A
REPORT

ON
"HEART DISEASE PREDICTION"

OF
T.E.(AI & DS Engineering)
(Academic Year: 2022-2023)

SUBMITTED BY
Isha Tadas
Amisha Sonone
Radhesh Khaire
Roll No :- T1411072
T1411071
T1411033

GUIDED BY
Prof.D.D.Sarpate

Department of AI & DS Engineering
Zeal Education Society's
Zeal College of Engineering & Research
Narhe, Pune-411041

A REPORT

ON

"HEART DISEASE PREDICTION"

Submitted By

Isha Tadas
Amisha Sonone
Radhesh Khaire

in partial fulfilment for the award of the degree of

Bachelor of Engineering
of
Savitribai Phule Pune University

IN

AI & DS Engineering

Zeal College Of Engineering and Research,Narhe, Pune
2022 - 2023

Zeal Education Society's

Zeal College of Engineering & Research
Department of AI & DS Engineering

CERTIFICATE

This is to certify that seminar entitled

"HEART DISEASE PREDICTION"

have successfully completed by **"Isha, Amisha, Radhesh"** of TE (AI & DS Branch) in the academic year 2022-2023 in partial fulfillment of the third Year of Bachelor degree in "AI & DS Engineering" as prescribed by the Savitribai Phule Pune University.

Prof.D.D.Sarpate	**Prof. S. A. Ubale**	**Dr.A. M. Kate**
Seminar Guide	HOD	Principal

Place: ZCOER, Pune.
Date: / /2022

Acknowledgement

We take this opportunity to thank our seminar guide **Prof. D.D.Sarpate** and Head of the Department **Prof. S. A. Ubale** for their valuable guidance and for providing all the necessary facilities, which were indispensable in the completion of this project report. We are also thankful to all the staff members of AI & DS the Department for their valuable time, support, comments, suggestions and persuasion. We would also like to thank the institute for providing the required facilities, Internet access and important books.

Abstract

This paper predicts the risk of suffering from heart disease among the elderly by exploring the feasibility of using logistic regression models. Through the technology of data mining, the main pathogenic factors of heart disease were found, and the incidence of heart disease was predicted by using the regression model. The accuracy of logistic regression model was compared with other explored algorithms, and I found that the logistic regression model was worthy of research in the field of heart disease prediction.

Keyword - Logistic regression, predict, heart disease

Contents

1 Introduction

The forecast of cardiovascular disease, one of the most common heart diseases, is considered to be one of the most significant topics in the analysis of clinical data. According to the World Health Organization (WHO), Cardiovascular diseases (CVDs) kills about 31 percent of the world's population each year, with older people at greater risk than other age groups. Through applying the technology of data mining, a new idea is provided for the prediction of heart disease, extracting clinical attributes and pathological data from large medical data sets, and generating biological hypotheses. At present, some studies have applied data mining technology to the prediction of heart disease, but there are limited studies on the important features of cardiovascular disease, while logistic regression can extract the risk factors of disease and predict the incidence probability of patients in real time. This study aims to determine the important characteristics and incidence probability of heart disease prediction, and compare the accuracy of the logistic regression algorithm used with other existing research algorithms, such as Naive Bayes, SVM and Neural Network, to determine the feasibility of the logistic regression algorithm in predicting heart disease.

2 Algorithm

Logistic regression model, a very common model in machine learning is selected by this paper, which is often applied in the actual manufacturing context the fields such as data mining, automatic disease diagnosis and economic prediction. For instance, this research discussed the risk factors for heart disease and forecast the probability of disease occurrence based on risk factors. Logistic regression is most frequently applied for classification, primarily two-category issues (that is, there are only two types of output, each representing one category), and can indicate the probability of occurrence of each classification event.

3 Implementation

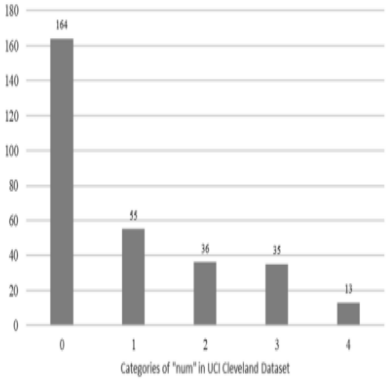
3.1 *Dataset*

The data from UCI machine learning repository was collected. The dataset contains 303 records, and 14 attributes. Thirteen parameters were used as the eigenvalues for the forecast of heart disease and one of the parameters is the output value or the forecast value of the patients with heart disease. ('num' means Numeric, and 'nom' means Nominal) 0 refers to no heart disease and 1-4stands for the number of patients with heart diseases, different numbers represent different degrees of illness (4 being the highest).

Table 1. Attributes from UCI Dataset.

Attribute	Description	Type
① Age	Age in years	num
② Sex	Sex (1 = male; 0 = female)	nom
③ Cp	chest pain type -- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic	nom
④ Trestbps	Resting blood pressure	num
⑤ Chol	Serum cholestoral in mg/dl	num
⑥ Fbs	Fasting blood sugar > 120 mg/dl	nom
⑦ Restecg	Resting electrocardiographic results -- Value 0: normal -- Value 1: having ST-T wave abnormality -- Value 2: showing probable or definite leftventricular hypertrophy by Estes' criteria	nom
⑧ Thalach	Maximum heart rate achieved	num
⑨ Exang	Exercise induced angina	nom
⑩ Oldpeak	ST depression induced by exercise relative to rest	num
11 Slope	The slope of the peak exercise ST segment Nominal -- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping	nom

12 Ca	Number of major vessels (0–3) coloured by fluoroscopy	num
13 Thal	3 = normal; 6 = fixed defect; 7 = reversable defect	nom
14 Num	Diagnosis of heart disease (angiographic disease status) -- Value 0: no heart disease -- Value 1-4: presence of heart disease	nom



3.2 *Data Analysis*

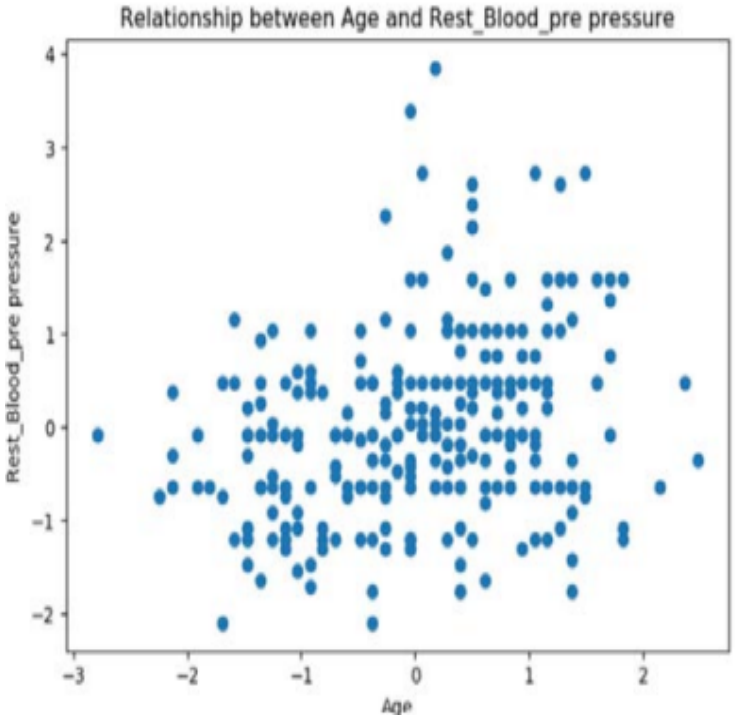
3.2.1 *Data Processing*

Since there are incomplete data in the data set and the output value is from 0 to 4 with different degrees of disease, while logistic regression corresponds to dichotomy, it is necessary to preprocess the data for subsequent analysis.The incomplete data were removed and the predictive value attributes for suffering from heart diseases in the dataset were converted from multicategory value to binary value.Diagnostic values are converted from 2 and 4 to 1.The final data set only includes 0 and 1; 0 stands for no heart disease, and 1 refers to the possibility of suffering from heart disease.

3.2.2 *Feature Selection*

In addition to predicting the probability of heart disease, the experiment also needs to dig the main factors of heart diseases.Therefore, characteristics should be extracted before data analysis, so as to find out the related pathogenic factors affecting heart problems and put forward suggestions for the prevention of physical health based on this.Figure 2 shows the relationship between age and blood pressure, and Figure 3 shows the correlation between each characteristic value.

According to the correlation between the eigenvalues, the combination with the most significant features was selected for data analysis, and different data mining techniques were used to test the selected combination.



3.2.3 *Classification modeling and performance testing*

After complete the feature selection, create logic regression model, then input data and initial parameters, the model of new parameters are calculated through gradient descent, the probability is more and more close to the real value, also is a process to make the error smaller and smaller, until it converges to a tend to a fixed value, it is concluded that eventually a set of parameters, according to the parameters into the model, and then verified with test data, if the error is not satisfied, you will need to adjust the initial parameters or data, to continue the above process, until a satisfactory parameters and test data error. After training, the final set of parameters is obtained, and the parameters are inserted into the model formula, and then the error is calculated with the test data. The smaller the error is, the closer the parameters are to the optimal one.If you are not satisfied with the obtained parameters, you need to adjust the super parameters and train again until you get a satisfactory set of parameters.The prediction accuracy of the final model is 84.98 percent .

3.2.4 *Contrastive Analysis*

Other data mining algorithms show different performance in building heart disease prediction models. The following table shows the accuracy rate and main influencing factors of different technologies.

After comparing logistic regression model with other models, it can be found that the accuracy of logistic regression model is more accurate than KNN and Neural Network algorithm in predicting heart disease probability, but SVM may be a better way to predict heart disease probability.This is mainly because logistic regression is very sensitive to extreme values, which can easily lead to problems of underfitting and low accuracy, while the generalization error rate of support vector machines is lower.

Technique	Accuracy	Combination
Support Vector Machine	86.87%	1, 2, 3, 5, 6, 9, 10, 11, 12
Naïve Bayes	85.86%	2, 3, 8, 9, 10, 12
Neural Network	84.85%	2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13
KNN	82.49%	2, 3, 6, 7, 10, 11, 12, 13
Logistic Regression	85.86%	1, 2, 3, 5, 6, 9, 10, 11, 12

4 Conclusion

Raw data is analyzed by using the technology of data mining, and new insights and accurate predictions are provided for disease prevention goals. In this paper, logistic regression models were used to explore the feasibility of predicting heart disease. Experiments were conducted using the data set provided by UCI and the results were evaluated.Significant features of logistic regression models affecting heart disease were found: Age,sex, cp, chol, restecg, oldpeak, slope, ca, thal. The main influencing factors and logistic regression technology are used to establish the prediction model, and the accuracy of the model is compared with the model proposed in the existing research. According to the test results, the classification model proposed is highly accurate and has certain research value.In the further study, new feature extraction methods and model parameters can be selected to further improve the accuracy.

5 References

- A. Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 704-706.

- A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.

- N. Mohan, V. Jain and G. Agrawal, "Heart Disease Prediction Using Supervised Machine Learning Algorithms," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 2021, pp. 1-3, doi: 10.1109/ISCON52037.2021.9702314.

- A. Chauhan, A. Jain, P. Sharma and V. Deep, "Heart Disease Prediction using Evolutionary Rule Learning," 2018 4th International Conference on Computational Intelligence Communication Technology (CICT), 2018, pp. 1-4, doi: 10.1109/CIACT.2018.8480271.