THE EFFECT OF FEATURE SELECTION TECHNIQUES ON THE ACCURACY OF HEART DISEASE PREDICTION USING MACHINE LEARNING

J. Lojana[1]

[1] Department of ICT, Faculty of Technological Studies, University of Vavuniya

Abstract

Artificial intelligence has recently had a significant impact, particularly on the healthcare sector. The use of machine learning has made it possible to predict a number of serious diseases that are now difficult to identify in the medical industry. In this study, the Heart Attack Analysis Prediction Dataset was considered for testing. This dataset was obtained from the Kaggle. The dataset contains 14 features and 303 patient records. To find the best classification algorithm with the highest accuracy, seven feature selection algorithms and eight classification algorithms were used. Simple logistic and Logistic Model Tree classification algorithms were found to be the best classification algorithms for the heart attack analysis and prediction dataset with 85.1485% accuracy. The accuracy of the classification was impacted with the number of features selected.
***Keywords**: Artificial intelligence, machine learning, feature selection, classification, simple logistic, logistic model Tree.*

1 INTRODUCTION

Heart diseases (HDs) are one of the major causes of life complications and subsequently leading to death. Heart Diseases are a more dangerous and risky health issue that prevailed throughout the world. [1]. Due to the uncommon availability of effective diagnostic instruments, a lack of qualified medical personnel, and other factors that affect patient prognosis and treatment, the diagnosis and treatment of cardiac disease are exceedingly challenging, especially in developing countries. The main causes are inadequate preventative measures and a shortage of qualified or skilled medical providers. [2].

These diseases are a class of disease that incorporates the heart and blood vessels. The HDs include coronary artery diseases (CAD) like angina and myocardial infarction and coronary heart disease (CHD).[3]These HDs would cause death, without the proper consultation of the doctors. HDs occur with the symptoms of Chest pain, Nausea, Pain in the Arms, Fatigue and Sweating. Cardiovascular problems can be prevented; the main reason they are still on the rise is a lack of effective preventive measures.

Several clinical decision support systems on heart disease prediction have been developed by various academics in the modern digital age to simplify and guarantee effective diagnosis. [4].Blood tests, electrocardiograms (ECG), exercise stress tests, echocardiograms (ultrasound), nuclear cardiac

stress tests, coronary angiograms, magnetic resonance imaging (MRI), and coronary computed tomography angiograms (CCTA) are among the medical tests that can be used to predict the heart's diagnosis manually[5].The doctors must personally review the testing results, evaluate each and every test result value, and determine whether or not a given patient has heart disease. It will take a while for a person to forecast it. It is difficult to manually determine the prevalence of heart disease based on risk factors[6]The mortality rate can be significantly decreased if the disease is identified in its early stages and preventative measures are implemented as soon as is possible.

To identify different types of metabolic syndromes, data mining and the perspective of medical research are used. Heart disease prediction and data analysis both greatly benefit from data mining with classification. [1] For prompt diagnosis of these disorders and effective treatment, a trustworthy, accurate, and practical system is required. In order to automate the examination of massive and complicated data, machine learning methods and techniques have been used on a variety of medical datasets. [2].

2 Related Works

Artificial intelligence has had a significant impact recently, particularly on the health care sector. It is now feasible to predict several serious diseases that are difficult to diagnose in the medical industry, including cancer [7]breast cancer [8], and thyroid [9], and melanoma skin cancer [10] with the help of the technologies. With the usage of the AI and ML techniques, the heart disease can be predicted with the effective manner, this research explored methods of machine learning for heart disease prediction.

The Authors in [1], performed heart disease classification with the Cleveland UCI dataset with 13 features using R studio rattle. They presented the HRFLM model to forecast the diagnosis of heart disease in comparison to other well-known classification methods like as Naive Bayes, Generalized Linear Model, Logistic Regression, Deep Learning, Decision Tree, Random Forest, Gradient Boosted trees, Support Vector machine, and VOTE. The HRFLM model was produced by combining the Random Forest and Linear Model and obtained 88.7% accuracy.

Another example by L. Yahaya et al. [2], Compared the heart disease classification with six machine algorithms such as Logistic Regression, Support Vector Machine, K Nearest Neighbors, Artificial Neural Network, Naive Bayes, and Random Forest) using the Cleveland dataset, with the help of six data mining tools (Orange, Weka, Rapid Miner, Knime, MATLAB, and Scikit-Learn) for the analysis and comparison. Based on the data analysis and the findings of the extracted performance measures, it was determined that MATLAB's Artificial Neural Network was the best performing technique. A technique was explored in [6], as Ensemble classification that performed with the combination of various classifiers to increase the precision of weak algorithms. The Cleveland dataset can be used to improve the prediction accuracy of weak classifiers using ensemble techniques like bagging and boosting. The use of ensemble classification resulted in an accuracy improvement of weak classifiers of up to 7%. The study [4] stated that the majority of studies employed the Cleveland heart disease dataset, which had just 303 instances and 14 features. The Authors came to the conclusion that additional numerous heart disease datasets from geographically different sources with more attributes should be investigated for constructing more effective machine learning models in order to achieve a more universal classification and prediction accuracy.

In [11], the authors addressed the issue to fix the problem of feature selection, and presented a new, Fast Conditional Mutual Information feature selection algorithm (FCMIM). The experimental findings demonstrate the validity of the suggested feature selection technique (FCMIM) for constructing a high-level intelligent system to detect heart disease using a classifier support vector machine. In

study [6], a summarization of data mining approaches (Decision Tree, C4.5, K-means algorithm, ID3 algorithm, SVM, Naive Bayes, ANN, CART, Random Forest, Regression, J48, A-Priori Algorithms, Fuzzy Logic, Association Rules), data mining tools (WEKA, Rapid Miner, TANAGRA, Apache Mahout, MATLAB, Java, C, and Orange) have performed to detect heart disease.

3 METHODOLOGY

The procedures used in this study were data collection, preprocessing, feature selection, classification, and performance and evaluation. Each procedure in the methodology will be discussed in this section.

3.1 Dataset Acquisition

"The Heart Attack Analysis Prediction Dataset" is considered for testing purposes in this study. This dataset was obtained from the Kaggle dataset. The dataset includes 303 patient records along with 14 features including the prediction label.

3.2 Pre-Processing

Using the WEKA data mining tool, the data was preprocessed using the "numerical to nominal" filter.

3.3 Features Selection

The following feature section mechanisms were used for this procedure to compare the different features for the analysis of the dataset: GainRatioAttributeEval, CFSSubsetEval, ClassifierAttributeEval, CorrelationAttributeEval, InfogainAttributeEval, ReliefAttributeEval, and SymmetricallUncert Attribute. The features are listed according to their rank of priority Shown in Table 2.

3.4 Classification

Datasets were tested using the eight classifier algorithms for this investigation, the cross-validation folds of 10 were used. The Table 3 lists accuracy against each classification algorithm.

3.5 Performance and Evaluation

The best classifier for this dataset was selected based on the "correctly classified instances" with the highest percentage of accuracy.

4 RESULTS AND DISCUSSIONS

4.1 Feature Selection

The study was compared against seven feature selection approaches based on the ranking of the selected attributes. The features were ranked by each and every feature selection method, described in Table 2 along with the feature selection methods.

By comparing the features ranked with the seven different feature selection methods, it was decided to test the classification algorithms with the following four cases.

- Case 1: Total 13 features were tested.

- Case 2: It was found that the feature age and FBS were within the last three positions for many times among seven feature selection methods, therefore age and FBS features were eliminated, there were no changes in the accuracy of Simple logistic, KStar, Attributed selected classifier and LMT classification algorithms (Table No:3, column:4).
- Case 3: The features age and sex were eliminated in order to test the correctness of the statement by the authors of [1]. The remaining 11 attributes are considered important while age and sex features were removed from this dataset. (Table No:3, column:5).
- Case 4: Only five features (cp, exng, Slp, Caa, thall) were selected by the CFS algorithm. All eight classification methods were tested against these five selected attributes. This testing confirmed that the accuracy was decreased. Therefore, it is impossible to forecast heart disease using at least these five chosen traits. (Table No:3, column:6).

4.2 Classification

The percentage of each classification algorithm tested with the above cases using the WEKA data mining tool were described in the Table 3. The classification algorithms used for the study were Simple Logistic, Voted Perceptron, KStar, AdaBoostM1, Attribute Selected Classifier, Iterative Classifier Optimizer, Logit Boost, and Logistic Model Tree (LMT). The selected classification algorithms were chosen among the 40 classification algorithms that exist in Weka tool, based on the performance metric such as accuracy.

4.3 Performance and Evaluation

The Voted perception and Attribute-selected classifier (81.1881%), KStar and Logit Boost (81.5182%), AdaBoostM1, and Iterative Classifier Optimizer (82.5083%) all achieved the same accuracy among the eight classification methods for the 13 features. Simple logistic and LMT classification algorithms' accuracy did not change when age fbs and age sex features were taken out of the classification testing. The accuracy was lower when classification procedures were used to the five features than it had been when classification techniques had been applied to the thirteen features. Therefore, to predict heart disease, it is better to have at least 11 features.

Finally, among these various classification techniques, simple logistic and LMT had the same and the maximum accuracy (85.1485%). The confusion matrix for the above classification algorithm is shown in Table 4 below. However, this accuracy percentage is seemed to be higher than the conventional machine learning algorithms used in previous studies such as: Support Vector Machine, Artificial Neural Networks, Random Forest.

5 Conclusion

Machine learning has proven to be useful in assisting with decision-making and prediction from the vast amount of data generated by the healthcare sector. There was a total of 14 features in the Heart Attack Analysis Prediction Dataset, including the output label. Age and sex were identified as the individual patients' personal data, so the other 11 variables were clinical data. Each feature was ranked using the features selection algorithm. The features selection process was executed with 4 different cases. In case 1, all features were considered for the classification. In order to do the classification for case 2, age and fbs features were eliminated. Age and sex features were taken out for the classification testing in case 3. The 13 features were ranked by six feature selection algorithms.

Table 1. The percentage of each classification algorithms

Classifier Name	Case 1	Case 2	Case 3	Case 4
Simple Logistic	85.1485%	85.1485%	85.1485%	84.4884%
Voted Perceptron	81.1881%	81.8484%	80.5282%	83.1683%
KStar	81.5182%	81.5182%	82.1782%	83.4983%
AdaBoostM1	82.5083%	82.8383%	81.8482%	83.4983%
Attribute selected classifier	81.1881%	81.1881%	81.5182%	82.1782%
Iterative classifier optimizer	82.5083%	82.8383%	81.8482%	82.8383%
Logit Boost	81.5182%	81.1881%	80.198%	84.4884%
LMT	85.1485%	85.1485%	85.1485%	84.4884%

Table 2. Confusion Matrix for Simple Logistic Algorithm

	Heart Disease	Not Heart Disease
Heart Disease	110	28
Not Heart Disease	17	148

However, the CFS algorithm only gave rankings to 5 features (cp, exng, Slp, Caa, thall). Therefore, in case 4, only those five features were selected to the classification process.

Eight different classification algorithms, including Simple Logistic, Voted Perceptron, KStar, AdaBoostM1, Attribute Selected Classifier, Iterative Classifier Optimizer, Logit Boost, and LMT, were tested in this study. The accuracy of the classification was impacted by the number of features used. It was found that the classification algorithms Simple Logistic and LMT outperforms the other classification algorithms with the highest accuracy percentage of 85.1485% for three cases (case 1, case 2 and case 3).

ACKNOWLEDGMENT

I would like to express my very great appreciation to Ms. Rukshani Puvanendran for her valuable and constructive suggestions during the planning and development of this wonderful research on the topic "The Effect of Feature Selection Techniques on the Accuracy of Heart Disease Prediction Using Machine Learning". Further, I would also like to thank the staff of the Department of ICT, Faculty of Technological Studies, University of Vavuniya for providing this golden opportunity.

REFERENCES

[1] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques". In: *IEEE access* 7 (2019), pp. 81542–81554.

[2] Lamido Yahaya, N David Oye, and Etemi Joshua Garba. "A comprehensive review on heart disease prediction using data mining and machine learning techniques". In: *American Journal of Artificial Intelligence* 4.1 (2020), pp. 20–29.

[3] Machine Learning. "Heart disease diagnosis and prediction using machine learning and data mining techniques: a review". In: *Advances in Computational Sciences and Technology* 10.7 (2017), pp. 2137–2159.

[4] R Radhika and S Thomas George. "Heart Disease Classification Using Machine Learning Techniques". In: *Journal of Physics: Conference Series*. Vol. 1937. 1. IOP Publishing. 2021, p. 012047.

[5] *Government of Western Australia.* 2022. URL: %E2%80%9CCommon%20medical%20tests%20to%20diagnose%20heart%20conditions.%E2%80%9D.

[6] C Latha and S Jeeva. *Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Informatics in Medicine Unlocked, vol. 16*. 2019.

[7] Aman Sharma and Rinkle Rani. "A systematic review of applications of machine learning in cancer prediction and diagnosis". In: *Archives of Computational Methods in Engineering* 28.7 (2021), pp. 4875–4896.

[8] R. Rawal. "Breast Cancer Prediction using Classification Techniques". In: *Int. J. Emerg. Trends Eng. Res.* 8.9 (2020), pp. 6074–6079.

[9] Gyanendra Chaubey, Dhananjay Bisen, Siddharth Arjaria, et al. "Thyroid disease prediction using machine learning approaches". In: *National Academy Science Letters* 44.3 (2021), pp. 233–238.

[10] MA Ahmed Thaajwer and UA Piumi Ishanka. "Melanoma skin cancer detection using image processing and machine learning techniques". In: *2020 2nd International Conference on Advancements in Computing (ICAC)*. Vol. 1. IEEE. 2020, pp. 363–368.

[11] Jian Ping Li, Amin Ul Haq, Salah Ud Din, et al. "Heart disease identification method using machine learning classification in e-healthcare". In: *IEEE Access* 8 (2020), pp. 107562–107582.