Using Machine Learning Technologies to Classify and Predict Heart Disease

Mohammed F. Alrifaie[1], Zakir Hussain Ahmed[2], Asaad Shakir Hameed[3], Modhi Lafta Mutar[4]
Computer Engineering Department, Faculty of Engineering, Karabuk University, Karabuk, Turkey[1]
Department of Information and Communications, Basra University College of science and technology, Basrah, Iraq[1]
Department of Mathematics and Statistics, College of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Kingdom of Saudi Arabia[2]
Faculty of Information and Communication Technology Universiti Teknikal Malaysia Melaka Hang Tuah Jaya Durian Tunggal, Melaka, Malaysia[3]
Department of Mathematics, General Directorate of Thi-Qar Education, Ministry of education, Thi-Qar, Iraq[3]
Faculty of Information and Communication Technology Universiti Teknikal Malaysia Melaka Hang Tuah Jaya Durian Tunggal, Melaka, Malaysia[4]
Department of Mathematics, General Directorate of Thi-Qar Education, Ministry of education, Thi-Qar, Iraq[4]

***Abstract*—The techniques of data mining are used widely in the healthcare sector to predict and diagnose various diseases. Diagnosis of heart disease is considered as one of the very important applications of these systems. Data is being collected today in a large amount where people need to rely on the device. In recent years, heart disease has increased excessively and heart disease has become one of the deadliest diseases in many countries. Most data sets often suffer from extreme values that reduce the accuracy percentage in classification. Extreme values are defined in terms of irrelevant or incorrect data, missing values, and the incorrect values of the dataset. Data conversion is another very important way to preconfigure the process of converting data into suitable mining models by acting assembly or assembly and filtering methods such as eliminating duplicate features by using the link and one of the wrap methods, and applying the repeated discrimination feature. This process is performed, dealing with lost values through the "Remove with values" methods and methods of estimating the layer. Classification methods like Naïve Bayes (NB) and Random Forest (RF) are applied to the original datasets and data sets with the feature of selection methods too. All of these operations are implemented on three various sets of heart disease data for the analysis of pre-treatment effect in terms of accuracy.**

Keywords—Classification; Naive Bayes (NB); (Support Vector Machine SVM); Random Forest; machine learning

I. Introduction

Nowadays one of the major causes of death is heart disease at the present time. The heart disease prediction system can support healthcare specialists in predicting heart condition based on the clinical data of patients that has been pre-entered into the system. There are several healthcare manufactures and hospitals which gather massive amounts of data for patients which are hard to deal with current systems [5]. There are a lot of tools that use prediction algorithms are available nonetheless they have several weaknesses [15,16]. Many of the tools cannot deal with large data. Actually, there are a lot of algorithms can be used to find and predict the heart disease such as the discrete differential evolution (DDE) algorithm [17]. Machine learning algorithm acts an important role in extracting hidden knowledge and information and analyzing it from these data sets. Actually, it improves speed and accuracy. Data extraction techniques have been used in many areas, including health care. This paper aims to check whether the prediction of heart disease can be depended on data mining and machine learning [9]. By using some techniques of data mining, Prediction helps detect if a patient suffers of heart disease or not. In addition, the prediction helps specialists to get to the appropriate diagnosis more quickly, not only that, but it increases the accuracy of diagnosis leading to better results may help to reduce or reduce heart attacks at the very least. Hidden relationships can untangle and diseases are diagnosed efficiently by the help of Data mining along with soft computing techniques [7,8]. The datasets are collected and gathered from the Machine Learning Repository (UCI). It now upholds 394 datasets copies with 14 attributes those names are sex, age, chest pain type, resting blood pressure, resting electrocardiographic results, fasting blood sugar>120 mg / dl, serum cholesterol in mg/dl, exercise induced angina, maximum heart rate achieved, the slope of the peak exercise ST segment, oldpeak = ST depression caused by exercise relative to rest, number of main vessels (0-3) colored by flourosopy, thal: 7 = reversible defect; 6 = fixed defect; 3 = normal. These features are used as a service package to the MLC (community of machine learning). There are 3 data bases in the Data Set of heart disease, these data bases namely Cleveland, Hungary, Switzerland. In this paper, we analyze cardiology data based on Dataset by using the link and one of the wrap methods, and applying the repeated discrimination feature. This process is performed, dealing with lost values through the "Remove with values" methods and methods of estimating the layer. However, the outline of this paper as follows, starts from the literatures to analyze the previous studies about classification and the used algorithms in this area. Then we discuss our methodology by elaborating the procedure of the work and the application of the algorithms. In result section we illustrate the obtained results and discuss it. Finally, we summarize our work in conclusion section and future work.

II. Literature Survey

There are a lot of common data mining algorithms, particularly the techniques of classification, each of them is distinguished by both excellence and weakness, for example three of which are: Decision Tree, k-Nearest Neighbor (KNN), and Naïve Bayes and [10]. Naive Bayes technique is a powerful, simple and good performance of classification. Basically, it depends on Paez's theory of the probability of P(c|x) from the previous possibility of P(c), the possibility of the given P(x|c) and the predictability probability as follows [10,11].

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{1}$$

The Bayesian Naive model is commonly used in many areas such as medical diagnosis, spam filtering, and even text classification. It gets a lot of attention among statisticians resulting in algorithm modifications. At the same time, Decision Tree algorithms (DT) can be, also it is fast in clear modeling and training. DT mechanism is by sorting the trained data in a form of tree. In the training phase, this tree is formed to determine the accuracy of the workbook for the test data. Then it will be categorized by using the tree [12].

On the other hand, we have k-nearest neighbors KNN. It is a simple technique of classification that is often used in several studies, particularly when there is little or no data distribution information. KNN is a non-deterministic algorithm, which means that it does not put assumptions about the data’s distribution that is used in the analysis. KNN fits into practical environments, since data are often not followed by real theoretical statistics such as natural distribution. Lazy algorithm is also another name of KNN, or it uses only a quick training stage. It does not make circular which means that it keeps all data of training [13].

Many of researches were conducted to focus on heart diseases ‘prediction and classification. Many data mining techniques are used for achieving and diagnosis various accuracy level of various methods [1]. The NB classifier algorithm uses conditional independence; the attribute value for a given category is believed to be independent of the values of other attributes. An example of an application of the Naïve Bayes algorithm is the proposed model for the prediction of high-risk heart disease. This application was introduced as part of the Web-based healthcare and detection package that is proposed by [2]. The data collected and prepared was considered as the training group as the work was based on two basic stages: the classification stage and the pre-processing which perform several tasks including: normalization, reduction, and cleaning of data, etc. On the other side, the second stage is the stage of prediction. In the prediction stage, the data are categorized into groups based on the type and nature of the disease, and then the establishment of a test group based on the questions of the disease. Finally send the results obtained from the prediction to the doctor.

The author in [2] proposed a Predicting heart disease model using neural network. The choice of feature is used for disease prediction. This method found a 100% accuracy of 15 features and precision of 92.5% for 13 features. There is an improvement of 7.5% after giving up two advantages of 15 to 13.

The proposed method by [14] is used associative classification and selection of a subset of the risk of disease. They used asymmetric uncertainty, information acquisition, and genetic algorithm as measures to select the feature. Their method got 95% accuracy with the choice of hybrid feature. Heart disease data set gathered for experimental analysis with 11 features.

The author in [3] examines the performance for heart disease diabetes dataset by using several classification algorithms of machine learning like Decision Tree (J48), Naive Bayes (NB), and Support Vector Machine (SVM) with bagging technique. In order to measure the efficiency of the Classification algorithms they depend on the precision, accuracy, specificity, sensitivity, and performance. All of tests are achieved in the "WEKA TOOL", the results shown that Decision Tree (C4.5) provides a high quality of accuracy around (95.06%).[4] applied some of imputation methods for dealing with lost data. Imputation methods are algorithms designed to recover lost values of data, depending on other entered data in the data base. The choosing of the imputation effects on the performance of the machine learning’s technology, for instance, it effects on the applied classification algorithm’s accuracy. Thus, applying and selecting the correct calculation method is very important and commonly requires a great deal of human intervention. In this article, they suggested to using genetic programming techniques in order to seek for the correct mix of classification and imputation algorithms. They build the work based on the TPOT library of Python, and integrating a heterogeneous set of computing algorithms as part of the pipeline search of ML. They have shown that genetic programming can routinely find gradually enhanced pipelines that contain the most effective blend of feature pre-processing, classifiers, and imputation methods for a varied of classification difficulties with lost data.

The author in [6] checked various classification techniques in analysis heart disease. The classifiers like Naive Bayes, KNN, and Decision Tree are used to divide dataset. After the classification and evaluation of the performance they considered that the Decision Tree as the best classifier and the most efficient for heart disease analysis from the dataset.

III. Methodology

Classification algorithms have been used in large data sets that are known for their study of layers and make prediction. Models store stored data sets in memory to predict. This model has the ability to predict the category label or the new data instance set [6]. So, this study used a classification algorithm under the supervision of Random Forest and Naïve Bayes to classify and predict heart disease.

In Random Forest for prediction, the accuracy is very important like healthcare fields especially when we talk about heart disease, the processing times can be used as a differentiation, while time-sensitive fields seek rapid predictions like the accuracy of the disaster prediction ratio that can also be used as a trade-off over time. This study shows the processing time and the accuracy percentage differences between the selected variables of chosen dataset. However, the second used algorithm in this study is Naïve Bayes. It is a probability classification based on the theory of Bayes. All classifiers of Naïve Bayes suppose that the value of any given feature is an independent value of any other value, given the variable category. Bayes theory is:

$$P(C|X) = P(X|C) * P(C)/P(X) \tag{2}$$

Where X is the dataset, C refers to the class so that P(X) is a constant for all classes. Although it supposes an unrealistic condition that attributes values be conditional independent, they work amazingly well in big data sets where this condition is supposed and suspended.

The datasets are collected and gathered from the Machine Learning Repository (UCI). It now upholds 394 datasets copies with 14 attributes those names are sex, age, chest pain type, resting blood pressure, resting electrocardiographic results, fasting blood sugar>120 mg / dl, serum cholesterol in mg/dl, exercise induced angina, maximum heart rate achieved, the slope of the peak exercise ST segment, oldpeak = ST depression caused by exercise relative to rest, number of main vessels (0-3) colored by fluoroscopy, thal: 7 = reversible defect; 6 = fixed defect; 3 = normal. These features are used as a service package to the community of machine learning (MLC). There are three data bases in the Data Set of heart disease, these data bases namely Cleveland, Hungary, Switzerland.

IV. Result and Discussion

In this study, three data sets of heart diseases are first addressed to pre-processing of data to address missing values. The Heart Disease Group in Switzerland contains 123 cases with 14 features. In this data set, the chol attribute contains 99% of the lost values, while the ca attribute contains 95% of the missing values, while the fbs attribute contains 61% of the lost values, as it is shown in Table I.

In this paper, more than 60% of the lost values are subtracted to be removed. Therefore, the three attributes of the data set are removed, as it shown in Fig. 1.

Similarly, the Hungarian Heart Disease Data Collection contains 294 cases with 14 features. In this data set, the slope attribute contains 64% of the lost values, and the ca attribute contains 99% of the lost values and the thal attribute contains 90% of the lost values, as it shown in Fig. 2.

Thus, more than three attributes are exposed to removal from the data set and some of the features contain less than 60% of the lost values.

The other method of selecting the Recursion Mode provides a subset of features that produce an accurate result. The RF algorithm is used at each frequency to assess the form. The algorithm is designed to find out all of the probable subsets of attributes. It has given a set of Cleveland data set features for thal, Ca, thalach, slope, oldpeak, exang, cp and sex. Similarly, it has given oldpeak, trestbps, cp, thal, thatlach, sex, exang, restecg and slope from the switzerland data set. Similarly, exang, oldpeak, cp, thalach, and sex were given from the Hungarian data collection of the classification. In pre-processing, 3 attributes are eliminated from Switzerland and also 3 attributes are removed from Hungarian datasets. In this research, two methods for selecting the feature, such as filter mode, are applied - the attributes associated with a high degree of removal and wrapping are identified. The way to remove the recurrence feature determines the best attributes for the classification. In the filtering method, a link matrix is generated from these attributes and high-linked attributes are selected to remove them. Thalach, Age, Thal, Slope, Exang, and oldpeack were identified from the Cleveland data set as greatly correlated and therefore could be removed. In general, remove and eliminate the absolute correlation attributes of 0.75 or greater, as it shown in Fig. 3. The other way of selecting the Recursion Mode provides a subset of features that produce an accurate result. The RF algorithm is used at each frequency to assess the form. The algorithm is designed to find out all of the probable subsets of attributes, as it shown in Table II. Ka, Thal, Oldbeck, CB, Thalach, Xanga, Slope and Sex have given a subset of features from the Cleveland Data Collection to the classification, as it shown in Fig. 4.

TABLE I. Evaluation Measures of Random Forest and Naïve Bayes with Eliminating Redundant Features and Preprocessing approaches (by using Filter Method)

Datasets	No of instances (NB / RF)	No of attributes (NB / RF)	Accuracy		Precession		Recall	
			NB	RF	NB	RF	NB	RF
Cleveland	296	7	98%	54%	0.9850	0.2888	0.9660	0.2766
Switzerland	123	5	84%	29%	0.7493	0.1731	0.6489	0.1923
Hungarian	294	7	95%	72%	0.9691	0.7110	0.9747	0.7146

*(NB) Naïve Bayes** (RF) Random Forest

TABLE II. Evaluation Measures of Random Forest and Naïve Bayes with Eliminating Redundant Feature and Preprocessing approaches (by Using Wrapper Method)

Datasets	No of instances (NB / RF)	No of attributes (NB / RF)	Accuracy		Precession		Recall	
			NB	RF	NB	RF	NB	RF
Cleveland	296	8	100%	57%	1.000	0.3452	0.9660	1.000
Switzerland	123	9	92%	44%	0.9468	0.2215	0.6489	0.8095
Hungarian	294	5	100%	82%	1.000	0.8139	0.8139	1.000

*(NB) Naïve Bayes ** (RF) Random Forest

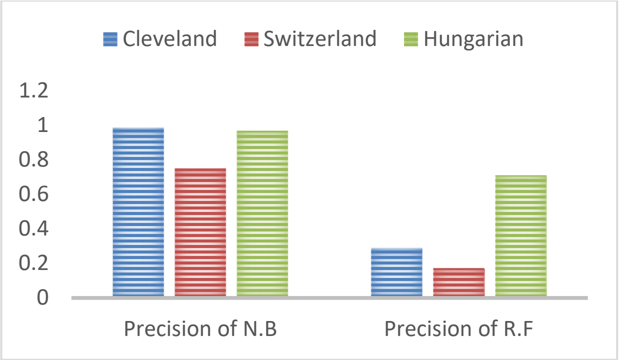

Fig. 1. Accuracy of Random Forest and Naïve Bayes with Eliminating Redundant Features and Preprocessing Approaches (by using Filter Method).

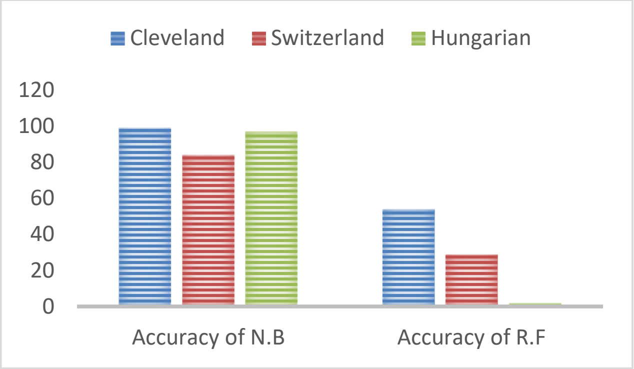

Fig. 2. Precession of Random Forest and Naïve Bayes with Eliminating Redundant Features and Preprocessing Approaches (by using Filter Method).

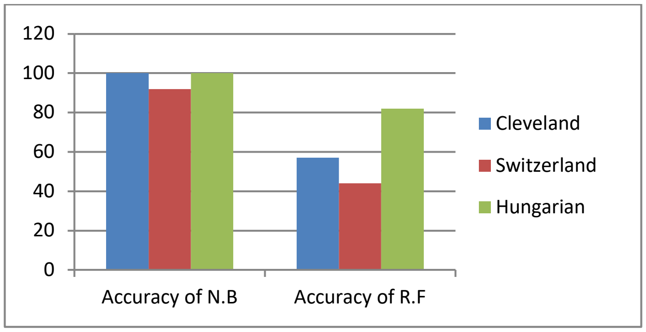

Fig. 3. Accuracy of Random Forest and Naïve Bayes with Eliminating Redundant Features and Preprocessing Approaches (by using Wrapper Method).

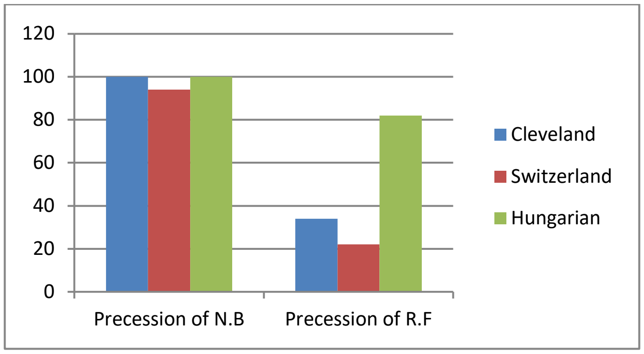

Fig. 4. Precession of Random Forest and Naïve Bayes with Eliminating Redundant Features and Preprocessing Approaches (Wrapper Method).

The following remarks are implemented between the primary data measures and post-pretreatment measures:

- For filter method when comparing Naïve Bayes with Random Forest, the obtained results indicate that Naïve Bayes gives more accurate results.
- For the feature selection wrap method, when comparing Naïve Bayes with Random Forest, the obtained results indicate that Naïve Bayes gives more accurate results.

From the observations above, we find that the Naive Bayes algorithm by using Wrapper method is appropriate for the classification of Hungarian, Switzerland, and Cleveland data sets of the Heart Disease Group.

V. Conclusion

In this paper, three datasets were used (Cleveland, Hungarian, and Switzerland) Heart Disease dataset, they are used for prediction and classification of heart diseases. In the data set, some significant features may contain lost values which may give it an effect on the superiority of the data set. Filling out lost value settings and features may be one of the most important steps in pretreatment.

After parsing the data set, the lost values are determined and changed with the average of the chapter. The subsequent process is to convert the data using the Max-Min normalization technique and then various approaches are applied to select the feature to frame the subset of the important properties of the classification, Recursive Feature Correlation and Elimination. Experimental result shows that dealing with lost values and feature selection approaches significantly boosts the classification's accuracy. The performance of both feature selection processes is evaluated with Naïve Bayes and Random Forest classifiers. Naïve Bayes method of Recursive Feature Elimination gives a better advantage to the three data sets of heart disease in terms of accuracy.

VI. Future Work

In the future work, various hybrid algorithms for optimization can be implemented as well for comparative analysis of several classification methods in addition to the possibility of using them as parameters in remote monitoring of patients by using the technology of M2M (machine-to-machine), particularly for patients those treated at remote clinics or home.M2M will be built then adding and embedding a prediction system as a new feature from one party to another.

References

[1] C. S.Dangare and S. S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications, vol. 47, no. 10, pp. 44-48, 2012. Available: 10.5120/7228-0076.

[2] C. S.Dangare and S. S. Apte, “A Data Mining Approach for Prediction of Heart Disease Using Neural Networks,” nternational Journal of Computer Engineering and Technology (IJCET), vol. 3, no. 3, pp. 30–40, Oct. 2012.

[3] T. T. Abiraami and A. Sumathi, “Analysis of Classification Algorithms for Diabetic Heart Disease,” International Journal of Pure and Applied Mathematics, vol. 118, no. 20, pp. 1925–1934, 2018.

[4] U. Garciarena, R. Santana, and A. Mendiburu, “Evolving imputation strategies for missing data in classification problems with TPOT,” arXiv, vol. 2, Aug. 2017.

[5] T. H. M. Prerana , N. C. Shivaprakash N, and N. Swetha, “Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS,” International Journal of Science and Engineering, vol. 3, no. 2, pp. 90–99, 2015.

[6] B. Bahrami and M. H. Shirvani, “Prediction and Diagnosis of Heart Disease by Data Mining Techniques,” Journal of Multidisciplinary Engineering Science and Technology (JMEST) , vol. 2, no. 2, Feb. 2015.

[7] R. Cincy, E. Philipsy, C. Siji, L. P. Suresh, and S. dE. E. P. A. Rajan, “A Survey on Predicting Heart Disease using Data Mining Technique,” Proc. IEEE Conference on Emerging Devices and Smart Systems (ICEDSS 2018)., 2018.

[8] D. Dua and T. Karra, Archive.ics.uci.edu, 2019. [Online]. Available: http://archive.ics.uci.edu/ml. [Accessed: 10- May- 2019].

[9] K. A. Enriko, M. Suryanegara, and D. Gunawan , “Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters,” Journal of Telecommunication, Electric, and Computer Engineering, vol. 8, no. 8, pp. 59–65, 2016.

[10] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, and D. Steinberg, “Top 10 algorithms in data mining,” Knowledge and Information Systems, vol. 14, no. 1, pp. 1–37, 2008.

[11] I. Rish, “An empirical study of the naive Bayes classifier,” IJCAI 2001 workshop on empirical, 2001.

[12] D. Lavanya and K. U. Rani, “Ensemble decision tree classifier for breast cancer data,” International Journal of Information Technology Convergence and Services (IJITCS), vol. 2, no. 1, pp. 17–24, 2012.

[13] L. E. Peterson, “K-nearest neighbor,” Scholarpedia, vol. 4, no. 2, p. 1883, 2009.

[14] M. A. Jabbar, B. L. Deekshatulu, and C. Priti , “Prediction of risk score for heart disease using associative classification and hybrid feature Selection,” . IEEE ISDA 2012, pp. 628–634, 2012.

[15] A.S. Hameed, B.M. Aboobaider, N.H. Choon, M.L. Mutar, and W. H. Bilal,. “Review on the Methods to Solve Combinatorial Optimization Problems Particularly : Quadratic Assignment Model,” Int. J. Eng. Technol., vol. 7, pp. 15–20, 2018.

[16] M. L. Mutar, M. A. Burhanuddin, A. S. Hameed, N. Yusof, M. F. Alrifaie, and A. A. Mohammed, “Multi-objectives ant colony system for solving multi-objectives capacitated vehicle routing problem,” J. Theor. Appl. Inf. Technol., vol. 98, no. 24, pp. 4014–4027, 2020.

[17] A.S. Hameed, B.M. Aboobaider, N.H. Choon, M.L. Mutar, ‘Improved Discrete Differential Evolution Algorithm in Solving Quadratic Assignment Problem for best Solutions’, International Journal of Advanced Computer Science and Applications, 9(12), pp. 434–439,2018. doi: 10.14569/ijacsa.2018.091261.