CARDIOVASCULAR DISEASE PREDICTION SYSTEM USING MACHINE LEARNING

S. Sundar Pandiyan[1], P. Hemanth Kumar[2], C.Chinna Babji Saheb[3], V.Anil[4], M.Chandru[5]
[1,2,3,4,5] Department of CSE, Annamacharya Institute of Technology & Sciences, Tirupati-517520

Abstract
Heart attack disease is one of the leading causes of death worldwide. In today's common modern life, deaths due to heart disease have become one of the major issues. Roughly one person lost his or her life per minute due to heart illness. Predicting the occurrence of disease at its early stages is a major challenge nowadays. Machine learning when implemented in health care is capable of early and accurate detection of disease. In this project, the arising situations of Heart Disease illness are calculated. The datasets used have attributes of medical parameters. The datasets are processed in python using an ML algorithm i.e. Random Forest Algorithm. This technique uses the past old patient records for getting predictions of new ones at early stages preventing the loss of lives. In this work, a reliable heart disease prediction system is implemented using a strong Machine Learning algorithm, which reads the patient record dataset in the form of a CSV file. After accessing the dataset the operation is performed and an effective heart attack level is produced.

Keywords: Machine learning, Random Forest Algorithm.

Introduction

Heart Disease affects the functioning of the heart. The World Health Organization had made a survey and concluded that 10 million people are affected by heart disease and lost their lives. In the existing system, researchers had initially proposed a system that is based on a self-applied questionnaire. In this system, the user needs to enter all the symptoms which he is suffering from, and based on that the result is predicted. Previously they used the technique of Vector Quantization which is one of the artificial intelligence techniques for classification and prediction purposes. Training of neural networks is performed using backpropagation to evaluate the prediction system. So to overcome this we are implementing a Random Forest Algorithm to achieve accurate results in less time. The program we use cannot process these non-numerical values, so it is mandatory to convert these values to numerical values. The approach followed is, the NaN values are replaced by the means of the column.

Related Work

1. Material and Methods:

This study uses a dataset of medical test results and records from 400 patients, which includes 250 patients with heart disease (HD) and 150 patients without CKD. The dataset was prepared in 2015 by Apollo Hospitals, Tamil Nadu, India, and is publicly available at the University of California, Irvine (UCI) machine learning repository. The dataset has 24 factors, such as blood pressure and blood flow etc, cell count, and a classification variable indicating whether the patient has HD or not. It's important to note that the dataset is imbalanced, with more patients having HD than not.

1.2 Information Gain:

Feature selection is an important step in building accurate predictive models. By removing attributes that are less useful or unrelated to the target variable, the computational cost is reduced, and the model's performance is improved. In this study, the information gain (IG) technique is used for feature selection. IG is a filter-based method that evaluates the ability of predictor variables to classify the dependent variable.

2. Algorithm

2.1 RANDOM FOREST

A Random Forest Algorithm is a supervised machine learning algorithm that is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher its accuracy and problem-solving ability. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

Algorithm 1: Conventional RF technique

Input: The input for the model consists of a dataset containing information about patients with various features such as age, sex, blood pressure, cholesterol levels, and other relevant medical history. Each patient is labeled as either having heart disease or not having heart disease. This dataset is split into training and testing sets.

Output: The output of the model is a binary classification of whether a patient has heart disease or not based on their input features. The output can be in the form of a probability score or a categorical label (0 or 1) depending on the threshold value set for the model. The model's performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC

Procedure:

1.Collect and preprocess the data: Gather a dataset with relevant features and preprocess the data to remove any missing values, normalize the data, and convert categorical variables into numerical values.

2.Split the dataset: Divide the dataset into training and testing sets. Typically, the training set should be larger than the testing set.

3.Build the Random Forest model: Build the Random Forest model using the training set. Specify the number of decision trees to use, the maximum depth of each tree, and the number of features to consider at each split.

4.Train the model: Train the Random Forest model using the training set. The model learns the patterns and relationships between the input features and the target variable.

5.Evaluate the model: Evaluate the performance of the model using the testing set. Use metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to assess the model's performance.

6.Tune the model: Fine-tune the model's hyperparameters to optimize its performance. Use techniques such as GridSearchCV to search over a range of hyperparameters and find the best combination.

7.Deploy the model: Once the model is trained and tuned, deploy it to make predictions on new data. You can use the model to predict whether a patient has heart disease or not based on their input features.

3. Result:

In order to establish a baseline for comparison with the proposed cost-sensitive AdaBoost (CS AdaBoost), this research also implemented several well-known classifiers, including logistic regression, decision tree, XGBoost, random forest, and SVM, along with the traditional AdaBoost presented in Algorithm 1. Both the complete and reduced feature sets were used to train the classifiers, in order to demonstrate the impact of feature selection.

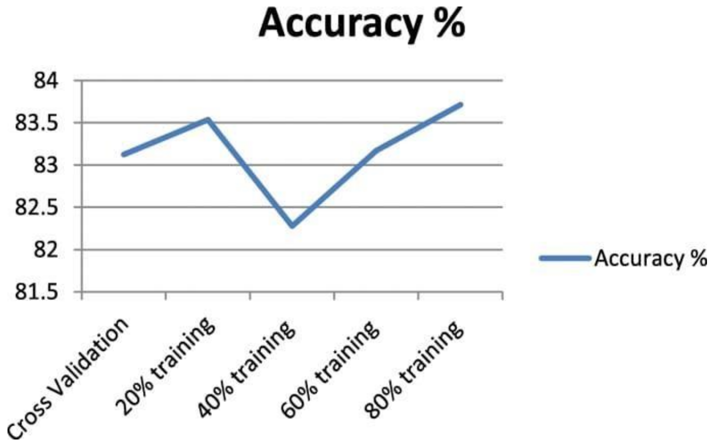
Table 2: Performance of the Classifiers without Feature Selection

Classifier	**ACC**	**SEN**	**SPE**	**AUC**
Logistic regression	0.940	0.948	0.935	0.940
Decision tree	0.902	0.932	0.790	0.910

XGBoost	0.963	0.974	0.942	0.960
Random forest	0.952	0.955	0.940	0.960
SVM	0.927	0.943	0.930	0.940

4. Conclusion:

The conclusion that we found is that machine learning algorithms performed better in the analysis. Many researchers have previously suggested that we should use ML where the dataset is not that large, which is proved in this paper. The Random Forest algorithm is an efficient algorithm that is an ensemble learning method for regression and classification techniques. The algorithm constructs N of Decision trees and outputs the class that is the average of all decision tree output. So the accuracy of prediction at early stages is achieved effectively. Processing of healthcare data i.e., data related to the heart will help in the early detection of heart disease or abnormal conditions of the heart which results in saving long-maths.



5. References:

1 C. C. Aggarwal, “Context-sensitive recommender systems,” in *Recommender Systems*. Cham, Switzerland: Springer, 2016, pp. 255–281.

2 A. Al-Molegi, M. Jabreel, and B. Ghaleb, “STF-RNN: Space-time features-based recurrent neural network for predicting people’s next location,” in *Proc. IEEE Symp. Series Comput. Intell. (SSCI)*, Athens, Greece, 2016, pp.
1–7.

3 M. Abdi, G. Okeyo, and R. Mwangi, “Matrix factorization techniques for context-aware collaborative filtering recommender systems: A survey,” *Comput. Inf. Sci.*, vol. 11, no. 2, p. 1, 2018.

4 I. Baytas, C. Xiao, X. Zhang, F. Wang, A. Jain, and J. Zhou, “Patient subtyping via time-aware LSTM networks,” in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Disc. Data Min. (KDD)*, 2017, pp. 65–74.

5 H. Cheng *et al.*, “Wide & deep learning for recommender systems,” in *Proc. 1st Workshop Deep Learn. Recommend. Syst. (DLRS)*, 2016, pp. 7–10.

6 M. Chen, Y. Liu, and X. Yu, “NLPMM: A next location predictor with Markov modeling,” in *Proc. Adv. Knowl. Disc. Data Min.*, 2014, pp. 186–197.

7 E. Cho, S. Myers, and J. Leskovec, “Friendship and mobility: User movement in location-based social networks,” in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min. (KDD)*, 2011, pp. 1082–1090. [8] C. Cheng, H. Yang, I. King, and M. Lyu, “Fused matrix factorization with geographical and social influence in location-based social networks,” in *Proc. 26th AAAI Conf. Artif. Intell. (AAAI)*, 2012, pp. 17–23.