VISUALIZING AND PREDICTING HEART DISEASES WITH AN INTERACTIVE DASH BOARD

TEAM ID:

PNT2022TMID15823

CONTENT

VISUALIZING AND PREDICTING HEART DISEASES WITH AN INTERACTIVE DASHBOARD

1.INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to heart disease. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project aims to predict future heart disease by analysing data of patients which classifies whether they have heart disease or not using many algorithms.

1.1 PROJECT OVERVIEW

The main objective of this project is to predict the heart disease. Among all fatal disease, heart attacks diseases are considered as the most prevalent. Medical practitioners conduct different surveys on heart diseases and gather information of heart patients, their symptoms and disease progression. Data analysis proves to be crucial in the medical field. It provides a meaningful base to critical decisions. It helps to create a complete study proposal. One of the

most important uses of data analysis is that it helps in keeping human bias away from medical conclusion with the help of proper statistical treatment. By use of data mining for exploratory analysis because of nontrivial information in large volumes of data. The health care 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.

1.2 PURPOSE

The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set. Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

- Provides new approach to concealed patterns in the data.
- Helps avoid human biasness.
- Reduce the cost of medical tests.

2. LITERATURE REVIEW

2.1 EXISTING PROBLEM

Healthcare industries generate enormous amount of data, so called big data that accommodates hidden knowledge or pattern for decision making. The huge volume of data is used to make decision which is more accurate than intuition. Exploratory Data Analysis (EDA) detects mistakes, finds appropriate data, checks assumptions and determines the correlation among the explanatory variables. In the context, EDA is considered as analysing data that excludes inferences and statistical modelling. Analytics is an essential technique for any profession as it forecast the future and hidden pattern. Data analytics is considered as a cost-effective technology in the recent past and it plays an essential role in healthcare which includes new research findings, emergency situations and outbreaks of disease. The use of analytics in healthcare improves care by facilitating preventive care and EDA is a vital step while analysing data.

2.2 REFERENCES

TITLE: PREDICTION AND ANALYSIS THE OCCURRENCE OF HEART DISEASE USING DATA MINING TECHNIQUES-2018

AUTHOR: CHALA BAYENE

DESCRIPTION: The main objective is to predict the occurrence of heart disease for early

automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organisation with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex is some of the attributes are included to identify if the person has heart disease or not. Analyses of dataset are computed using WEKA software.

TITLE:NON-LINEAR CLASSIFICATION ALGORITHM FOR HEART DISEASE PREDICTION -2018

AUTHOR: R. SHARMILA

DESCRIPTION: It is proposed to use bigdata tools such as Hadoop Distributed File System (HDFS), MapReduce along with SVM for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting heart diseases. It suggests to use HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. SVM is used in parallel fashion which yielded better computation time than sequential SVM.

TITLE: C45 RULES AND PARTIAL TREE TECHNIQUE TO PREDICT HEART DISEASE -2015

AUTHOR: SHARMA PURUSHOTTAM

DESCRIPITION: This paper can discover set of rules to predict the risk levels of patients based on given parameter about their health. The performance can be calculated in measures of accuracy classification, error classification, rules generated and the results. Then comparison has done using C4.5 and partial tree. The result shows that there is potential prediction and more efficient.

TITLE: THE DATA MINING TECHNIQUES AND MACHINE LEARNING TO PREDICT HEART DISEASE-2017

AUTHOR: JAYMIN PATEL

DESCRIPTION: There are two objectives to predict the heart system. 1. This system not assume any knowledge in prior about the patient’s records. 2. The system which chosen must be scalar to run against the large number of records. This system can be implemented using WEKA software. For testing, the classification tools and explorer mode of WEKA are used.

2.3 PROBLEM STATEMENT DEFINITION

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 cost for surgical treatment and other expensive. The overall objective of my work will be to predict accurately with 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.

3.IDEATION & PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS

The empathy map is about is describing the project with a gain and pain in this regard the depicts

about the outcome of our heart disease prediction project.

3.2 IDEATION & BRAINSTORMING

3.3 PROPOSED SOLUTION

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Analyze the heart disease dataset to explore the machine learning algorithms and build decision tree model to predict the disease.
2.	Idea / Solution description	To clean the model and pre-processing the model and to test and train the model using decision tree with the given dataset.
3.	Novelty / Uniqueness	We proposed a method for heart disease prediction using machine learning techniques, these results showed a great accuracy standard for producing a better estimation result.
4.	Social Impact / Customer Satisfaction	By using this method, we can separate the people those who can affected vs normal people, and it will play a vital role combining both medical and technology field. Customer(patients) can get benefit through saving financial cost (spending medical test), and by collecting dataset of their detailed condition,we can say that whether they get affected or not.

5.	Business Model (Revenue Model)	We can make revenue from this by making our developed model or a product form which can be modified into software kit, application or a webpage where they can interact easily. This all comes and developed under data analytics. We can get profited by selling or giving access with permission to our clients (Doctors).
6.	Scalability of the Solution	It is based on the number of users who maintaining the software or a system according to its performance like work flow, increase or decrease in efficiency, response time etc.By this a good quality of product is determined. If you suffer from a heart condition that interferes with your ability to work, you may qualify for disability benefits.

3.4 PROBLEM SOLUTION FIT

4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENT

Following are the functional requirements of the proposed solution.

	No. Functional equirement (Epic)	**Sub Requirement (Story / Sub-Task)**
FR-1	User Registration	Enables user to make registration for theapplication through Gmail
FR-2	User Confirmation	Once after registration, the user will getconfirmation mailvia Email platform
FR-3	Visualizing Data	User can visualize the trends on the heart disease throughDashboard created usingIBM Cognos Analytics
FR-4	Generating Report	User can view his/her health report and canmakedecisions accordingly

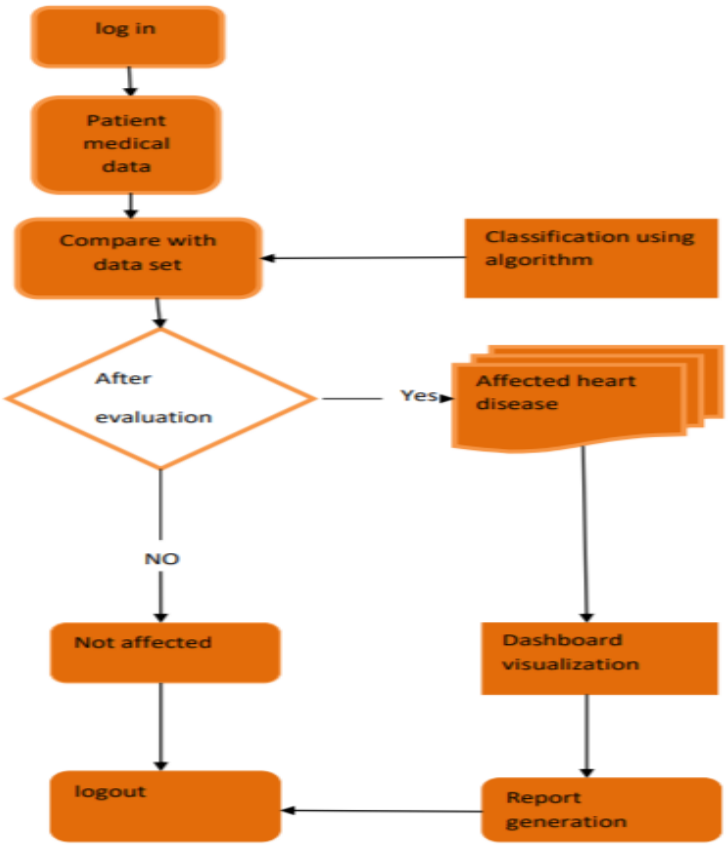
4.2 NON-FUNCTIONAL REQUIREMENT

	. Non-Functional Requirement	**Description**
NFR-1	**Usability**	The application will have a simple and user friendly graphical interface. User can understandthe application easily. Any action has to be performed with just a few clicks
NFR-2	**Security**	User's Dataprivacy is implemented using the login verification process. After verification only the user is granted permission to access theIBM Cognos Analytics Tool
NFR-3	**Reliability**	The application has to be consistent at every scenario and has to work without failure in any environment
NFR-4	**Performance**	Performance of the application depends on the response time and the speed of the datasubmission. The response time of the application is direct and faster which depends on the efficiency of implemented algorithm

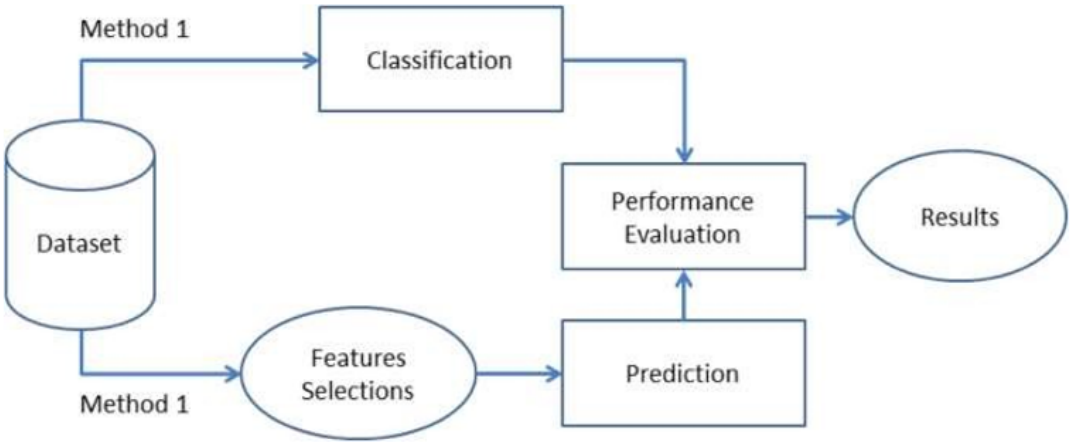
NFR-5	**Availability**	The application has to be available 24 x 7 forusers without any interruption. The user can access the application anytime at anywhere to get the desired / accurate results

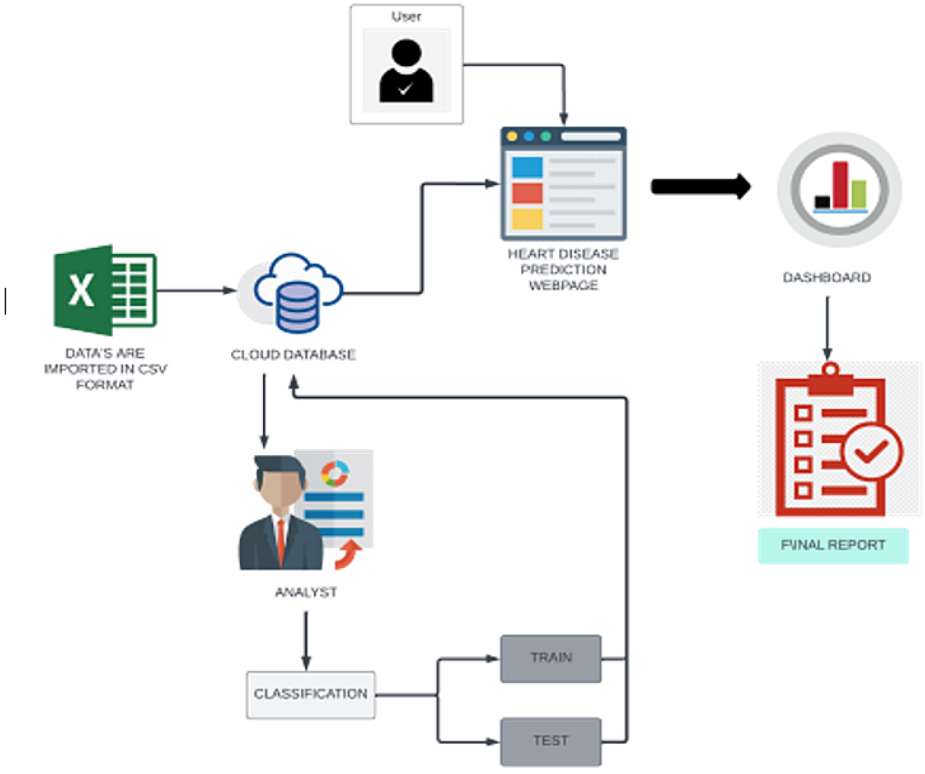
5. PROJECT DESIGN

5.1 DATA FLOW DIAGRAMS



5.2 SOLUTION & TECHNICAL ARCHITECTURE





5.3 USER STORIES

User Type	Functional Requirement (Epic)	UserStory Number	User Story / Task	Acceptancecriteria	Priority	Release
Customer (Mobile user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account/ dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Gmail	I can receive conformation email	Medium	Sprint-1
Customer (Web user)	Login	USN-4	As a user, I can log into the application by entering email & password	I can access my account using my details	High	Sprint-1
	Dashboard	USN-5	User can view his/her complete medical analysis and accuracy of disease prediction	I can view my medical analysis and accuracy	High	Sprint-2

	Dashboard	USN-6	User can view the accuracy of occurrence of heart disease through report generation	I can view the accuracy of heart disease in the dashboard	high	Sprint-2
Customer Care Executive	Helpdesk	USN-7	As a customer care executive, he/she can view the customer queries.	I can post my queries in the dashboard	Medium	Sprint-3
		USN-8	As a customer care executive, he/she can answer the customer queries	I can get support from helpdesk	High	Sprint-3
User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Administrator	User profile	USN-9	As an admin, he/she can update the health details of users.	I can view my updated health details	High	Sprint-4
		USN-10	As an admin, he/she can add or delete users.	I can access my account / Dashboard when logged in	High	Sprint-4
		USN-11	As an admin, he/she can manage the user details.	I can view the organized data of myself.	High	Sprint-4

6. PROJECT PLANNING & SCHEDULING

6.1 SPRINT PLANNING & ESTIMATION

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	As a user, I can registerfor the application by entering my email, password, and confirming my password.	2	High	1
Sprint-1		USN-2	As a user, I will receive confirmation email once I haveregistered for the application.	1	High	2
Sprint-2		USN-3	As a user,I can register for the applicationthrough Facebook	2	Low	4
Sprint-1		USN-4	As a user,I can register for the applicationthrough Gmail	2	Medium	3
Sprint-1	Login	USN-5	As a user,I can log into the application by entering email &password	1	High	2
Sprint-2	Dashboard	USN-6	Profile - view & update your profile	2	High	5
Sprint-1		USN-7	ChangePassword - user can changethe password	1	High	2
Sprint-1		USN-8	Home - Analyze your Heart	2	High	5

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members

Sprint-3		USN-9	The user will have to fill in the below13 fields for the systemto predict a disease -Agein Year -Gender -Chest Pain Type -Fasting Blood Sugar -Resting Electrographic Results(Restecg) -Exercise InducedAngina(Exang) -The slope of the peak exerciseSTsegment -CA – Numberof major vesselscolored by fluoroscopy -Thal -Trest Blood Pressure -Serum Cholesterol - Maximum heartrate achieved(Thalach) -ST depression induced by exercise(Oldpeak)	2	High	5
		USN-10	View Doctors - view doctor detail bysearching by names or filter by specialty	1	Medium	4
Sprint-3	System Requirment	USN-11	A. Hardware Requirement i. Laptop or PC I5 processor system or higher	2	High	2

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
			4 GB RAM or higher 128 GB ROM or higher Android Phone (12.0 and above)			
Sprint-3		USN-12	Software Requirement Laptop or PC Windows 10 or higher Android Studio	2	Medium	2
Sprint-4	Dashboard	USN-13	Query	1	High	1
		USN-14	Toll Free	1	High	1
		USN-15	Ratings	2	Medium	2
		USN-16	Verification	2	High	2
		USN-17	Validation	1	High	2
		USN-18	Feedback – send feedbackto the Admin	2	Medium	3

6.2 SPRINT DELIVERY SCHEDULE

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	SprintRelease Date(Actual)
Sprint-1	20	6Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	18	06 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	11 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	19	19 Nov 2022

7 CODING AND SOLUTIONING

7.1 MACHINE LEARNING

```
models = GetBasedModel()
names,results = BasedLine2(X_train, y_train,models)
```

```
LR_L2: 0.858874 (0.045985)
LDA: 0.854113 (0.052426)
KNN7: 0.777922 (0.087742)
KNN5: 0.749567 (0.097254)
KNN9: 0.777922 (0.085119)
KNN11: 0.749567 (0.099558)
CART: 0.731169 (0.063378)
NB: 0.839610 (0.057009)
SVM Linear: 0.840043 (0.051116)
SVM RBF: 0.763853 (0.102216)
AB: 0.783333 (0.061148)
GBM: 0.802165 (0.074146)
RF_Ent100: 0.830087 (0.056553)
RF_Gini100: 0.820779 (0.049480)
ET100: 0.820130 (0.067026)
ET500: 0.810606 (0.067851)
MLP: 0.741991 (0.191860)
SGD3000: 0.674675 (0.118784)
XGB_2000: 0.792641 (0.050455)
XGB_500: 0.797403 (0.049569)
XGB_100: 0.816017 (0.044422)
XGB_1000: 0.792641 (0.054765)
ET1000: 0.820130 (0.063553)
```

Comparing with all the algorithms

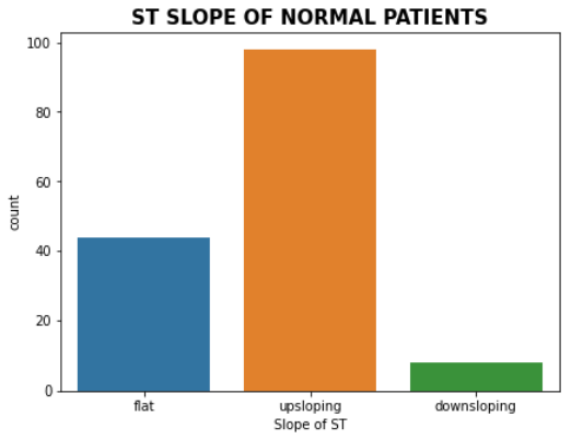
	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	Random Forest	0.754717	0.727273	0.695652	0.800000	0.711111	0.747826	8.471866	0.498551
1	MLP	0.792453	0.772727	0.739130	0.833333	0.755556	0.786232	7.168501	0.575812
2	KNN	0.792453	0.772727	0.739130	0.833333	0.755556	0.786232	7.168501	0.575812

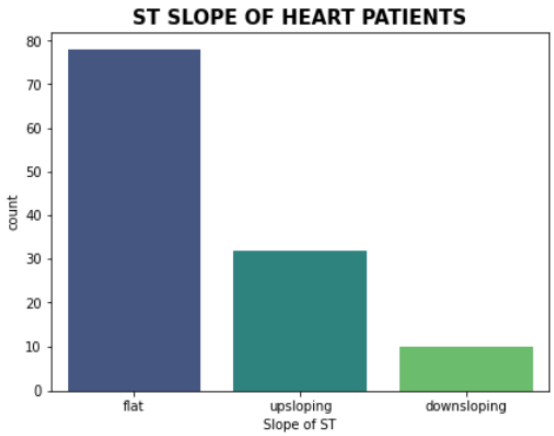
ks/IBM FINAL.ipynb#

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
3	EXtra tree classifier	0.773585	0.761905	0.695652	0.833333	0.727273	0.764493	7.820176	0.536023
4	XGB	0.811321	0.782609	0.782609	0.833333	0.782609	0.807971	6.516826	0.615942
5	SVC	0.811321	0.809524	0.739130	0.866667	0.772727	0.802899	6.516811	0.613857
6	SGD	0.433962	0.433962	1.000000	0.000000	0.605263	0.500000	19.550703	0.000000
7	Adaboost	0.735849	0.695652	0.695652	0.766667	0.695652	0.731159	9.123556	0.462319
8	CART	0.792453	0.750000	0.782609	0.800000	0.765957	0.791304	7.168516	0.580092
9	GBM	0.773585	0.720000	0.782609	0.766667	0.750000	0.774638	7.820206	0.545338

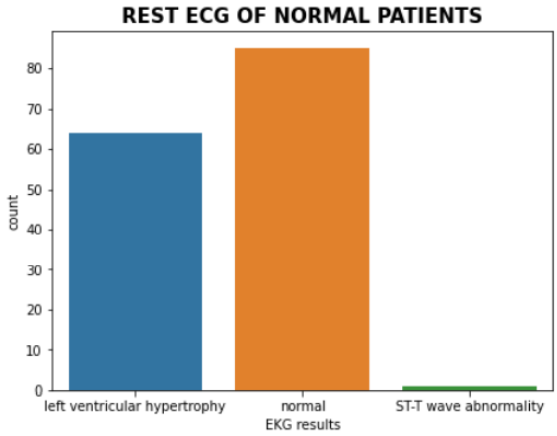
7.2 DASHBOARD

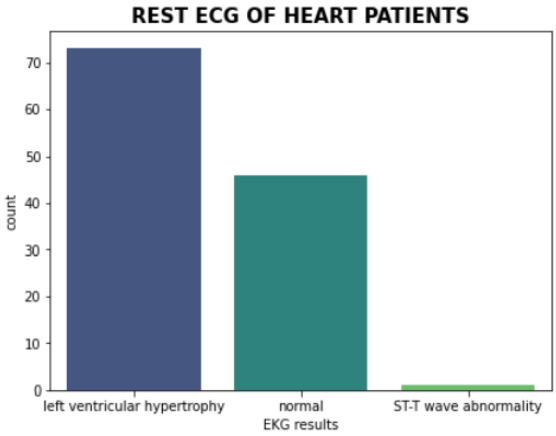
Exploration of ST Slope of Normal Patients vs Heart Patients



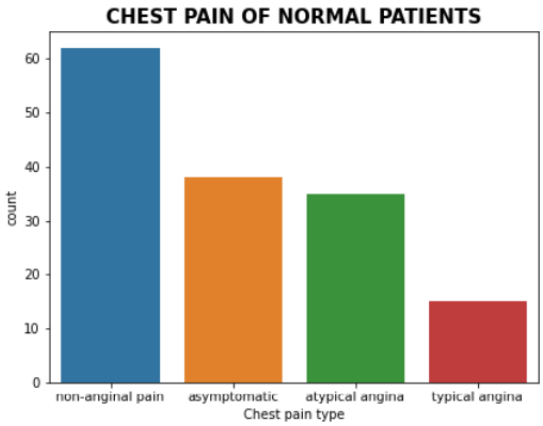


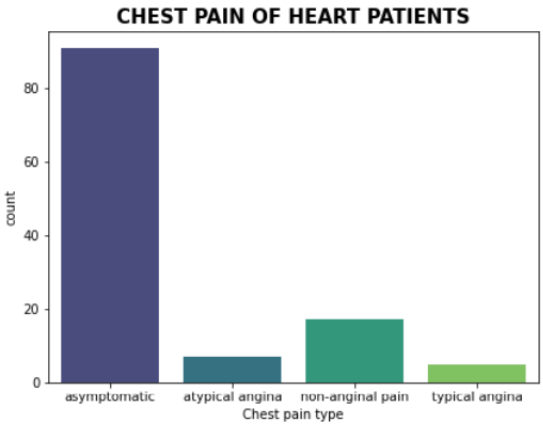
Exploration of Rest ECG of Normal Patients vs Heart Patients



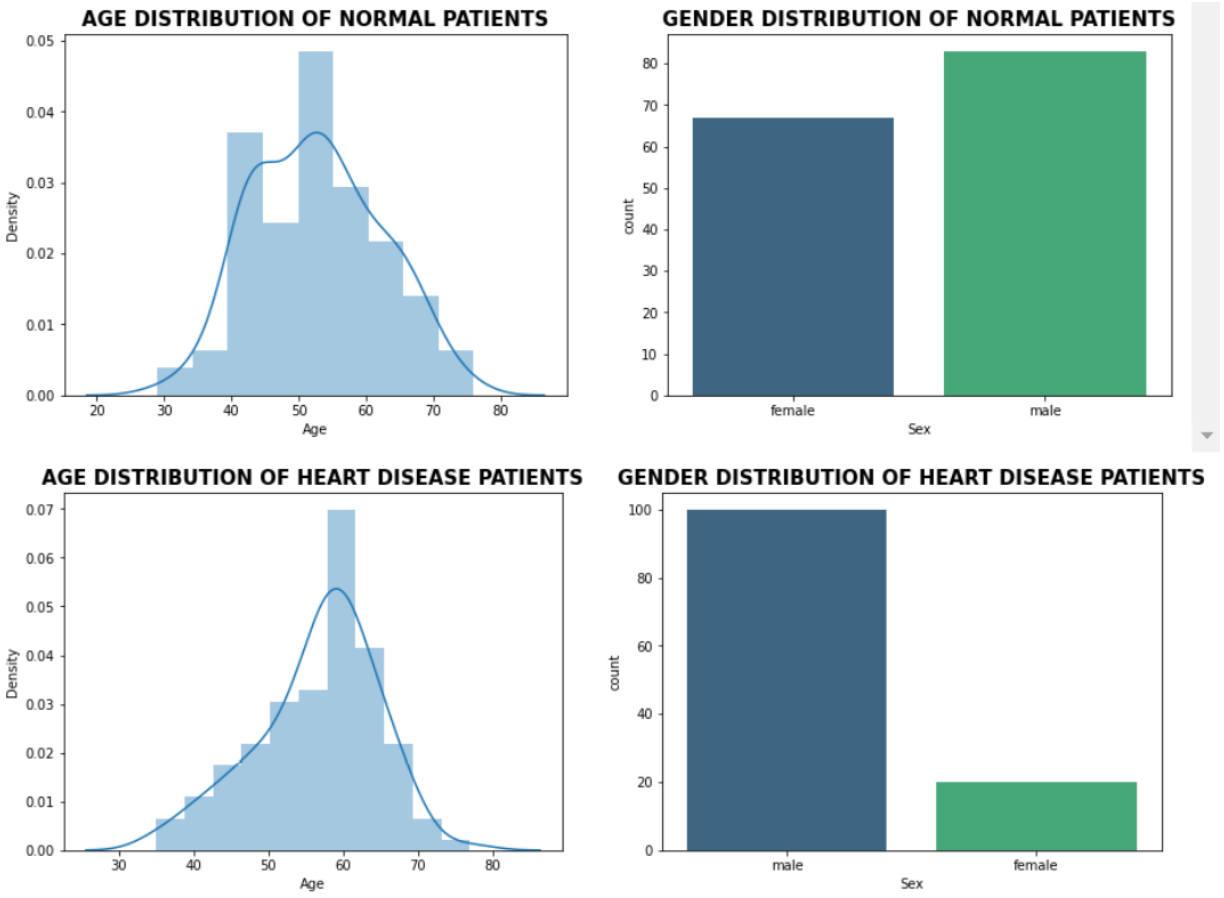


Exploration of Chest Pain of Normal Patients vs Heart Patients

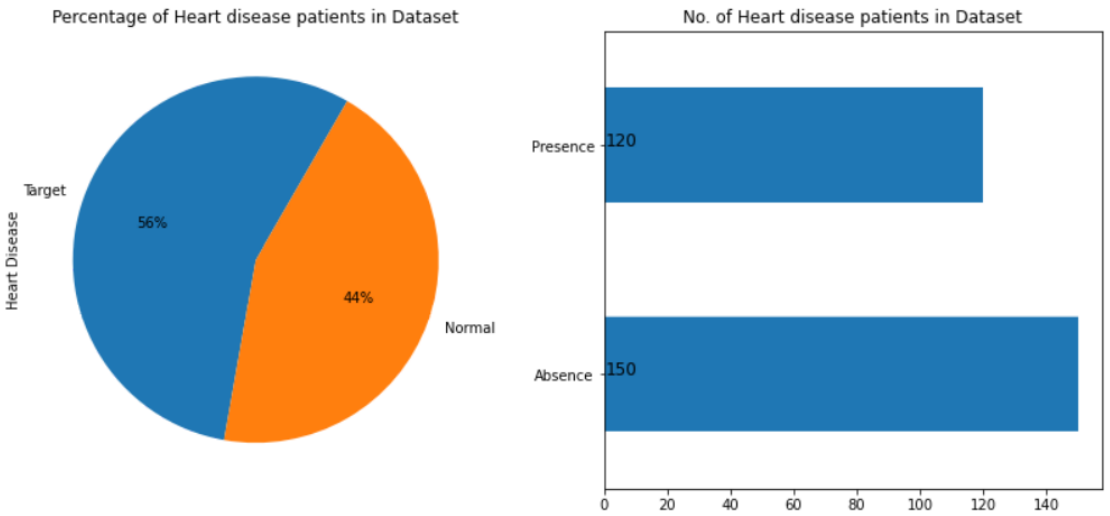




Distribution of normal patients and heart patients by Age and Gender



EDA of Heart Diseases Patients by target value



8. TESTING

8.1 TEST CASES TESTING THE DATA MODEL FOR VARIOUS INPUT VALUES

```
: Y={0.584906,0.320896,0.458955,1.14521}
  if(Y==1):
      print("The Person having a data diseases")
  else:
      print("The person not having a heart diseases")
```

```
The person not having a heart diseases
```

8.2 USER ACCEPTANCE TESTING

Testing a case where the output is given

```
: Y={0.584906,0.320896,0.458955,1.14521}
  if(Y==1):
      print("The Person having a data diseases")
  else:
      print("The person not having a heart diseases")
```

```
The person not having a heart diseases
```

9. RESULT

9.1 PERFORMANCE METRICS

The confusion matrix below shows the performance metrics of the machine learning model.

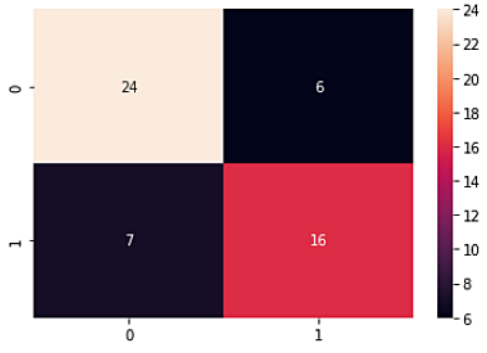
```
CM=confusion_matrix(y_test,y_pred_rfe)
sns.heatmap(CM, annot=True)

TN = CM[0][0]
FN = CM[1][0]
TP = CM[1][1]
FP = CM[0][1]
specificity = TN/(TN+FP)
loss_log = log_loss(y_test, y_pred_rfe)
acc= accuracy_score(y_test, y_pred_rfe)
roc=roc_auc_score(y_test, y_pred_rfe)
prec = precision_score(y_test, y_pred_rfe)
rec = recall_score(y_test, y_pred_rfe)
f1 = f1_score(y_test, y_pred_rfe)

mathew = matthews_corrcoef(y_test, y_pred_rfe)
model_results =pd.DataFrame([['Random Forest',acc, prec,rec,specificity, f1,roc, loss_log,mathew]],
               columns = ['Model', 'Accuracy','Precision', 'Sensitivity','Specificity', 'F1 Score','ROC','Log_Lo

model_results
```

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	Random Forest	0.754717	0.727273	0.695652	0.8	0.711111	0.747826	8.471866	0.498551



10. ADVANTAGES & DISADVANTAGES

ADVANTAGES:

- Increased accuracy for effective heart disease diagnosis.
- Reduce the time complexity of doctors.
- Cost effective for patients.
- This is one of the fastest ways to determine if a person is likely to suffer from a heart disease or not.
- Useful for medical practitioners to easily classify their patients.
- User Friendly.
- Easy to understand.
- Secure.
- Dashboard provides insightful informations.

DISADVANTAGES:

- Needs work.
- Users need to know all the fields.
- Does Not take null value as input.
- Does not provide suggestions to user.

11. CONCLUSION

The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio-patients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system This project resolved the feature selection i.e., backward elimination and

behind the models and successfully predict the heart disease, with 85% accuracy using logistic regression algorithm.Further for its enhancement, we can train on models and predict the types of cardiovascular diseases providing recommendations to the users, and also use more enhanced models.

12.FUTURE SCOPE

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

13. APPENDIX

Source code

https://github.com/IBM-EPBL/IBM-Project-16911-1659624749/tree/main/Final%20Deliverables

Demo Video Link

https://drive.google.com/file/d/1lLMM82414zSPLx7zKha0qkPIW6q8n8f0/view?usp=drivesdk