LOVELY PROFESSIONAL UNIVERSITY

PHAGWARA, PUNJAB

Submitted in partial fulfillment of the requirements for the award of degree of

BTech Computer Science Engineering

TOPIC

Heart Disease Prediction

ACKNOWLEDGEMENT

Primarily I would like to thank you my college and my teacher **Dr. Pande sir** for guiding continuously throughout the project .Then I would like to express my special thanks for college who provided such an opportunity for students to gain hands on experience on machine learning by learning and building projects which helps to get career ready.

I would like to again thank my own college Lovely Professional University for offering such a opportunity which not only improve my programming skill but also taught me other new technology.

Then I would like to thank my parents and friends who have helped me with their valuable suggestions and guidance in completion of this project

Date: 27/09/2021

ABSTRACT

Heart acts a major role in corporeal body part. Heart diseases or Cardiovascular Diseases (CVDs) are some of the main reason for a huge number of death in the world and some of the reports said that over the last few decades heart disease or CVD has emerged as the most life-threatening disease, not only in India but in the whole world. The objective of this paper is to develop simple, effective method for heart disease prediction, where user or patient can check their heart Status by self diagnosis and can get clear cut to cut report about their heart.

The diseases of heart wants more perfection and exactness for diagnose and analyse. This disease occurs due to various problems such as over pressure, blood sugar, Chest pain,high blood pressure, Cholesterol etc. in human body . Heart is the next major organ comparing to the brain which has more priority in the Human body. It pumps the blood and supplies it to all organs of the whole body.

In the health care sector, Machine Learning plays an important role in the health care Industry.So, there is a need fora reliable, accurate, and feasible system to diagnose such diseases in time for proper treatment.In these paper ,I have used various machine learning techniques and it compare the result with various Performance metrices.

In these paper dataset is used is Public Health Dataset to perform different comparative analysis of heart disease prediction. This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient.

By performing various machine learning technique like K-NN,RF,GB,ADA B The maximum accuracy I achieved is 97.82% .For simplicity and this project will Accessible I have made this as webapp using flask and deploy this on herouko so That everyone can access it.

Keywords- : KNN, RF, confusion Matrix, Jupyter-Notebook, FLASK,HEROKU, Analysis, dataset,heatmap,Gradient boosting

INTRODUCTION

Heart is one of the most indispensable organs in the human body. It is a organ that serves as a pump to circulate the blood. The heart is a muscular organ about the size of a fist, located just behind and slightly left of the breastbone. The heart pumps blood through the network of arteries and veins called the cardiovascular system . Oxygen is distributed through the circulatory system of the body in the blood, and if the heart does not function correctly, the entire circulatory system of the body will fail. So if the heart doesn't work properly, it could even lead to death.

According to the World Health Organization (WHO), in the last 15 years, an estimated 17 million people die each year from cardiovascular disease, particularly heart attacks and strokes [1]. Heart disease and stroke are the biggest killers. To predict heart disease, Machine Learning can be used for identifying unseen patterns and providing some clinical insights that will assist the physicians in planning and providing care

Most common Symptoms of heart disease are:

- Chest pain:
  It is the most common symptom of heart attack. If someone has a blocked artery or is having a heart attack, he may feel pain, tightness or pressure in the chest.
- Nausea,Indigestion:
  Heartburn and Stomach Pain These are some of the often overlooked symptoms of heart attack. Women tend to show these symptoms more than men.
- Pain in the Arms :
  The pain often starts in the chest and then moves towards the arms, especially in the left side.
- Feeling Dizzy and Light Headed :
  Things that lead to the loss of balance.
- Fatigue
  Simple chores which begin to set a feeling of tiredness should not be ignored.
- Sweating:
  Some other cardiovascular diseases which are quite common are stroke, heart failure, hypertensive heart disease, rheumatic heart disease, Cardiomyopathy, Cardiacarrhyth

Heart disease is common among both men and women in most countries around the world. Therefore, people should consider heart disease risk factors. Although it plays a genetic role, some lifestyle factors significantly affect heart disease . The known risk factors for heart disease; radiation therapy for age, gender, family history, smoking, some chemotherapy drugs and cancer, malnutrition, high blood pressure, high blood cholesterol levels, diabetes, obesity, physical mobility, stress, and poor hygiene. These are the various risk factors in which the patient's exposure towards developing a CVD.

The most common type is coronary artery disease, which can cause a heart attack. Other types of heart disease may involve the valves in the heart, or the heart may not pump well and cause heart failure. Some people are born with heart disease. Anyone, including children, can develop heart disease. It happens when a substance called plaque builds up in your arteries. Smoking, unhealthy eating and lack of exercise increase your risk of heart disease. High cholesterol, high blood pressure or diabetes can also increase your risk of heart disease.

the main heart disease risk factors such as physiological factors (age, sex, and menopausal status), lifestyle factors (smoking, physical activity, alcohol, stress), metabolic syndrome factors (insulin resistance), dyslipidemia, abdominal obesity, high blood pressure) and dietary factors. A heart disease risk factor is defined as a factor in which the patient's exposure to this factor increases the risk of developing a CVD. In contrast, the removal or improvement of this factor decreases this risk. The risk factor's importance is defined by the association's strength with the disease (expressed by the relative risk observed in the exposed subjects compared to the unexposed) and the gradual association (parallel to the risk factor).

There are several types of heart disease which include :

- Coronary Artery Disease (CAD)
- Heart Arrhythmias.
- Heart Failure.
- Heart Valve Disease.
- Pericardial Disease.
- Cardiomyopathy (Heart Muscle Disease)

Congenital Heart Disease.

A machine-learning system is trained rather than the explicitly programmed. Machine learning could be a better choice for achieving high accuracy for detection of heart diseases. This paper is dedicated for wide scope survey in the field of machine learning technique in prediction of heart disease.

To deal with this disease, there are several methods of prevention, such us natural methods, like stoping smoking, maintaining a healthy weight, adopting a healthy diet and practicing sports regularly.We also have the scientific methods such as drugs and surgeries. The prediction of this disease before being infected is part of the prevention

Machine learning (ML) plays a significant role in disease predicting [9]. It predicts whether the patient has a particular disease type or not based on an efficient learning technique. In making of these project I have used several supervised learning techniques for predicting the early stage of heart disease by providing them risk factor Whether they have any chance of getting heart disease or not. Some of the techniques which I have used in these project are :

- K-nearest neighbor (KNN)
- Random forest (RF)
- Ada Boosting With Random Forest
- Gradient Boosting
- XG BOOST

With using all these techniques the highest accuracy of 97.82% I got in K-nearest neighbor (KNN).

The rest of this paper is structured as follows: Section 2 describes the literature review of the current research proposed in this field. Section 3 describes the machine learning techniques, proposed architecture and methodology. In Section 4, results and the comparison between classification methods and algorithm are presented. Finally, Section 5 describes the conclusion of the paper.

LITERATURE REVIEW

There are various work done by various researcher or scientist on heart diease using the UCI dataset which tend to predict heart disease . Using different data mining methods, various levels of accuracy have been achieved. Typically, heart is unable to push the necessary amount of blood to other areas of the body in order to satisfy the normal functioning of the body in this disease, and because of this, heart failure eventually occurs. The prevalence of heart disease is very high in the United States. Symptoms of heart disease include shortness of breath, physical body fatigue, swollen feet, and tiredness with associated signs, such as increased jugular venous pressure and peripheral edoema due to functional or non-functional cardiac irregularities. The early-stage investigation approaches used to detect heart dis- ease have been difficult, and the resulting difficulty is one of the key factors affecting the standard of living. Diagnosis and treatment of heart disease is very difficult, especially in developing countries, owing to the rare availability of diagnostic instruments and the shortage of doctors and other services affecting the proper prediction of heart disease. The precise and correct detection of heart disease is important to reduce the associated risk of serious heart complications and to improve heart safety. Approximately 3 percent of the health care financial budget is impacted by the costs of heart disease management.

There are various machine learning techinques that I have used in these project which are K-nearest neighbor (KNN) , Random forest (RF), Ada Boosting With Random Forest , Gradient Boosting, XG BOOST. After the result shows that the K-nearest neighbor (KNN) has achived the highest accuracy of 97.82 % .To Which RF achived 86 % accuracy , Ada Boosting With Random Forest achieved accuracy of 91 %, Gradient Boosting achieved accuracy of 89 % , XG BOOST achieved accuracy of 91 %. Heart attack must be diagnosed in a timely and effective way due to its high prevalence.Therefore proceeding with maximum accuracy can lead to provide actual status of the heart to patient whether they are having any chances of heart disease or not.

Objectives of this research are as follows:

- Data collection from new features about heart disease.
- Prediction and classification of incidence of heart disease using the proposed method.
- Using new feature selection algorithms for the first time.
- Providing a new combined approach with higher accuracy

METHODOLOGY

1. Machine learning techniques :-

Using Machine Learning computers can learn and act like humans, and improve their learning by feeding them the data and information in the form of observations." There are various machine learning techniques available. In this study I have used four different algorithms for analysis and comparison.

k-nearest neighbors (KNN):-

KNN The k-nearest neighbors (KNN) algorithm is a simple and easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

Random Forest :-

Random forest It is a supervised classification algorithm. As a name suggests, algorithm creates the forest with a number of trees and higher the number of trees in the forest higher will be the accuracy. It will handle the missing values. If there are more trees in the forest, random forest classifier won't overfit the model

Gradient Boosting :-

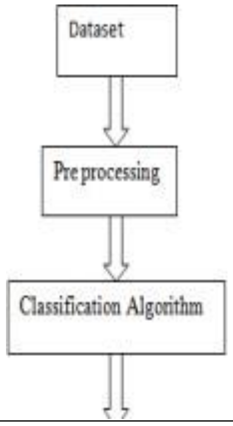
Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels.

There is a technique called the **Gradient Boosted Trees** whose base learner is CART (Classification and Regression Trees).

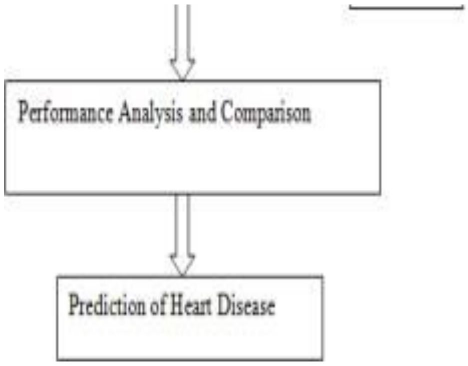
Ada Boosting With random forest :-

AdaBoost is a boosting ensemble model and works especially well with the decision tree. Boosting model's key is learning from the previous mistakes, e.g. misclassification data points. AdaBoost learns from the mistakes by increasing the weight of misclassified data points.

2. PROPOSED WORK :-



1. **K-NN**
2. **Random Forest**
3. **Ada Boosting With random forest**
4. **Gradient Boosting**


Performance Analysis and Comparison
Prediction of Heart Disease

Above figure shows the proposed work in heart disease prediction where :-

- The first step is to select the dataset for analysis and comparison.

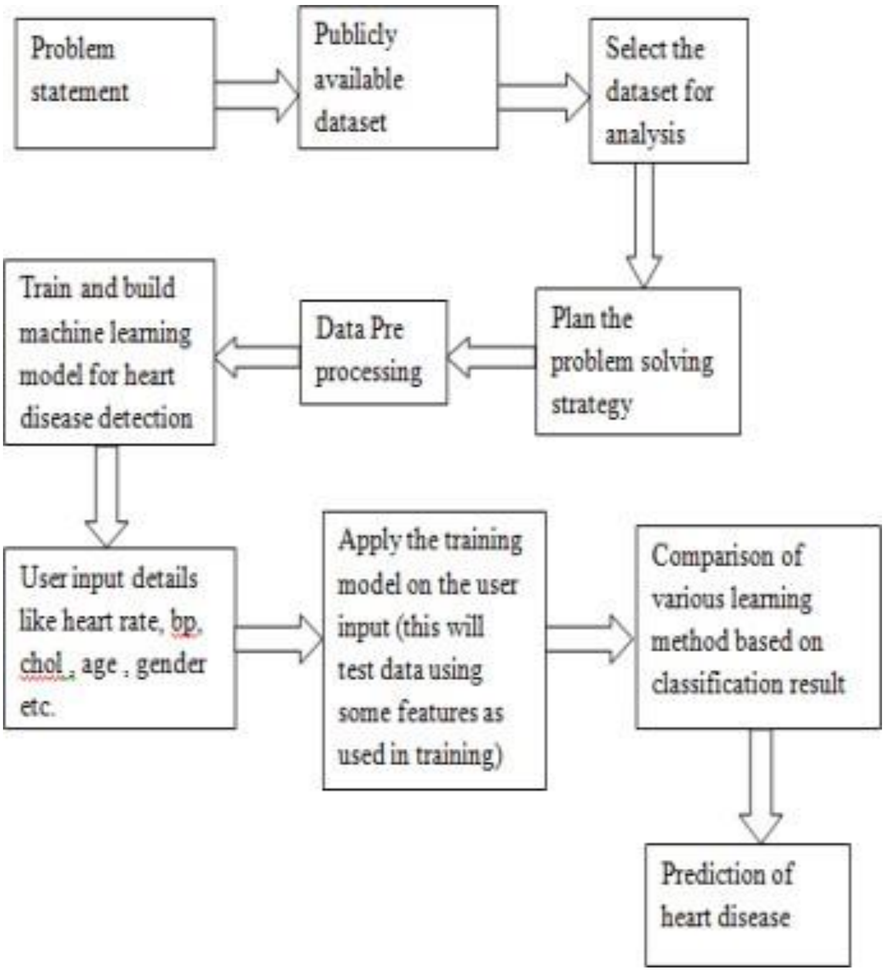
- Second step shows of Pre-Processing the dataset . Preprocessing is a process of removing unwanted and missing data.

- Third steps shows the classification of various machine learning algorithm like K-NN,RF , Gradient Boosting,Ada Boosting, which Are used for comparative analysis.

- In fourth step performance is analyzed and all the above classification algorithm are compared and pick the best algorithm whose having maximum accuracy among all.

- In Final Step We will predict whether the person Is having any chances of Heart disease or not by giving and filling required inputs from the patient.

3. METHODOLOGY:-


Problem statement
Publicly available dataset
Select the dataset for analysis
Plan the problem solving strategy
Data Pre processing
Train and build machine learning model for heart disease detection
User input details like heart rate, bp, chol , age , gender etc.
Apply the training model on the user input (this will test data using some features as used in training)
Comparison of various learning method based on classification result
Prediction of heart disease

Above Figure shows the methodology for heart disease prediction where :

A. Problem Statement:

The problem statement state that with the help of dataset predict whether a patient has any chance of getting heart disease or not. If the person has heart disease it will return 1 and if the person doesn’t have heart disease it will return 0.

B. Selecting DataSet:

dataset is used is Public Health Dataset to perform different comparative analysis of heart disease prediction. This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient.

The attributes are as follows-:

1. Age
2. Sex
3. CP
4. Trestbps
5. Chol
6. Fbs
7. Restecg
8. Thalach
9. Exang
10. Old peak
11. Slope
12. Ca
13. Thal
14. Target

C. Problem Solving Approach/Method:-

Machine learning techniques has already contributed in the field of healthcare .

The machine learning Technique used in these project are mainly four i.e

- K-NN
- RF
- Gradient Boost
- Ada Boost with Random Forest

D. Data Preprocessing:

It is a process of removing all the unwanted and missing data from the data set.

E. Data Train and Build model:

In this step the dataset is divided into two parts: training dataset and testing dataset. Training dataset contains 60% and testing dataset contains 40% which are selected randomly.

F. Input Details:-

Patient or user has to input all the required input which has been asked in form to get accurate result whether they have any chance of getting heart disease or not.

G.Comparison of different machine learning algorithms:

It this step the comparison is done between the classifiers. Different classifiers such as svm, random forest, knn, naive bayes, decision tree, logistic regression are compared based on the accuracy, precision, recall and f1 score.

H. Prediction of heart disease :-

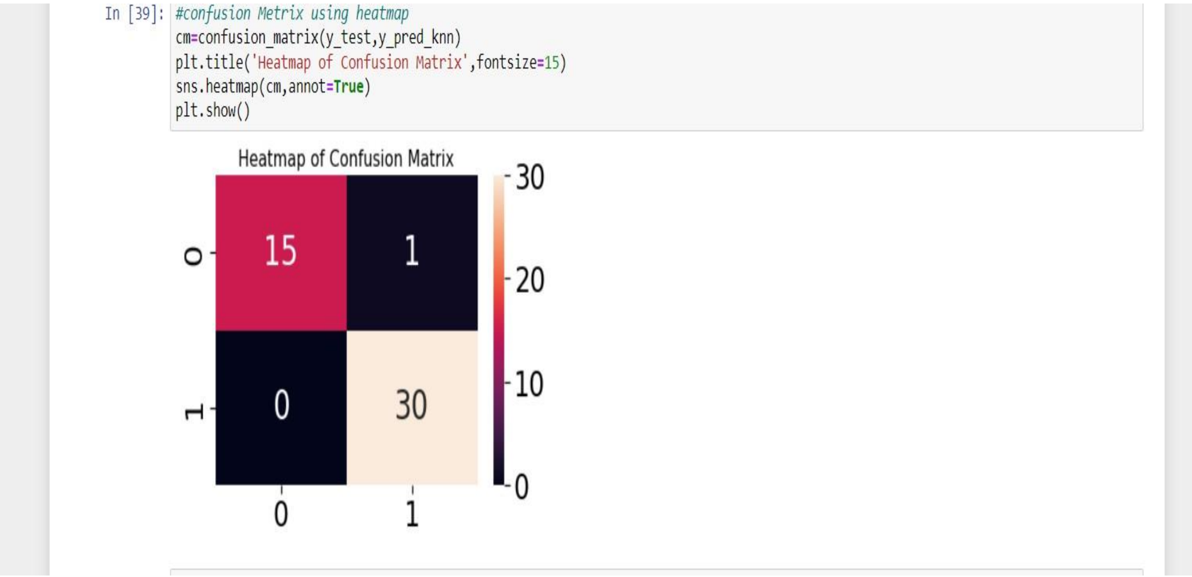
A web app has been developed using flask where frontend has been developed using HTML and CSS and backend with Python.Where User has to fill all the required inputs which has been asked in form and they shall been get the output as whether they have heart disease or not.

I. Checking accuracy of all the machine learning models:

Sr no	Model Name	Accuracy %
0	K - Nearest Neighboor	97.82%
1	Random Forest	86.95 %
2	Ada Boost With Random Forest	93.47 %
3	Gradient Boosting	89.91%

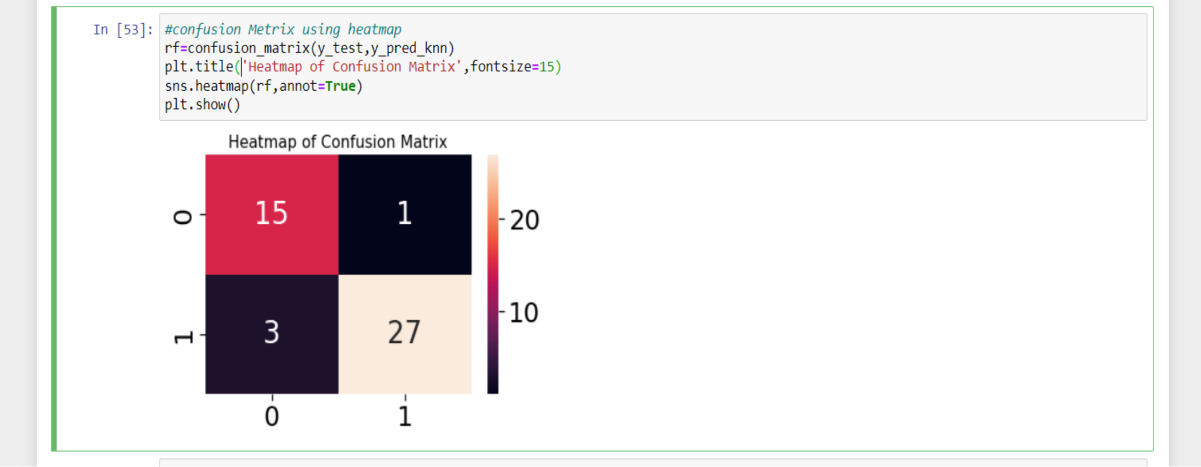
J. Confusion matrix of all the machine learning models:

1. K-NN



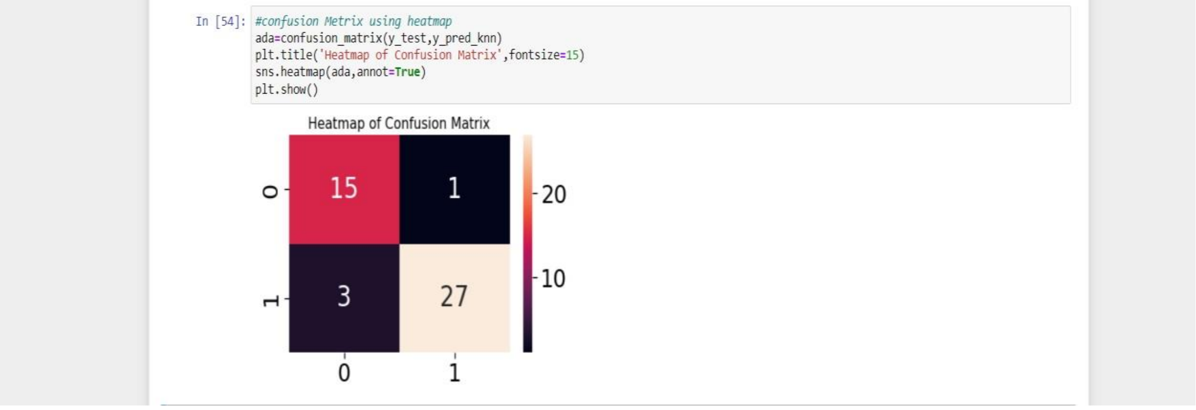
```
In [39]: #confusion Metrix using heatmap
         cm=confusion_matrix(y_test,y_pred_knn)
         plt.title('Heatmap of Confusion Matrix',fontsize=15)
         sns.heatmap(cm,annot=True)
         plt.show()
```

2. Random Forest:



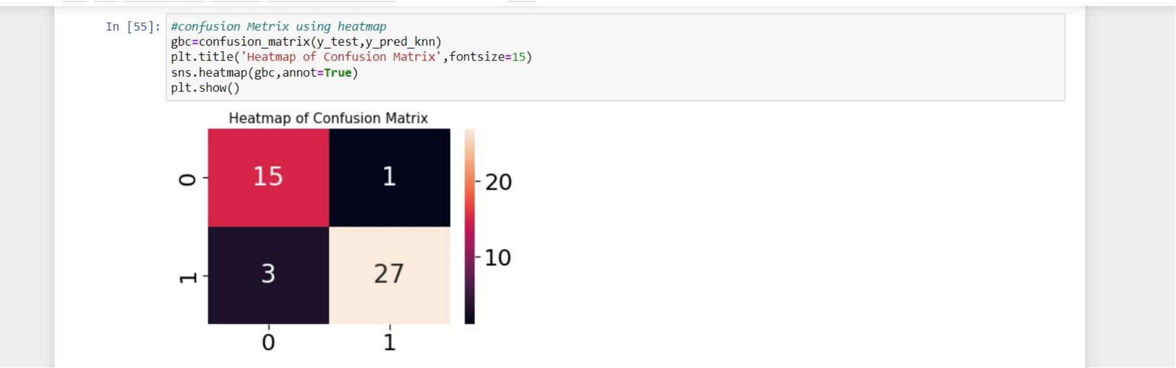
In [53]:
```
#confusion Metrix using heatmap
rf=confusion_matrix(y_test,y_pred_knn)
plt.title('Heatmap of Confusion Matrix',fontsize=15)
sns.heatmap(rf,annot=True)
plt.show()
```

3. Ada Boost with Random Forest:



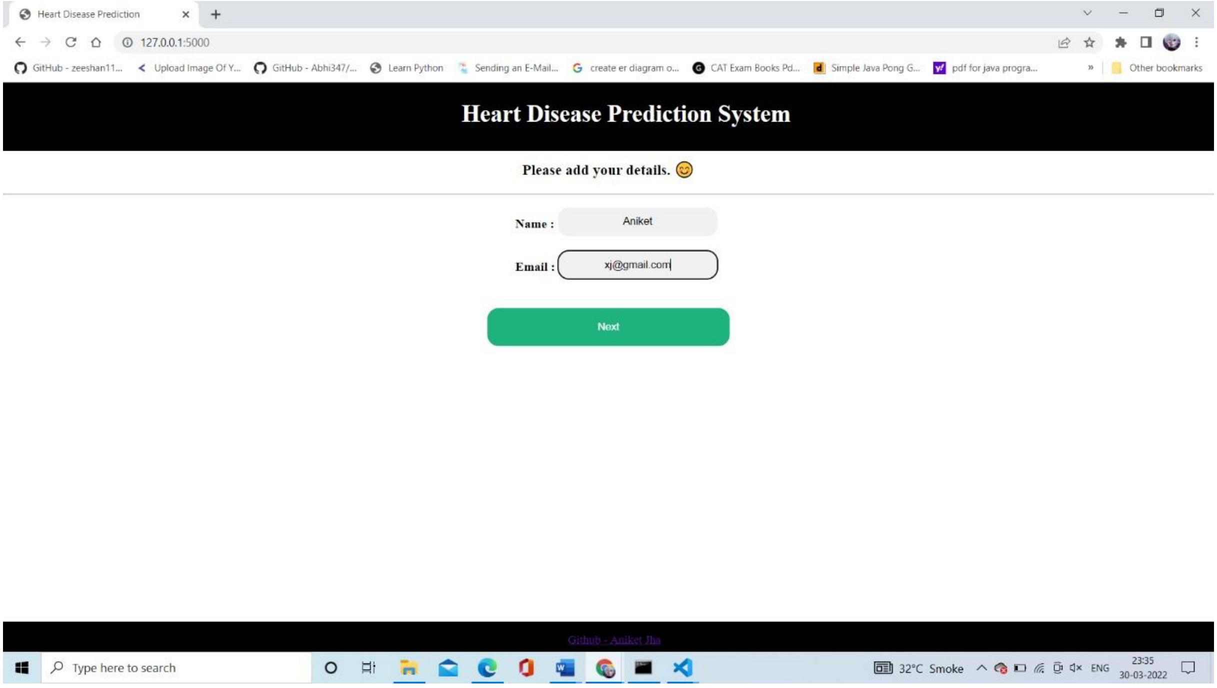
In [54]:
```
#confusion Metrix using heatmap
ada=confusion_matrix(y_test,y_pred_knn)
plt.title('Heatmap of Confusion Matrix',fontsize=15)
sns.heatmap(ada,annot=True)
plt.show()
```

4. Gradient Boost



In [55]:
```
#confusion Metrix using heatmap
gbc=confusion_matrix(y_test,y_pred_knn)
plt.title('Heatmap of Confusion Matrix',fontsize=15)
sns.heatmap(gbc,annot=True)
plt.show()
```

K. Result:



Heart Disease Prediction System

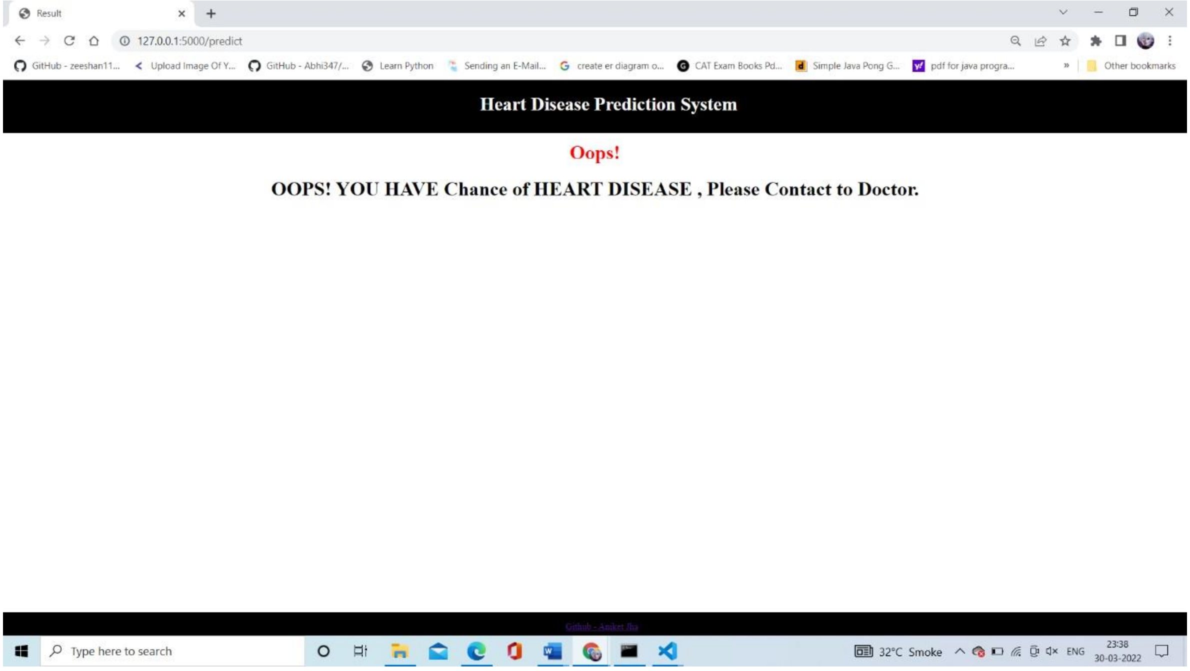
Please add your details. 😊

Name : Aniket

Email : xj@gmail.com

Next

Github - Aniket Jha



Heart Disease Prediction System

Oops!

OOPS! YOU HAVE Chance of HEART DISEASE , Please Contact to Doctor.

Github - Aniket Jha

FUTURE SCOPE:-

Today’s, world most of the data is computerized, the data is distributed and it is not utilizing properly. By Analyzing the available data we can also use for unknown patterns. The primary motive of this research is the prediction of heart diseases with high rate of accuracy.

The future scope of this system aims at giving more sophisticated prediction models, risk calculation tools and feature extraction tools for other clinical risks. This method can also be used to select the proper treatment methods for a patient in future, instead of just predicting the chances of developing a heart disease among the patient.In future my plan is to add more disease prediction like diabetes, symptoms,Pneumonia detection and many more.With that to provide a platform where user can self diagnosis and can check their health status ,In addition to that I’m also thinking to add doctor detail’s where user can get all the info related to that particular doctor in addition this I’ am also willing to add appointment and give user a functionality where they can get the best hospital near them with just giving PIN code which will extract provide that.To pack all these and to eradicate communication difference I am also willing to add a Voice assistant which user can just do all these with just voice.

CONCLUSION

Heart acts a major role in corporeal organism. The diseases of heart wants more perfection and exactness for diagnose and analyses.In real time heart diseases may not be detect in early stage. Due to heart disease, there is an increased in the number of deaths, day by day. The implementation of a method to efficiently and reliably predict heart diseases has become compulsory. The main motivation of this study is to find a powerful ML algorithm for detection of heart disease.

This project provides the deep insight into machine learning techniques for classification of heart diseases. The role of classifier is crucial in healthcare industry so that the results can be used for predicting the treatment which can be provided to patients. The existing techniques are studied and compared for finding the efficient and accurate systems. Machine learning techniques significantly improves accuracy of cardiovascular risk prediction through which patients can be identified during an early stage of disease and can be benefitted by preventive treatment

In these project I have used **K - Nearest Neighbour, Random Forest, Gradient Boost Ada Boost With Random Forest and the highest accuracy achived is as follows:**

Sr no	Model Name	Accuracy %
0	**K - Nearest Neighboor**	**97.82%**
1	**Random Forest**	**86.95 %**
2	**Ada Boost With Random Forest**	**93.47 %**
3	**Gradient Boosting**	**89.91%**

The outcome of this analysis shows that the **K - Nearest Neighbor** algorithm is the most powerful algorithm for heart disease prediction, with an accuracy score of 100%
It can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart related diseases.

REFERENCES

1. https://archive.ics.uci.edu/ml/datasets/heart+disease
2. https://en.wikipedia.org/wiki/Cardiovascular_disease.
3. https://stackoverflow .com
4. https://www.cdc.gov/heartdisease/index.htm#:~:text=Heart%20disease%20is%20the%20leading,can%20lead%20to%20heart%20attack.
5. https://www.medicalnewstoday.com/articles/237191
6. https://flask.palletsprojects.com/en/2.1.x/