**CFD: Cardiovascular disease prediction through Feature engineering and Data mining**

122003023 - Anandalakshmi R

122003050 - Bhargavi T

122003096 - Jaisaiarun P Srinivasan

Cardiovascular diseases (CVDs) are one of the important causes of death worldwide. According to the Centers for Disease Control and Prevention (CDC), a heart attack occurs every 40 seconds in the United States. Prediction of CVDs plays a crucial role in the healthcare sector. Data mining techniques are used be operate upon the massive set of data collected from the healthcare sector to produce impeccable results in the prediction of CVDs [1].

We can find existing methodologies to predict CVDs in clinical research. Following are some of those techniques. The One Dependency Augmented Naïve Bayes classifier (ODANB) and naive credal classifier 2(NCC2) are an extension of naive Bayes used for CVDs prediction [2]. The Radial basis function (RBF) network, which is also present in weka classifiers functions [3]. The combination of Principal Component Analysis (PCA) for feature reduction and SVM for prediction is also used [4]. In Neural Network (NN) with Fuzzy, a genetic algorithm (GA) is developed using recurrent fuzzy neural networks (RFNN) for the prediction of CVDs [5]. In Decision Tree with Gain Ratio, a wide range of discretization techniques and voting methods were applied on different types of Decision Trees seeking better performance in heart disease diagnosis [6]. In NN with GA, Information Gain is used in determining the attributes and classifying the diagnosis of patients [7]. Other existing works are Extreme Learning Machine [8], and Classification and Regression Tree CART [9].

Unlike existing methodologies, the base paper proposed a model that concentrates on finding the significant features and data mining techniques in predicting CVDs. A brute force approach was used to identify a combination of three or more crucial features from the available set of features. The identified feature sets were used in combination with the data mining techniques for classification modeling. Classification techniques considered are k-NN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network, and Vote (a hybrid technique with Naïve Bayes and Logistic Regression). The Cleveland database from the UCI machine learning repository was used to identify the crucial features and the top-performing models. The performance metrics for all combinations of features and classification models were obtained. The performance metrics used are accuracy, f-measure, and precision. The best feature set and the corresponding model were validated using a different dataset, the UCI Statlog Heart Disease dataset. This study identifies nine significant features and three data mining techniques that produce accurate results.

**Keywords:** Data mining, Prediction model, Feature selection, Cardiovascular disease prediction.

**REFERENCES:**

[1] Amin, M.S., Chiam, Y.K. and Varathan, K.D., 2019. Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, pp.82-93.

[2] Srinivas, K., Rani, B. K., Govrdhan, A., 2010. Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE). 2(02), 250-255.

[3] Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. 2010. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology, 2(10), 5370-5376

[4] Dey, A., Singh, J., Singh, N., 2016. Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis. Analysis. 140(2), 27-31.

[5] Paul, A. K., Shill, P. C., Rabin, M. R. I., & Akhand, M. A. H. 2016. Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. In 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 145-150. IEEE.

[6] Shouman, M., Turner, T., Stocker, R. 2011. Using decision tree for diagnosing heart disease patients. Proceedings of the Ninth Australasian Data Mining Conference (AusDM’11), Darlinghurst, Australia, Volume 121, pp. 23-30.

[7] Khemphila, A., & Boonjing, V. 2011. Heart disease classification using neural network and feature selection. In 21st International Conference on Systems Engineering (ICSEng), Las Vegas, pp. 406-409. IEEE.

[8] Ismaeel, S., Miri, A., Sadeghian, A., Chourishi, D., 2015. An Extreme Learning Machine (ELM) Predictor for Electric Arc Furnaces' vi Characteristics. IEEE 2nd International Conference on Cyber Security and Cloud Computing (CSCloud), New York, pp. 329-334.

[9] Chaurasia, V., Pal, S., 2013. Early prediction of heart diseases using data mining techniques. Carib. J. SciTech. 1, 208-217.



**Dr. A. Joy Christy**