Heart Disease Prediction Using Machine Learning and Data Mining Techniques: Application of Framingham Dataset

Walaa Adel Mahmoud [a], Prof. Dr. Mohamed Aborizka [a],Prof. Dr. Fathy Ahmed Elsayed Amer2[b]

[a]Information System Department, College of Computing & Information Technology, Arab Academy for Science, Technology and Maritime Transport, Cairo, Egypt
[b]Computer Science Department, Cairo University, Oct.6 University, Cairo, Egypt
walaaadel_mm@yahoo.com, m.aborizka@aast.edu.eg
dr_fathi_amer@yahoo.com

Abstract: Introduction: Heart problems have gained a lot of interest in medical research because of their impact on human health where early diagnosis is critical to delaying the development of heart disease, the world's leading cause of death. thus, it is much needed to predict the possibility of occurrence of heart disease based on their attributes.

Objective: This research aims into a variety of machine learning classification algorithms for predicting heart disease.

Methods: The 10-fold cross-validation resampling is used to validate the prediction model. Aim and the prediction scores of each algorithm are evaluated with performance metrics such as prediction accuracy, confusion matrix, F1-Meuser, and suggested geometric mean.

Results: It was revealed that classifying the HD dataset using different classification algorithms produces extremely promising results, with a classification accuracy of 83.95, 84.5, 84.82, 84.89, and 85.05 % for the KNN, SVM, DT, LR, and RF algorithms, respectively. The RF algorithm successfully predicts 85.05 % (true positive rate) of the deceased cases correctly.

Conclusion: This study suggests that the RF method predicts Framingham possibilities better than other algorithms for the smaller (4240 records) dataset, based on the findings of other machine learning classification techniques on the Framingham dataset.

Keywords: Framingham dataset, Classification, Recall, Precision, Accuracy, F1-Meuser, Geometric Mean, Confusion Matrix, Supervised, Unsupervised and Reinforcement learning, Outlier data, Missing values.

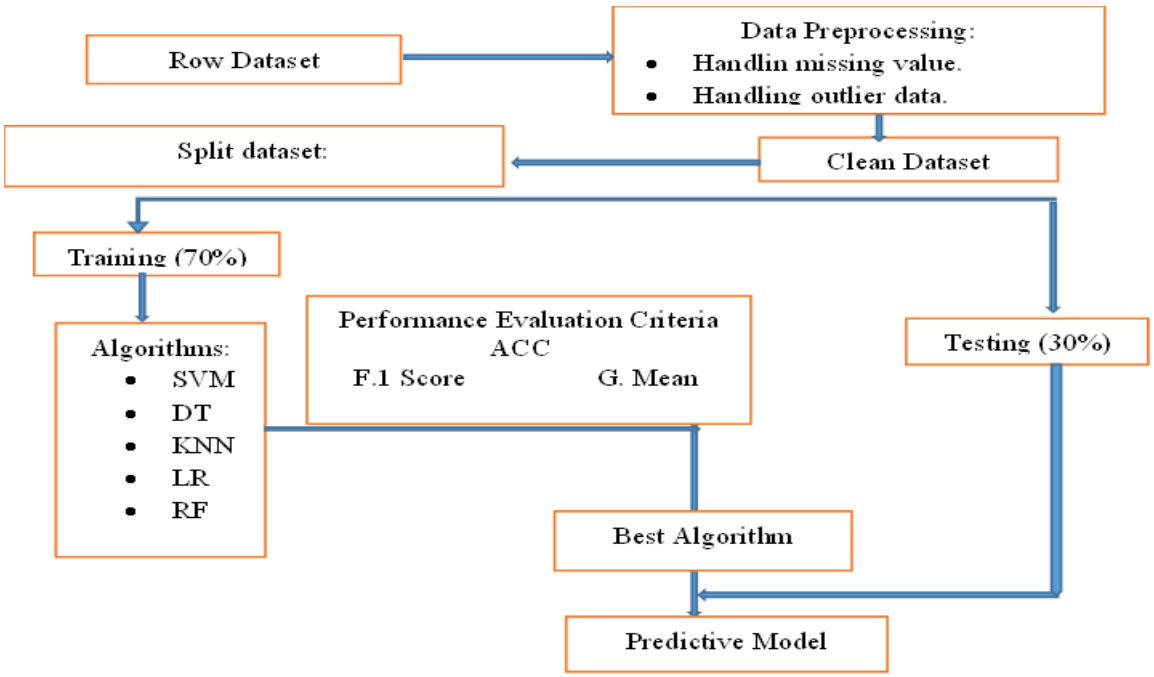
1. Introduction

Data mining (DM) is the process of extracting implicit, previously unknown, and potentially helpful information from data. It refers to a set of automated processes for uncovering and retrieving hidden data from large datasets**(Kumar et al., 2013)**. The three types of machine learning (ML) are supervised machine learning, unsupervised machine learning, and reinforcement learning**(Singh and Samagh, 2020)**. ML is a broad area that incorporates ideas from computer science, statistics, cognitive science, engineering, optimization theory, and a variety of other math and science fields. An overall workflow of our study has been shown in Figure1.

In ML and DM, classification is one of the most researched issues. One of the most exciting and difficult areas in which to create DM applications is predicting the result of a disease. In ML, there are several algorithms. For data classification, SVM, RF, KNN, DT, and LR were utilized.

Cleaning and preparing data. Data cleaning is the process of cleansing data in a systematic and precise manner so that it may be analyzed. Most of the time, there will be inconsistencies in the data collected, such as incorrect data formats, missing data values, and mistakes during data collection. The data must be separated into two sections for the algorithm's training and testing. The training set includes well-known categorized output, and the model is trained on this data before being applied to new data**(Swain et al., 2020)**.

Figure.1 Methodology of the research work



The organized of the paper is organized as follows. The next section introduces the different machine learning Algorithms. While in Section 3, introduces the dataset description. In Section 4, we illustrate the performance metrics. A real data set is analyzed in Section 5 and concluding remarks are included in Section 6.

2. Algorithms

The various classification machine learning algorithms used for the prediction of the Framingham dataset are reviewed below:

2.1 Logistic Regression:LR models the probabilities of the target belonging to a certain category. The basic model works on binary classification problems divides each sample into two groups (Yes/No) **(Rahim et al,.2021)**.

2.2 Decision Tree: When the relationship between features and outcome is nonlinear or where features interact with each other, A simple solution to these situations is decision trees. Decision trees work by splitting the data multiple times based on a measure such as information gain that determines how much information can be gained by that split **(Kumar et al., 2013).**

2.3 Random Forest: This is an ensemble algorithm that performs well with massive datasets with high dimensionality and is based on the DT algorithm. RF is used to process massive datasets and is sluggish in comparison to other methods. This approach can be used for both regression and classification problems, but it performs best in the latter. This method is predicated on the idea that if there were more trees, they would all come to the same conclusion. The mean of all the outputs of each of the DTs is used in regression, whereas classification employs a voting mechanism to determine the class **(Kumar et al., 2019)**.

2.4 K-Nearest Neighbor: This is a lazily supervised ML method for prediction and classification that is simple to construct and comprehend, needs little training time, and uses the entire training set. This nonparametric approach measures the distance between two sets of data to predict and categorize unknown data from known data. The distance metric is used to calculate the distance between each point in the testing data and each point in the training data **(Reddy et al., 2019)**.

2.5 Support Vector Machine:SVM works by employing an optimization strategy that can be solved with quadratic programming to identify a hyperplane that divides classes with the largest gap (greatest margin) on each side. In a two-dimensional space, this hyperplane would be a line the separates the plane into two parts **(Ha et al. 2020).**

3.Dataset

The Framingham dataset included in this research work there is 16 columns, 15 independent variables, and one dependent target variable in this table. It contains a total of 4240 rows. Table 1 includes a summary of the variables.

Table.1. Selected Framingham dataset attributes.

Attribute & Description		**Number of outlier and missing data**	
		Missing	**Outlier**
Age	(32-70)	0	0
Sex	0=Female, 1=Male	0	0
Education	It takes values as: 1=High School, 2=High School or GED, 3=College or Vocational School, 4=College	105	0
Current Smoker	0=No 1=Yes	0	0
Cigs Per Day	Number of Cigarettes smoked Per Day (0-70)	29	0
BP Meds	0=No 1=Yes	53	0
Prevalent Stroke	0=No 1=Yes	0	0
Prevalent Hyp	0=No 1=Yes	0	0
Diabetes	0=No 1=Yes	0	0
Tot Chol	Serum Cholesterol (107-696) (mg/dl)	50	196
Sys BP	(83.5-295) (mm/hg)	0	302
Día BP	(48-142.5) (mm/hg)	0	248
Body Mass Index (BMI)	(15.54-56.8)	19	240
Heart Rate	Heart Rate achieved (44-143)	1	271
Glucose	(40-394) (mg/dl)	388	522
10-year CHD	0=Healthy, 1=Diseases	0	0

4.Performance Evaluation Criteria

Commonly, the performance of ML prediction systems is frequently assessed using metrics based on the classification algorithm. The accuracy, confusion matrix, and F-measure geometric mean are used to evaluate the prediction findings in this research.

4.1.Confusion Matrix

A confusion matrix is used to evaluate algorithms in ML. A matrix for a binary classification problem is a square of two by two as illustrated in Table 2, where the column represents the algorithm's prediction and the row reflects the real value of the class label, where true-positive (TP) is numerous positive samples accurately predicted. False-negative (FN) refers to numerous positive samples that were anticipated incorrectly. False-positive (FP) refers to a situation in which numerous negative samples are incorrectly classified as positive. A true-negative (TN) is a collection of negative samples that have been accurately anticipated **(Bekkar et al., 2013)**.

Table.2. Confusion Matrix

Actual	**Predicted**	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

4.2.Accuracy

The equation of accuracy as follows:

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN).$$

4.3.F-measure

Isdefined as the harmonic mean of precision and recall **(Saleh et al., 2020)**. And computed with the next equation:

$$\text{F1 Score} \quad 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}).$$

Were,

$$\text{Precision} = TP/(TP+FP)$$

and

$$\text{Recall} = TP/ (TP + FN)$$

4.4.Geometric Mean

$$\text{GM} = \sqrt[2]{\text{Accuracy} \times \text{F1 score}}$$

5.Application

The classification approaches are implemented using R programming. The missing and outlier data values in the dataset are handled by the data preparation and cleaning processes (data imputation-mean approach).

- **Outlier** is considered as noise in the data and affects the accuracy of the algorithms. We have used Boxplot to find out and treat outliers. Figure. 2 shows the boxplot of 6 attributes namely; age, totChol, sysBP, diaBP, BMI, heart rate, and glucose.
- **Missing values**in data can happen for a variety of causes, including measuring device malfunctions, human mistakes, or incorrect measurement units, among others. Missing values should be addressed before training the model since they impair the learning algorithm's accuracy. **(Rahim et al,.2021)**. Thus, the first step in cleaning the data is to look for missing data or null values and deal with them to improve the model's performance. As seen in the figure.3, there are 645 missing values out of 4240 rows, accounting for almost 12% of the total data.

Figure.2 Boxplots of the dataset

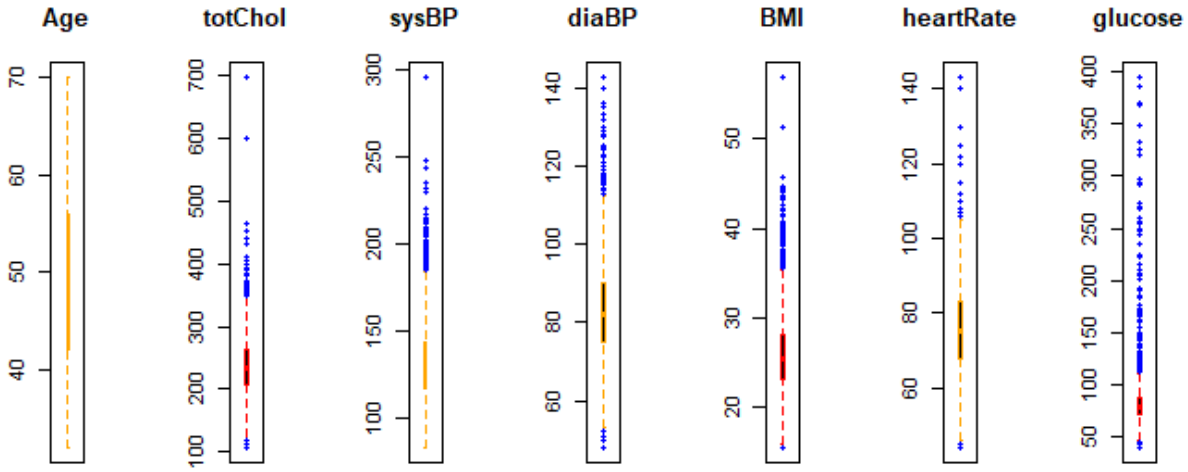


Figure.3 Missing value.

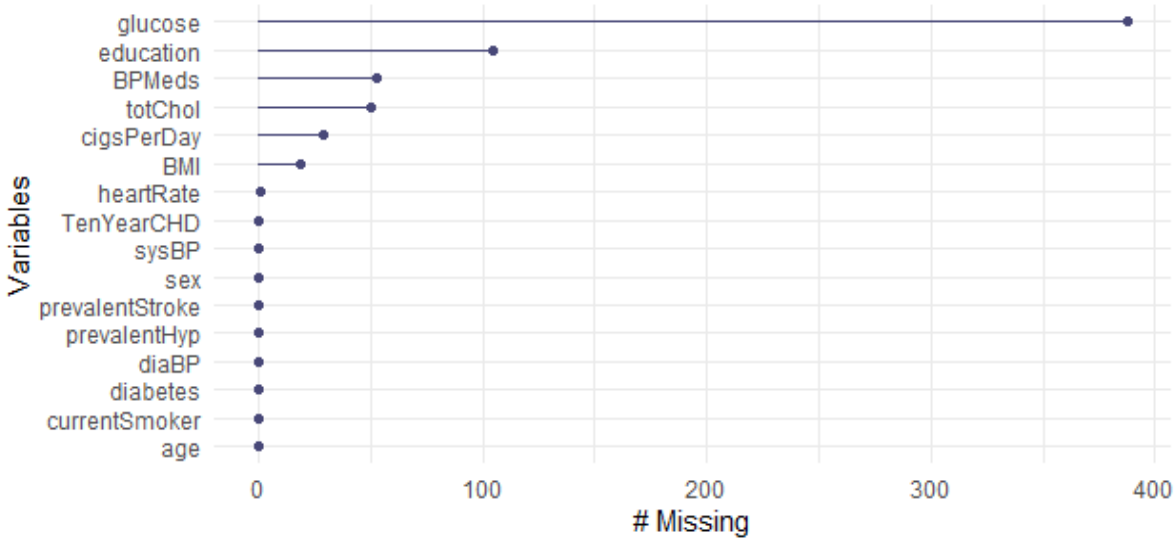


Table 1 shows the variables as well as the missing values for everyone. All missing values and outliers of a single property (in the dataset) are replaced by the median of all the values of the corresponding variable, as specified in our approach. The median substitution is used since it increases the number of samples in our data without contributing any more information. As a result, it assists in making better-educated predictions/decisions.Figure.4 and Figure.5 show that no outliers and missing values are present in the dataset.

Figure.4 Handling Outlier data

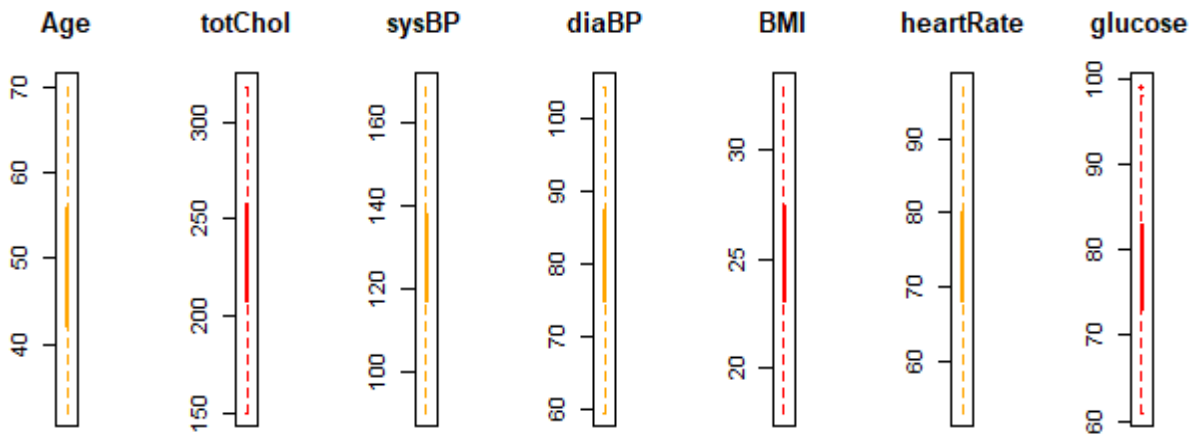
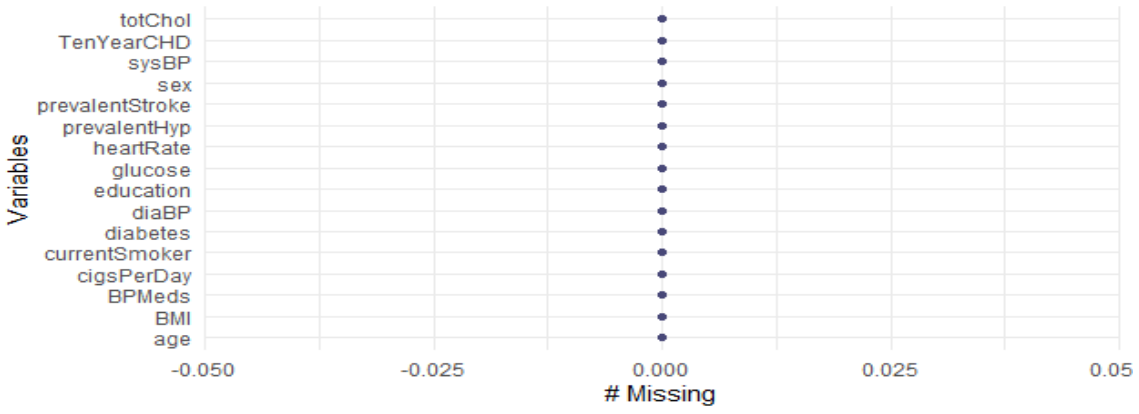


Figure.5 Handling Missing Values.



After imputation to build a classification model, the combined dataset with 16 attributes is divided into training and testing data with a percentage split of 70–30%. In this case, data is split below into two subsets: training (70%) and testing (30%). The confusion matrix obtained by five different supervised machine learning algorithms is given below.

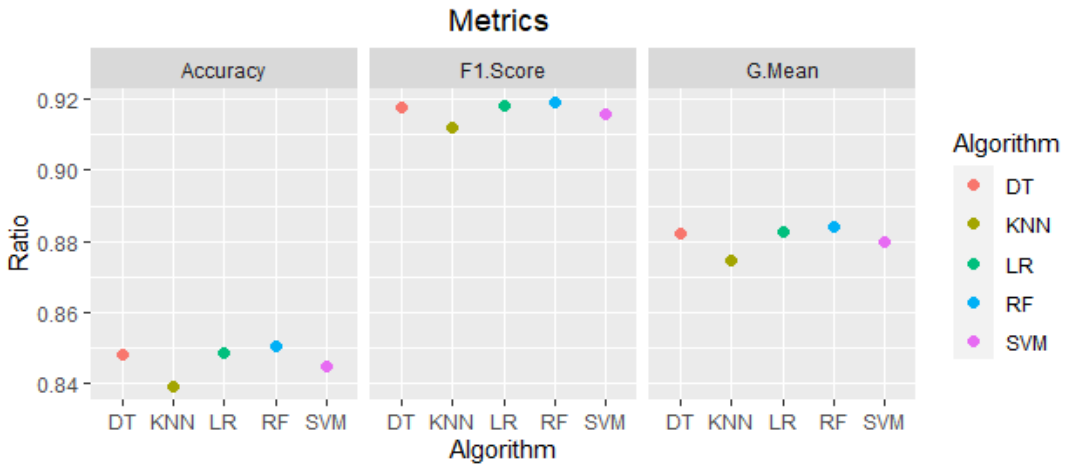
The performance measures are by the accuracy of each classification algorithm. The classification models' performance was evaluated using 10-fold cross-validation. The complete dataset is partitioned into 10 subsets and processed 10 times in this method, with 9 subsets serving as testing sets and the remaining subset serving as training sets. After then, the results are averaged every 10 iterations.

Table.3. Split data to 70–30% with the set. seed (5020).

	Algorithms	**Metric**		
		ACC	**F1**	**GM**
Performances	SVM	0.845	0.9159	0.8797
	DT	0.8482	0.9178	0.8823
	KNN	0.8395	0.9119	0.8749
	LR	0.8489	0.9180	0.8827
	RF	**0.8505**	**0.9190**	**0.8840**

Table 3 compares the classification metrics of algorithms. In Table 3, the dataset is classified, the accuracy rates of SVM, DT, KNN, LR, and RF are found in the range of 83.95%–85.05.

Figure.6 Evaluation of Performance.



The performance parameter measurement in Table 4 gives a very promising result by RF algorithm in our dataset. This algorithm reaches 91.90% for F-measure, 88.40% for geometric mean, and 85.05% for accuracy. Figure.6 also confirms this conclusion, while the other algorithms have fewer different metrics.

6.Conclusion

In this paper, it may be stated that ML algorithms have a lot of potential in predicting heart diseases and heart-related diseases. In this research, various popular ML algorithms have been discussed with their basic working mechanism where are applied to various real-world datasets (in this case Framingham dataset) and a study is carried out to find out the classifier which can perform well in the real-world data sets. Finally, several binary classification techniques for predicting heart diseases or heart-related diseases are investigated. With an accuracy of 85.05 percent, RF has shown to be the best classification algorithm for predicting the risk of HD.

References

Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013). Evaluation measures for models' assessment over imbalanced data sets. J Inf EngAppl, 3(10).

Ha, D. T., Loan, P. T. T., Giap, C. N., & Huong, N. T. L. (2020). An Empirical Study for Student Academic Performance Prediction Using Machine Learning Techniques. International Journal of Computer Science and Information Security (IJCSIS), 18(3).

Kumar, A., Sushil, R., and Tiwari, A. K. (2019). Comparative study of classification techniques for breast cancer diagnosis. International Journal of Computer Sciences and Engineering, 7(1), 234-240.

Kumar, G. R., Ramachandra, G. A., and Nagamani, K. (2013). An efficient prediction of breast cancer data using data mining techniques. International Journal of Innovations in Engineering and Technology (IJIET), 2(4), 139.

Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A., and Muzaffar, A. W. (2021). An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases. IEEE Access, 9, 106575-106588.

Reddy, N. S. C., Nee, S. S., Min, L. Z., and Ying, C. X. (2019). Classification and feature selection approaches by machine learning techniques: heart disease prediction. International Journal of Innovative Computing, 9(1).

Saleh, B., Saedi, A., Al-Aqbi, A., and Salman, L. (2020). Analysis of Weka Data Mining Techniques for Heart Disease Prediction System. International Journal of Medical Reviews, 7(1), 15-24.

Swain, D., Ballal, P., Dolase, V., Dash, B., and Santhappan, J. (2020). An Efficient Heart Disease Prediction System Using Machine Learning. In Machine Learning and Information Processing (pp. 39-50). Springer, Singapore.

Singh, D., and Samagh, J. S. (2020). A comprehensive review of heart disease prediction using machine learning. Journal of Critical Reviews, 7(12), 281-285.