# Logistic Regression to Predict Heart Disease

**Vardhini Gujrati**

*Student,Centre of Excellence,AIT Management,Chandigarh University*

**Mr Soham Joshi**

*Namo RIMS Jr College, Pune, hamjosh106@gmail.com*

**Sachi Nagpal**

*Student,Centre of Excellence,AIT Management,Chandigarh University*

**ABSTRACT:**
The early diagnosis of cardiovascular disorders can help high risk individuals make decisions about lifestyle adjustments, which can lessen their problems. Using homogeneous data mining approaches, research has sought to identify the most significant risk variables for heart disease as well as reliably estimate the total risk. Recent studies have looked into combining these methods utilizing strategies like hybrid data mining algorithms. In order to offer an accurate model for predicting heart disease, this study suggests a rule-based model to evaluate the accuracy of applying rules to the individual findings of logistic regression, decision trees, and support vector machines on the Cleveland Heart Disease Database.

***KEYWORDS***- Heart disease, support vector machine (SVM), logistic regression, decision trees, and rule based approach

## 1. INTRODUCTION

The study of cardiovascular illnesses utilizing data mining has been continuing and involves very accurate risk score analysis, prediction, and therapy. The Cleveland Heart Clinic data collection is the most well-known of the several CVD surveys that have been carried out. The de facto database for heart disease research has therefore been regarded as the Cleveland Heart Disease Database (CHDD). This study offers a framework to employ logistic regression, support vector machines, and decision trees to get individual predictions that are then used in rule-based algorithms, recommending the parameters from this database. On the basis of accuracy, sensitivity, and specificity, the output of each rule from this system is then contrasted. The technique attempts to achieve two objectives: the first is to primarily give a prediction framework for heart disease, and the second is to assess how effective it is to combine the results of many models as compared to utilizing only one.This study employs various data mining methods, including logistic regression, support vector machines, and decision trees, to develop individualized predictions for heart disease risk. These predictions are then incorporated into rule-based algorithms, and the parameters for these algorithms are derived from the CHDD.By contrasting the output of each rule from the rule-based algorithms and evaluating their performance in terms of accuracy, sensitivity, and specificity, this study aims to provide valuable insights into the effectiveness of these predictive models and the advantages of combining them. Ultimately, the goal is to improve the diagnosis and management of cardiovascular diseases through data-driven approaches, potentially leading to more effective and personalized treatment strategies.

## 2. LITERATURE REVIEW

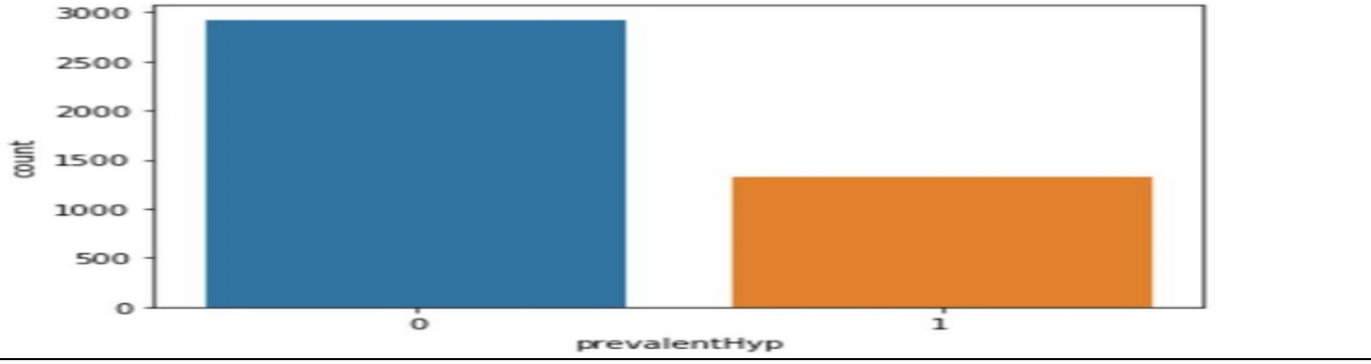
Over the past 20 years, there has been continuing research into utilizing data mining techniques to predict cardiac disease. On several patient databases from throughout the world, the majority of the publications have applied techniques like SVM, neural networks, regression, decision trees, nave Bayesian classifiers, etc.The choice of parameters that the approaches have been applied to is one of the basis on which the publications differ from one another. For assessing the accuracy, several writers have defined various metrics and datasets. A study by Xing et al. On 1000 patients revealed that SVM had an accuracy rate of 92.1%, artificial neural networks had a rate of 91.0%, and decision trees had an accuracy rate of 89.6% when TNF, IL6, IL8, HICRP, MPO1, TNI2, sex, age, tobacco, hypertension, diabetes, and survival were taken into account. The accuracy of SVM, neural networks, Bayesian classification, decision trees, and logistic regression were similarly compared by Chen et al. SVM had the greatest accuracy of 90.5% out of 102 examples, followed by neural networks with 88.9%, Bayesian with 82.2%, decision trees with 77.9%, and logistic regression with 73.9%.When comparing the accuracies of several data sets with various parameters, different results are obtained that do not give a fair foundation for comparison. Soni et al. recognized this and classified the most important factors as gender, smoking, being overweight, drinking alcohol, eating a diet heavy in salt and saturated fat, exercising, leading a sedentary lifestyle, genetic factors, blood pressure, cholesterol, fasting blood sugar, and heart rate. More recently, Shouman et al. Highlighted age, blood pressure, cholesterol, smoking, total cholesterol, diabetes, hypertension, heredity, obesity, and lack of physical exercise as the statistically determined risk factors. The Cleveland Heart Disease Database was described in the same study as the recognized standard database for heart disease research.

## 3. METHODOLOGY

Since the World Health Organization's founding, 12 million fatalities per year globally have been attributed to heart disease. In the United States and other affluent nations, cardiovascular illnesses account for fifty percent of fatalities. High-risk individuals can make decisions regarding lifestyle adjustments that will lower problems with advanced cardiovascular disease prognosis. Using data preparation logistic regression, this study tries to categorise the most important risk variables for heart disease as well as forecast the total risk. A variety of predictors or independent variables are employed in the statistical technique of logistic regression to forecast the result of a categorical dependent variable. The dependant variable in logistic regression is always binary. The major applications of logistic regression are prediction and calculation.The dataset used to create comes from ongoing cardiovascular research of people living in the Massachusetts town of Framingham. Determining if the patient has a 10-year risk of developing coronary heart disease (CHD) is the categorization objective.

## 4. RESULTS

```
sns.countplot(x="prevalentHyp", data=heart_data)
<matplotlib.axes._subplots.AxesSubplot at 0x7efd59adec50>
```



It is evident from the aforementioned statistics that the model is far more specific than sensitive. Compared to women, men appear to be more prone to heart disease. Age, daily cigarette consumption, and systolic blood pressure all indicate an increased risk of developing heart disease. The chances of CHD do not significantly alter with total cholesterol. The presence of HDL in the total cholesterol value may be the cause of this. Additionally, glucose only slightly affects chances (0.2%). The model foresaw using 0.87 accuracies. The model is more sensitive than it is specific. More data and the use of more machine learning models might enhance the overall model.

```
Accuracy Score : 0.8702830188679245
Precision Score : 0.8
Recall Score : 0.10084033613445378
F1 Score : 0.17910447761194032
```

## 5. CONCLUSION

To enhance the accuracy of predicting the early onset of cardiovascular illnesses, it is evident from the literature review that more advanced and combined models are required. This research proposes a system that integrates three prominent machine learning techniques: support vector machines, logistic regression, and decision trees. The objective is to create a precise forecasting tool for cardiac diseases by harnessing the strengths of these algorithms. This study outlines a methodology for training and testing the system, ultimately identifying the most effective model among various rule-based combinations. The research leverages the Cleveland Heart Disease database for this purpose and aims to compare the outcomes of these models, including sensitivity, specificity, and accuracy. Additionally, the study seeks to determine which model carries the highest efficiency and importance in making accurate predictions.

## REFERENCES

1. Avinash Galande, Pavan Kumar T. Heart disease prediction using effective machine learning techniques.
2. The Lancet Global Health. The changing patterns of cardiovascular diseases and their risk factors in the states of India: The global burden of disease study 1990-2016.
3. Himanshu Sharma, M A Rizvi. Prediction of heart disease using machine learning algorithms: A survey.
4. World health ranking.
5. Himanshu Sharma, M A Rizvi. Prediction of heart disease using machine learning algorithms: A survey.
6. Sana Bharti,2015. Analytical study of heart disease prediction compared with different algorithms; International conference on computing, communication, and automation (ICCA2015).
7. Monika Gandhi,2015. Prediction in heart disease using techniques of data mining, International conference on futuristic trend in computational analysis and knowledge management (ABLAZE- 2015)
8. SarathBabu, 2017.Heart disease diagnosis using data mining technique, international conference on electronics, communication and aerospace technology (ICECA2017)
9. A H Chen, 2011. HDPS: heart disease prediction system; 2011 computing in cardiology