VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Jth Component Project Report – I
ITA5007 - Data Mining and Business Intelligence
WINTER Semester 2022-23

TITLE

HEART DISEASE PREDICTION USING ML

Submitted by-

22MCA0137- KANISHK SHARMA

22MCA0174 - UDDASHAY KUMAR GUPTA

Submitted to-

PROF. EPHZIBAH E.P

Abstract

Machine Learning has found widespread applications in various industries worldwide, including the healthcare sector. With its potential to analyze large datasets and make accurate predictions, Machine Learning can be utilized in predicting the presence or absence of locomotor disorders, heart diseases, and other medical conditions. Early prediction of such diseases can provide valuable insights to healthcare professionals, enabling them to tailor their diagnoses and treatments on a per-patient basis. In this research project, our focus is on predicting possible heart diseases using Machine Learning algorithms. We conduct a comparative analysis of several popular classifiers, including decision tree, Naïve Bayes, logistic regression, support vector machine (SVM), and random forest. We also propose an ensemble classifier that combines strong and weak classifiers to improve accuracy and predictive performance. The proposed ensemble classifier can be trained and validated on multiple samples of data, which enhances its ability to make accurate predictions. Furthermore, we perform an in-depth analysis of existing classifiers, such as Ada-Boost and XG-Boost, to identify their strengths and weaknesses in predicting heart diseases. These classifiers are known for their ability to handle imbalanced datasets, and we investigate their performance in comparison to other classifiers. Our research aims to identify the most effective classifier for heart disease prediction, considering factors such as accuracy, predictive analysis, and potential for practical implementation in a healthcare setting. The results of our research have the potential to contribute to the field of cardiovascular health by providing insights into the performance of different classifiers and proposing an ensemble classifier that can enhance the accuracy of heart disease prediction. The findings of this study may have important implications for clinical decision-making and patient care, ultimately improving health outcomes for individuals at risk of heart diseases. Further research can be carried out to explore other Machine Learning techniques and optimize the ensemble classifier for real-world applications in healthcare settings.

Keywords: SVM; Naive Bayes; Decision Tree; Random Forest; Logistic Regression; XG-boost; confusion matrix;

Introduction

According to the World Health Organization, every year 12 million deaths occur worldwide due to heart disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section
of data analysis. 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 reduces the complications. Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. 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 algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analyzing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

Literature Survey

With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers.

S.No	Title of the paper	Authors	Year of publication	Journal name	The technical content of the paper (algorithms used, accuracy/ performance of the model)
1	Heart disease prediction	Harshit Jindal1 , Sarthak Agrawal 1 , Rishabh	2020	IOP Conference Series: Materials	Data is directly retrieved from electronic records

	using machine learning algorithms	Khera1 , Rachna Jain2 and Preeti Nagrath		Science and Engineering	that reduce the manual tasks. Different algorithms of machine learning such as logistic regression and KNN are implemented to predict and classify the patient with heart disease. The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naive bayes etc
2	Prediction of Heart Disease Using Machine Learning Algorithms	Santhana Krishnan J. , Geetha S.	2019	IEEE	Implemented decision tree and Naive Bayes algorithm for prediction of heart disease. In decision tree algorithm the tree is built using certain conditions which gives True or False decisions. The algorithms like SVM, KNN are results based on vertical or horizontal split conditions depends on dependent variables. But decision tree

					for a tree like structure having root node, leaves and branches base on the decision made in each of tree Decision tree also help in the understating the importance of the attributes in the dataset. They have also used Cleveland data set. Dataset splits in 70% training and 30% testing by using some methods. This algorithm gives 91% accuracy. The second algorithm is Naive Bayes, which is used for classification. It can handle complicated, nonlinear, dependent data so it is found suitable for heart disease dataset as this dataset is also complicated, dependent and nonlinear in nature. This algorithm gives an 87% accuracy.
3	Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques	Senthilkumar Mohan; Chandrasegar Thirumalai; Gautam Srivastava	2019	IEEE	their main objective is to improve exactness in cardiovascular problems. The algorithms used are KNN, LR, SVM, NN to produce an improved exhibition level

					with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with linear model (HRFLM).
4	Machine Learning Techniques for Heart Disease Prediction	A. Lakshmanarao,Y.Swathi, P.Sri Sai Sundareswar	2020	INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH	] Lakshmana Rao et al,proposed "Machine Learning Techniques for Heart Disease Prediction" in which the contributing elements for heart disease are more. So, it is difficult to distinguish heart disease.To find the seriousness of the heart disease among people different neural systems and data mining techniques are used.
5	Heart Disease Prediction Using Effective Machine Learning Techniques	Avinash Golande, Pavan Kumar T	2019	International Journal of Recent Technology and Engineering (IJRTE)	" In this paper, few data mining techniques are used that support the doctors to differentiate the heart disease. Usually utilized methodologies are k-nearest neighbour, Decision tree and Naïve Bayes. Other unique characterization-based strategies utilized are packing calculation, Part thickness, consecutive

					negligible streamlining and neural systems, straight Kernel selfarranging guide and SVM (Bolster Vector Machine).

Proposed System

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format.
The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is
obtained by testing the system using the testing data. This system is implemented using the following modules.

1.) Collection of Dataset
2.) Selection of attributes
3.) Data Pre-Processing
4.) Balancing of Data
5.) Disease Prediction

Deployment Of Machine Learning Model in AWS Using Amazon SageMaker

Building, training, and deploying machine learning models can all be done with one integrated set of tools thanks to Amazon SageMaker, a fully managed service. The procedures for deploying an ML model using Amazon SageMaker are as follows: We Develop our model and make sure our machine learning model is performance-optimized by training and testing it on a dataset. Create an inference script to specify the inputs and outputs of our model and package it using a framework like TensorFlow, Scikit-Learn, or MXNet. Then we upload it to Amazon S3: Put our packed model in a bucket on Amazon S3.Build a SageMaker model on the Amazon SageMaker console by indicating the location of our model data in Amazon S3.Endpoint configuration creation: Provide the resources to be

used for hosting the model in an endpoint configuration, including the kind and number of instances. Launch the model: Create an Amazon SageMaker endpoint with the endpoint settings we created in s to deploy the model. Then we Analysed the model: Once the endpoint has been constructed, we may use the SageMaker runtime API to make predictions to test the model .Machine learning model deployment and management are made simple with a number of capabilities offered by Amazon SageMaker. To optimise our model's hyperparameters, utilise SageMaker's automatic model tuning function. SageMaker's automatic scaling feature lets us automatically change the number of instances needed to serve predictions based on incoming traffic.

Team members:

S.No	Reg. No	Name
1	22MCA0137	KANISHK SHARMA
2	22MCA0174	UDDASHAY KUMAR GUPTA