IEEE SEM

Enhanced Detection of Heart Diseases Using Data Mining Techniques

Reem Reda Mohamed [1], Mona Mohamed Nasr[2], Ahmed I.B ElSeddawy [3]

[1]Business Information Systems Department, Faculty of commerce and business, Helwan University.Cairo,Egypt; [2]Information Systems Department, Faculty of Computers and Artificial Intelligence, Helwan University. Cairo, Egypt; [3] Information Systems Department, Arab Academy for Science, Technology and Maritime Transport (AAST). Cairo, Egypt.

[1]Email: Reem.reda21@commerce.helwan.edu.eg
[2]Email: m.nasr@helwan.edu.eg
[3]Email: ahmed.bahgat@aast.edu

ABSTRACT

Data mining techniques nowadays is considered to be reliable tool on health sector, specifically in diagnosing and detecting different diseases. In this paper, we shed the light on the detection of heart disease, seeking to reduce the development of the disease into serious and complicated situation, and to minimize the costly health care. We used fifteen important attributes that considered to be essential assessment factors to any cardiologist, such as Age, Sex, Cp, Trest bps, Chol, etc. By using these attributes, we predict the possibility of having heart disease at an early stage to group of individuals, in reach to know the expected number of heart patients, and help decision-makers take the procedure of early treatment steps and recovery, which will achieve more efficiency of both diseases prediction and medical treatment services. We used a data set from two resources, UCI university AI archive, and a survey that built on group of people responses. By using RapidMiner Studio, a framework of three different classification algorithms was used, to determine the best accuracy among them. The results show that the best accuracy are Random Forest (RF) Algorithm, k-Nearest Neighbor (K-NN) Algorithm, and Decision tree (DT) Algorithm respectively. Random forest accuracy was 91.71%, and a Cluster K-mean Algorithm was used, and the result of Avg. within centroid distance: 0.250.

Keywords : Data mining, Classification algorithms, Random forest, Decision tree, k-Nearest Neighbor, clustering algorithm, K-mean

1 INTRODUCTION

THE Data mining is discovering a data set to extract patterns and relationships to gain knowledge that is difficult to discover in other ways [1]. It is one of the technological developments used in many sectors including the health care, which many works on and used to achieve better health care for patients. It will help reaching high accuracy diagnostic procedures, where many diseases that affect us severely, including heart disease, that many medical reports mentioned as the main cause of death around the world [2]. Also, it provides a great help of gaining the high benefits of early detection through a faster diagnosis in a specific time, represented for both (patient better health care, and doctors efficient medical service). The aim of this research is to detect heart disease at an early stage, which can be done through coordinating medical information (the attributes) and identifying people who suffer or do not suffer from heart disease [3]. This gives the patient the opportunity to obtain better health care and more treatment accuracy in the case of detecting heart disease, and this leads to a rapid response to treatment in less time. According to what mentioned, this will reduce the cost of the health care, because detection at early stages makes the cost reasonable, and patient can reach full recovery through proper diagnosis earlier. The reason we chose heart disease is because it is very difficult to predict without conducting any medical examinations. There are different attributes by which we can have an analysis that provides an accurate prediction, and reduces the risk of having serious heart disease. The most important causes that we can take into consideration, which have an important effect of having heart disease, are (age, smoking, diabetes, obesity, genetics, and blood pressure) [4]. The research investigates the problem of early detection of heart disease, identifying it and knowing whether the patient needs early care or not. It will also propose an approach based on early detection and decision making process. After the completion of Data Collection and Data Preprocessing stages, to achieve the research objective, processing will be completed using one of Data Mining Techniques, for us, the tool that will be used is RapidMiner Studio [5].

2 Literature Review

Data mining techniques help in early detection of heart disease, and through research, the use of techniques is enhanced.

M. Alex, and Sh. P. Shaji. [1] This project was presented to provide insight into the detection and treatment of heart diseases using data mining. Data was collected from Thrissur Hospital on 2200 patients. Data were collected by interaction with patients. It was applied to 20 attributes. The data was taken and subjected to data mining algorithms. In order to reach and predict the possibility of heart disease. This is trained on the following algorithms, which had the best result with the highest accuracy ANN 92,21% and then Random forest 85,88% accuracy. Then SVM 85,88% accuracy. Then KNN 83,21% accuracy.

G. Kalaiarasi, M. Maheswari, M. Selvi, R. Yogitha, and P. Devadas. [6] This is used to create a dataset appropriate for data mining by taking into consideration various data cleaning and mining techniques. The overall goal of the endeavor is to use data mining methods to predict cardiovascular disease events with greater accuracy. Currently, the same research of three calculations, such as random forest calculation performs best with 81% accuracy, decision trees, and naive Bayes, is being played out using the UCI data repository. As opposed to decision trees and naive Bayes, random forest is known to produce perfect results.

J. Thomas, and R.T. Princy. [7] The health sector holds a wealth of valuable hidden facts and information that, particularly in the realm of medicine, might be used to make predictions. Data mining is a method or technique used to analyze huge datasets and then produce substantial and practical outcomes using exceptional AI-based tools. This article aims to forecast heart disease using three of these AIbased methods: Decision Tree, Naïve Bayes, and Neural Network. All of these techniques will be predicated on many special & parameters with improvements for greater accuracy. The accuracy based on different factors of each approach will then be compared. Then, the most reliable method is used to determine whether a man or woman will develop coronary heart disease.

S.K. Yadav, Y. Chouhan and m. Choubisa. [8] In this paper, a different set of algorithms are presented, such as Naïve Bayes classifier, Decision Tree Classifier, Random Forest Classifier, K Neighbors Classifier, Logistic Regression Classifier, voting model classifier, and based on the results reached by the research, it was concluded that the voting model classifier is higher, with an accuracy of 88%. In this research, the application was on various neural networks to reach the highest achievement of accuracy in the best and most efficient way. In this research, a group of attributes of heart disease was carried out. The research also confirmed that by changing the data set, the accuracy of the results differs. The application of the approach was searched ANN (Artificial neural network) to predict heart disease, as its accuracy is 91%, and its performance is more powerful compared to other algorithms.

A. Bharadwaj, D.Yadav, and A.K. Yadav. [9] This research is based on prediction of heart disease using machine learning, and the proposed method is to use a new hybrid classifier design consisting of these classifiers by combining two classifiers, namely KNN and SVM. KNN is implemented to use the features from the dataset and SVM is implemented for the final prediction and result rendering. The research proved that the proposed classifiers perform better in terms of accuracy and implementation. By changing the attributes in the study, better predictions were made for the patient.

S. Ouf, and A. I. B. ElSeddawy. [10] In paper, data mining methods are used to forecast cardiac problems. oblivious to the value of using the other cross-validation techniques and repeating them randomly to increase prediction accuracy when using the cross-validation techniques with data mining algorithms to identify heart diseases early. The paper's scientific uniqueness also comes from a large-scale comparison of the cross validation with the most significant. data mining classification techniques (Linear Discriminant Analysis, Logistic regression, Support Vector Model, KNN, Decision Tree, Naïve Bayes, Random Forest, and Neural Network). The optimal prediction model with the highest accuracy should be chosen by partitioning datasets using the four ways, according to the prediction models. done on two datasets, Kaggle and UCI Cleveland, to identify the most accurate prediction models. We choose the most accurate prediction models, and then The findings demonstrated that the two data mining classification methods with the greatest accuracy for heart disease prediction. are Holdout crossvalidation with a neural network and logistic regression in a large dataset. Repeated Random with Random Forest and holdout with KNN, have the highest accuracy in the small dataset.

Name	Year	Technique (method)
M.Alex; Sh.P. Shaji.	2019	ANN 92,21% accuracy; RF 85,88% accuracy; SVM 85,88% accuracy; KNN83.21%accuracy[1].

**G. Kalaiarasi;
R.Yogitha;
M.Maheswari; M.Selvi;**
	2022	Random forest 81% accuracy; decision trees; naive Bayes [6].

J.Thomas; R.T. Princy.	2016	Decision Tree; Naïve Bayes; Neural Network[7].
S.K.Yadav; Y.Chouhan	2022	Voting mode classifier; ANN 91% accuracy[8].

m.Choubisa.		
A.Bharadaj; D. Yadav A.K. Yadav	2022	SVM; KNN[9].
S. Ouf; A.I.B.ElSeddawy.	2021	neural network; logistic regression; Repeated Random; Random Forest; KNN[10].

Table. 1 literature review summary.

3 Proposed Method

The framework for this study consists of these important components namely data collection, data preprocessing, and data classification. to classify the data, three algorithms were used: random forest (rf), decision tree (dt), and k-nearest neighbor (knn). the final step was clustering in which we used the k-mean algorithm.

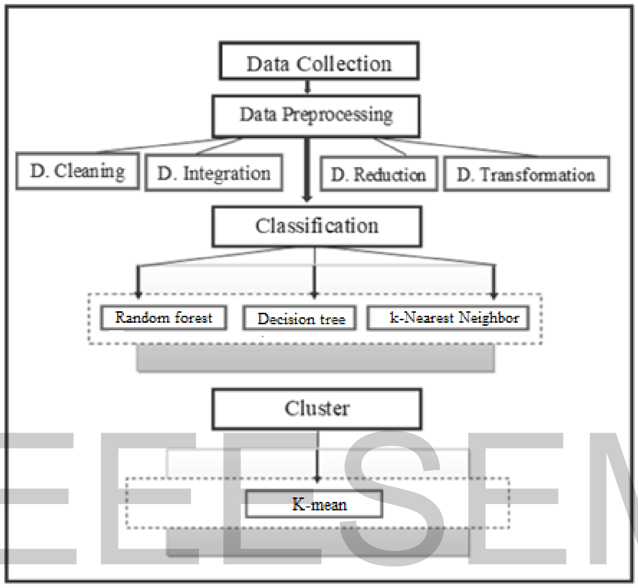


Fig.1 Research Framework.

3.1 Data Collection

In this study, the data set will be built using these attributes and will be collected from different sources. We chose these attributes according to what some previous studies used, adding to meetings with heart disease specialists and doctors. The main objective is to form a group of attributes that considered to have a significant relation to early detection of heart disease.

No	Attribute
1	Age
2	Sex
3	Cp
4	Trest bps
5	Chol
6	Fbs
7	restecg
8	thalach
9	exang
10	oldpeak
11	Slop
12	Ca
13	Thal
14	Smoker
15	Target

Table.2 All attributes to heart disease detection.

3.1.1 Data Set Descriptions

The data for this research is a data set consisting of 15 attributes, and were collected from 2 sources. The first source was a data set that have been used for experimentation from Data mining repository of the University of California, (UCI) [11]. and the second source was a registering data set that have been collected via Google Forms [12]. The data set in this work consists of a total of (3499), and our objective is to indicate the presence of heart disease at early stages, in a way that helps achieving a lot of benefits, whether medical or nonmedical.

No	Attribute	Description	Value
1	Age	Age in years	28 to 80
2	Sex	Gender of the patient	1=male; 0=female
3	Cp	Chest Pain Type	1 = typical angina; 2 = atypical angina; 3 =non-angina pain; 4=asymptomatic
4	Trest bps	Resting blood pressure in mm Hg on admission to the hospital.	92 to 200 Mg Hg
5	Chol	Serum cholesterol	126 to 564 mg/dl
6	Fbs	Fasting blood sugar	> 120 mg/dl (1= true; 0= false)

7	Restecg	Resting electrocardiographic results	0= normal;
			1= having ST-T wave abnormality;
			2= probable or definite left ventricular hypertrophy.

8	Thalach	Maximum heart rate achieved.	71 to 202
9	Exang	Exercise-induced angina	1=yes; 0=no
10	Oldpeak	ST (sports test) depression induced by exercise relative to rest.	0 TO 6.2
11	Slope	The slope of the peak exercise ST segment.	1= upsloping; 2= flat; 3= down sloping.
12	Ca	Number of major vessels colored by fluorosopy.	0–3 Value
13	Thal	Thalassemia- Defect type	3= normal; 6= fixed defect; 7= reversible defect
14	Smoker	I believe this is (is or not to a smoker)	1=yes; 0=no
15	Target	Heart disease	0= healthy; 1= have heart disease

Table.3 summarizes heart disease detection.

3.1.2 Data Preprocessing

Data Preprocessing is the way to represent the data in format as per mining techniques. Data Preprocessing is classified into four categories (Data Cleaning, Data Integration, Data Reduction, and Data Transformation) [13]. These four steps will be applied respectively.

3.1.3 Data Cleaning

Data cleaning steps is the first step of data pre-processing it removes inconsistencies and redundancies from the selected data sets [14]. data attributes transformation from inconsistent data, noisy data, and Missing values to a data set can be used efficiently.

ID	Age	Sex	Cp	Trestbps	Chol
1	63	Male	typical angina	145	233
2		Male		160	286
3	67	Male	asymptomatic	120	229
4	37	Male	non-anginal		250
5	41	Female	atypical angina	130	

Table.4 Missing values.

3.1.4 Data Integration

Data Integration combines the data which gathered from different sources at one place, for example, multiple databases and files [15]. Fig 2 shows these steps.

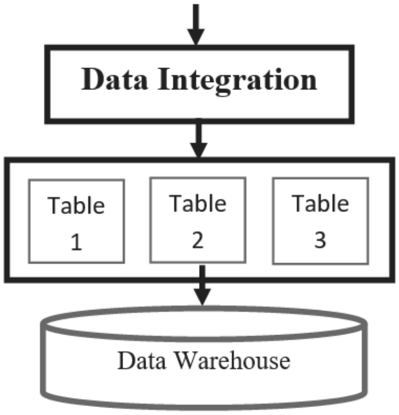


Fig.2 Data Integration.

3.1.5 Data Reduction

Data reduction is defined as a technique that can be applied to obtain a reduced representation of the data set, which is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient, yet produce similar (or almost the same) analytical results.

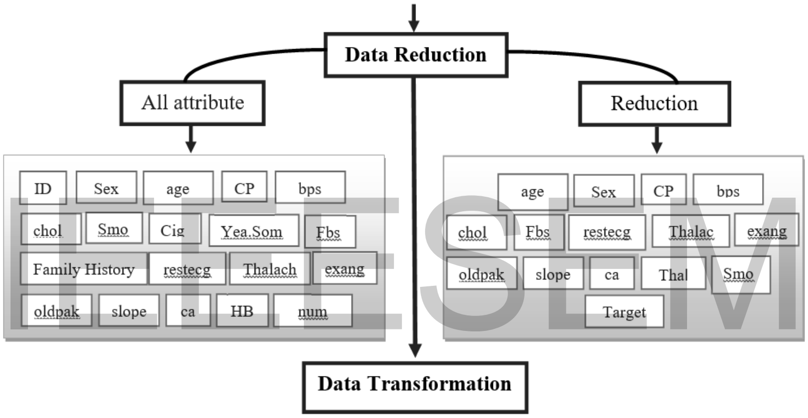


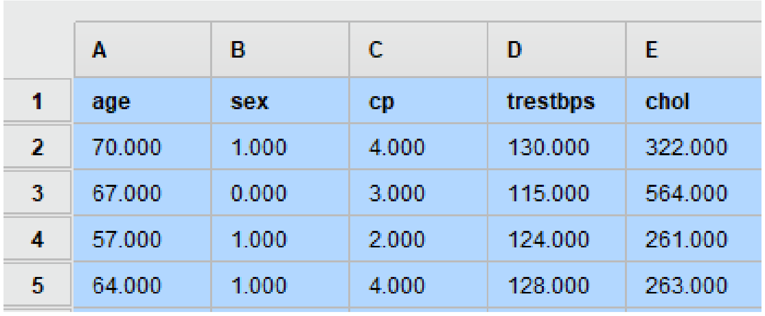
Fig.3 Data Reduction and Data Transformation.

3.1.6 Data Transformation

In this phase, we convert the data into a form that is required for the research purpose, for example, in this study the required form of data is the "Early Detection Claims database for Heart Disease". So, data transformation will be done by applying normalization to the data, and creating the database. data attributes are converted to numerical data.

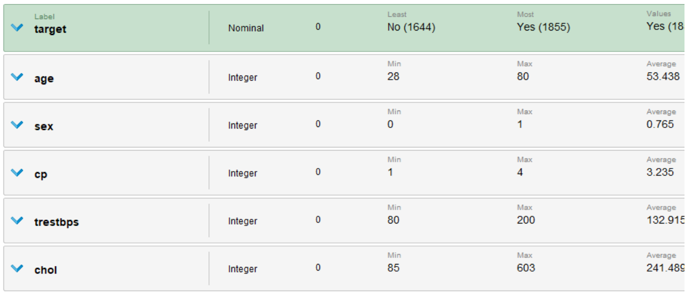
3.1.7 Data Set Run

After completing the data collection phase and data preprocessing phase, the data set must be tested and ensure that the data is valid for use to work on. This has been done using **Rapid miner studio**, and **fig 4 & 5** shows an example of the extracted results.



	A	B	C	D	E
1	age	sex	cp	trestbps	chol
2	70.000	1.000	4.000	130.000	322.000
3	67.000	0.000	3.000	115.000	564.000
4	57.000	1.000	2.000	124.000	261.000
5	64.000	1.000	4.000	128.000	263.000

Fig.4 Add dataset and Select the cells.



Label					
target	Nominal	0	Least No (1644)	Most Yes (1855)	Values Yes (18
age	Integer	0	Min 28	Max 80	Average 53.438
sex	Integer	0	Min 0	Max 1	Average 0.765
cp	Integer	0	Min 1	Max 4	Average 3.235
trestbps	Integer	0	Min 80	Max 200	Average 132.915
chol	Integer	0	Min 85	Max 603	Average 241.489

Fig.5 Data Statistics.

3.1.8 Training and Testing for Accuracy

Here, the model will be trained and tested using the data set to know the accuracy of the model. The improvement of the model accuracy results from training using 70% of data set volume, and testing the other 30% of the data volume. Then, the chosen classification algorithms that have been built – or choose– s by building algorithms and make predictions on the data to produce a result from training for the training data set and be 70% of the data volume, which is then compared to the target. The test dataset comes in and is 30% of the data volume. It is for evaluation and is according to the training data set [16].

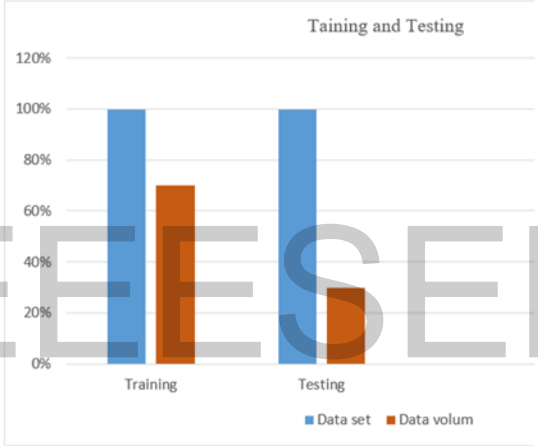


chart.1 data splitting Training 70% and Testing 30%.

3.2 CLASSIFICATION

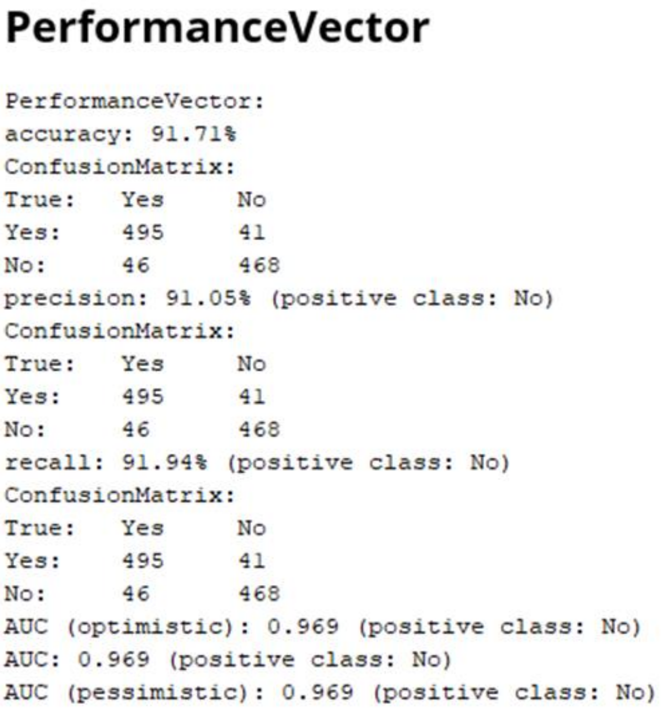
The process of learning a function that can classify data items into a subset of a specified class set is known as classification. Classification is considered to be one of the most well-known difficulties in data mining. Some classification goals include many steps, on top of them is to identify a good generic that can accurately forecast the class of unknown data objects for each of the classes [17]. To reach the results we seek in this research, three algorithms were used: Random Forest(RF), Decision Tree(DT), and K-Nearest Neighbors(K-NN).

3.2.1 RANDOM FOREST

Generally, random forest is an assistance tool that generates a final result that considered to be a final result that have been aggregated from collection of decision trees. It performs admirably in a variety of real-world prediction issues & situations, for example in health care. The reasons are simply because it is unaffected by noise in the data set, isn't over fitted in any way, also is constructed by merging the forecasts of many trees and it works quickly. Many other tree-based algorithms, such as decision tree, usually show a large performance improvement [1].

	True Yes	True No	Class precision
Pred. Yes	495	41	92.35%
Pred. NO	46	468	91.05%
Class recall	91.50%	91.94%	

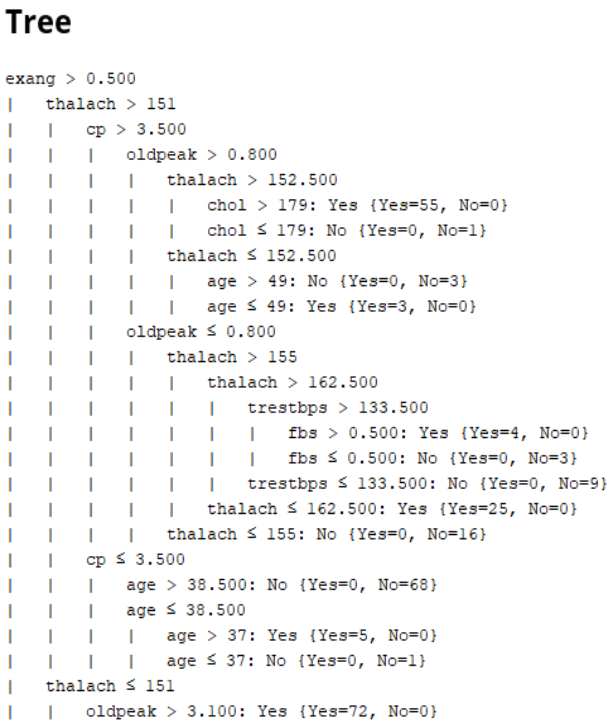
Table.5 RF Algorithm Accuracy: 91.71%.



PerformanceVector

```
PerformanceVector:
accuracy: 91.71%
ConfusionMatrix:
True:   Yes     No
Yes:    495     41
No:     46      468
precision: 91.05% (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    495     41
No:     46      468
recall: 91.94% (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    495     41
No:     46      468
AUC (optimistic): 0.969 (positive class: No)
AUC: 0.969 (positive class: No)
AUC (pessimistic): 0.969 (positive class: No)
```

Fig.6 Performance Results of RF Algorithm.



Tree

```
exang > 0.500
|   thalach > 151
|   |   cp > 3.500
|   |   |   oldpeak > 0.800
|   |   |   |   thalach > 152.500
|   |   |   |   |   chol > 179: Yes {Yes=55, No=0}
|   |   |   |   |   chol ≤ 179: No {Yes=0, No=1}
|   |   |   |   thalach ≤ 152.500
|   |   |   |   |   age > 49: No {Yes=0, No=3}
|   |   |   |   |   age ≤ 49: Yes {Yes=3, No=0}
|   |   |   oldpeak ≤ 0.800
|   |   |   |   thalach > 155
|   |   |   |   |   thalach > 162.500
|   |   |   |   |   |   trestbps > 133.500
|   |   |   |   |   |   |   fbs > 0.500: Yes {Yes=4, No=0}
|   |   |   |   |   |   |   fbs ≤ 0.500: No {Yes=0, No=3}
|   |   |   |   |   |   trestbps ≤ 133.500: No {Yes=0, No=9}
|   |   |   |   |   thalach ≤ 162.500: Yes {Yes=25, No=0}
|   |   |   |   thalach ≤ 155: No {Yes=0, No=16}
|   |   cp ≤ 3.500
|   |   |   age > 38.500: No {Yes=0, No=68}
|   |   |   age ≤ 38.500
|   |   |   |   age > 37: Yes {Yes=5, No=0}
|   |   |   |   age ≤ 37: No {Yes=0, No=1}
|   thalach ≤ 151
|   |   oldpeak > 3.100: Yes {Yes=72, No=0}
```

Fig.7 Trees attributes of the RF algorithm.

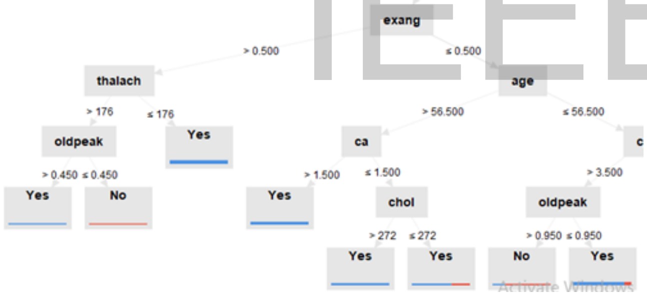


Fig.8 trees of the RF Algorithm.

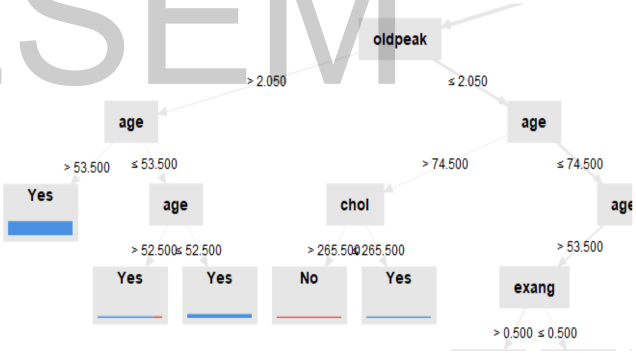


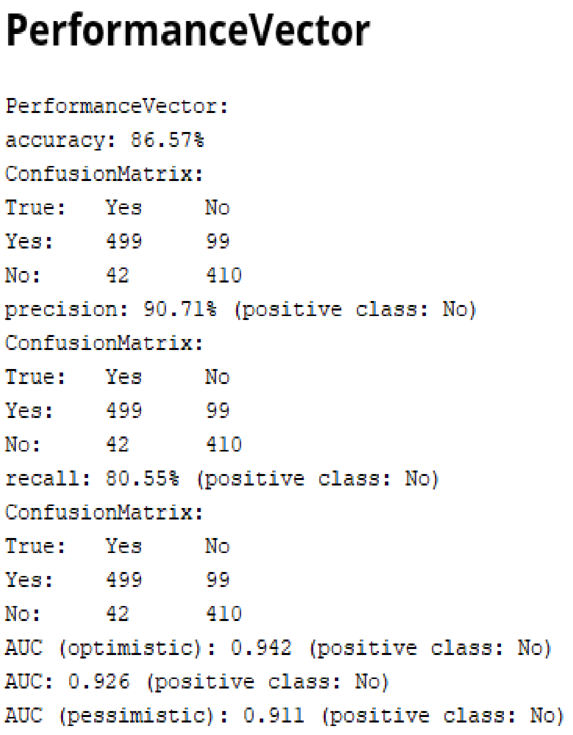
Fig.9 trees of the RF Algorithm.

3.2.2 Decision Tree

The knowledge is represented in a tree diagram in this method of classification. To portray the decisions, the schematic representation will be in the shape of a tree. These selections categorize the data frame and the rules. To begin, the data is organized into root nodes, which are followed by the attribute's terminal node. Nodes have attribute names, positive values represent edges, and different classes represent the trees leaves [18].

	True Yes	True No	Class precision
Pred. Yes	499	99	83.44%
Pred. NO	42	410	90.71%
Class recall	92.24%	80.55%	

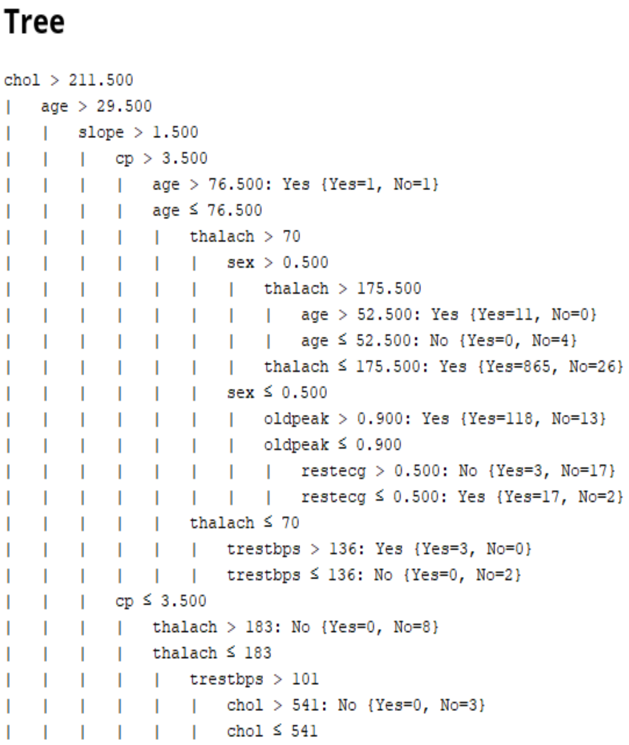
Table.6 DT Algorithm Accuracy: 86.57%.



PerformanceVector

```
PerformanceVector:
accuracy: 86.57%
ConfusionMatrix:
True:   Yes     No
Yes:    499     99
No:     42      410
precision: 90.71% (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    499     99
No:     42      410
recall: 80.55% (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    499     99
No:     42      410
AUC (optimistic): 0.942 (positive class: No)
AUC: 0.926 (positive class: No)
AUC (pessimistic): 0.911 (positive class: No)
```

Fig.10 Performance Results of DT Algorithm.



Tree

```
chol > 211.500
|   age > 29.500
|   |   slope > 1.500
|   |   |   cp > 3.500
|   |   |   |   age > 76.500: Yes {Yes=1, No=1}
|   |   |   |   age ≤ 76.500
|   |   |   |   |   thalach > 70
|   |   |   |   |   |   sex > 0.500
|   |   |   |   |   |   |   thalach > 175.500
|   |   |   |   |   |   |   |   age > 52.500: Yes {Yes=11, No=0}
|   |   |   |   |   |   |   |   age ≤ 52.500: No {Yes=0, No=4}
|   |   |   |   |   |   |   thalach ≤ 175.500: Yes {Yes=865, No=26}
|   |   |   |   |   |   sex ≤ 0.500
|   |   |   |   |   |   |   oldpeak > 0.900: Yes {Yes=118, No=13}
|   |   |   |   |   |   |   oldpeak ≤ 0.900
|   |   |   |   |   |   |   |   restecg > 0.500: No {Yes=3, No=17}
|   |   |   |   |   |   |   |   restecg ≤ 0.500: Yes {Yes=17, No=2}
|   |   |   |   |   thalach ≤ 70
|   |   |   |   |   |   trestbps > 136: Yes {Yes=3, No=0}
|   |   |   |   |   |   trestbps ≤ 136: No {Yes=0, No=2}
|   |   |   cp ≤ 3.500
|   |   |   |   thalach > 183: No {Yes=0, No=8}
|   |   |   |   thalach ≤ 183
|   |   |   |   |   trestbps > 101
|   |   |   |   |   |   chol > 541: No {Yes=0, No=3}
|   |   |   |   |   |   chol ≤ 541
```

Fig.11 Tree attributes of the DT algorithm.

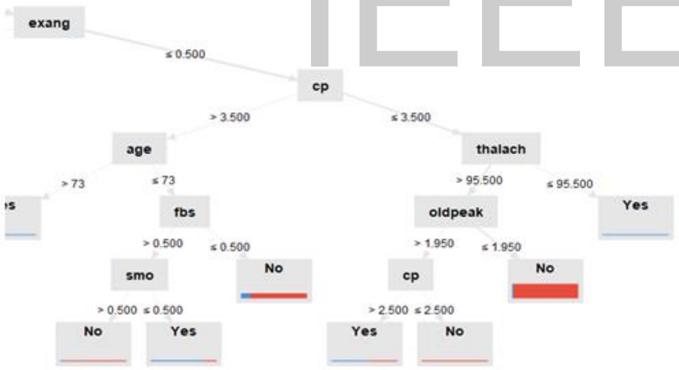


Fig.12 The tree of the DT algorithm.

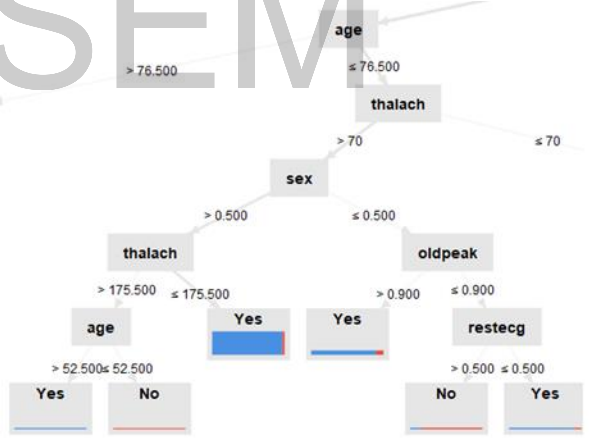


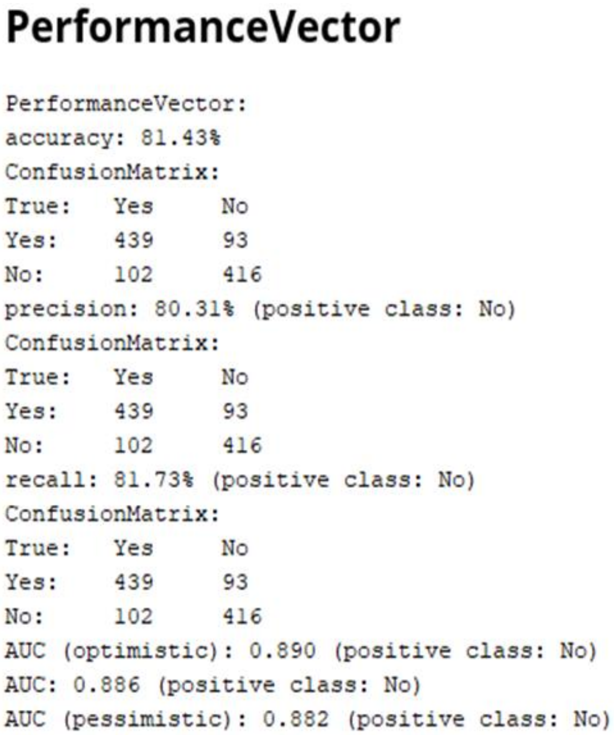
Fig.13 Tree exang, cp used in the attributes of the DT algorithm.

3.2.3 K-Nearest Neighbors

When we seek pattern recognition, the k-NN classifier is commonly employed. Learning via relationship, or more clearly contrasting a given test tuple and preparing tuples that are similar, is how K-neighbor classifiers works. In this process, (n) attributes represents the preparation of tuples. All of the prepared tuples are stored in a n dimensional example space since each tuple represents a point in a n dimensional space. K-nearest neighbor classifier searches the example space for the K preparing tuples that are closest to the obscure tuple when given an obscure tuple [19].

	True Yes	True No	Class Precision
Pred. Yes	439	93	82.52%
Pred. NO	102	416	80.31%
Class recall	81.15%	81.73%	

Table.7 K-NN Algorithm Accuracy: 81.43%.



PerformanceVector

```
PerformanceVector:
accuracy: 81.43%
ConfusionMatrix:
True:   Yes     No
Yes:    439     93
No:     102     416
precision: 80.31% (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    439     93
No:     102     416
recall: 81.73% (positive class: No)
ConfusionMatrix:
True:   Yes     No
Yes:    439     93
No:     102     416
AUC (optimistic): 0.890 (positive class: No)
AUC: 0.886 (positive class: No)
AUC (pessimistic): 0.882 (positive class: No)
```

Fig.14 Performance accuracy of K-NN Algorithm.

3.2.4 Accuracy Results

By building algorithms on Rapid miner studio, predictions were made for access to discover knowledge, and it was verified that the best result is RF, then DT and K_NN.

Algorithms	Accuracy
Random forest	91.71%
Decision tree	86.57%
K-NN	81.43%

Table.8 Classification Algorithm Accuracy results.

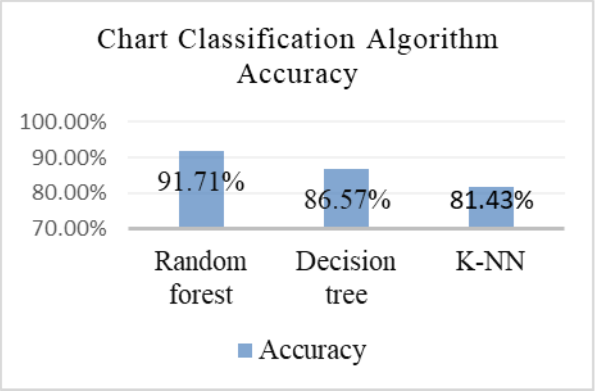


Chart.2 Classification Algorithm Accuracy.

3.3 Cluster

Cluster is the process of grouping objects according to specific criteria. For example, data elements into different groups of similarity or classes. Each and every near object to centroid point is a neighbor object. Inter class cluster means cluster distance is maximized, while intra cluster means cluster distances are minimized [20].

3.3.1 K-Mean

We used K-mean algorithm for many reasons, basically because it is faster than other clustering algorithms and it works great if clusters are spherical. Where other clustering algorithms can be applied, K-mean becomes a great solution for pre-clustering and reducing the space into disjoint smaller sub-spaces [20]. This algorithm is a method of vector quantization, its job is to partition number (n) of observations, depends on the size of data set, into number (k) of clusters. In addition, each observation after clustering will belong to a specific cluster centroid.

The importance of used equation below is represented in the set of k clusters that minimizes the error, through the following steps [20]:

1) Arbitrarily choose k objects as the initial cluster centers,
2) Repeat,
3) Assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster,
4) Update the cluster means, calculate the mean value of the objects for each cluster,
5) Until no change.

This algorithm function known as squared error function given by:

$$E =\sum_{i=1}^{k} \sum_{P\epsilon C_i}(P - C_i)^2$$

and when we applied on our data, the Avg. within centroid distance equals (0.250).

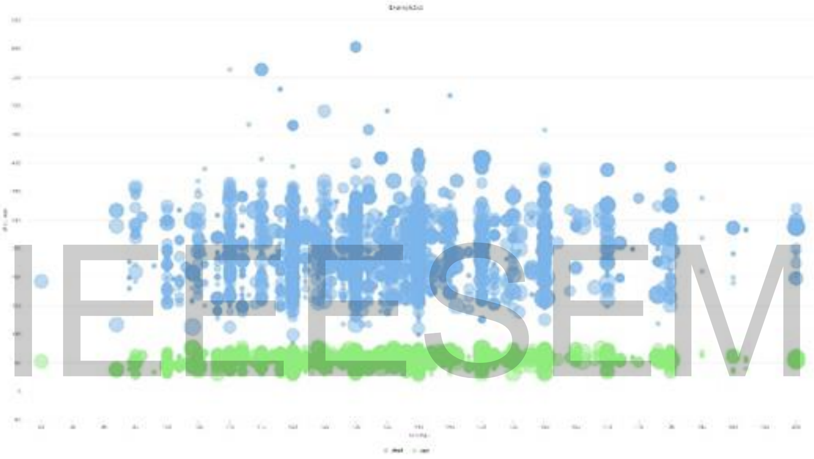


Fig.15 K-mean clustering for heart disease patients.

4. Conclusion

The purpose of this work was to compare the best algorithm that can be used to enhance detecting heart diseases. The data collected and the attributes went through four stages of data preprocessing: data cleaning, data integration, data reduction, and data transformation. It was tested using different classification algorithms, in specific (Random Forest, K-Nearest Neighbors and Decision Tree), and highest accuracy was 91.71% for (RF), then Decision tree, and K- nearest neighbors. We choose K-mean algorithm in data clustering, and Avg. with centroid distance was 0.250.

According to previous studies that depended on different numbers of attributes and size of data, we can conclude that more attributes to be used in predicting health conditions is preferable. Suggested future work could focus more on improving this work by integrating the methods and forming a hybrid model, where better results can be achieved compared to the separated methods.

5.Reference

[1] S. P. Shaji, "Predictionand diagnosis of heart disease patients using data mining technique," in 2019international conference on communication and signal processing (ICCSP), 2019: IEEE, pp. 0848-0852.

[2] S. Babu et al., "Heart disease diagnosis using data mining technique," in 2017 international conference of electronics, communication and aerospace technology (ICECA), 2017, vol. 1: IEEE, pp. 750-753.

[3] T. Puyalnithi and M. Vankadara, "Performance Analysis of Classification Algorithms on a Novel Unified Clinical Decision Support Model for Predicting Coronary Heart Disease Risks."

[4] A. Shahnaz, U. Qamar, and A. Khalid, "Using blockchain for electronic health records," IEEE Access, vol. 7, pp. 147782-147795, 2019.

[5] R. Miner, " rapidminer studio,"[Online]. https://rapidminer.com/studio/,2021.

[6] G. Kalaiarasi, M. Maheswari, M. Selvi, R. Yogitha, and P. Devadas, "Detection of Heart Disease Using Data Mining," In Biologically Inspired Techniques in Many Criteria Decision Making, 2022, pp. 627-637. Springer, Singapore.

[7] J. Thomas, and R.T. Princy, "Human heart disease prediction system using data mining techniques," In 2016 international conference on circuit, power and computing technologies (ICCPCT), 2016, March, pp. 1-5. IEEE.J .

[8] S. K. Yadav, Y. Chouhan, and M. Choubisa, "Predictive Hybrid Approach Method to Detect Heart Disease," Mathematical Statistician and Engineering Applications, vol. 71, no. 1, pp. 36–47-36–47, 2022.

[9] A. Bharadwaj, D. Yadav, and A. K. Yadav, "Heart Disease Prediction Using Hybrid Classification Methods," in International Conference on Innovative Computing and Communications, 2022: Springer, pp. 565-573.

[10] S. Ouf, and A. I. B. ElSeddawy, ‘’A Proposed Paradigm For Intelligent Heart Disease Prediction System Using Data Mining Techniques,’’ 2021: Journal of Southwest Jiaotong University, 56(4).

[11] University of California, UCI Machine Learning Repository,[Online]. https://archive.ics.uci.edu/ml/datasets/heart+disease

[12] Google Forms, https://docs.google.com/forms/d/1bdtyYWyghvDoY_iuo8sQuFZno37bLzlghi_nZ9wjI0U/edit.

[13] S. Sharma, and A. Bhagat, "Data preprocessing algorithm for web structure mining," in 2016 Fifth International Conference on Eco-friendly Computing and Communication Systems (ICECCS), 2016: IEEE, pp. 94-98.

[14] P. Guo, S.-S. Chen, and Y. He, "Study on data preprocessing for daylight climate data," in International Conference on Information Computing and Applications, 2012: Springer, pp. 492-499.

[15] P.-T. Chung and S. H. Chung, "On data integration and data mining for developing business intelligence," in 2013 IEEE Long Island Systems, Applications and Technology Conference (LISAT), 2013: IEEE, pp. 1-6.

[16] S. Manoj and B.N Yuvaraju, " Design & Implementation of Heart Disease Prediction using Machine Learning," in 2020 International Research Journal of Engineering and Technology (IRJET), 2020: IRJET, pp. 1-5.

[17] C. Sowmiya and P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques," in 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017: IEEE, pp. 1-5.

[18] N. Priyanka and P. R. Kumar, "Usage of data mining techniques in predicting the heart diseases—Naïve Bayes & decision tree," in 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT), 2017: IEEE, pp. 1-7.

[19] N. Maleki, Y. Zeinali, and S. T. A. Niaki, "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection," Expert Systems with Applications, vol. 164, p. 113981,2021.

[20] S. Babu, E. M. Vivek, K. P. Famina, K. Fida, P. Aswathi, M. Shanid, & M. Hena, “Heart disease diagnosis using data mining technique,” In 2017 international conference of electronics, communication and aerospace technology (ICECA) (2017, April). (Vol. 1, pp. 750-753). IEEE.