Development of Prediction Model for Heart Disease by Combining Clustering and Classification Techniques

Reenah K Uthama Seelan[1], Ganthan Narayana Samy[2], Mahiswaran Selvananthan[3], Nurazean Maarop[4], Sundresan Perumal[5] & David Lau Keat Jin[6]

[1,2,4,6]Advanced Informatics Department, Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, 54100, Malaysia
[3]Perdana Department, Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, 54100, Malaysia
[5]Faculty of Science and Technology, Universiti Sains Islam Malaysia, Nilai, 71800, Malaysia

reenah@graduate.utm.my

Article history

Received:
26 Oct 2023

Received in revised form:
10 Nov 2023

Accepted:
16 Nov 2023

Published online:
18 Dec 2023

*Corresponding author
reenah@graduate.utm.my

Abstract

The concerning trends in deaths related to heart disease some measures need to be in place to ensure early treatment and diagnosis of the disease. Therefore, one of the way can be done is by leveraging the abundance of medical data available. Advancement in technology today has improved the availability and accessibility huge amounts of valuable data and it only makes sense for us to explore the opportunities that lie in the data that could possibly save lives and reduce costs. Thu, this study aims to do that with the help of classification and clustering data mining techniques to predict heart disease based on some key indicators of the disease. Studies show that applying classifiers on clustered data can improve the performance of algorithms. Hence, this method will be explored in this study using the Naïve Bayes, Decision Tree and Random Forest classifiers together with both K-Means Clustering and Density-Based Clustering on the data analysis using tool WEKA. The performance of the each model will be measured and compared against each other using accuracy, precision, recall, specificity, AUC and model build time. Thus, this paper will focused on development of prediction model for heart disease by combining clustering and classification techniques in detail.

Keywords: *Heart Disease, Clustering and Classification Techniques, Prediction Model Development*

1. Introduction

The advancement of technology throughout the years, we are able to access and store abundance of patient data. These data can consist of patient's medical history, physician notes, clinical reports, biometric data and other medical data related to health [1]. Without doubt, the analytics of healthcare big data has a lot of potential in adding value to the healthcare system as a whole. Big data analytics in healthcare, if done efficiently could lead to an annual savings of over 25% in the years to come [2]. With the presence of large amounts of valuable healthcare data, there lies huge opportunity for discovering patterns and trends within the data which could increase the potential to improve care, save lives and lower costs [3]. One way the big data in healthcare can help in the battle against heart diseases is through predictive analysis. Numerous studies have been conducted in predicting heart disease with a focus on identifying the most accurate and efficient models as well significant attributes that should be used to build the prediction models. Several data mining techniques and algorithms have been explored in coming up with a heart disease prediction model.

There were a few studies which used clustering techniques in the predictions [4][5][6] but classification techniques are a more popular and preferred technique in the prediction of heart disease [7][8][9][10][11][12][13]. All the models resulting from these studies have been evaluated with various metrics to measure the performance and accuracy of the models and numerous algorithms with high performance and accuracies were identified. However, there are very limited studies on the use of clustering techniques together with classification techniques in the prediction of heart disease. Researchers have found that combining the two techniques together could lead to improvement in the performance of classification algorithms [14][15][16][17]. Therefore, a prediction model combining clustering techniques and classification techniques could be developed for the prediction of heart disease as an effort towards improved accuracy and efficiency.

2. Literature Review

This section will discuss about definition of key concepts, explanation about clustering and classification techniques and related works.

2.1. Definition of Key Concepts

2.1.1. Heart Disease: Heart disease and cardiovascular disease are often used interchangeably; however, they are not the same. Cardiovascular disease refers to all medical conditions that affect the heart or blood vessels. Heart disease on the other hand is a type of cardiovascular disease and is a commonly used term to refer to coronary heart disease also known as coronary artery disease. This disease is caused by plaque build-up in the artery walls. The build-up of plaque could narrow the artery over time and reduce the amount of oxygenated blood to the heart. The plaque could also lead to the blockage of blood flow, causing a heart attack. There are many risk factors that are usually associated with the disease which include high blood pressure and cholesterol levels, unhealthy diets, high stress levels, smoking as well as the lack of physical activities [18].

2.1.2. Prediction: Prediction is one of the goals in data mining and refers to using attributes or variables in a dataset to predict unknown or future values or variables of interest. Predictive models can be used to forecast values or variables based on patterns identified from a set of data. In this research, the prediction of heart disease based on a selected set of attributes will be tested using different clustering and classification techniques [19].

2.2. Clustering and Classification Techniques

Both clustering and classification are data mining techniques which is the process of extracting useful information from vast amounts of data [20]. Data mining is valuable in the health sector as hidden patterns, anomalies and correlations can be identified in medical data which could be used for better diagnosis and treatment of patients. Clustering is an unsupervised data mining technique which refers to the grouping of data into groups so that the data share similar characteristics, trends and patterns [21]. The goal of the algorithm is to identify and segregate all similar sets of data systematically. Classification is a supervised learning data mining technique and is used to classify each item in a dataset into a predefined set of classes in a training dataset. A classifier algorithm analyses the training data set and predicts the class labels for each item where the goal is to correctly predict the class label for each item in the dataset [22].

2.2. Related Works

Various studies have been conducted throughout the years on the prediction of heart disease using data mining techniques. [11] uses WEKA in the prediction of heart disease on the dataset obtained from University of California Irvine data repository which contains 14 features and 270 samples. The indicator for the presence of heart disease is based on cardiac catherization, where a diameter narrowing of more than 50% is diagnosed as having heart disease. Summary descriptive statistics of the dataset was generated using SPSS. For feature selection, Pearson's correlation is used for numerical data and Chi Squared attributes evaluation is used for categorical data so that the data can be narrowed down where attributes with smaller correlations were removed. The classification algorithms used on the dataset were KNN, Linear SVM, Naïve Bayes, J48, Ada Boost, Bagging, Stacking and Bayesian Network where k-folds of three, five and 10 were used. The study found that Naïve Bayes produced the highest accuracy of 87.41% with k-fold of three. Stacking has the worst performance at 55.56% which indicates that it might not be suitable for the dataset where the effect of k-fold 5 is worse than k-fold 3 and 10. Bayesian Network also resulted in good accuracy of at 83.70%. Bayesian Network and J48 can also provide useful insights using WEKA's visualization.

[13] have studied the prediction of heart disease using various tools and algorithms based on user experience in data mining. The dataset used in this research is the Cleveland heart disease dataset from the UCI Machine Learning Repository. The data set contains 303 instances with 13 input features and one output feature. The six data mining tools used in this research are Orange, WEKA, RapidMiner, Knime, Matlab and Scikit-Learn. Using each of these tools the algorithms, Logistic Regression, SVM, KNN, ANN, Naïve Bayes and Random Forest were applied to the dataset. The 10-fold cross validation technique was used to sample the dataset. As for the evaluation of the performance, accuracy, sensitivity

and specificity performance measures are extracted and compared. From the research it was found that ANN is the best model for heart disease classification and was the best in terms of accuracy and sensitivity on the Matlab tool. RapidMiner's SVM was the most specific model with 94.38% while Knime's KNN model had the lowest accuracy of 63.64%, and was the least sensitive with 56.93%, and the least specific with 69.38% concurrently with RapidMiner.

The study [23] acknowledges that the diagnosis and prediction of heart disease requires a higher degree of precision and calculates the accuracy of a few different machine learning algorithms in the prediction. Python was used on Jupyter Notebook to work with the heart disease data from the UCI repository. The dataset then underwent attribute selection, the replacement of null values and the labelling of numerical data. Since the heart disease dataset is an imbalanced dataset, the data is balanced to produce more accurate results. Linear Regression, Decision Tree, SVM and KNN algorithms were used and the accuracy was calculated based on the output of the confusion matrix. The study found that K-Nearest Neighbour or KNN is the best performer with 87% followed by Support Vector Machine (SVM).

[4] focused on identifying the best algorithm in heart disease prediction. The paper compares and discusses different unsupervised clustering algorithms which include Simple K-means, Hierarchical Clustering, OPTICS, Filtered Cluster and Farthest First. The said algorithms were tested on a heart disease dataset obtained from the UCI Machine Learning Repository with 303 samples, 14 input features and one output feature. The performance of each algorithm was evaluated and compared using the time taken to assemble each of the clusters. The study found that Filtered Cluster and Farthest First Algorithms took the least amount of time to assemble the clusters both at 0.02 seconds followed by Simple K-Means taking up 0.05 seconds. The algorithms that performed the worst was OPTICS and Hierarchical Clustering at 0.22 seconds and 0.23 seconds respectively.

[24] aims to discover valuable patterns and information that will be beneficial in clinical diagnosis. To realise this, a Smart Heart Disease Prediction (SHDP) models is built to predict risk factors that are often associated with heart disease. The study is also focused on identifying an approach that is cost efficient and effective. With the dataset from the UCI repository, the model is built using Sequential Minimal Optimisation (SMO), Bayes Net, ANN Multilayer Perceptron and Naïve Bayes. The performance of the models was then evaluated using accuracy and build time. The study found Naïve Bayes had the best performance with an accuracy of 89.77% and a build time of 0.01 seconds. This is followed by Sequential Minimal Optimisation (SMO) with an accuracy of 84.07% and build time of 0.02 seconds.

[25] acknowledges that there is a vast amount of data that need to be explored in the field of healthcare and choose to narrow down their focus to predict the possibilities of heart disease occurring in patients. They aim to do this using python. First, the UCI Heart Disease dataset is split into training and testing data and then the pre-processing stage. After that Decision Tree and Naïve Bayes algorithms are applied to the dataset. This study found Decision Tree to be the better performer with an accuracy of 91%.

3. Methodology and Model Development Discussion

The procedures in this research have been divided into four phases which are (1) Initial Phase, (2) Data Collection & Preparation, (3) Building the Model and (4) Evaluation and Reporting as illustrated in Figure 1. Each phase will be discussed in detail in this section.

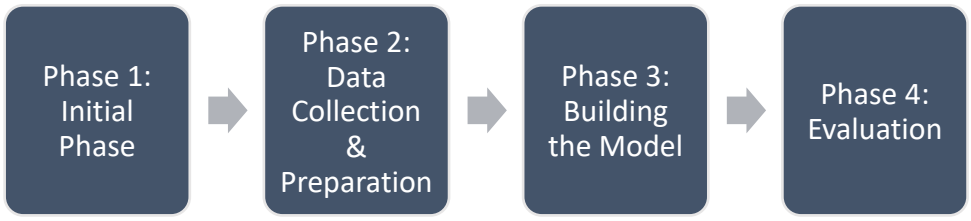

Figure 1. Research Design and Procedure

3.1. Phase 1: Initial Phase

In the initial phase of this research, literature review was conducted on the related work for heart disease prediction using clustering and classification data mining techniques. From the studies reviewed, the methods and outcomes in the predictions were synthesized and compared. The synthesis from the literature will allow a smoother process for the identification of gaps. Based on the literature and other studies viewed, the prominence of studies related to the prediction of heart disease is rather popular. The classification technique was also used more often as compared to clustering technique and can be attributed to the fact that classification is a supervised learning technique where the input and possible outputs are known. Clustering on the other hand merely groups data into clusters such that the similarities among data members within the same cluster are maximal while similarities among data members from different clusters are minimal [15].

3.2. Phase 2: Data Source & Preparation

3.2.1. Data Source: The data for this study which was obtained from Kaggle is originally a dataset from the Centre of Disease Control in the USA more commonly known as the CDC. The data is part of an annual telephone survey to gather data on the health status of American citizens conducted in the year 2020. The dataset contains responses from survey respondents about their health status with 319,795 records and 18 attributes as illustrated in Table 1. Of the 18 attributes, 17 are input attributes and one is an output attribute. Basically, most of the researcher, used Kaggle dataset due to its reliability and availability [43].

3.2.2. Data Preparation: This phase is where the dataset is checked, cleaned and transformed to improve its usability and to ensure that the dataset is free from missing and invalid values. The dataset obtained from Kaggle did not have any missing or invalid values, hence the data does not need to undergo the cleansing process. For the data transformation process, all nominal data were assigned a number as a label as demonstrated in the "Value Range" column in Table 1. However, the numbers assigned to the nominal data, does not have any mathematical meaning.

Table 1. Dataset Attributes

No	Attribute	Attribute Description	Value Range	Data Type
1	HeartDisease	Reported having coronary heart disease	1: Yes 2: No	Nominal
2	BMI	Body Mass Index	12 – 94.8	Numerical
3	Smoking	Smoked at least 100 cigarettes in entire lifetime	1: Yes 2: No	Nominal
4	AlcoholDrinking	Having more than 14 drinks per week for men and more than 7 drinks per week for women	1: Yes 2: No	Nominal
5	Stroke	Have experienced a stroke	1: Yes 2: No	Nominal
6	PhysicalHealth	Frequency of physical illness and injury in the past 30 days	1 – 30	Numerical
7	DiffWalking	Difficulty walking or climbing stairs	1: Yes 2: No	Nominal
8	Sex	Gender of respondent	1: Male 2: Female	Nominal
9	Age	Age-Category	1: 18-24 2: 25-29 3: 30-34 4: 35-39 5: 40-44 6: 45-49 7: 50-54 8: 55-59 9: 60-64 10: 65-69 11: 70-74 12: 75-79 13: 80 or older	Ordinal
10	Race	Race or ethnicity	1: White 2: Hispanic 3: Black 4: Other 5: Asian 6: American Indian/ Alaskan Native	Nominal
11	Diabetic	Has been diagnosed with diabetes	1: Yes 2: No	Nominal
12	PhysicalActivity	Physical activity in the past 30 days other than regular job	1: Yes 2: No	Nominal
13	GenHealth	Rating of respondent's general health	1: Excellent 2: Very Good 3: Good 4: Fair 5. Poor	Nominal
14	SleepTime	Average hours of sleep in 24 hours	1-24	Numerical
15	Asthma	Diagnosed with asthma	1: Yes 2: No	Nominal
16	KidneyDisease	Diagnosed with kidney disease	1: Yes 2: No	Nominal
17	MentalHealth	Poor mental health in past 30 days	0-30	Numerical
18	SkinCancer	Diagnosed with skin cancer	1: Yes 2: No	Nominal

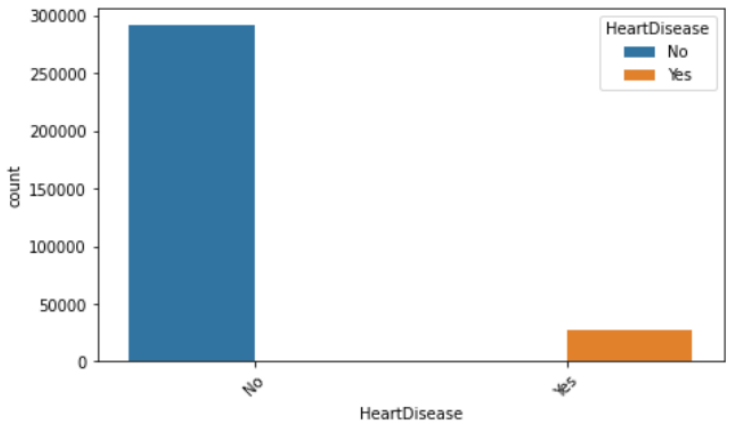

Figure 1. Composition of Heart Disease Patients in the Dataset

The bar chart in Figure 1 shows us the composition of 'Yes' and 'No' responses on whether or not the respondent has heart disease in the dataset used. It is evident that the dataset in use is imbalanced as the number of records which has 'Yes' for heart disease is very much lower compared to the number of records which has 'No' for heart disease. This can be problematic when building a prediction model, as the accuracy of the model is more representative of identifying the majority class. The model might be successful in predicting the majority class but not as successful in predicting the minority class and yet the accuracy will be high due to the imbalanced dataset. Hence, this will be handled by applying undersampling to the dataset. Undersampling will delete or merge samples in the majority class so that the majority and minority class are balanced. The original records in the dataset of 319,795 had been reduced to only 54,746 by applying undersampling.

Once undersampling has been applied, the data is the split to training and test sets through cross validation. This step is essential to prevent the overfitting of data. Overfitting occurs when a prediction model fits exactly against its training data. The train dataset is used to fit to the model and the test dataset is the input for the model to produce predictions. If the same datasets are used for both training and test, the accuracy of the predictions would be at a 100%.In order to use the dataset in the mining tool called WEKA, the dataset was converted from a .csv file to .arff using the ArffViewer tool that is available on WEKA.

3.3. Phase 3: Building the Model

3.3.1. **Classification and Clustering:** Classification is used to classify each item in a set of data into a predefined set of classes or groups and is a supervised learning algorithm as well as predictive algorithm [22]. Classification involves predicting a certain outcome based on input which is already given [26]. Classification is often a two-step process where the classification algorithm is first applied on training data and then followed by testing a predefined set of data using the extracted model from the first step to evaluate the performance of the model trained [27]. The classification algorithms that will be used in this study are Decision Tree, Naïve Bayes and Random Forest classifiers. The classifiers will be applied on the dataset which has been clustered. Clustering on the other hand is a descriptive data mining algorithm and an unsupervised learning algorithm [26]. The main purpose of a clustering algorithm is to classify the data into groups that share similar features [28] and to cluster data which are different from each other into separate clusters. Clustering is considered to be more difficult than supervised classification since with clustering there are no predefined labels associated with patterns [29]. The clustering algorithms used in this study will be K-Means Clustering and Density-Based Clustering. These clustering algorithms will be applied to the dataset before the classifiers are applied.

3.3.2. Algorithms Used: As mentioned in the previous section, three classifiers will be used in this study. They are Decision Tree, Naïve Bayes and Random Forest. Each of these algorithms will be used together with K-Means Clustering and Density-Based Clustering. The algorithms are explained below:

3.3.2.1 Decision Tree: A decision tree is made up of decision nodes and leaf nodes where each decision node corresponds to a test X over a single attribute of the input data. The decision node each have branches which handles an outcome of the test X. The leaf nodes on the other hand represent a class or the "final decision" (Stein et al., 2005). It is a decision support system that uses a tree-like graph decision and their possible consequences [30] learned trees can also be represented as sets of if-then rules to improve comprehension of the user [31]. The decision tree is highly desired algorithm and has been widely used in various real-life cases [32].

3.3.2.2 Naïve Bayes: Naïve Bayes is a supervised learning algorithm that uses the Bayes' theorem with a strong assumption that the attributes are conditionally independent between every pair of features. Despite its "naïve" assumptions, the algorithm is still a popular option due its computational capabilities and low variance [33]. The three types of Naïve Bayes common Naïve Bayes Classifiers are Multinomial Naïve Bayes, Bernoulli Naïve Bayes and Gaussian Naïve Bayes.

3.3.2.3 Random Forest: Random Forests are a collection of classification and regression trees which use binary splits for prediction [34]. In simple terms, random forest is a machine learning algorithm that combines multiple decision trees. Each tree in random forest is trained on a randomly selected dataset and random features. The predictions of the individual trees are then combined to make the final prediction. Because of this, Random Forest is also known to be one of the best-performing algorithms [35]. The algorithm is also beneficial due to its ability to handle large datasets [36].

3.3.2.4: K-Means Clustering: K-means Clustering is a numerical, unsupervised, non-deterministic iterative method [4]. It is a portioning clustering algorithm where the algorithm clusters data points into a fixed number of clusters through an iterative process, converging to a local minimum, resulting in compact and independent clusters [16]. The level of similarity between members in a cluster is measured by the proximity of the object to the mean value on the cluster or usually referred to as

the centroid cluster. This is a popular algorithm due to its ability to group a large amount of data efficiently in a comparatively fast build time [37].

3.3.2.5: Density Based Clustering: Density-based clustering is a method that find clusters of various shapes and sizes based on density of points in a given space. This is in contrast with other algorithms which assume spherical and convex shapes for clusters. Clusters are identified in this algorithm as spaces with higher density than the rest of the data space [38]. One of the benefits of this algorithm is that it works efficiently with the presence of noise in the data [39].

3.3.3. Building the model using Weka:

3.3.3.1. Models Built: The model will be built on Weka using the Explorer interface. For this study cross-validation of 10 folds will be used to split the training and test data. A total of 9 models will be built as demonstrated in Table 2. It includes three of the standalone classifier models and six models which combined clustering and classification.

Table 2. Models Built

No	Model
1	Decision Tree
2	Naïve Bayes
3	Random Forest
4	K-Means Clustering + Decision Tree
5	K-Means Clustering + Naïve Bayes
6	K-Means Clustering + Random Forest
7	Density-Based Clustering + Decision Tree
8	Density-Based Clustering + Naïve Bayes
9	Density-Based Clustering + Random Forest

3.3.3.2. Steps to build basic models: Decision Tree, Naïve Bayes and Random Forest.

1. Load the dataset into Weka's Explorer interface.
2. Select the "Classify" tab and choose the Naïve Bayes classifier.
3. The “HeartDisease” variable is used as the target variable.
4. Click "Start" to build the model.
5. Steps i to iv are repeated for the Decision Tree and Random Forest classifiers.

3.3.3.3. Steps to build combined models of clustering algorithms and classifiers

1. Apply K-Means Clustering as a filter on the “Preprocess” tab on Weka.
2. Navigate to the “Classify” tab and select Naïve Bayes.
3. Cluster labels generated from preprocessing will be used as the target variable.
4. Click “Start” to build the model.
5. The steps are repeated for Decision Tree and Random Forest classifiers.
6. Steps i to v are repeated by replacing K-Means Clustering with Density-based Clustering.

3.4. Phase 4: Evaluation

In this phase, the performance of the prediction models built will be measured and compared against each other. The evaluation metrics that will be used include accuracy, build time, precision, specificity, sensitivity, recall and AUC. The summary of the evaluation metrics is enlisted in Table 3. Accuracy, precision, specificity, recall and sensitivity can be derived from the confusion matrix which

will be generated by WEKA when the model is tested on the dataset. The components of the confusion matrix are demonstrated in Table 4.

Accuracy is the most commonly used evaluation metric in any study evaluating classifiers. Accuracy measures the ratio for the number of correct predictions against the total number of predictions made. While it seems like a straightforward and simple evaluation metrics that could be easily understood by everyone, the metric does not come without its weaknesses. The weaknesses include less distinctiveness, less discriminability, less informativeness and bias to majority class data [40].

AUC or area under the curve was introduced as an evaluation metrics in this study because it was unpopular amongst the literature reviewed. The AUC evaluation metric reflects the overall performance of a classifier [40]. AUC is also deemed to be a better measure than accuracy both theoretically and empirically and is found to be more discriminating that the accuracy metric [41]. AUC is also a more suitable evaluation metric when dealing with highly imbalance datasets in comparison with accuracy [42].

Build time is another evaluation metric used which is looking into efficiency aspect of the classifiers in terms of how long it takes for each model to be built. The lesser time it takes to build the model, the better it is. The lesser time taken means that the model will be a good option to be deployed in real-life scenarios. However, the performance (Accuracy, precision, recall, specificity, AUC) of the model is not discussed detail in this paper because this paper focused on development of prediction model.

Table 3. Evaluation Metrics

No	Metrics	Formula	Definition
1	Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Measures ratio of correct predictions over total number of instances
2	Precision	$\frac{TP}{TP + FP}$	Measures the ratio of correct positive identifications from total positive identifications
3	Recall	$\frac{TP}{TP + TN}$	Measures ratio of positive identifications which were classified correctly
4	Specificity	$\frac{TN}{TN + FP}$	Measures ratio of negative identifications which were correctly classified
5	AUC	$\frac{S_p - n_p(n_n + 1)/2}{n_p n_n}$	Measures overall ranking performance of a classifier
6	Build Time		Time taken to build model

Table 4. Components of the Confusion Matrix

Component	Definition
True Positive (TP)	When patients predicted to have heart disease actually have heart disease
True Negative (TP)	When patients predicted to not have heart disease actually do not have heart disease
False Positive (FP)	When patients predicted to have heart disease do not actually have heart disease

False Negative (FN)	When patients predicted to not have heart disease actually have heart disease

4. Conclusion

By leveraging the information and insights from the literature review, we developed six different combination models of clustering and classification algorithms. The models built are, K-Means Clustering with Naïve Bayes, Decision Tree and Random Forest respectively and Density-Based Clustering with Naïve Bayes, Decision Tree and Random Forest respectively. The algorithms used to build the models were based on the popularity of the models from the literature review. The performance of the six models were compared against one another and also compared with the performance of the standalone models of Naïve Bayes, Decision Tree and Random Forest. However, the model with the best performance amongst all will not be discussed in this paper. Since the dataset used in this study is a heart disease dataset and is imbalanced, under sampling was applied to the dataset before the models were trained and tested. This resulted in a more balanced dataset and reduced the total records. In brief, the models built by combining clustering and classification demonstrated a promising approach towards the prediction of heart disease and further emphasized the potential of combining clustering and classification techniques.

Acknowledgement

The authors would like to acknowledge the Universiti Teknologi Malaysia for providing the facilities in writing the manuscript.

References

[1] Priyanka, K. and Kulennavar, N. (2014). A survey on big data analytics in health care. International Journal of Computer Science and Information Technologies 5(4), 5865-5868.
[2] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma and Sandeep Kaushik. (2019). Big data in healthcare: management, analysis and future prospects. Journal of Big Data 6 (54), 1-25.
[3] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems 2(1), 1-10.
[4] Kodati, S., Vivekanandam, R., and Ravi, G. (2019). Comparative analysis of clustering algorithms with heart disease datasets using data mining weka tool. Advances in Intelligent Systems and Computing, 900, 111–117.
[5] Mirpouya Mirmozaffari, Alireza Alinezhad and Azadeh Gilanpour. (2017). Heart Disease Prediction with Data Mining Clustering Algorithms. International Journal of Computing, Communication and Instrumentation Engineering, 4(1), 16-19.
[6] Reetu Singh and E. Rajesh. (2019). Prediction of Heart Disease by Clustering and Classification Techniques. International Journal of Computer Sciences and Engineering, 7(5), 861-866.
[7] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82–93.
[8] Gavhane, A., Pandya, I., Kokkula, G., and Devadkar, K. (2018). Prediction of Heart Disease Using Machine Learning. 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA). 29-31 March 2018, Coimbatore, India.
[9] Hlaudi Daniel Masethe, Mosima Anna Masethe (2014). Prediction of Heart Disease using Classification Algorithms. Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA.
[10] Singh, P., Singh, S., and Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. International Journal of Nanomedicine, 13, 121–124.
[11] Sipail, H. S., Ahmad, N., and Noor, N. M. (2021). Heart Disease Prediction Using Machine Learning Techniques. 1st National Biomedical Engineering Conference, NBEC 2021, 9 - 10 November 2021, Virtual, Online.
[12] Tarawneh, M., and Embarak, O. Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques. In Lecture Notes on Data Engineering and Communications Technologies, Springer Science and Business Media Deutschland GmbH, Vol. 29, (2019), pp. 447–454.
[13] Tougui, I., Jilbab, A., & el Mhamdi, J. (2020). Heart disease classification using data mining tools and machine learning techniques. Health and Technology, 10(5), 1137–1144.
[14] B.Madasamy and J.Jebamalar Tamilselvi (2013). Improving classification Accuracy of Neural Network through Clustering Algorithms. International Journal of Computer Trends and Technology, 4(9), 3242-3246.

[15] Deelers, Sirichai and Auwatanamongkol, Surapong (2007). Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance. International Journal of Physical and Mathematical Sciences, 1(11), 518-523.
[16] Na, S., Xumin, L., and Yong, G. (2010). Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 63–67.
[17] Zhao, S., Xiao, Y., Ning, Y., Zhou, Y., and Zhang, D. (2021). An Optimized K-means Clustering for Improving Accuracy in Traffic Classification. Wireless Personal Communications, 120(1), 81–93.
[18] NHLBI. (2021). Know the Difference - Cardiovascular Disease, Heart Disease, Coronary Heart Disease. 1st National Biomedical Engineering Conference, NBEC 2021, 9 - 10 November 2021, Virtual, Online.
[19] Srinivas, K., and Raghavendra Rao, G. (2011). Survey on Prediction of Heart Morbidity Using Data Mining Techniques. International Journal of Data Mining & Knowledge Management Process, 1(3), 14–34.
[20] Sarangam Kodati and R. Vivekanandam. (2018). Analysis of Heart Disease using in Data Mining Tools Orange and Weka. Global Journal of Computer Science and Technology, 18(1), 1-6.
[21] Arun K. Pujari. (2001). Data Mining Techniques. Universities Press (India) Private Limited. 2001. ISBN: 81-7371-380-4.
[22] Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 4-6 July 2013, Tiruchengode, India.
[23] Singh, A., and Kumar, R. (2020). Heart Disease Prediction Using Machine Learning Algorithms. 2020 International Conference on Electrical and Electronics Engineering (ICE3). 14-15 February 2020, Gorakhpur, India.
[24] Repaka, A. N., Ravikanti, S. D., and Franklin, R. G. (2019). Design and Implementing Heart Disease Prediction Using Naives Bayesian. 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). 23-25 April 2019, Tirunelveli, India.
[25] Santhana Krishnan J. and Geetha S. (2019). Prediction of Heart Disease Using Machine Learning Algorithms. 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT). 25-26 April 2019, Chennai, India.
[26] Umadevi, S., and Jeen Marseline, K. S. (2017). A Survey on Data Mining Classification Algorithms. 2017 International Conference on Signal Processing and Communication (ICSPC). 28-29 July 2017, Coimbatore, India.
[27] Nikam, S. S. (2015). A Comparative Study of Classification Techniques in Data Mining Algorithms. Oriental Journal of Computer Science & Technology, 8(1), 13-19.
[28] Mathivanan, N. M. N., Nor, N. A., and Janor, R. M. (2018). Improving classification accuracy using clustering technique. Bulletin of Electrical Engineering and Informatics, 7(3), 465–470.
[29] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T. (2017). A review of clustering techniques and developments. Neurocomputing, 267, 664–681.
[30] Mathuria, M., Alam, L., an Haq, N. F., Mamun, T., Bhargava, N., Sharma, G., and Bhargava, R. (2013). Decision Tree Analysis on J48 Algorithm for Data Mining. International Journal of Advanced Research in Computer Science and Software Engineering, 3(6), 1114-1119.
[31] S.A. Ali, N. Sulaiman, A. Mustapha and N. Mustapha (2009). K-Means Clustering to Improve the Accuracy of Decision Tree Response Classification. Information Technology Journal, 8(8), 1256-1262.
[32] A. Bar-Or, R. Wolff, A. Schuster, and D. Keren. Decision Tree Induction in High Dimensional, Hierarchically Distributed Databases. Proceedings of 2005 SIAM International Conference on Data Mining (SDM'05), April 2005, Newport Beach, CA.
[33] Claude Sammut, Geoffrey I. Webb. 92017). Encyclopedia of Machine Learning and Data Mining. Springer 2017, ISBN: 978-1-4899-7685-7.
[34] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (2017). Classification And Regression Trees. 1st Edition. Routledge. 2017. ISBN: 9781315139470.
[35] Schonlau, M., and Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal: Promoting Communications on Statistics and Stata, 20(1), 3–29.
[36] Speiser, J. L., Miller, M. E., Tooze, J., and Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications, 134, 93–101.
[37] Ahmar, A. S., Napitupulu, D., Rahim, R., Hidayat, R., Sonatha, Y., & Azmi, M. (2018). Using K-Means Clustering to Cluster Provinces in Indonesia. Journal of Physics: Conference Series, 1569, International Conference on Science and Technology 2019 17-18 October 2019, Surabaya, Indonesia.
[38] Panthadeep Bhattacharjee and Pinaki Mitra (2021). A survey of density based clustering algorithms. Frontiers of Computer Science, 15(1), 151308.
[39] Sander, J., Ester, M., Kriegel, H.-P., and Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Mining and Knowledge Discovery, 2(2), 169–194.
[40] Hossin, M., and Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1–11.
[41] Jin Huang, and Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 17(3), 299–310.
[42] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake News Detection on Social Media. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.
[43] L. Quaranta, F. Calefato and F. Lanubile. (2021). "KGTorrent: A Dataset of Python Jupyter Notebooks from Kaggle," 2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR), Madrid, Spain, 550-554.