Heart Disease Prediction Based on Age Detection using Novel Logistic Regression over Decision Tree

C.B.M Karthi, A. Kalaivani

Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode: 602105.

ABSTRACT

Aim: To improve the accuracy in Heart Disease Prediction using Novel Logistic Regression and Decision tree

Materials and Methods: This study contains 2 groups i.e Novel Logistic Regression and Decision tree Each group consists of a sample size of 10 and the study parameters include alpha value 0.01, beta value 0.2, and the Gpower value 0.8. Results: The Novel Logistic Regression (91.60) achieved improved accuracy than the Decision Tree (89.42) in Heart Disease Prediction. The statistical significance difference is 0.01 (p<0.05). Conclusion: The Novel Logistic Regression model is significantly better than the Decision Tree in Heart Disease Prediction. It can be also considered as a better option for Heart Disease Prediction.

Keywords

Novel Logistic Regression, Decision Tree, Heart disease prediction, Accuracy, Machine Learning.

Imprint


INTRODUCTION

Heart Complaint describes a range of conditions that affect your heart. Heart conditions include Blood vessel complaints, similar to coronary roadway complaints (Lee et al. 2020)). Heart-meter problems (arrhythmias) Heart blights you are born with (natural heart blights) Coronary roadway complaints, arrhythmia, heart stopcock complaints, and heart failure are the four most common types of heart complaints (Gondkar 2015). The major challenge that the Healthcare assiduity faces currently is the superiority of installations (Shah et al. 2021). Diagnosing the complaint rightly & furnishing effective treatment to cases will define the quality of service. Poor opinion causes disastrous consequences that aren’t accepted (Mohamadlou et al. 2018). A heart complaint is a kind of complaint that affects the functioning of the heart. In a moment, heart complaints are the primary reason for death. WHO-World Health. The association has anticipated that 12 million people die every time because of heart conditions Habitual Disease Prediction plays a vital part in healthcare informatics (Sung, Hung, and Hu 2021). It’s pivotal to diagnose the complaint at an early stage. This paper presents a check on the application of point selection and bracket ways for the opinion and vaccination of habitual conditions (Baig 2020). The system uses 15 medical parameters similar to age, coitus, blood pressure, cholesterol and rotundity for the vaccination.
The research gap in Heart Disease Prediction is the availability of real-time data sets is limited and the accuracy is to be improved (Lee et al. 2020). The selection of the algorithm also plays a vital role in Heart Disease Prediction. The studied plant that was largely sensitive to troponin tests was especially good at prognosticating cardiovascular events when added to the results of a special equation generally used to calculate a person's 10- time threat of having a heart attack or stroke. So this research focuses on improved accuracy in Heart Disease Prediction Using Novel Logistic Regression over a Decision tree.

MATERIALS AND METHODS

This work is carried out at Saveetha School of Engineering, Department of Information Technology in the Data Analytics Lab. The study consists of two sample groups Logistic Regression and Decision Tree. Each group consists of 10 samples with a pre-test power of 0.18. The sample size kept the threshold at 0.05, G power of 80%, confidence interval at 95%, and enrolment ratio at 1.

Data Preparation

To perform Heart Disease Prediction, the real-time data sets are used which are collected from the heart disease patients. The input data sets used as heart.csv collected from The data sets consist of six attributes and 500 instances.

To perform Heart Disease Prediction the real-time data sets used are heart data. The input data sets for the proposed work in heart data.csv collected from GitHub.com (“Git Hub: Your Machine Learning and Data Science Community”). The data sets consist of the attributes are age, chest pain type, and resting blood pressure are dependent attributes, and fasting blood sugar, resting electrocardiographic results, and maximum heart rate achieved are independent attributes that do not affect the results removed from the csv file.

NOVEL LOGISTIC REGRESSION

Novel Logistic Regression analysis is used to predict the value of a variable predicated on the value of another variable. The value of the variable you want to predict is known as the dependent variable. The novel Novel Logistic Regression algorithm is represented in the graphs showing the difference in the value between the attributes. From the training data, we have to estimate the swish and approximate measure and represent it.

Medical lives of conditions analogous as commerce, age, hypertension, blood sugar, and other symptoms are used for vaccination. The model is designed to predict the possibility of cases of heart complaints. Novel Logistic Regression analysis is used to predict the value of a variable predicated on the value of another variable. The variable you are using to predict the value of another variable's value is called the independent variable. Novel Logistic Regression is used to predict the class (or order) of individualities predicated on one or multiple predictor variables (x). It's used to model a double outgrowth, a variable that can have only two possible values 0 or 1, yes or no, diseased or non-diseased.

The regression portions are generally estimated using maximum liability estimation. The maximum liability rate helps to determine the statistical significance of independent variables on the dependent variables. The liability-rate test assesses the donation of individual predictors (independent variables). Also, the probability (p) of each case is calculated using the odds rate. Is Calculated using equation 1.

\[
P/(1-P) = e^Y - eq. \tag{1}\]

From this p-value is found. This gives the probability or chance for the individual to have coronary heart disease. Pseudocode and Accuracy Values for the regression model has been mentioned in Table 1 and Table 3.

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pseudocode for Novel Logistic Regression</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input: Heart Symptoms dataset records</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Import the required packages.</td>
</tr>
<tr>
<td>2. Convert the Data Sets into numerical values after the extraction feature.</td>
</tr>
<tr>
<td>3. Assign the data to X_train, y_train, X_test, and y_test variables.</td>
</tr>
<tr>
<td>4. Using the train_test_split() function, pass the training and testing variables.</td>
</tr>
<tr>
<td>5. Give test_size and the random_state as parameters for splitting the data using the Logistic training model.</td>
</tr>
<tr>
<td>6. Compiling the model using metrics as accuracy</td>
</tr>
<tr>
<td>7. Calculate the accuracy of the model.</td>
</tr>
</tbody>
</table>

**OUTPUT: Accuracy**
DECISION TREE

A decision tree is a flowchart-suchlike structure in which each internal knot represents a test on a trait (e.g. whether a coin flip comes up heads or tails), each branch represents the outgrowth of the test, and each splint knot represents a class marker (decision taken after calculating all attributes) as given in equation 2.

\[
\text{Entropy}(s) = \text{P(yes)} \log_2 \text{P(yes)} - \text{P(no)} \log_2 \text{P(no)} \quad (2)
\]

Where,
- \(S\) = Total number of samples
- \(\text{P(yes)}\) = probability of yes
- \(\text{P(no)}\) = probability of no

The Stoner precedes the processes by checking the specific detail and symptoms of the heart complaint. The decision tree (ID3) and naive Bayes ways in data mining are used to recoup the details associated with each case. Grounded on the accurate result vaticination, the performance of the system is anatomized. The Decision Tree algorithm belongs to the family of supervised literacy algorithms. The thing of using a Decision Tree is to produce a training model that can be used to prognosticate the class or value of the target variable by learning simple decision rules inferred from previous data (training data). Pseudocode and Accuracy Values for the regression model have mentioned in Table 2 and Table 4.

STATISTICAL ANALYSIS

The minimum requirement to run the software used here is Intel Core i3 Dual-Core CPU clocked @3.2 GHz, 4GB or above memory of RAM, more than 512MB space is required and Software specification includes Windows 7/8/10/11 Professional 64-bit OS, Jupyter Notebook Version 6.30 with Python3, and MS-Office. Statistical Package for the Social Sciences Version 26 software tool was used for statistical analysis. An independent sample T-test was conducted for accuracy. Standard deviation, and standard mean errors were also calculated using the SPSS Software tool. The significance values of proposed and existing algorithms contain group statistical values of proposed and existing algorithms.

RESULTS

The group’s statistical analysis of the two groups shows novel Novel Logistic Regression (group1) has more mean accuracy than novel Decision tree (group 2) and the standard error mean is slightly less than novel Novel Logistic Regression. The Novel Logistic Regression scored an accuracy of 91.79% and the novel Decision tree has scored 89.42%. The accuracies are
recorded by testing the algorithms with 10 different sample sizes and the average accuracy is calculated for each algorithm.

In SPSS, the datasets are prepared using 10 as the sample size for novel Novel Logistic Regression and novel Decision tree. Group is given as a grouping variable and Heart Disease is given as the testing variable. Group is given as 1 for Novel Logistic Regression and 2 for Decision tree. Descriptive Statistics is applied for the dataset in SPSS and shown in Table 5, Group statistics are shown in Table 6, and Two Independent Sample T-Tests in Table 7.

**Table 5**
Descriptive Statistic analysis, representing Novel Logistic Regression and Decision tree

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group1</td>
<td>20</td>
<td>1.00</td>
<td>2.00</td>
<td>1.500</td>
<td>0.51299</td>
</tr>
<tr>
<td>Accuracy</td>
<td>20</td>
<td>79.48</td>
<td>91.89</td>
<td>87.9025</td>
<td>4.43044</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>8.11</td>
<td>20.52</td>
<td>12.0975</td>
<td>4.43044</td>
</tr>
</tbody>
</table>

**DISCUSSION**

From the results of this study, novel Novel Logistic Regression is proved to be having better accuracy than the Decision tree model. Novel Logistic Regression has an accuracy of 91.60% whereas the novel Decision tree has an accuracy of 89.42%. The group statistical analysis on the two groups shows that Novel Logistic Regression appears to perform significantly better than Decision tree with the value of p=0.18.

**Table 6**
Group Statistic analysis, representing Novel Logistic Regression (mean accuracy 91.65%, standard deviation 0.08600, 0.09333) and Decision Tree (mean accuracy 91.59%, standard deviation 0.08600, 0.09333)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>10</td>
<td>91.6730</td>
<td>14622</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>10</td>
<td>84.1320</td>
<td>3.13442</td>
</tr>
<tr>
<td>Error</td>
<td>10</td>
<td>8.3270</td>
<td>14622</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>10</td>
<td>15.8680</td>
<td>13442</td>
</tr>
</tbody>
</table>

**Table 7**
Independent Sample Tests results with a confidence interval of 95% and a level of significance as 0.01 (Novel Logistic Regression appears to perform significantly better than Decision tree with the value of p=0.18).

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Levene's Test for Equality of Variances</th>
<th>T-test for Equality of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig.</td>
</tr>
<tr>
<td>Accuracy</td>
<td>19.449</td>
<td>0.01</td>
</tr>
<tr>
<td>Equal variances assumed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 1.** Comparison of Novel Logistic Regression and Decision tree in terms of accuracy. The mean accuracy of Novel Logistic Regression is greater than a novel Decision tree and the standard deviation is also slightly higher than a novel Decision tree. X-axis: Novel Logistic Regression vs Decision tree. Y-axis: Mean accuracy of detection + 1 SD.
Regression (group 1) has more mean accuracy than Decision tree (group 2) and the standard error mean including standard deviation mean is slightly less than Novel Logistic Regression.

Heart Disease Prediction using Machine learning is now becoming widely used as a methodology (Wang et al. 2021). Citizens who have employed machine learning algorithms to address problems based on their own industry data. Industry professionals have used machine learning to perform classification jobs and diagnose malfunctions (Narayan and Gobal 2018). People in the field of business frequently used machine learning algorithms in financial research. The paper focuses on the accuracy of A Khemphila, V Boonjing – 2010. evaluating housing prices in each of the 50 states using Classification Function Algorithms (CFA) and Long Short-Term Memory (LSTM), two different forecasting techniques developed and motivated by (Bavani et al. 2019). The strategies take into account all of the $K = 2^m$ distinct model combinations in each time period $t$ when there are $m$ predictors available The limitation of the proposed work is due to inconsistent data and difficulty in getting the right datasets for analysis (Sivarajanji, Naresh, and Murthy 2019; Purushottam et al. 2015). Future work can be concentrated on effective data preprocessing techniques and the usage of ensemble machine learning algorithms can be focused.

CONCLUSION

Based on the experimental results, Novel Logistic Regression (91.60%) has been proved to predict Heart Disease more significantly than a Novel Decision tree (89.42%) model mentioned in Fig.1. The quality of datasets formed with prediction value and accuracy is improved.

DECLARATIONS

Conflicts of Interest

No conflicts of interest in this manuscript.

Author Contributions

Author CBMK was involved in data collection, data analysis, data extraction, and manuscript writing. Author AK was involved in the conceptualization, data validation, and critical review of the manuscript.

Acknowledgment

The authors would like to express their gratitude to Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES


