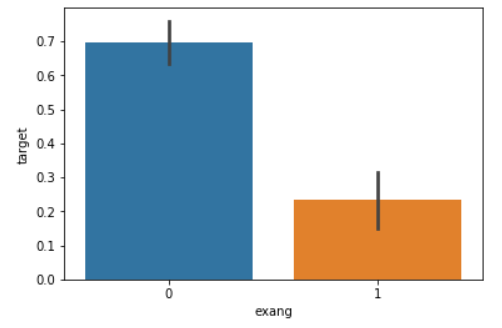
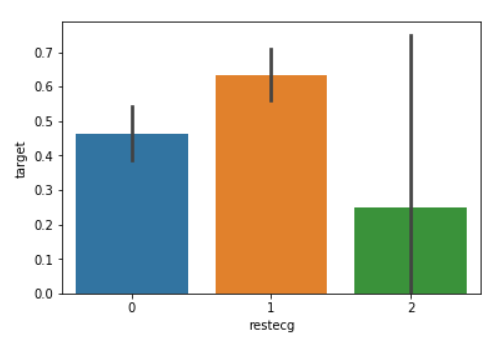
Github code link

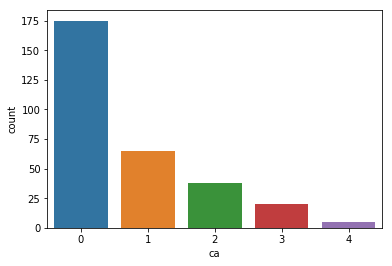
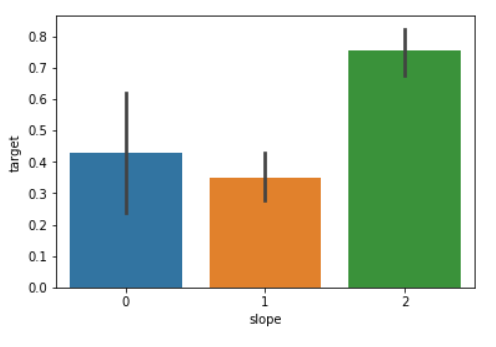
https://github.com/CHIHNITA-REDDY-B/AI-for-DS/blob/main/Heart\_disease\_prediction.ipynb

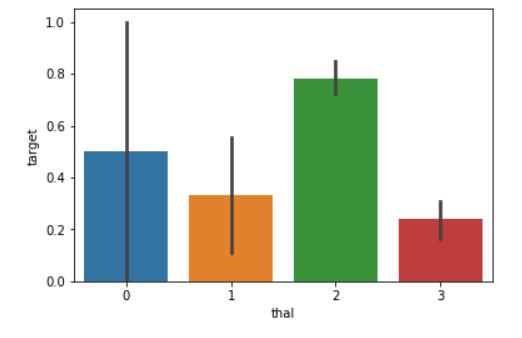
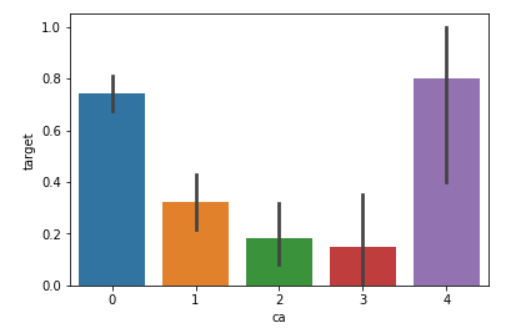
**OUTPUTS**

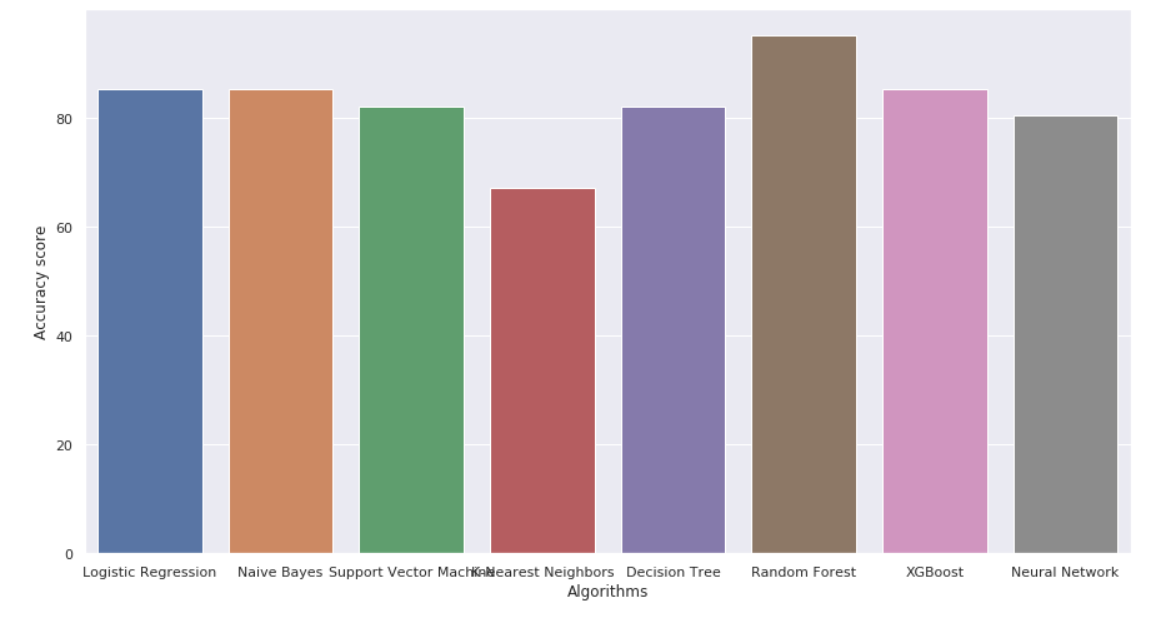












**CONCLUSION**

**& Future Work**

The proposed system is a GUI-based, user-friendly, scalable, reliable, and expandable system. The proposed working model can also help reduce treatment costs by providing Initial diagnostics on time. The model can also serve the purpose of a training tool for medical students and will be a soft diagnostic tool available for physicians and cardiologists. General physicians can utilize this tool for the initial diagnosis of cardio-patients. Many possible improvements could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in the future we can use this system for the analysis of different data sets. The performance of health diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In the 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.