Optimal Prediction of Heart Disease Using Machine Learning Techniques with Logistic Regression Model

Ghulab Nabi Ahmad[1], Hira Fatima[2] and Shafiullah[3]

{ghulamnabiahmad@gmail.com[1],
hirafatima2014@gmal.com[2], shafi.stats@gmail.com[3]}

Institute of Applied Sciences, Mangalayatan University, Aligarh, U. P, India.[1, 2]
Department of Mathematics, K.C.T.C College, Raxual, BRA, Bihar University Muzaffarpur, India.[3]

Abstract. One of the most difficult challenges in the health-care sector is to anticipate coronary artery disease. Heart disease seems to be more common in males than in women. The quantity of smokers smoked each day, as well as systolic and diastolic blood pressure, all increase the risk of heart disease. As a result, we propose to create an application that can forecast the risk of heart disease based on fundamental symptoms such as age, sex, pulse rate, and so on. The proposed solution makes use of the machine learning methodology logistic regression, which has been proved to be the most accurate and reliable. The model's performance is assessed using publicly available datasets such as the Cleveland Heart Disease Dataset (CHD), with logistic regression achieving the highest accuracy of 89.52 %. And an accuracy of 93.54 % for ROC_ AUC. We describe a predictive analytics-based technique for detecting heart disease in this research.

Keywords: forecast; heart disease; symptoms; logistic regression.

1. Introduction

There are several publications available about the medical symptoms of patients who have had a heart attack. Their ability to predict comparable results in otherwise healthy people, on the other hand, has largely gone unnoticed. Consider the following illustration: Half of all heart strokes occur in adults under the age of 50, and a quarter of all heart strokes occur in those under the age of 40, according to the Indian Heart Association. In cities, heart attacks strike three times as many people as they do in rural regions. [1] As a result, we recommend collecting relevant data on all aspects of our field of study, training the data with the proposed machine learning method, and predicting the likelihood of a patient contracting a cardiac ailment. We advocate assessing basic aspects with widely available sensors like watches and mobile phones for the aim of patients contributing data

The following is a breakdown of the structure of the paper: Section II examines previous heart disease research using a variety of machine learning approaches, Section III explains the database we use and our recommended model's approach analysis, and Section IV concludes with detailed results and comparisons to other methods. Finally, in Section V, the paper's conclusion and future study potential are discussed.

ICIDSSD 2022, May 24-25, New Delhi, India

DOI 10.4108/eai.24-5-2022.2318558

2. Research Work

To begin, we have begun gathering data in all aspects of the system in order to achieve the system's purpose. First and foremost, the research focused on the primary causes or other factors that have a big influence on heart health. Some characteristics, such as age, sex, and family history, cannot be changed, while others, such as blood pressure and heart rate, can be controlled by adhering to specific guidelines [2].

Many experts prescribe a healthful diet and regular exercise to keep the heart healthy. The parameters that are investigated for the study in building the system that have a high-risk percentage in terms of CAD are mentioned below. Age, gender, hypertension, pulse rate, obesity, metabolic disorders, and BMI [3] are all factors to consider. The following stage was to gather data. We utilised the CHD dataset from the Kaggle for this. The dataset comprises up to 76 factors that describe the heart's overall health. Expensive clinical procedures, such as an ECG or a CT scan, are used to collect these values. The classic heart disease prediction system [4-5] employs 13 primary factors out of these. Because determining ECG, chest pain type, ST depression, and other characteristics necessitates costly lab testing, to prevent these issues and make the system less complex, we chose the characteristics listed above, which can be readily monitored using a variety of sensors available on the market. The following research paper provides a brief overview of the most recent sensors available on the market for monitoring various factors.

a. Alive Core Inc

It's available as a touchpad or a bracelet that connects to using a Wi-Fi data connection your mobile The touchpad uses Bluetooth to mimic the patient's ECG on his phone. As a result, as well as all of the critical metrics, such as heart rate and blood heaviness, are simply accessible. Its pulse function on the bracelet's dial, but from the other hand, is displayed through finger touch. It might indicate the start of atrial fibrillation [6].

b. My heart

A variety of on-body devices are employed in this system to collect sensor parameters, which is then relayed electronically to a PDA. The dataset is examined, and the user is provided medical suggestions derived from the findings [7].

c. Fitbit

This detector is expected to protect record of one's healthcare and includes functions such as heart rates, plasma heaviness, and fatty expended.

Following conducting this study, we came to the conclusion that we should use Fitbit to gather data that is readily accessible and so much less costly and Health Gear for all other aspects.

3 Proposed Mechanism

Logistic Regression: The issue can arise, "Why logistic, but not linear?" Because linear regression is unlimited and the classifier might make mistakes, logistic regression is utilised instead. The simplest interpretation of logistic regression [8], which is between 0 and 1. From the available data, the model predicts the probability of our random variable, which is our goal value [9]. The sigmoid function is the cost function used in logistic regression. The linear function is essentially used as an input to another function, such as f in the equation, in logistic regression.

$$\hat{y} = h_\theta(x) = f(\theta^T x) \quad \text{where } 0 \le h_\theta(x) \le, \text{ Where } \hat{y} = \text{predicted value} \tag{1}$$

$$X^T = [\,1\,, x_1, x_2, x_3\,, \dots\dots\dots x_n], \quad \theta^T = [\beta_0, \beta_1, \beta_2, \beta_3, \dots\dots\dots\dots \beta_n] \text{ Then} \Rightarrow \hat{y} = \theta^T X \tag{2}$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \beta_1 x_1 + \cdots \dots\dots + \beta_n x_n = \sum_{i=0}^{n} \beta_n x_n = Z \tag{3}$$

x=independent variable and are coefficient to be learnt. But, because of the linear regression stuff, we need to figure out a method to convert the logistic regression issue in a way that allows us to utilise at least the expression above. For example, if we computed the probabilities of the result as,

$$Odds = \frac{P}{1-P}\,, 0 \le p \le 1 \Rightarrow P = \frac{Odds}{1+Odds} \tag{4}$$

We can move a step closer to casting the problem in a continuous linear manner but this is still just having positive values we need a range of $(-\infty, +\infty)$ .That can be done by getting the (natural) logarithm of the odds as:

$$logit(P) = log\frac{P}{1-P} = \hat{y} = \theta^T X, \quad log\frac{P}{1-P} = \theta^T X \Rightarrow \quad \frac{P}{1-P} = e^{\theta^T X}$$

$$\Rightarrow\ P = (1-P)e^{\theta^T X}\ \Rightarrow\ P = \frac{1}{1+e^{-\theta^T X}} \tag{5}$$

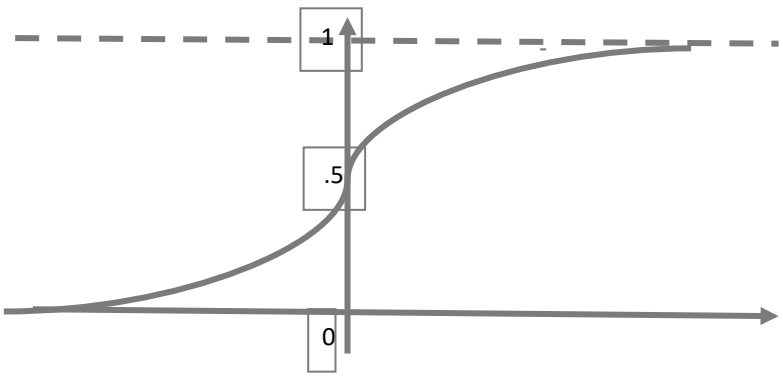


Fig. 1 Logistic Regression (sigmoid function) of the graph

The system is created with Jupiter notebook and python code. The system is constructed using the sci-kit learn python module. The following are the standard parameters settings for the MLP function:

Algorithm of logistic regression

```
1. from statsmodels. tools import odd constant as odd constant
2. heart_df_constant = odd constant(heart_df)
3. heart_df_constant.head()
4. st. chisqprob = lambda chisq, df: st.chi2.sf(chisq, df)
5.   st.chisqprob = lambda chisq, df: st.chi2.sf(chisq, df)
6.  st.chisqprob = lambda chisq, df: st.chi2.sf(chisq, df)
7.   cols=heart_df_constant.columns[:-1]
8.  model=sm.Logit(heart_df.target,heart_df_constant[cols])
9.  result=model.fit()
10.  result.summary()
```

Visit the sci-kit learn library [10] for further information on the variables. To do logistic regression, different python libraries such as SciPy, NumPy, and Panda are employed.

4. Results

4.1 Exploratory Data Analysis (Cleveland data set)

Kaggle heart disease at UC Irvine Dataset from Cleveland Figures 2 and 3 show data analysis to see how characteristics connect to the outcome, feature encoding, model fitting, and the resulting histogram, as well as the frequency of heart illness by age group and gender, and the greatest or lowest number of heart patient diseases or no heart patient diseases. Out of 303 participants in our data research, 165 people (54.5%) had heart illness, while 138 people (45.56%) had no such abnormalities, according to machine learning algorithms.

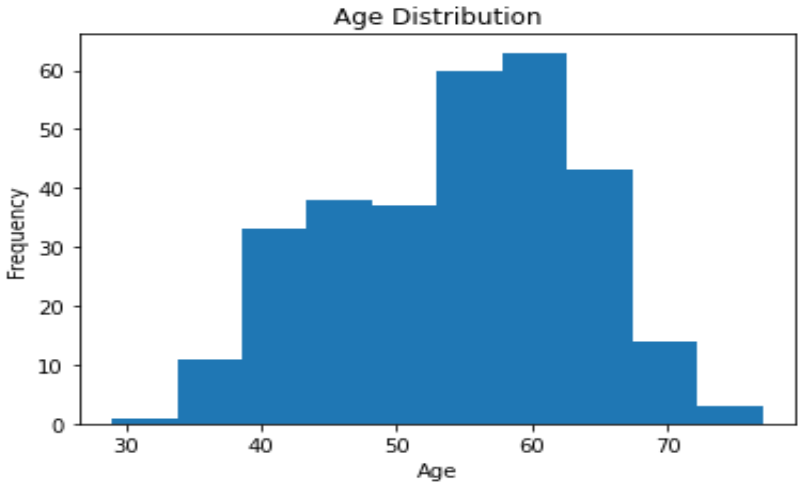


Fig. 2 Histogram frequency of heart disease by age group

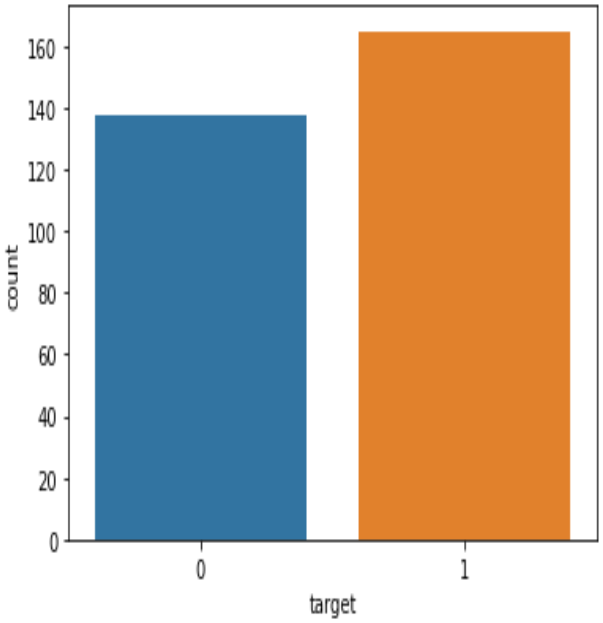


Fig. 3 Heart disease (target) histogram

4.2 Experimental result with logistic regression model (Cleveland data set)

Logistic regression is a sort of statistical regression analysis that uses a collection of classifier or relationship between the independent variable to predict the result of a categorical dependent variable. The dependant variable in logistic regression is always binary. The major applications of logistic regression are predictions and assessing the probability of success.

Table-1 statistically significant relationship with the probability of heart disease.

Features	**Coeff.**	**Std. err.**	z	Z >|p|	[0.026	0.975]
const	3.4505	2.571	1.342	0.180	-1.590	8.490
age	-0.0049	0.023	-0.212	0.832	-0.050	0.041
sex_male	-1.7582	0.469	-3.751	0.000	-2.677	-0.839
cp	0.8599	0.185	4.638	0.000	0.496	1.223
trtbps	-0.0195	0.010	-1.884	0.060	-0.040	0.001
chol	-0.0046	0.004	-1.224	0.221	-0.012	0.003
fbs	0.0349	0.529	0.066	0.947	-1.003	1.073
restecg	0.4663	0.348	1.339	0.181	-0.216	1.149
thalach	0.0232	0.010	2.219	0.026	0.003	0.044
exang	-0.9800	0.410	-2.391	0.017	-1.783	-0.177
oldpeak	-0.5403	0.214	-2.526	0.012	-0.959	-0.121
slp	0.5793	0.350	1.656	0.098	-0.106	1.265
ca	-0.7733	0.191	-4.051	0.000	-1.147	-0.399
thal	-0.9004	0.290	-3.104	0.002	-1.469	-0.332

The results above show some of the attributes with P value higher than the preferred alpha (5%) and thereby showing low statistically significant relationship with the probability of heart disease. Backward elimination approach is used here to remove those attributes with highest P value one at a time followed by running the regression repeatedly until all attributes have P Values less than 0.05.

Table-2: Feature Selection: Backward elimination (P-value approach)

features	**Coeff.**	**Std. err.**	z	Z >|p|	[0.025	0.0975]
sex_male	-1.3898	0.405	-3.431	0.001	-2.184	-0.596
cp	0.7861	0.174	4.509	0.000	0.444	1.128
thalach	0.0261	0.004	5.905	0.000	0.017	0.035
exang	-1.0130	0.376	-2.695	0.007	-1.750	-0.276
oldpeak	-0.7262	0.176	-4.130	0.000	-1.071	-0.382
ca	-0.7053	0.173	-4.087	0.000	-1.043	-0.367
thal	-0.8674	0.259	-3.351	0.001	-1.375	-0.360

The probabilities of being identified with heart disease for men (sex male = 1) over females (sex male = 0) are exp(0.469) = 0.24912 in this fitted model, holding all other characteristics constant. We may state that the odds for men are 83.9 percent greater than the odds for females in terms of percent change (use table-1&3).

Table-3: Interpreting the results: Odds Ratio, Confidence Intervals and P values				
featurs	CI 95%(2.5%)	CI 95 %(97.5%)	Odds Ratio	P value
sex_male	0.112623	0.551073	0.249126	0.001
cp	1.559575	3.088655	2.194764	0.000
thalach	1.017567	1.035326	1.026408	0.000
exang	0.173839	0.758508	0.363123	0.007
oldpeak	0.342750	0.682775	0.483757	0.000
ca	0.352232	0.692750	0.493973	0.000
thal	0.252918	0.697612	0.420046	0.001

Predicted Probabilities with a default classification threshold of 0.5, the test data were classified as 0 (heart disease: No) and1 (heart disease: Yes).

Table-4 Prob of no heart disease (0) & Prob of Heart Disease (1)		
	Prob of no heart disease (0)	**Prob of Heart Disease (1)**
0	0.075252	0.924748
1	0.971773	0.028227
2	0.990804	0.009196
3	0.464527	0.535473
4	0.220590	0.779410

Lower Boundary We may deduce from the confusion matrix that a high number of False Negatives (FN) (Type II mistake) is potentially harmful since it includes dismissing the risk of disease when only one OR True option is available. As a consequence, the sensitivity may be improved by lowering the threshold.

Table-5: confusion matrix , I-Type & II-Type error ,sensitivity, specificity				
Confusion matrix	**I-Type error(correct)**	**II-Type error(incorrect)**	**sensitivity**	**Specificity**
[[17 13] [1 30]]	47	14	96.77%	56.67
[[22 8] [2 29]]	51	10	93.54%	73.34
[[23 7] [2 29]]	52	9	93.54%	76.66

Table-6: Comparison of the Model from previous studies for Cleave land Dataset				
[[23 7] [2 29]]	52	9	93.54%	76.66
[[25 5] [2 29]]	54	7	93.54%	83.33

The size of the area under the ROC curve reflects how accurate the model is at recognising members of the train set; the larger the area, the greater the gap between true and false positives, and the better the model is at classifying individuals in the training data. Because a model with an area of 0.5 performs no better than random classification, a competent classifier strives to avoid it as much as possible. A one-square-foot surface area is ideal. The AUC should be as near to 1 as possible.

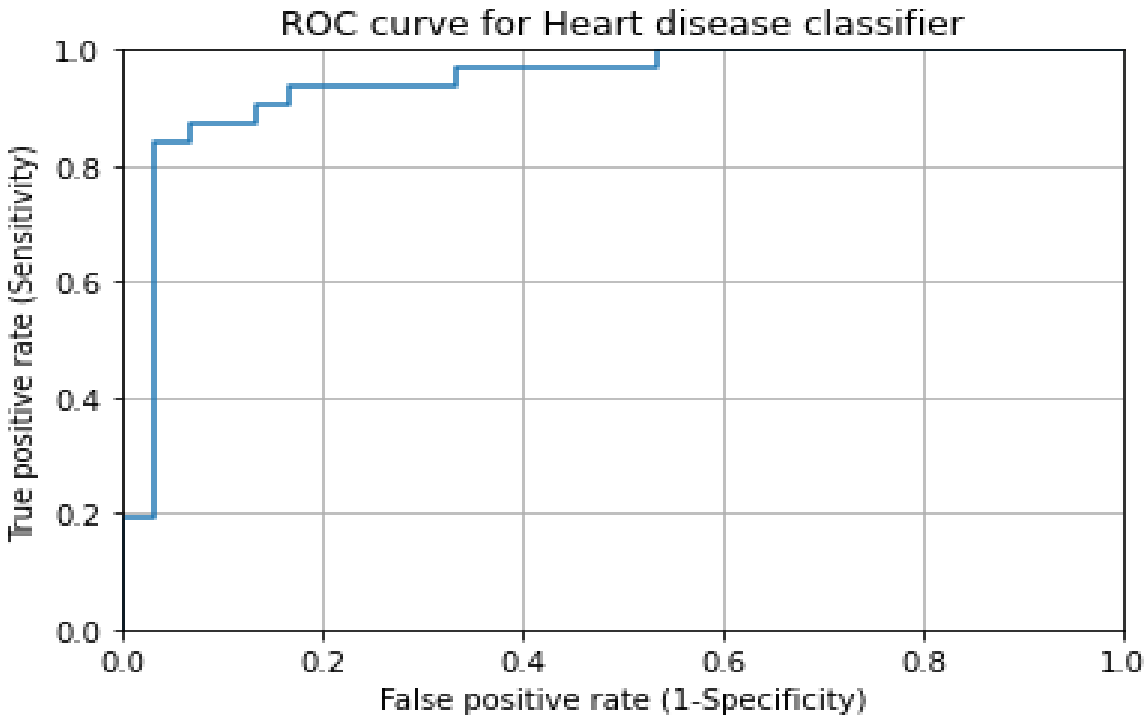


Figure 4. ROC curve for heart disease classifier

Authors	Methods	Results
Long et al [11]	CFARS-AR	88.3%
Chadha and Mayank [12]	DT, NB	88.03%, 85.86%
Soni et al [13]	Association rules	81.51%
Kumari and Godara [14]	SVM	84.15%
Leema et al [15]	DE + BP	86.6%
Amin et al [16]	Hybrid (NB + LR)	87.41%
A.K. Dwivedi [17]	SVM, LR	82.00% ,85.00%
Saqlain et al [18]	MFSFSA + SVM	81.19%
Latha and Jeeva [19]	Naïve Bayes + BN + Random Forest + MLP	85.48%
Mohan et al [20]	RF + Linear Model	88.4%
Ayon et al [21]	RF	87.45%
Proposed model	Logistic Regression model	**88.52**%
	Sensitivity	87.09%
	Specificity	**90%**
	Roc-Auc	**93.52%**

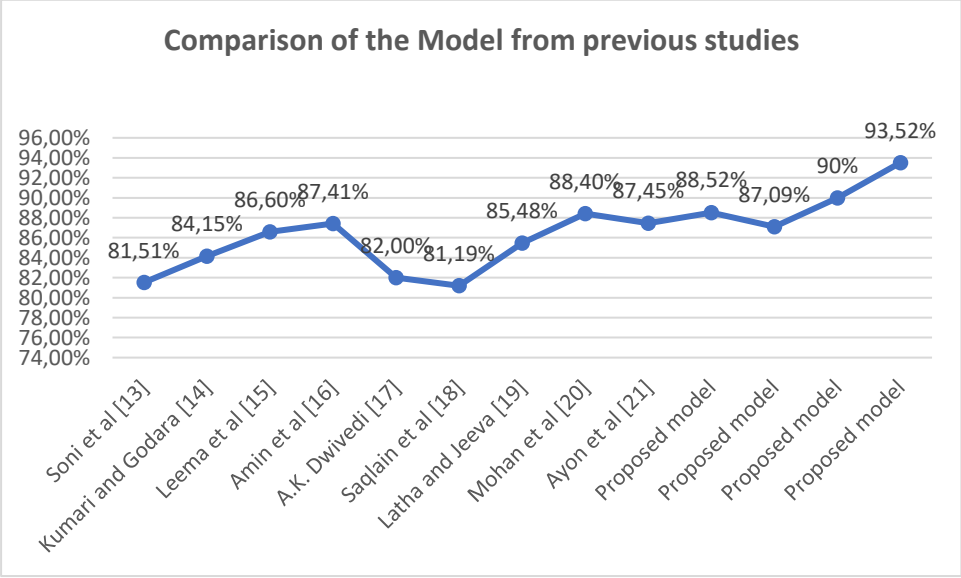


Figure 5. Comparison of the Model from previous studies

We also compared our technique to various previously published methods offered by other researchers. For example, Mohan et al [20] employed a combination of the RF and Linear Model to get a high-accuracy classification, Table-8 shows the exact comparative findings.

5. Conclusion

Identifying heart disease is one of the most challenging issues in the health-care industry. Males appear to have a higher rate of heart disease than females. The number of cigarettes smoked each day, as well as systolic and diastolic blood pressure, all contribute to an increased risk of heart disease. As a consequence, we propose developing an app that can predict the risk of heart disease based on basic symptoms like age, sex, pulse rate, and other factors. The suggested approach employs the machine learning technique of logistic regression, which has been shown to be the most accurate and dependable. Using datasets that are publicly available like the CHD, the model's performance is evaluated, with logistic regression attaining the highest accuracy of 89.52 percent. ROC_ AUC has an accuracy of 93.54 percent. In this fitted model, the odds of being diagnosed with heart disease are exp (0.469) = 0.24912 for men (sex male = 1) and exp (0.469) = 0.24912 for females (sex male = 0). In terms of percent change, the odds for men are 83.9 percent greater than the odds for females (use table-1,3 & table-7,8), and all attributes selected after the elimination process have P-values less than 5%, implying that the attributes chosen have a significant role in the prediction of heart disease. We describe a predictive analytics-based technique for detecting heart disease in this research.

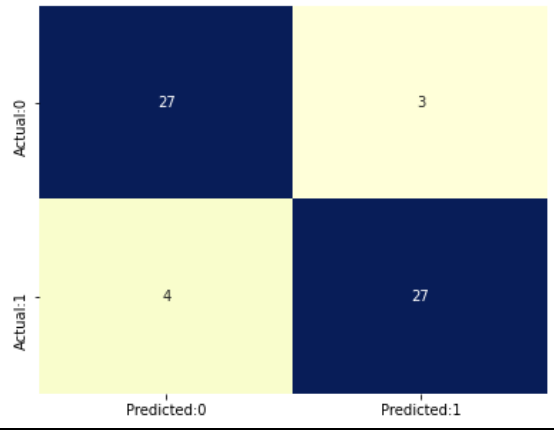
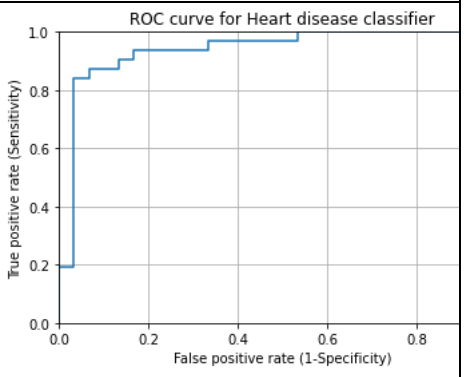
Table-7

	Accuracy
Logistic Regression	**89.52%**
The Miss classification	11.475%
Sensitivity	87.09%
Specificity	**90%**
Positive Predictive value	**90%**
Roc-Auc	**93.54%**
Negative Predictive Value	87.09%
Positive Likelihood Ratio	8.70%

Negative Likelihood Ratio	14.03%

Figure 6. result of logistic regression model

Table-8: For Confusion Matrix and ROC_Curve(93.54%)

CONFUSION MATRIX	ROC_CURVE(93.54%)
	

6. Future Work

Heart disease is predicted using logistic regression The World Health Organization (WHO) claims that, heart attacks account for four out of every five fatalities caused by cardiovascular diseases (CVD). In the future, the work could be improved by creating an internet application that supports logistic regression, as well as using a larger dataset than the one used in this analysis, which would help to provide better results and aid health professionals in effectively and efficiently predicting gut disease [22-23].

References

1. Prerana T H M, Shivaprakash N C et al "Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS",Vol 3, PP: 90-99 ©IJSE, 2015.
2. S. Ismaeel, M .Ali et al "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis", IEEE Canada International Humanitarian Technology Conference, DOI:10.1109/IHTC.2015.7238043, 03 September 2015.
3. ScikitLearn,'MLPCLassifier',[Online]Available:http://scikitlearn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.
4. Prediction System for heart disease using Naïve Bayes *Shadab Adam Batakari and Asma Parveen Department of Computer Science and Engineering Khaja Banda Nawaz College of Engineering.
5. A. Kor, [Online]Available:https://www.alivecor.com/how-it-works.
6. Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Online: 25 March 2017 DOI: 10.1007/s10462-01
7. X. Zou, Hu, Y., Tian, Z., & Shen, K. (2019, October). "Logistic regression model optimization and case analysis," In 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT) (pp. 135-139), (2019, October). IEEE.
8. T.Minka. (2001). Algorithms for maximum-likelihood logistic regression. Statistics Tech Report, 758.
9. S.Learn,'MLPCLassifier',[Online]Available:http://scikitlearn.org/stable/modules/generated/sklearn.neural_network.MLPClassifi er.html.
10. N.C. Long, P. Meesad, H. Unger, A highly accurate firefly-based algorithm for heart disease prediction, Expert. Syst. Appl. 42 (2015) 8221–8231.
11. R. Chadha, Sudhakar Mayank, Prediction of heart disease using data mining techniques, CSI Trans. ICT 4 (2-4) (2016) 193–198.
12. J. Soni, U. Ansari, D. Sharma, S. Soni, Intelligent andeffective heart disease prediction system using weightedassociative classifiers, Int. J. Comput. Sci. Eng. 3 (6) (2011) 2385–2392.
13. M. Kumari, S. Godara, Comparative study of datamining classification methods in cardiovascular disease prediction, IJCST 2 (2) (2011) 304–308.
14. N. Leema, H. Khanna Nehemiah, A. Kannan, Neural network classifier optimization using differential evolution with global information and back propagation algorithm for clinical datasets, Appl. Soft. Comput. 49 (2016) 834–844.
15. M. S. Amin, Y. K. Chiam, and K. D. Varathan, ''Identication of signicant features and data mining techniques in predicting heart disease,'' Telematics Inform., vol. 36, pp. 82_93, Mar. 2019, doi: 10.1016/j.tele.2018.11.007.

16. Ashok Kumar Dwivedi, Performance evaluation of differentmachine learning techniques for prediction of heartdisease, Neural Comput. Appl. 29 (10) pp. 685–693, (2018).
17. S. M. Saqlain, M. Sher, F. A. Shah, I. Khan, M. U. Ashraf, M. Awais, and A. Ghani, ‘‘Fisher score and Matthew’s correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines,’’ Knowl. Inf. Syst., 58(1), pp. 139_167, Jan. 2019, doi:10.1007/s10115-018-1185-y.
18. C. B. C. Latha and S. C. Jeeva (2019), ‘‘Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques,’’ Inform. Med. Unlocked, 16, Jan. 2019, Art. no. 100203, doi: 10.1016/j.imu.2019.100203.
19. Senthil Kumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access 7 (2019) 81542–81554, https://doi.org/10.1109/ACCESS.2019.2923707.
20. S. Islam Ayon, Md. Milon Islam, Md. Rahat Hossain (2020), “Coronary Artery heart disease prediction: a comparative study of computational intelligence techniques,” IETE J. https://doi.org/10.1080/03772063.2020.1713916
21. M. N. R. Chowdhury, E. Ahmed, M. A. D. Siddik and A. U. Zaman, (2021), "Heart Disease Prognosis Using Machine Learning Classification Techniques," 6th International Conference for Convergence in Technology (I2CT), pp. 1-6, doi: 10.1109/I2CT51068.2021.9418181
22. G. N. Ahmad, Shafiullah, A. Algethami, H. Fatima, and S.M.H.Akhter (2022), “Comparative study of Optimum Medical Diagnosis of Human Heart Disease using Machine Learning Technique with and without Sequential Feature Selection,” IEEE Access,doi:10.1109/ACCESS.2022.3153047