PROJECT REPORT ON

HEART DISEASE PRIDICTION

Submitted by: Medha Bhat (03)
Yashvi Das (08)
Aakanksha Deshmukh (11)

Of

Bachelor's of Technology, Information Technology from USHA MITTAL INSTITUTE OF TECHNOLOGY

Under the Guidance of: **Prof. Anita Morey**

ACKNOWLEDGEMENT

We would like to express our gratitude to our professor Anita Morey for providing us with an opportunity to work on a Artificial intelligence project on Heart Disease Prediction System. With your guidance and support we were able to complete this project.

This project helped us in enhancing our skills in coding and working on different software environments. The research work done for this project helped us in acquiring knowledge in various aspects. The sincere efforts and coordination from our group members helped us in completing this project on time.

Medha Bhat (03)

Yashvi Das (08)

Aakanksha Deshmukh (11)

ABSTRACT

The Health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. In this study, a Heart Disease Prediction System (HDPS) is developed using Naives 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 etc. 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 diseases.

Chapter 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 analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithms.

1.2:**PROBLEM DEFINITION**

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and over all complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more

sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

1.3:MAIN OBJECTIVES

The main objective of this research is to develop a heart prediction system. The system can discover an 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 prediction of heart diseases.

1.4:SPECIFIC OBJECTIVE

- Provides new approach to concealed patterns in the data.
- Helps avoid human biasness.
- To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.
- Reduce the cost of medical tests

1.5:JUSTIFICATION

Clinical decisions are often made based on doctor's insight and experience rather than on the knowledge rich data hidden in the dataset. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed system will integrate clinical decision support with computer-based patient records (Data Sets).This will 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.There are voluminous records in medical data domain and because of this, it has become necessary to use data mining techniques to help in decision support and prediction in the field of healthcare. Therefore, medical data mining contributes to business intelligence which is useful for diagnosing of disease.

1.6: 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.

1.7:LIMITATIONS

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor's intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

CHAPTER 2: DATASET

The dataset is publicly available on the Kaggle Website at [4] which is from an ongoingcardiovascular study on residents of the town of Framingham, Massachusetts. It provides patientinformation which includes over 4000 records and 14 attributes. The attributes include: age, sex,chest pain type, resting blood pressure, serum cholesterol, fasting, sugar blood, restingelectrocardiographic results, maximum heart rate, exercise induced angina, ST depression induced by exercise, slope of the peak exercise, number of major vessels, and target ranging from 0 to 2,where 0 is absence of heart disease. The data set is in csv (Comma Separated Value) format whichis further prepared to data frame as supported by pandas library in python

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

The education data is irrelevant to the heart disease of an individual, so it is dropped. Further withthis dataset pre-processing and experiments are then carried out.

1.12 Classifiers Used for Experiments.

#	Attributes	Description	Values
1	Age	Patient's age in years	Continuous Value
2	Sex	Sex of Patient	1 = Male 0 = Female
3	Cp	Chest pain	Value 1: typical angina Value 2: atypical angina Value 3: non-angina pain Value 4: asymptomatic
4	Trestbps	Resting blood pressure	Continuous value in mm/Hg
5	Chol	Serum cholesterol in mg/dl	Continuous value in mg/dl
6	Fbs	Fasting blood sugar	1 ≥120 mg/dl 0 ≤ 120 mg/dl
7	Restcg	Resting electrocardiographic results	0 = normal 1 = having_ST_T wave abnormal 2 = left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved	Continuous value
9	Exang	Exercise induced angina	1: yes 0: no
10	Oldpeak	ST depression induced by exercise relative to rest	Continuous value
11	Slope	the slope of the peak exercise ST segment	1: upsloping 2: flat 3: down sloping
12	Ca	number of major vessels colored by fluoroscopy	0-3 value
13	Thal	defect type	3 = normal 6 = fixed defect 7 = reversible defect
14	num	diagnosis of heart disease	no_heart_disease have_heart_disease

CHAPTER 3: METHODS AND ALGORITHMS USED

The main purpose of designing this system is to predict the ten-year risk of future heart disease.We have used Logistic regression as a machine-learning algorithm to train our system and variousfeature selection algorithms like Backward elimination and Recursive feature elimination. These algorithms are discussed below in detail.

3.1 Logistic Regression

Logistic Regression is a supervised classification algorithm. It is a predictive analysis algorithm based on the concept of probability. It measures the relationship between the dependent variable(TenyearCHD) and the one or more independent variables (risk factors) by estimating probabilitiesusing underlying logistic function (sigmoid function). Sigmoid function is used as a cost functionto limit the hypothesis of logistic regression between 0 and 1 (squashing) i

.e. 0 ≤ h

θ

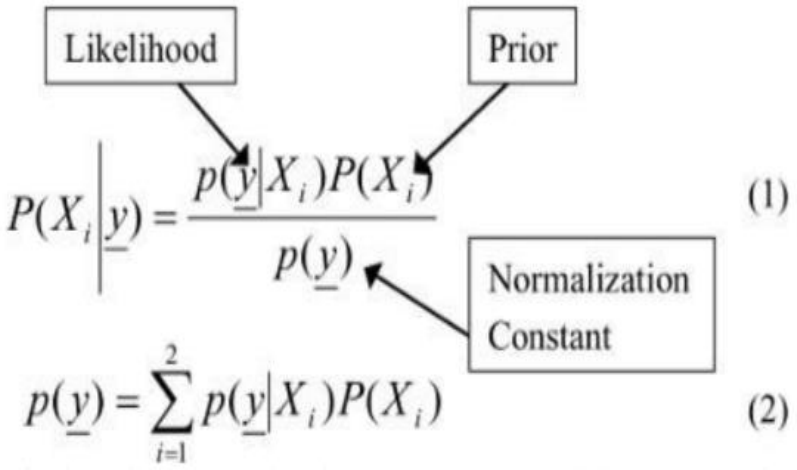
(x) ≤ 1.

$$Cost(\mathrm{h}\theta(\mathrm{x}),\mathrm{y}) = \begin{cases} -\log\big(h\theta(x)\big) & if\ y = 1 \\ -\log\big(1-h\theta(x)\big) & if\ y = 0 \end{cases}$$

Logistic Regression relies highly on the proper presentation of data. So, to make the model more powerful, important features from the available data set are selected using Backward eliminationand recursive elimination techniques

3.2: Naïve Bayesian

It is a probabilistic classifier based on Bayes' theorem specified by the prior probabilities of its root nodes. The Bayes theorem is given in Equation 1 and normalization constant is given in Equation 2. It proves to be an optimal algorithm in terms of minimization of generalized error. It can handle statisticalbased machine learning for feature vectors and assign the label for feature vector based on maximal probable among available classes {XX1, X2..., XM}. It means that feature "y" belongs to Xiclass, when posterior probability is maximum ie Max. The Bayesian classification problem may be formulated by aposterior probabilities that assign the class label ωi to sample X such that is maximal. The Bayesian classification problem may be formulated by a-posterior probabilities that assign the class label ωi to ample X such that is maximal.



Application of Bayes' rule with the mutual exclusivity in diseases and the conditional independence in findings is known as the Naïve Bayesian Approach. It is a probabilistic classifier based on Bayes' theorem with strong independence assumptions between the features. Naïve Bayesian classifier despite its simplicity, it surprisingly performs well and often outperforms in complex classification. Simple Naïve Bayesian can be implemented by plugging in the following main Bayes' formula:

P (X1, X2…, Xn | Y) = P (X1 | Y) P (X2 | Y) … P (Xn | Y) (3)

The above-mentioned Naïve Bayesian network produces a mathematical model, which is used for modeling the complicated relations of random variables of disease attributes and decision outcome. The algorithm uses the formula to calculate conditional probability with respect to disease condition attributes value and decision attribute value. Based on prior knowledge, the algorithm classifies the decision attribute into labels assigned, and hence the conditional support is computed for each variable

3.3: Ensemble DM approach.

In order to have more reliable and accurate prediction results, ensemble method is a well-proven approach practiced in research for attaining highly accurate classification of data by hybridizing different classifiers. The improved prediction performance is a well-known in-built feature of ensemble methodology. This study proposes a weighted vote-based classifier ensemble technique, overcoming the limitations of conventional DM techniques by employing the ensemble of two heterogeneous classifiers: Naive Bayesian and classification via decision tree

3.4: Decision Trees

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 these J48 algorithm is used for this system. J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications. 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 balance of flexibility & accuracy.

CHAPTER 4: EXPERIMENTS

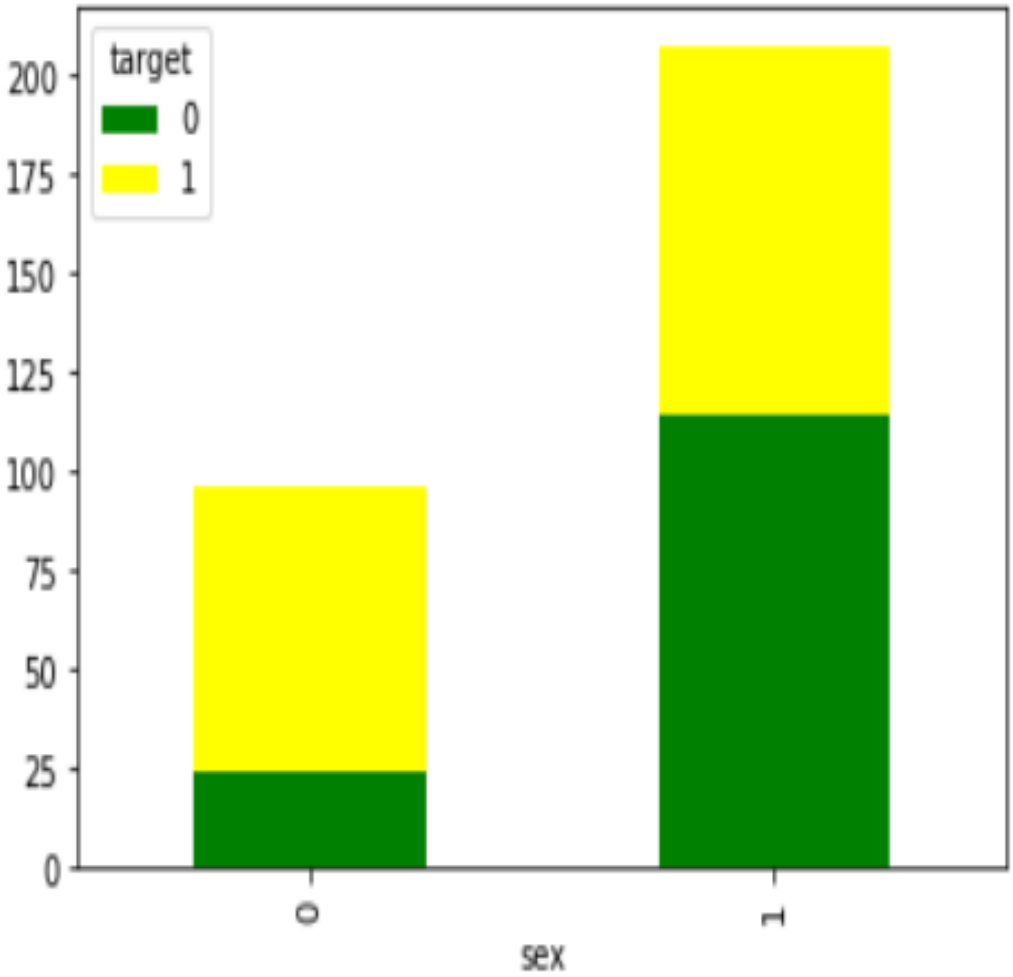
4.1: Data Preparation

Since the dataset consists of 303 observations with 114 men having no heart disease and 93 men with heart disease. Similarly 24 women having no heart disease and 72 women with risk of having heart disease. So,we progressed with imputation of data with the mean value of the observations and scaling themusing SimpleImputer and StandardScaler modules of Sklearn.

```
In [47]: gen= pd.crosstab(df['sex'], df['target'])
         print(gen)
```

target	0	1
sex		
0	24	72
1	114	93

Fig2:-



These two shows the numbers having no heart disease and having risk of heart disease.

Fig3:-

```
In [6]: df.columns

Out[6]: Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
               'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'],
              dtype='object')

In [7]: ##we can see the column names here

In [8]: df.describe()
```

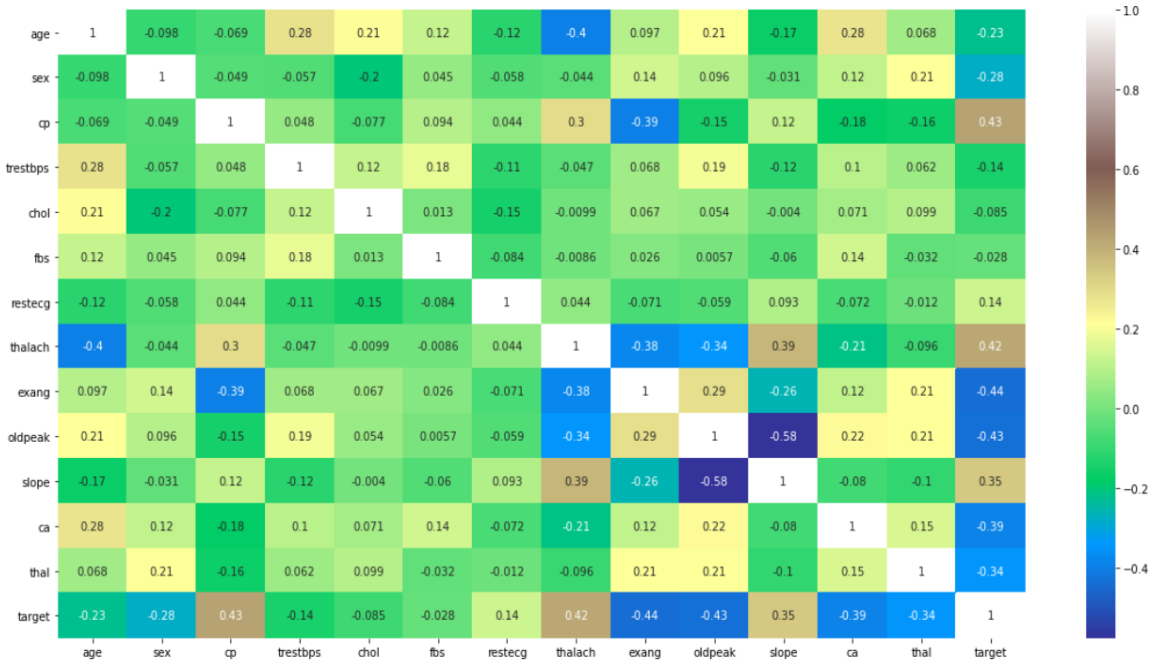
Out[8]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.(
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.3
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.(
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.(
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.(
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.(
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.(
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.(

The above figure shows the colums which shows :- (age,sex,cp,trestbps,chol,fbs,resteg,exang, oldpeak, slope,ca,thal,target)

Along with describing the the mean,std ,min and count values of the given dataset

Fig4:-



Above figure shows the positive correlation between Target and Cp,Thalach ,slope and also negative correlation between Target,sex,exang,ca ,Thal,oldpeak

Fig5:-

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
StandardScaler = StandardScaler()
columns_to_scale = ['age','trestbps','chol','thalach','oldpeak',]
df[columns_to_scale] = StandardScaler.fit_transform(df[columns_to_scale])
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	0.952197	1	3	0.763956	-0.256334	1	0	0.015443	0	1.087338	0	0	1	1
1	-1.915313	1	2	-0.092738	0.072199	0	1	1.633471	0	2.122573	0	0	2	1
2	-1.474158	0	1	-0.092738	-0.816773	0	0	0.977514	0	0.310912	2	0	2	1
3	0.180175	1	1	-0.663867	-0.198357	0	1	1.239897	0	-0.206705	2	0	2	1
4	0.290464	0	0	-0.663867	2.082050	0	1	0.583939	1	-0.379244	2	0	2	1

Above fig shows the data output after scaling the data

Fig6:-

```
TP=cm[0][0]
TN=cm[1][1]
FN=cm[1][0]
FP=cm[0][1]
print('Testing accuracy:',(TP+TN)/(TP+TN+FN+FP))
```

```
Testing accuracy: 0.9230769230769231
```

Testing accuracy of the dataset

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test,prediction1)
```

```
0.9230769230769231
```

```
from sklearn.metrics import classification_report
print(classification_report(y_test, prediction1))
```

```
              precision    recall  f1-score   support

           0       0.92      0.90      0.91        40
           1       0.92      0.94      0.93        51

    accuracy                           0.92        91
   macro avg       0.92      0.92      0.92        91
weighted avg       0.92      0.92      0.92        91
```

Predicting the data and getting output using classification

Fig7

```
In [75]: #Random forest

In [76]: from sklearn.ensemble import RandomForestClassifier

         rfc=RandomForestClassifier()
         model3=rfc.fit(X_train, y_train)
         prediction3= model3.predict(X_test)
         confusion_matrix(y_test, prediction3)

Out[76]: array([[36,  4],
                [ 6, 45]], dtype=int64)

In [77]: accuracy_score(y_test, prediction3)

Out[77]: 0.8901098901098901

In [78]: print(classification_report(y_test, prediction3))

                    precision    recall  f1-score   support

                 0       0.86      0.90      0.88        40
                 1       0.92      0.88      0.90        51

          accuracy                           0.89        91
         macro avg       0.89      0.89      0.89        91
      weighted avg       0.89      0.89      0.89        91
```

Predicting the data and getting the output using Random Forest classifier

Fig8:-

```
In [88]: print('KNN:', accuracy_score(y_test,prediction6))
         print('lr:', accuracy_score(y_test,prediction1))
         print('dtc:', accuracy_score(y_test,prediction2))
         print('rfc:', accuracy_score(y_test,prediction3))
         print('KB:', accuracy_score(y_test,prediction4))
         print('SVC:', accuracy_score(y_test,prediction5))

         KNN: 0.8351648351648352
         lr: 0.9230769230769231
         dtc: 0.7362637362637363
         rfc: 0.8901098901098901
         KB: 0.9010989010989011
         SVC: 0.8791208791208791

In [89]: #Best accuracy is given by Logistic Regression: 92
         #Followed by NB and Decision tree: 90
```

Accuracy score and the best accuracy is given by logistic regression follow by NB and Decision tree.

CHAPTER 5: Training and testing

Finally, this resulting data split into 80% train and 20% test data, which was further passed to theLogisticRegression model to fit, predict and score the model.

CHAPTER 6: EVALUATION METRICS

For the evaluation of our output from our training the data, the accuracy was analyzed

6.1:Accuracy

The accuracy is calculated as:Accuracy =

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

•True Positive (TP) =Observation is positive, and is predicted to be positive.

•False Negative (FN) = Observation is positive, but is predicted negative.

•True Negative (TN) = Observation is negative, and is predicted to be negative.

•False Positive (FP) =Observation is negative, but is predicted positiveThe obtained accuracy during training the data after feature selection using backward eliminationwas 86 % and during testing was 83%.The obtained accuracy during training the data after

feature selection using REFCV method was86 % and during testing was 85 %.

CHAPTER 7: DISCUSSION ON RESULTS

When performing various methods of feature selection, testing it was found that backwardelimination gave us the best results among others. The various methods tried were BackwardElimination with and without KFold, Recursive Feature Elimination with Cross Validation. Theaccuracy that was seen in them ranged around 85% with 85.5% being maximum. Though bothmethods gave similar accuracy but it was seen that in Backward Elimination we found that thenumber of misclassifications of True Negative was more and it was observed that the accuracy hadmore variance compared to RFEV. The precision of Backward Elimination and RFEV are 84%and 86% respectively. And the recalls are 0.99 and 1 respectively. The precision and recall alsoshows that the number of misclassifications is less in RFECV than in Backward Elimination.Evaluation Metrics Backward Elimination RFECVAccuracy 83% 85%Recall 0.99 0.99Precision 0.84 0.86

Table 3: Comparison between the feature selection models after training and testing through LogisticRegression model

Evaluation Metrics	Backward Elimination	RFECV
Accuracy	83%	85%
Recall	0.99	0.99
Precision	0.84	0.86

Comparison between the feature selection models after training and testing through LogisticRegression model

CHAPTER 8: CONCLUSION

The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes inhigh risk patients and in turn reduce the complications, which can be a great milestone in the fieldof medicine. This project resolved the feature selection i.e. backward elimination and RFECV behind the models and successfully predict the heart disease, with 85% accuracy. The model usedwas Logistic Regression. Further for its enhancement, we can train on models and predict thetypes of cardiovascular diseases providing recommendations to the users, and also use more enhanced models

REFERENCES

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

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

[3] K. Bhanot, "towarddatascience.com," 13 Feb 2019. [Online]. Available:https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c. [Accessed 2 March 2020].

[4] [Online]. Available: https://www.kaggle.com/ronitf/heart-disease-uci#heart.csv.. [Accessed05 December 2019].[5] M. A. K. S. H. K. M. a. V. P. M Marimuthu, "A Review on Heart Disease Prediction usingMachine Learning and Data Analytics Approach