International Journal of Synergy in Engineering and Technology
IJSET

Data Analytics Approach for Heart Disease Prediction

Ziti Fariha Mohd Apandi[1,2], Nurul Haslinda Ngah[1,2], Nurul Jannah Mahat[1]

[1]Department of Computer Science, Faculty of Computer, Media and Technology Management, University College TATI, 24000 Kemaman, Terengganu, Malaysia.
[2]Terengganu Big Data Institute, University College TATI, 24000 Kemaman, Terengganu, Malaysia.
*Corresponding author: ziti@tatiuc.edu.my

KEYWORDS

Data Analytics
Heart Disease
Machine Learning
Classification

ABSTRACT

Recent advancement in technology has cast a strong impact on the utilization of the available data in the healthcare systems. This include employing the enormous data for prediction of heart disease in patient. Nevertheless, predicting heart disease has become one of the main challenges in the medical industry due to the accuracy and performances of prediction and diagnosis issues. One of the approaches to use enormous health care data is the use of data analytics to facilitate the prediction process. However, choosing the best analytics techniques is the most important process because it will influence the prediction and diagnosis result. Thus, this study aims to apply data analytics approach in collection of databases to evaluate the performance of the techniques for prediction of heart disease. In this study, a methodology to analysis the heart disease data will presented. Five algorithms based on the data analytics technique is implemented to observe the performance in heart disease data. Based on the result the decision tree technique shows better performance compared with other techniques with the highest accuracy, 98.53%. The results show the selected features and suitable analytics technique will influence the accuracy and performance of prediction process.

Received 30 March 2023; Revised 09 July 2023; Accepted 28 July 2020; Published 31 October 2023.

1.0 INTRODUCTION

Heart diseases have become one of the leading causes of the current mortality globally [1]. As reported by the World Health Organization (WHO), heart disease accounted for an estimate of 17.9 million deaths every year, representing 32% of all global deaths [2]. This proved by the number of deaths by causes according to statistics of global burden of disease [3] and it is forecasted to occupy the same ranking up to 2030 [2]. The early prediction of the heart disease is significantly important because it can prevent complicated risk to the patient. However, predicting heart disease has become one of the main challenges in the medical industry due to the accuracy and performances of prediction and diagnosis system.

Recent advancement in technology has cast a strong impact on the utilization of the available data in the healthcare systems [4]. The healthcare data can be employed to develop a health prediction system that can improve monitoring and diagnose of heart diseases [5]. One of the approaches to use enormous health care data is the use of data analytics to facilitate the advance diagnosis of heart problem [5] [6]. Data analytics approach is a method to analysed enormous data using advanced techniques to find hidden pattern and enhance the insight of a given data to develop analysis model that can be used for prediction process, improving diagnosis, and analysing symptoms.

Choosing the best analytics techniques is the most important process of data analytics approach [5] [7]. The technique cannot efficiently determine the most effective model for diagnosis implementation without a full conceptual comprehension of each algorithm and selected of the input [6]. To overcome the issues, this study aims to apply data analytics approach in collection of different databases of heart disease data and investigate the performance of the techniques in order to predict the heart disease in patient.

In this study, a methodology to analyze the data by using different data analytics techniques is presented. Five algorithms based on the data analytics technique are implemented to observe their performance in heart disease data. Each model will evaluate to see their influence on the final decision according to its performance on the training data. Many factors such as physical examination, early symptoms, and signs of the patient that can influence the prediction of heart status were investigated.

2.0 MATERIALS AND METHOD

2.1 Dataset

The data set used in this study is the heart disease directory from UCI Machine Learning Repository collected in Kaggle website [8]. The directory consists of a data from four databases which are from Hungarian Institute of Cardiology, Budapest, University Hospital Zurich, Switzerland, Cleveland Clinic Foundation and Medical Centre, Long Beach. It contains 1025 patients' data with 76 attributes which describe the individual health factors and diagnosis of heart disease. In this study, only 14 attributes have been selected in the dataset included the target field refers to the presence of heart disease in patient as shown in Table 1.

Table 1: UCI Machine Learning Repository Dataset Attributes (Heart Disease UCI data, 2022)

Number	Attribute Name	Description	Domain of Value
1	Age	Age of the patients in years	
2	Gender	Sex of the patient	1 = Male, 0 = Female
3	Type of Chest Pain	Chest Pain Type of the patient.	1 = Typical angina, 2 = Atypical angina, 3 = Non-anginal pain, 4 = Asymptomatic
4	Resting Blood Pressure	Resting blood pressure (in mm Hg on admission to the hospital	The normal range is 120/80
5	Cholesterol	Serum cholesterol in mg/dl	
6	Fasting Blood Sugar	Fasting blood sugar larger than 120 mg/dl	1 = True 0 = False

7	Resting Electrocardiogram (ECG)	Resting electrocardiogram results	0 = Normal 1 = ST-T wave abnormality 2 = Probable of definite left ventricular hypertrophy
8	Maximum Heart Rate	Maximum heart rate achieved	
9	Exercise Induced Angina		1 = Yes 0 = No
10	OldPeak	ST depression induced by exercise relative to rest	
11	Slope	Slope of the peak exercise ST segment	1 = Upsloping 2 = Flat 3 = Down sloping
12	Major Vessels	Number of major vessels (0-3) coloured by fluoroscopy	
13	Thalassemia		3 = Normal 6 = Fixed Detect 7 = Reversable Defect
14	Target	Diagnosis of heart disease	1 = Yes 0 = No

2.2 Method

The methodology for data analytics approach will be described in the following sections and illustrated in Figure 1. In this approach, five basic steps are involved to analysis the data which is 1) Data Collection and Preparation; 2) Data Pre-Processing; 3) Feature Selection 4) Classification analysis; and 6) Evaluation as described below:

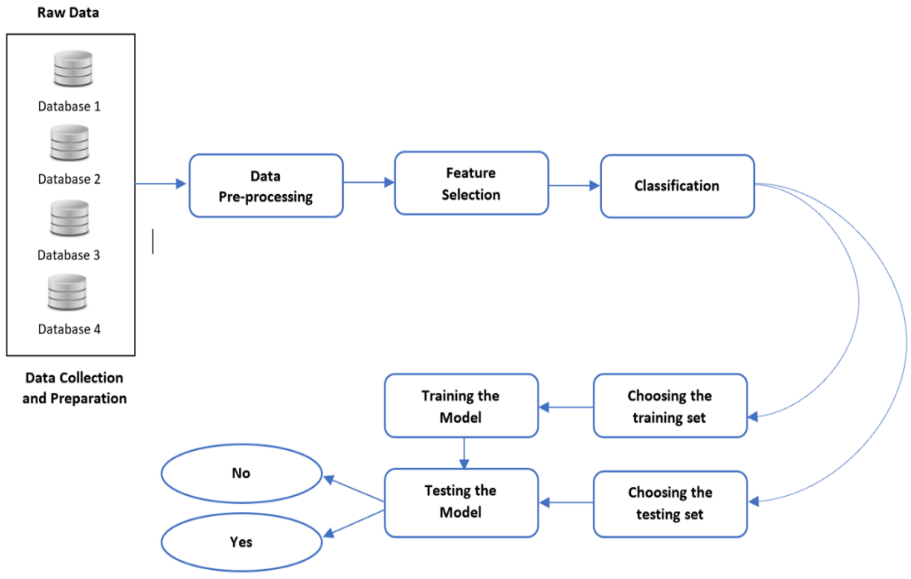

Figure 1: Methodology for Data Analytics Approach

Data Collection and Preparation

Data collection is the process of gathering and measuring information on variables of interest to evaluate the outcomes. In this section, the raw data related with heart disease is collected and prepared from the databases. The data from four database that include 76 attributes was collected for further process.

Data Pre-Processing

Data pre-processing process will be used to improve the quality of raw data. The raw data will be processed to reduce the noises and missing values that influence the analysis result. The output from pre-processing will were subsequently used in the next stages. In this study, the raw data will be filtered, and the missing values and noise is removed from the dataset.

Feature Selection

Feature selection also known as variable selection or attribute selection is the process of selecting a subset of relevant features for developing the model. The feature selection process is important in data analytics approach because the selection of the features will influence the analysis process and performance of the prediction model. In this study, 14 attributes were selected for further analysis. The selection process is done based on review from other previous work that related with heart disease [9].

Classification

Classification analysis is a data analysis task by using machine learning technique, that identifies and assigns target to a collection of data to allow for more accurate analysis. Classification can be used to make a decision or predict the pattern through the use of algorithm. In this study, five classification techniques were selected to analyse the data. To select the best technique, we tested several machine learning algorithms namely: Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbours (K-NN), Decision Tree and Naïve Bayes.

Support Vector Machine (SVM) is a pattern recognition supervised learning algorithm to classify both linear and non-linear data. The primary concept of SVM is to determine separators that can best distinguish the distinct classes in the search space [10]. The data points which are closest to the hyperplane or points of the data set that, if deleted, would change the position of dividing the hyperplane are known as support vectors. As a result, they might be regarded as essential components of the data set. The margin is the distance between hyperplane and nearest data point from either collection. SVM's main objective is to find a hyperplane in N-dimensional space which will classify all the data points.

Logistic regression is often used a lot of times in machine learning for predicting the likelihood of response attributes when a set of explanatory independent attributes are given [11]. It's a statistical technique to predict classes which are binary. It is used when the target attribute is also known as a dependent variable having categorical values like yes/no or true/false, etc. It is widely used for solving classification problems and falls under the category of supervised machine learning. It efficiently solves linear and binary classification problems and one of the most used and easy to implement algorithm.

k-NN is intuitive classification algorithm wherein a sample is classified based on a common majority rule of its k closest neighbours [12]. It is a very effective classification with limited training samples. It works better with a small number of inputs features. k-NN is a supervised machine learning algorithm. It assumes similar objects are nearer to one another. When the parameters are continuous in that case k-NN is preferred. In this algorithm it

classifies objects by predicting their nearest neighbour. It is simple and easy to implement and has high speed because of which it is preferred over the other algorithms when it comes to solving classification problems. The algorithm classifies whether the patient has disease by taking the heart disease dataset as an input.

In machine learning, decision tree is one of the well-known classification algorithms and one of the most widely used inductive learning method. It can handle training data with missing values and can handle both continuous and discrete attributes. Decision trees are built from labelled training data using the concept of information entropy [13]. Their robustness to noisy data and their capability to learn disjunctive expressions seem suitable for text classification [14].

Naïve Bayes is regarded as one of the most efficient and effective inductive learning algorithms and has been used as an effective classifier in several social media studies [10]. Since 1950s, Naïve Bayes classification for text has been commonly used in document categorization assignments and has ability to classify any type of data from text, network features, phrases, and so on [14]. This technique is a generative model, it refers to how dataset is generated based on probabilistic model. By sampling from this model, it can generate new data like the data on which the model is being trained. In our study, we used the most basic version of Naïve Bayes classifier for textual features and word embeddings.

2.3 Evaluation

The last stage of data analytics approach involved the evaluation process. The preparation of training and testing dataset for evaluation and validation need to be done in this phase. The training dataset will contribute to evaluation while testing dataset adopted for validation process. The prediction model will be trained to evaluate the detection performance. Then the new model will be tested using testing dataset to validate the performance of the model. In this study, 80% from dataset has been used for training while 20% for testing. The evaluation process is observed, and the result is presented in Section 4.0.

3.0 MODEL EVALUATION

To validate the performance, each data was categorized as true positive (TP), false positive (FP) and false negative (FN). TP denotes the total number of correctly predict a disease presence outcome, FP denotes the total number of incorrectly predict a diseases outcome, FN represents the total number of incorrectly predict a no diseases outcome and TN denotes the total number of correctly predict a no disease presence outcome. Four evaluation metrics which used such as accuracy, recall, precision, and F1-score were calculated using Equation (1) to (4), respectively [11]. The accuracy is used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input. While F1-score is used to the weighted average of precision and recall values.

		Predicted Condition	
	Total	No	Yes
Actual Condition	No	True Negative (TN)	False Positive (FP)
	Yes	False Negative (FN)	True Positive (TP)

Figure 2. The confusion matrix terminology

$$\text{Accuracy} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \times 100\% \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \quad (2)$$

$$\text{Precsion} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \quad (3)$$

$$\text{F1} - \text{Score} = 2 * \frac{\text{Recall} * \text{Precison}}{\text{Recall} + \text{Precision}} \times 100\% \quad (4)$$

4.0 RESULT AND DISCUSSION

In this section, relationship between selected attributes with heart diseases is analysed. Figure 3 shows the relationship between Age of the patient with the diagnosis of heart disease. As shown in the Figure 3, heart disease is very common among adults which composed of age group 29 and above. The highest number of patients with heart disease is from age group 54 years old. Relationship between the gender and frequency of heart disease shown in Figure 4. From the analysis of the data, both of gender have group of diagnosis heart disease and not. For the female patient, there is more group that diagnosed with heart disease compared to not. However, the result show different for the male group where the total number that diagnosis with not heart diseases is higher with the group with the disease.

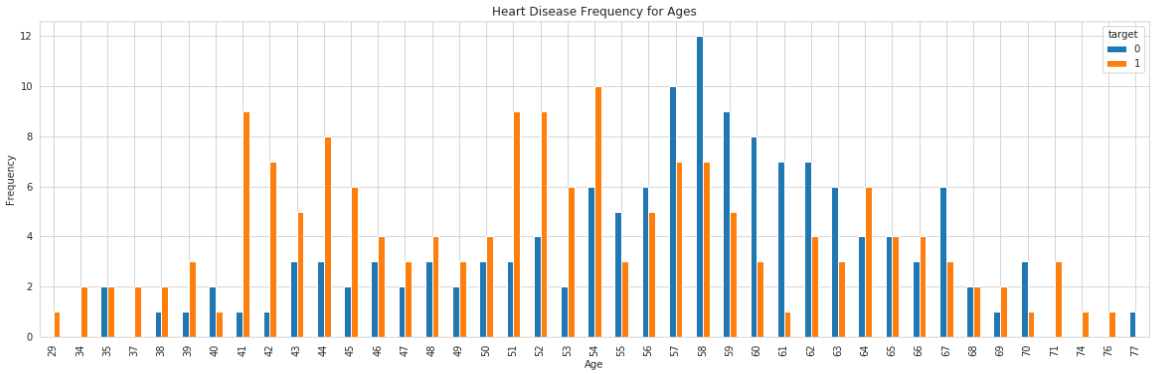

Figure 3: Relationship between heart disease frequency with Ages

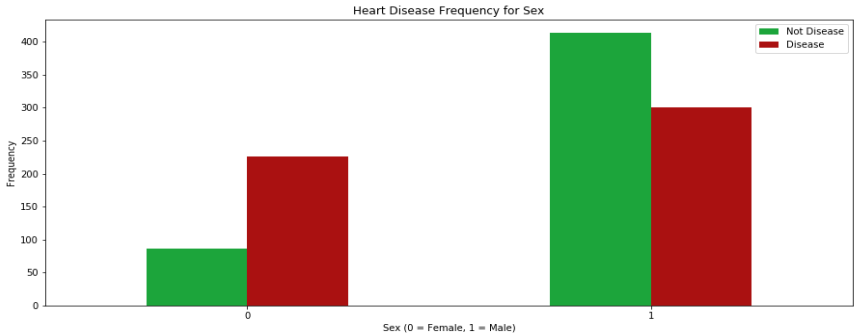

Figure 4: Relationship between heart disease frequency with Gender

Maximum heart rate is important attributes that can facilitate interpret and diagnosis heart disease. Maximum heart rates can vary from person to person and can be indicator for planning the exercise or physical work especially with heart disease patient. Age is one of the important attributes that can affects maximum heart rate [1]. Figure 5 show the relationship between maximum heart rate and age for heart disease patient. As shown in Figure 5, the maximum heart rate for the patient is 200 beats per minutes and it quite dangerous for people with age 30. As can be seen, mostly patient that have heart disease is with highest maximum heart rate is from age group 30 to 50. This shows that the maximum heart rate has high relationship between age to predict the diagnosis result.

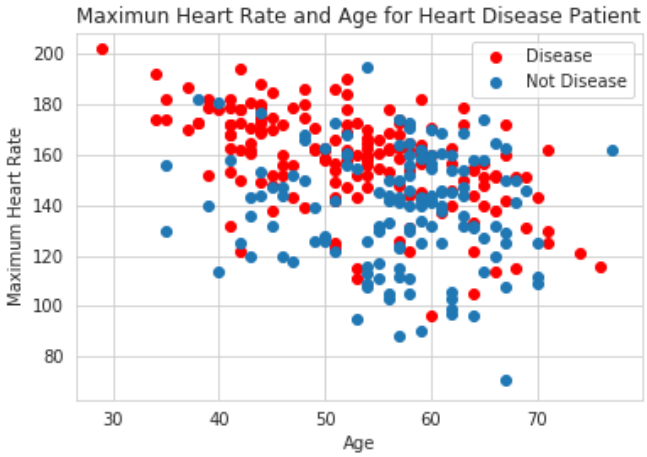

Figure 5: Relationship between maximum heart rate and age for heart disease patient

The performance results of classification technique are compared and presented in this section. As explained before, 80% from the dataset is used for evaluation and another 20% for performance validation. As shown in Table 2, five classification technique was implemented to obtain the best prediction model. Among all algorithms, the decision tree technique achieved the best performance with 98.54% accuracy compared to others. The classification report of the model shows that 97.14% of recall values that shows the prediction of heart disease was predicted correct. These results show the decision tree is suitable to analysis the heart disease dataset and the 14 features selected able to influence the decision tree model performance. The worst performances results from the analysis are SVM and Logistic Regression with both 80.98%.

Table 2: Comparative Analysis of five classification techniques

Method	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)
SVM	80.98	88.88	70.59	78.69
Logistic Regression	80.98	87.95	71.56	78.91
K-NN	84.39	86.45	81.37	83.83
Decision Tree	98.53	97.14	100	98.55
Naïve Bayes	93.65	97.09	93.66	98.50

5.0 CONCLUSION

In this study, we presented the data analytics approach for prediction of heart disease. Five basic steps were involved to analysis the data, which is data collection and preparation, data pre-processing, feature selection, classification analysis and evaluation. The heart disease dataset from four databases includes from Hungarian Institute of Cardiology, Budapest, University Hospital Zurich, Switzerland, Cleveland Clinic Foundation and Medical Centre, Long Beach has been used in this study. The analytics techniques have been implemented by using 14 features that influence the diagnosis of heart disease. The decision tree technique shows better performance compared with other techniques with the highest accuracy, 98.53%. Based on analysis, features selection has contributed to the accuracy and performances of the prediction of the classification model, thus improve the data analytics approach.

Author Contribution

Z.F.M. Apandi: Conceptualization, methodology, analysis, and writing. N. H. Ngah: reviewing, editing, and formatting. N. J. Mahat: visualisation, reviewing, editing, and formatting.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the University College TATI for support which makes this work possible.

REFERENCES

[1] Almustafa, K. M. Prediction of Heart Disease and Classifiers Sensitivity Analysis. BMC Bioinformatics, vol. 21, p. 278, 2020.

[2] World Health Organization (WHO). Cardiovascular Diseases. [Online]. 2022. Available: https://www.who.int/health-topics/cardiovascular-diseases/.

[3] Burden of disease by Cause. Our World in Data. 2019. [Online]. 2019. Available: https://ourworldindata.org/grapher/burden-of-disease-by-cause?country=~OWID_WRL.

[4] Silverio A, Cavallo P, De Rosa R, and Galasso G. Big Health Data and Cardiovascular Diseases: A Challenge for Research, an Opportunity for Clinical Care. Frontiers in Medicine, vol. 6, no. doi:10.3389/fmed.2019.00036, p. 36, 2019.

[5] Thimbleby H. Technology and the Future of Healthcare. J Public Health Res, Vols. 2, no.3, p. 28, 2013.

[6] Belle A, Thiagarajan R, Soroushmehr S. M. R, Navidi F, Beard D A, Najarian K. Big Data Analytics in Healthcare. K. BioMed Research International, no. https://doi.org/10.1155/2015/370194, 2015.

[7] Nazir S, Nawaz M, Adnan A, Shahzad S, and Asadi S. Big Data Features, Applications, and Analytics in Cardiology - A Systematic Literature Review. IEEE Access, vol. 7, pp. 143742-143771, 2019.

[8] Heart Disease UCI data. [Online]. 2022. Available: https://www.kaggle.com/ronitf/heart-disease-uci.

[9] Ranganathan I, Poongodi T, Jena S. Heart Disease Prediction using Exploratory Data Analysis. Procedia Computer Science, vol. 173, pp. 130-139, 2020.

[10] Hazra A, Mandal S K, and Gupta A. Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms. International Journal of Computer Applications, vol. 145, pp. 39-45, 2016.

[11] Jabbar M A, and Samreen S. Heart disease Prediction System Based on Hidden Naïve Bayes Classifier. International Conference on Circuits, Controls, Communications and Computing (I4C), no. doi: 10.1109/CIMCA.2016.8053261, pp. 1-5, 2016.

[12] C C. Prediction of Heart Disease using Different KNN Classifier. 5th International Conference on Intelligent Computing and Control Systems (ICICCS), vol. doi: 10.1109/ICICCS51141.2021.9432178., pp. 1186-1194, 2021.

[13] Sivakami K. Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model. International Journal of Scientific Engineering and Applied Science (IJSEAS), vol. 1(5), 2019.

[14] Maheswari S, Pitchai R. Heart Disease Prediction System Using Decision Tree and Naive Bayes Algorithm. Current Medical Imaging, vol. 15(8), no. doi: 10.2174/1573405614666180322141259, pp. 712-717, 2019.

[15] Ciu T, and Oetama R. Logistic Regression Prediction Model for Cardiovascular Disease. International Journal of New Media Technology, vol. 7(1), no. https://doi.org/https://doi.org/10.31937/ijnmt.v7i1.1340, pp. 33-38, 2020.

