

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Library Philosophy and Practice (e-journal)

Libraries at University of Nebraska-Lincoln

August 2020

A REVIEW PAPER: ANALYSIS OF WEKA DATA MINING TECHNIQUES FOR HEART DISEASE PREDICTION SYSTEM

Basma Jumaa Saleh

Al-Mustansiriyah University, eng.basmaj@uomustansiriyah.edu.iq

Ahmed Yousif Falih Saedi

Al-Mustansiriyah University, ahmed.yousif@uomustansiriyah.edu.iq


Ali Talib Qasim al-Aqbi

Al-Mustansiriyah University, lali.al_aqbi@uomustansiriyah.edu.iq

Lamees abdalhasan Salman

Al-Mustansiriyah University, lameesiteng2013@gmail.com

Follow this and additional works at: <https://digitalcommons.unl.edu/libphilprac>

 Part of the [Biomedical Engineering and Bioengineering Commons](#), [Digital Communications and Networking Commons](#), [Library and Information Science Commons](#), and the [Robotics Commons](#)

Saleh, Basma Jumaa; Saedi, Ahmed Yousif Falih; al-Aqbi, Ali Talib Qasim; and Salman, Lamees abdalhasan, "A REVIEW PAPER: ANALYSIS OF WEKA DATA MINING TECHNIQUES FOR HEART DISEASE PREDICTION SYSTEM" (2020). *Library Philosophy and Practice (e-journal)*. 4032. <https://digitalcommons.unl.edu/libphilprac/4032>

A REVIEW PAPER: ANALYSIS OF WEKA DATA MINING TECHNIQUES FOR HEART DISEASE PREDICTION SYSTEM

Basma Jumaa Saleh ¹, Ahmed Yousif Falih Saedi ²,

Ali Talib Qasim al-Aqbi ³ and Lamees abdalhasan Salman⁴

¹Department of Computer Engineering, Al-Mustansiriyah University, Baghdad, Iraq
eng.basmaj@uomustansiriyah.edu.iq

²Department of Computer Engineering, Al-Mustansiriyah University, Baghdad, Iraq
ahmed.yousif@uomustansiriyah.edu.iq

³Department of Computer Engineering, Al-Mustansiriyah University, Baghdad, Iraq
lali.al_aqbi@uomustansiriyah.edu.iq

⁴Department of Computer Engineering, Al-Mustansiriyah University, Baghdad, Iraq
lameesiteng2013@gmail.com

Abstract

Data mining is characterized as searching for useful information through very large data sets. Some of the key and most common techniques for data mining are association rules, classification, clustering, prediction, and sequential models. For a wide range of applications, data mining techniques are used. Data mining plays a significant role in disease detection in the health care industry. The patient should be needed to detect a number of tests for the disease. However, the number of tests should be reduced by using data mining techniques. In time and performance, this reduced test plays an important role. Heart disease is a cardiovascular disease that causes death. Health problems are enormous in this recent situation because of the prediction and the classification of health problems in different situations. The data mining area included the prediction and identification of abnormality and its risk rate in these domains. Today the health industry holds hidden information essential for decision-making. For predicting heart problems, data extraction algorithms like K-star, J48, SMO, Naïve Bayes, MLP, Random Forest, Bayes Net, and REPTREE are used for this study (Weka 3.8.3) software. The results of the predictive accuracy, the ROC curve, and the AUC value are combined using a standard set of data and a collected dataset. By applying different data mining algorithms, the patient data can be used for diagnosis as training samples. The main drawbacks of the previous studies are that they need accurate and the number of features. This paper surveys recent data mining techniques applied to predict heart diseases. And Identifying the major risk factors of Heart Disease categorizing the risk factors in an order which causes damages to the heart such as high blood cholesterol, diabetes, smoking, poor diet, obesity, hypertension, stress, etc. Data mining functions and techniques are used to identify the level of risk factors to help the patients in taking precautions in advance to save their life.

Keywords:

Data mining, Heart Disease, Prediction, Classification, Dataset, WEKA tool, Prediction techniques.

INTRODUCTION

Heart attack complications tend to be the world's leading causative agent, with early diagnosis, to stop attacks. Physicians produce information with rich hidden material, and it is not efficiently used for forecasting. For this reason, the work turns the unused data onto a sculpting dataset utilizing various data mining strategies. People are dying with signs which have not been taken into account. Medical professionals must foresee heart disease before it occurs in any patient (1). The characteristics that make heart incidents more likely are:

Smoking: damages the arteries lining by creating a fat-type material, known as atheroma that reduces the arteries that cause heart attack,

High cholesterol: cholesterol is a waxy substance which is contained in the blood vessels ' fatty deposits. High cholesterol does not permit enough blood to reach the lungs that cause a heart disease,

Inappropriate diet: eating so much junk food raises cholesterol and blood pressure, which can lead to heart disease,

Lack of physical activity, raise cholesterol levels in the veins, which contribute to a risk of heart problems.

Harmful consumption of alcohol: a psychoactive sequence of use that causes health damage. The determination may be physiological (e.g. chlamydia) or emotional (e.g. primary to serious drinking depressed episodes). Harmful locational but not always have negative psychological implications; however, the social effects are not enough with themselves to warrant a diagnosis of damaging use (3),

High sugar rates: blood sugar readings greater than 180 mg / dL or any measurements above your normal range are excessive. A 300 mg / dL or higher blood sugar test can be harmful. Contact your doctor when you have two readings in a series of 300 or more. Practiced polyuria, increasing hunger and polyphagia (4), are usually seen in high insulin cases.

Overstress: the body's unspecific reaction to any demand is a commonly occurring phenomenon throughout its entire lifetime. It has been felt by all people in their culture and history. Stress is now one of the extra features of life and its nature has been emphasized so that it has been discussed in all eras of fine arts and fiction (5),

Blood pressure: a painful condition under which the blood heavy enough for the artery's walls may ultimately cause health conditions (6), ages, sex, the history of cardiac disease in your family, etc. A common condition. These causes can be used as risk factors for cardiac disease prediction (7).

The word cardiac disease involves many types of heart injury disorders. Heart diseases are common in:

1. Coronary heart disease: Coronary Heart disease is the most common type of heart disease in the world. It is also known as Heart disease. Statue residues obstruct the coronary channel, creating a decreased blood and oxygen flow to the cardiovascular.
2. Arrhythmias: it is linked to a condition of the heartbeat's repetitive movement. A low, rapid or abnormal heart beat can be. In addition to abnormal heart beats, the cardiac system has a malfunction.
3. Heart failure: it is a disease in which the heart cannot transfer sufficient blood to the human body. It is usually referred to as heart failure.
4. Congenital cardiac disease: it is also referred to as a congenital cardiac disorder and refers to the development and operation of an irregular carbonaceous state. Kids are also born with a congenital disease.
5. Cardiomyopathy: Undermining the muscle of the heart or altering the musculature because of an insufficient heart beating. Cardiomyopathy. High blood pressure, alcohol use, bacterial infections, and genetic disorders are the most common forms of cardiomyopathy.
6. Angina pectoris: is a medical treatment for angina that arises because of insufficient blood supplies to the heart; it is a warning for a cardiac arrest. Chest pain lasts several seconds or minutes.
7. Myocarditis: It is an infectious, fungal, and bacteria heart inflammation typically affecting the heart. It is a cardiac irritation. It's a rare illness that does not have direct connections with discomfort, inflammation of the arms, or temperature (8). The primary causes of death for individuals around the world are all these illnesses. WHO, and CDC, have confirmed that cardiovascular disease is the leading cause of death (7) and disease Prevention Center's.

Data mining in medical care is becoming ever more common in today's world because it provides a great deal of complex information involving hospital services, medications, medical equipment, patients, the diagnosis of diseases, etc. Such complicated data have to be processed and evaluated for information retrieval, which is both price-effective and helpful in decisions. In 2011, 17.5 million people died in heart disease, which is 31% of all international deaths in the World Health Organization. Of these, 7.4 million were affected by coronary heart disease and 6.7 million by cerebrospinal disorders. Almost 23,6 million individuals, approximately by World Health Organization 2030, will die of heart disease (9). Many hospitals electronically store their patient information via certain hospital system management systems. Such devices produce vast amounts of data every day. Such data can be organized in unlimited text or in picture form as in servers. Such data may collect valuable information for choice-making purposes. This assumption leads to the use of Information Development in Data sets, which transforms small-level data into heavy-level choice-making information. The findings can be used in successful decisions and research and can be further analyzed. Data mining is graded, clustered, analyzed and identified by interaction (8). Data mining applications predict future trends by knowledge-based choice-making. Detection of cardiovascular disease involves an enormous number of details, too difficult and massive to process and interpret modern techniques. Experts use a variety of data mining methods. Our goal is to figure out the data mining software, algorithmically effective.

The remainder of the text is organized accordingly. Section II provides a description of Literature and associated works. The technique for data mining is outlined in Section III. Section V eventually brings our results to a conclusion.

LITERATURE REVIEW and RELATED WORKS

Heart disease is a term used to refer to a wide range of heart-related health circumstances. Such medical issues define the pathological disorders of the heart as well as all portions of it explicitly. Heart disease is a significant health issue. The number of people who have heart disease has increased over the years (10). Several research based on the treatment of heart disease were performed. Various diagnostic data mining strategies were implemented and different probabilities were obtained. Several studies are being performed to test K-star, J48, SMO, Naïve Bayes, MLP, Random Forest, Bayes Net and REPTREE algorithms inefficiency. The tests so far show that Bayes Net outperforms random forests only rarely. To decrease the number of features while losing accuracy and effectiveness, the optimizing process using genetic algorithms is also expected to diagnostics of heart disease. To treat cardiac disorders, there are many potential techniques (11):

- A. Naïve Bayes: Naïve Bayes classification mathematical methods are used. The membership class is calculated using the principle of the probability of the principle of the Bayes. The algorithm of Naïve Bayes is built on implicit independence. It is the autonomy of feature values for a particular class with other feature values. The posterior likelihood of the response parameter is determined for the basic training database. Conditional possibility calculations are also carried out for the other parameters. Then for all the cases of the response variable, the probability of occurrence is calculated on behalf of each test dataset sample. The response variable which is having the highest probability of occurrence is then selected (12). With the Temporary Association, Orphanou et al. (13) developed a diagnostic system for heart disease depending on the Naïve Bayes algorithm. The horizontal support and the average feature display are compared for each TAR. The research used a strong sample. Through creating a feature development process, Alizadehsani et al. (14) had been diagnosed with coronary heart disease. The database used by the researchers is Z-Alizadeh with 303 samples for patients and 54 characteristics, data for the dataset is compared with the SMO method, Bagging, Naïve Bayes algorithm, and ANN results. With the Weka method, the accuracy of Naïve Bayes was 85.39% and the accuracy of 72.83% was achieved in SPSS Programmer. Through creating a feature development process, Alizadehsani et al. (14) had been diagnosed with coronary heart disease. The database used by the researchers is Z-Alizadeh with 303 samples for patients and 54 characteristics, data for the dataset is compared with the SMO method, Bagging, Naïve Bayes algorithm, and ANN results. With the Weka method, the accuracy of Naïve Bayes was 85.39% and the accuracy of 72.83% was achieved in SPSS Programmer. A. K. Ramotra et al. (15) suggested a model using a WEKA teaching tool to estimate heart diseases. There were 303 data and 76 features in the database. 297 data records with 13 input attributes were taken into account for the analysis after database re-processing and the elimination of the absent values. The authors estimated that the total accuracy was 85% using Naïve Bayes.

- B. Decision Trees (J48): The inner structure of the decision tree is just like a tree in which all inner nodes are a check on the features, in which all roots reflect the exam results available. Various choice measures of features help choose the feature for the class division of the data. Different types of decision trees are used in data mining. The major variance is the mathematical formula used in law-extraction to pick an attribute type. C4.5 is the predecessor to the algorithm for the ID3 decision tree. C4.5 software implements a greedy tree-build method with the split-and-conquer approach at its edge-down stage. The benefit rate is used by C4.5, an extension of knowledge gain, as a feature selection metric. ID3 is used for the attribute selection calculation by gaining information (16). A risk management model with several rates for loss of heart prediction was produced by Aljaaf et al. (17). The model will predict the risk of extremely high risk at five rates as a reference point without the use of C4.5 algorithms and 10 fold cross-validations. The Cleveland Clinic Foundation uses the data with 297 cases. The authors suggested that their overall accuracy is 86.30%. Alexopoulos et al. (18) suggested a model using C4.5 classificatory to predict heart disease using a data mining method. The writers used 1000 patient data with 114 features and five separate conditions to identify them. The writers said that the average accuracy was 74.15%. A. K. Ramotra et al. (15) suggested a model using a WEKA teaching tool to estimate heart diseases. There were 303 data and 76 features in the database. 297 data records with 13 input attributes were taken into account for the analysis after database re-processing and the elimination of the absent values. The authors estimated that the total accuracy was 74.15% using J48.
- C. K-star: The algorithm in the immediate vicinity operates on analogy techniques by contrasting the data for research with the data specified by similarity-based learning data known under the name of the memory-based strategy. The KNN method searches the sequence of the nearest K learning samples for the specified exam sample. The distance is usually computed using a Euclidean length method (8). Masetic et al. (19) suggested the PTB Clinical ECG datasets of BIDMC Congestive Heart Failure system to distinguish healthy and congestive heart attacks using data mining. As a method of extracting functions, Auto-regressive data is used. The classification was used in the SVM criteria, k-NN method, random forest, C4.5, and NN. The KNN classifier is used to achieve a precision of 82.02 percent in Weka A. K. Ramotra et al. (15) suggested a model using a WEKA teaching tool to estimate heart diseases. There were 303 data and 76 features in the database. 297 data records with 13 input attributes were taken into account for the analysis after database re-processing and the elimination of the absent values. The authors estimated that the total accuracy was 82.02% using 1BK.
- D. MLP: is a metaphor of the human brain used for information processing. A linked set of input and output units is included in the neural network. The network is trained to detect a pattern utilizing input data. Those weights are related to the interrelationships. The network knows and can accurately estimate the group labels by implementing weight changes. Multilayer feed is the kind of ANN in which the back-propagation method performs the training. If there is no loop of links, a neural network is called the feed-forward network. There are three layers in it the first and last layers are respectively known as the initial and the exit layer, and a layer between them is known as the hidden layer. All these layers have links Training sample attributes are passed to the network as input (20).

The inputs are then sent to the hidden layers together with the weights the output layer is the weighted output from the previously hidden layer which describes the group labels expected. A model for predicting cardiac problems via neural networks was suggested by Gharehchopogh et al. (21). The scientists used 40 individuals in their clinical reports. Blood pressure, sex, age, and smoking are the conditions used in the detection. The model correctly predicted 85 percent of the cases. Using MLP on the heart disease dataset has exceeded 80.89% accuracy in Weka software. A. K. Ramotra et al. (15) suggested a machine-learning model to use the WEKA method to predict cardiovascular disease. There were 303 data and 76 characteristics in the dataset. 297 data with 13 input features are required for the analysis after data pre-processing and elimination of the absent values. The authors claimed 80.89% of accuracy. An Efficient Heart Disease Detection System was introduced by Purushottam et al. (22) using data mining. This device can assist physicians in making decisions based on the parameter efficiently. The device is trained and tested by a 10 fold model, and the accuracy of 86.3% during the test and 87.3% during the training process is demonstrated. The authors reported that the MLP classification got an overall accuracy of 74.85%.

- E. Random Forest: Random Forest is one of the most famous and influential techniques for data mining. It is a software of data mining named Bagging or Web application Hierarchical clustering. The bootstrap is a really effective mathematical method to an approximate value from a data set like a medium. Several tests of data are taken, the average measured and all sample variables are then summed to give the true mean value the best prediction. The same approach is used for labeling, but assumption trees are usually used instead of calculating the mean of each sample. Many samples of test data are taken and models for each data sample are generated. Each model provides the prediction, but these predictions are averaged to better estimate the real value of the output (23). Although a prediction of every data is necessary. R. Jothikumar et al. (24) suggested a model using a learning algorithm for predicting heart conditions with 295 samples and 13 features apply to the Naive Bayes Algorithm in Quick Miner. This has 78,24% accuracy The other similar measurements are Kappa linear 0.499, absolute error 0.247 percent, relative error 24.19 percent, and RMSE is 0.378. Sarangam Kodati et al. (25) Suggested the precession analysis is 77.9% in Orange and Recall 73.4% for heart disease results. In WEKA precession 81.8% and Recall 81.9%. the Comparison between Orange software and WEKA, Weka is the best Recall and precession.
- F. SMO: John C. Platt (26) suggested the SMO method in 1998 and has been the quickest method for optimizing algebraic programming. SMO is utilized for preparing the vector classification supporter with algebraic kernels or RBF kernels. This substitutes the missing values for all conditional features and converts them into binary. The same concept as an SVM is used for a standard hidden layer of a neural network. An approach to machine learning to predict cardiac conditions by using the WEKA tool is suggested by Aung Nway Oo et al. (27) A design that uses an SMO and Lazy classification prevention strategy. For the prediction of heart disease the Weka data mining method has been used. Of instruction, 66% of the data set (training) and 34% (testing) for research.
- G. Bayes Net: The Bayesian network is a probability, concept-based visual prediction model. Bayesian networks consist of deterministic distributions and use probability

regulations for forecasting and diagnostic purposes. All discrete and continuous samples are provided in the Bayesian network. The network is defined as a set of data with acyclic directed graphs representing the conditional connections. The edges between the nodes in a Bayesian network are conditional, while non-connected nodes are provisionally independent (28).

A system for the classification of different techniques in data mining was suggested by Mirpouya Mirmozaffari et al. (29) to estimate heart disease. A specific model of various filters and examination methods has been developed.

The superior method and the more accurate clinical resolution assistance systems for the diagnosis of diseases are used for multilayer pre-process filtering, as are varying assessment methods. There are 209 examples and 8 features in the database. The Bayes Net Algorithm achieves 80.83 percent accuracy in Weka.

- H. REPTREE: REPTREE utilizes the logic of the evaluation tree and in various iterations produces several trees. It then chooses one of all the trees that grows best. The delegate will be listed. The calculation of cutting the tree is the average square error in the tree's estimates. In essence, "REPT" is a fast-paced randomized tree that is a monitored classifier, it is a choice-making ensemble, and it is built upon an algorithm of learning that produces many people. Gain or minimize the variance of information. REPTREE is used for a random dataset to generate a quick decision-making tree that creates a decision-making tree. That node decision tree with knowledge gain as a regular tree divides all parameters by the best division. Each node is divided by the better among the cuttings in dividing criteria, using decreased error of random forest. All numerical values are sorted once. A randomly selected subset of predictors at this node. Lost values are resolved by the variable instances utilizing an approach of C4.5. Leo Breiman presented the illustration of trees and a method for the UCI database is used and the Adele Cutler uncertainty is used. For classes of sex with 6 classification and reversion problems, the classifier may address both matrices (30). Mirpouya Mirmozaffari et al. (29) suggested a model of classifying different data mining prediction algorithms to estimate cardiovascular disease. A unique system consists of various filters and assessment methods. To find better algorithms and more reliable medical action support structures to treat diseases of the multilayer processing of the pre-process and multiple assessment methods are used. The database is made up of 209 samples and eight features. In Weka, the REPTREE classification achieves 82.77% accuracy.

DATA MINING APPLICATIONS

Data mining is utilized in different fields, such as food service, telecommunications, medical care, economic, network monitoring, sports, and student achievement analysis (2).

1. FOODSERVICE INDUSTRY: data mining is a wonderful retail industry application because it gathers extensive information, including transport, sales and use of commodities and service providers.
2. INDUSTRY TELECOMMUNICATION: Telecoms industry is the world's most advanced sector, offering a range of fax, internet, pager or e-mail services. Telecommunications services have been united with and work more essentially with the advancement of computers. Data mining helps determine trends of telecommunications, cheating, efficient use of resources, and enhance service quality.

3. **MEDICAL-CARE INDUSTRY:** Database extraction is very useful in the health service and in the treatment of heart disease, diabetes, and cancer. Data mining is very helpful in the healthcare industry. This leads to the detection of patient history patterns and variations with the same potential cause and assists in choice-making.
4. **Market data Evaluation:** dependable and high-quality financial data in the banking system that enables system-based data analyzes. It supports the prediction of mortgage payments and the evaluation of the consumer loan strategy. It also supports the grouping of aim of marketing consumers.
5. **INTERFERENCE DETECTION:** interference is any action that endangers internet bandwidth from any confidentiality or honesty. Interference monitoring has always been an important topic in network management, due to the growing use of the internet and the proliferation of resources and techniques for interference and breaching network. Data mining assistances to develop an intrusion prevention data mining method and to analyze flow data in order to prevent attacks to the invasion.
6. **SPORTS:** during sports, each person, team, match and seasonal of statistics are collected. For the prediction of the player's performance, team selection, and upcoming events, data mining is utilized.
7. **STUDENT'S EFFICIENCY:** data mining is used to determine the efficiency of the students by using a data classification tool. The student records collect attendees, tests, seminars, and qualifications to predict the student's efficiency at the end of the course.

Methodology

The goal of this analysis is to predict future heart problems successfully from the medical data collection. A model has been established using a prediction technique to assess the cardiac disease's features by certain features. Data mining is used to construct class prediction models based on chosen features in this work. Because of their capacity for exploration, study, and prediction of trends (31), Waikato Environment for Knowledge Study (WEKA) was used for prediction. The entire process can usually be divided into four steps:

4.1 TOOL Used

The Waikato Environment for Knowledge Analysis (WEKA) (32) is an accessible-source toolkit and software for machine learning launched by Waikato University, New Zealand. The structure is developed in Java and is distributed below the GNU General Public License. It operates on nearly every device and has been evaluated on Linux, Windows, Macintosh and a digital device. This offers a standard application for too many learning methods, pre-and post-processing techniques, and tests the effects of learning strategies on a single dataset. There are also various methods, such as formulas, for converting data sets. WEKA supports a number of specific data mining activities such as data pre-processing, grouping, classification, correlation, simulation, and functional choice. We may also use new WEKA algorithms for proven data mining and machine. WEKA provides a number of outlets since the loading of files, URLs and databases is included. This allows file formats to include WEKA's own ARFF format, CSV format, Lib SVMs, C4.5 and WEKA's other evaluation criteria such as uncertainty, accuracy, recall, true negative and positive, etc. It also includes many evaluation variables. Some benefits of the WEKA software contain Open Source, a standalone, portable

platform, and a graphical UI with a very large collection of various data mining methods. The Weka toolbox is a compilation of innovative algorithms and tools for data pre-processing. It is structured so that current approaches can be flexibly evaluated on novel datasets easily. It offers active support for the entire research data mining cycle such as preparing data input, the statistical evaluation of learning strategies and the visualization of information and learning outcomes. It contains a wide variety of pre-processing methods and several machine learning. A standard interface enables users to compare the different techniques and determine the methods best suited to the issue at hand, to access this robust and varied toolkit (30). A graphical user interface named Explorer as shown in Figure1 makes it easier to using Weka. The WEKA 2018 choice system and form completion (version 3.8.3) allow access to all its services.

Main Features

- data preprocessing tools
- classification/regression algorithms
- clustering algorithms
- algorithms for finding association rules
- 15 attribute/subset evaluators + 10 search algorithms for feature selection

Main GUI

- Three graphical user interfaces
- “The Explorer” (exploratory data analysis)
- “The Experimenter” (experimental environment)
- “The Knowledge Flow” (new process model inspired interface)

We have introduced five classifiers to predict the algorithms based on their normal use. The following are the classifiers (33):

Table 1. Commonly used Algorithms

Generic Name	WEKA Name
Bayesian Network	Naïve Bayes(NB)
Support Vector Machine	SMO
C4.5 Decision Tree	J48
K-Nearest Neighbor	1Bk

Which is shown in Fig 1 the GUI Chooser comprises five icons (41):

Explorer: A data discovery framework with the WEKA.

Experimenter: a research area and a statistical testing system.

Knowledge Flow: This framework supports the same features as the Explorer, but has an ability to push and drop. One benefit is that sped up learning is assisted.

Workbench: The workbench of Weka is an integrated set of advanced algorithms for the learning of machines and pre-processing computer instruments.

Simple CL: provides an easy battalion-line interface that enables WEKA functions to be executed directly on systems not having their own battalion-line interface. In many operation areas, this Java version is used, especially for academic and scientific purposes. Weka has different benefits:

1. Underneath the Public Authorization, it is available freely.
2. It is scalable because it is implemented fully and operates on almost all architectures in Java language.
3. This is an extensive set of re-processing and simulation approaches.
4. The user interface makes it much easier to use. Including re-processing, clustering, categorizing, analysis, simulation and function choice, Weka support multiple various data mining activities.



Figure 1. Weka GUI chooser.

- Exploring the Data (42):
Figure 2 shows the most important graphical user interface, the Explorer. There are six tabs, accessible from the top tabs corresponding to the different supporting data mining functions. Data can be downloaded from a folder or retrieved from a server using a query of an SQL within the "reprocess" window shown in Figure 2. The folder might be in the CSV or ARFF format of the process. Java Server Connectivity provides data accessibility to enable SQL queries to be put in any server with an appropriate driver. After a dataset is read different post-processing methods known as "filters" can be used numerical data can be discretized, for example. The consumer has downloaded a data file in Figure 2 and focuses on one particular feature, minimized losses, and a histogram.

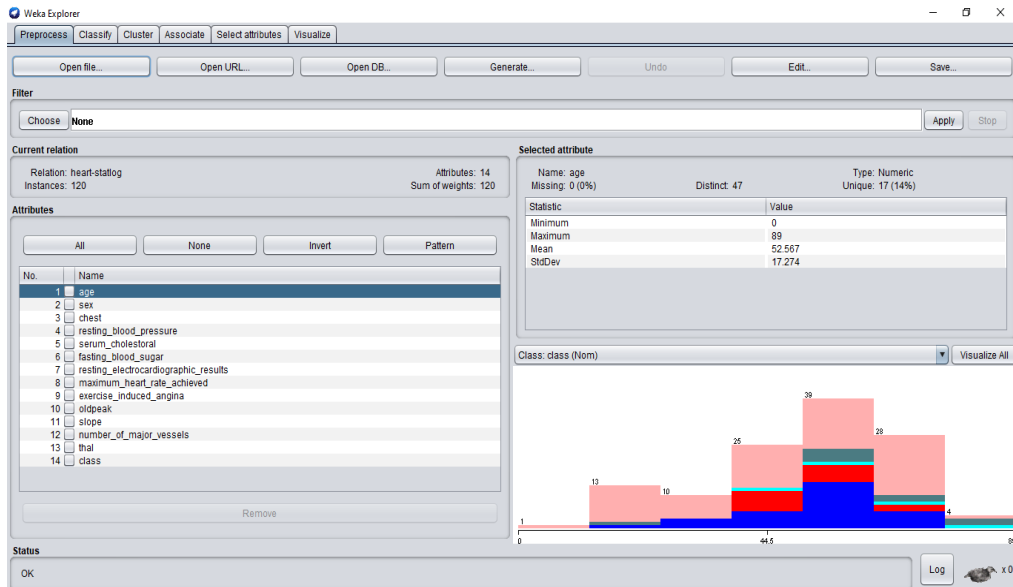


Fig. 2. The Explorer Interface.

The "classification, and simulation algorithms" for re-processed data can be implemented via a second function of the explorer. This table allows users to test the resultant models including numerically by statistical calculations and graphically by visualizing the data and by analyzing the model. Consumers could load and store models.

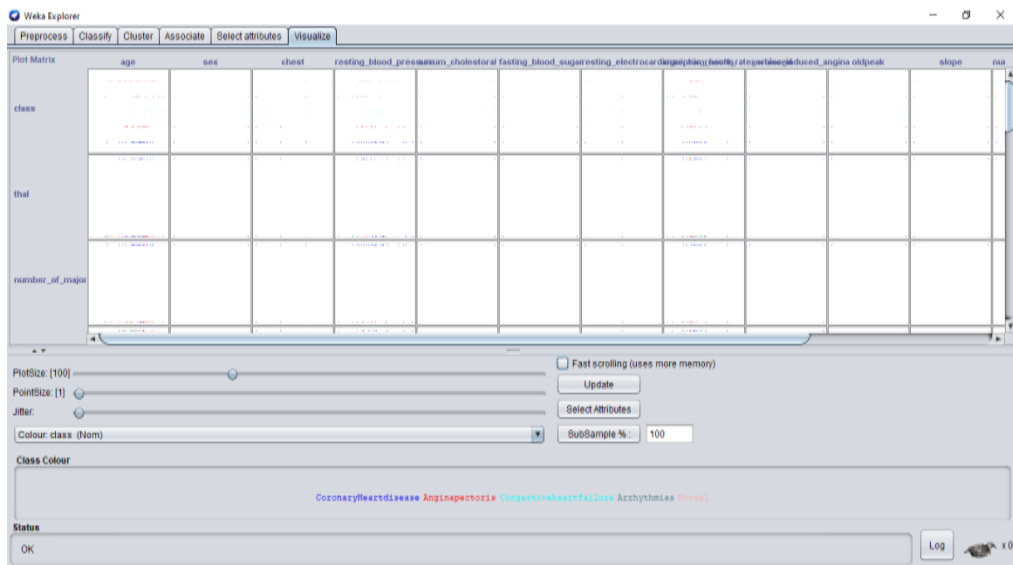


Fig. 3 The Knowledge Flow Interface.

The third section, "cluster," allows users to choose the set of data for grouping algorithms. The findings can again be visualized and analyzed based on the statistical probability of the data if the fragments represent intensity predictions. Clustering is one of two data analysis approaches that need to be estimated without a specific goal attribute. The other involves association laws allowing users to evaluate the data in a business-basket sort. The fourth section, 'Associate,' provides exposure to training association principles algorithms. The next section supports the collection of attributes, a significant data mining activity. It includes access to different methods

for calculating the efficacy of parameters and for the search for data descriptive subgroups of attributes. The final panel "Show" helps users who want to analyze the data digitally, which includes a color-coded matrix of the dispersion graphics, and users can choose and extend individual graphs. Sections of the data can be zoomed through, the specific record behind a specific data point can be retrieved and so on. The exponential learning cannot be done with the Explorer framework since the modular Panel loads the set of data in their entirety. This shows that only low to the medium of sized challenges can be used. Other exponential algorithms can process large data sets. The command-line manager provides access to all device functionality is one way of implementing this. The second main graphical user interface, known as 'Knowledge Flow,' is an alternative method that is more practical. Figure 4 shows that this allows users to enter a data flow through graphically linking components comprising sources of data and tools for pre-processing, training algorithms, evaluation processes, and viewing tools. Information can be loaded as in the Explorer in collections, or incrementally downloaded and manipulated by the filters and training algorithm which can be learned in increments.

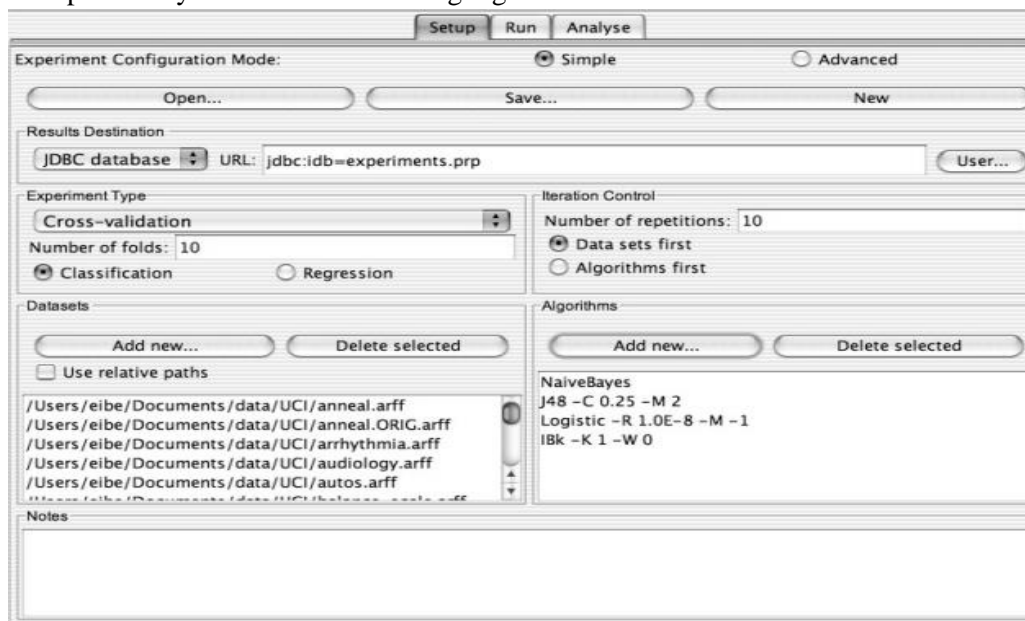


Fig. 4. The Experimenter Interface.

4.2 PERFORMANCE METRICS

In this section (34) the metrics utilized in research are described:

1. PRECISION

Precision is the part of major instances between both the instances collected. The Eq. of precision is set in Eq. (1)

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (1)$$

2. RECALL

The recall is the small part of the suitable instances in the total number of specific instances. The Eq. of recall is set in Eq. (2).

$$\text{Recall} = TP / (TP + FN) \quad (2)$$

3. F-MEASURE

Depending on the 2-times precision reminder time separated by the amount of precision and reminder, the f-measure is investigated. The equation of F-Measure is set in Eq. (3).

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (3)$$

4. ROC AREA

Roc formulas are widely used as graphic displays for any cut off for an examination or a combination of tests, the relations, and trade-off including a clinical sensitiveness and precision.

5. PRC AREA

Curves for precise recall are not influenced by the number or lower test scores of patients without a diagnosis. To obtain the whole picture during evaluation and comparison, it is highly recommended to just use accuracy recording curves to complement the ROC formulas. The product of the binary classification (35), as illustrated in Table 3.

Table (3): different outcomes of a two-class prediction

Actual class	Predicted class	
	YES	No
YES	True Positive (TP)	False Negative (FN)
No	False Positive (FP)	True Negative (TN)

- True Positive (TP): Patients were reasonably predicted to be positive (the patients are supposed to need heart disease and need heart catheterization).
 - True negative (TN): If TP and TN are approximately 100 percent, the model is ideally predicted to be negative, because they are not supposed to have cardiac catheterization.
 - TN is a true negative: healthy individuals are classified correctly as healthy.
 - FN is False Negative: Mistakenly identified as stable (36) patients with heart disease.
 - Correct Classified Instances (CCI): It reflects the proportion of patients diagnosed appropriately who need and do not need heart catheterization.
- Accuracy (37) is also known by eq. (4).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

- Mean absolute error (MAE): a prediction-to-outcomes measure. It is possible to compute 1- ACC. A good system has an absolute mean error (38) very high.

- Kappa: Kappa tests prediction alignment with a true class. The kappa statistical result is a value in the range 0-1. A value is higher than 0 means that the classifier is better than average. (39) Kappa is a predictor of the real class.
- Root mean Squared Error (RMSE): the variance between the expected value and observable value (40) is the root mean Squared Error.

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n \left(\frac{\rho_{i,j} - \tau_j}{\tau_j} \right)^2} \quad (5)$$

$\rho_{i,j}$ = predicted value.

i = fitness.

τ_j = fitness applicability target value for j .

Conclusion

Heart disease has now become one of the main causal issues. Only by taking into account characteristics is it possible to predict heart disease. This approach can be accomplished by integrating data extraction methods. This technique can be analyzed. Data mining method includes K-star, J48, SMO, Naïve Bayes, MLP, Bayes Net REPTREE, etc. More must be done to predict heart disease in a large data framework. The analysis of certain published papers by the researcher found that various methods used show various accuracies based on the number of features taken and the resources used to execute by software (Week 3.8.3).

Acknowledgments

The authors would like to thank Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad- Iraq for its support in the present work.

References

1. *A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level*. Ashish, Kumar, Sen, Shamshe, r Bahadur, Patel and D., P., Shukla. 9, 2013, International Journal of, Vol. 2, pp. 1663–1671.
2. *REVIEW ON PREDICTION SYSTEM FOR HEART DIAGNOSIS USING DATA MINING TECHNIQUES*. Divya, Kundra and Er., Navpreet, Kaur. 5, 2015, International Journal of Latest Research in Engineering and Technology (IJLRET), Vol. 1, pp. 09-14.
3. *Initial preference for drinking goal in the treatment of alcohol problems: II. Treatment outcomes*. Adamson, SJ, et al., et al. 2, 2010, Alcohol and Alcoholism, Vol. 45, pp. 136-142.
4. Lal, B Suresh. DIABETES: CAUSES, SYMPTOMS AND TREATMENTS. *Public Health Environment and Social Issues in India*. 1. s.l. : Serials Publications, 2016, 5.

5. *Stress: Facts and Theories through Literature Review*. **Amir, Mohammad, Shahsavarani, Esfandiar, Azad, Marz, Abadi and Maryam, Hakimi, Kalkhoran.** 2, 2015, International Journal of Medical Reviews, Vol. 2, pp. 230-241.
6. **Control, South Carolina Department of Health and Environmental.** What Is High Blood Pressure. <https://dc.statelibrary.sc.gov/handle/10827/25131>. [Online] 2017. [Cited: December 15, 2019.]
7. *Analysis of data mining techniques for heart disease prediction*. **Marjia, Sultana, Afrin, Haider and Mohammad, Shorif, Uddin.** Dhaka, Bangladesh : IEEE, 2016. 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT).
8. **Sujata, Joshi and Mydhili, K., Nair.** Prediction of Heart Disease Using Classification Based Data Mining Techniques. [book auth.] L. Jain, et al., et al. *Computational Intelligence in Data Mining - Volume 2 Smart Innovation, Systems and Technologies*. s.l. : Springer, New Delhi, 2014, Vol. 32, pp. 503-511.
9. *A Data Mining Approach for Prediction of Heart Disease Using Neural Networks*. **Chaitrali, S., Dangare and Sulabha, S., Apte.** 3, 2012, International Journal of Computer Engineering and Technology (IJCET), Vol. 3.
10. Heart Disease—General Info and Peer reviewed studies. <http://www.aristoloft.com>. [Online] [Cited: December 15, 2019.]
11. *Heart Disease Diagnosis Using Predictive Data mining*. **B., Venkatalakshmi and M., V., Shivsankar.** 3, 2014, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, pp. 1873-1877.
12. *Heart disease prediction system based on hidden Naïve Bayes classifier*. **Jabbar, M., A. and Samreen, S.** s.l. : IEEE, 2016, 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), pp. 1-5.
13. *Incorporating repeating temporal association rules in Naïve Bayes classifiers for coronary heart disease diagnosis*. **Orphanou, K., et al., et al.** s.l. : Elsevier, 2018, Journal of Biomedical Informatics, Vol. 82, pp. 74-82.
14. *A data mining approach for diagnosis of coronary artery disease*. **Alizadehsani, R., Habibi, J., Hosseini, M.J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A.,.** 1, s.l. : Elsevier, 2013, Computer Methods and Programs in Biomedicine, Vol. 111, pp. 52-61.
15. *Comparative Analysis of Data Mining Classification Techniques for Prediction of Heart Disease Using the Weka and SPSS Modeler Tools*. **Atul, Kumar, Ramotra, Amit, Mahajan and Rakesh, Kumar.** s.l. : Springer, 2020, Smart Trends in Computing and Communications. Smart Innovation, Systems and Technologies, Vol. 165, pp. 89-97.
16. *Data Mining Concepts and Techniques*. **Han, J. and Kamber, M.** 2006, Morgan Kaufmann Publishers.

17. *Predicting the likelihood of heart failure with a multi level risk assessment using decision tree.* **Aljaaf, A.J., Al-Jumeily, D., Hussain, A.J., Dawson, T., Fergus, P., Al-Jumaily, M.** Beirut : s.n., 2015. 2015 Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAECE). pp. 101-106.
18. *Medical diagnosis of stroke using inductive.* **Alexopoulos, E., Dounias, G. and Vemmos, K.** 1999. In Proceedings of Workshop on Machine Learning in Medical Applications, Advance Course in Artificial Intelligence-ACAI99. pp. 20-23.
19. *A congestive heart failure detection using random forest classifier.* **Masetic, Z. and Subasi, A.** s.l. : Elsevier, 2016, Computer Methods and Programs in Biomedicine, Vol. 130, pp. 54-64.
20. *Effective data mining using neural networks.* **Lu, H., Setiono, R. and Liu, H.** s.l. : IEEE, 1996, IEEE Trans. Knowl. Data Eng. , pp. 957–961.
21. *Neural network application in diagnosis of patient: a case study.* **Gharehchopogh, F., .S. and Khalifelu, Z., A.** Abbottabad : IEEE, 2011. International Conference on Computer Networks and Information Technology. pp. 245–249.
22. *Efficient Heart Disease Prediction System.* **Purushottama, Kanak, Saxenab and Richa, Sharma.** s.l. : Elsevier, 2016, Procedia Computer Science, Vol. 85, pp. 962 – 969.
23. *Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System.* **Shadman, Nashif, et al., et al.** 2018, World Journal of Engineering and Technology, Vol. 6, pp. 854-873.
24. *Analysis of Classification Algorithms for Heart Disease Prediction and its Accuracies.* **R., Jothikumar and R., V., Sivabalan.** 2016, Middle-East Journal of Scientific Research, pp. 200-206.
25. *Analysis of Heart Disease using in Data Mining Tools Orange and Weka.* **Sarangam, Kodati and R., Vivekanandam, Sri.** 1, 2018, Double Blind Peer Reviewed International Research Journal, Vol. 18, pp. 16-22.
26. **Platt, John.** *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.* s.l. : Technical Report MSR-TR-98-14, 1998.
27. *SMO AND LAZY CLASSIFIERS FOR HEART DISEASE PREDICTION.* **Aung, Nway, Oo and Thin, Naing.** 2, 2019, IJARIIIE, Vol. 5, pp. 2395-4396.
28. *Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques.* **C., Beulah, Christalin, Latha and S., Carolin, Jeeva.** s.l. : Elsevier, 2019, Informatics in Medicine Unlocked , Vol. 16, pp. 1-9.
29. *Data Mining Apriori Algorithm for Heart Disease Prediction.* **Mirpouya, Mirmozaffari, Alireza, Alinezhad and Azadeh, Gilanpou.** 1, 2017, International Journal of Computing, Communication and Instrumentation Engineering, Vol. 4, pp. 20-23.

30. *Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News.* **Sushilkumar, Kalmegh.** 2, 2015, IJISSET - International Journal of Innovative Science, Engineering & Technology, Vol. 2, pp. 438-446.
31. **Witten, I. H. and Frank, E.** *Data Mining Practical Machine Learning Tools and Techniques.* s.l. : Morgan Kaufman Publishers, 2005.
32. **Kirkby, R.** *WEKA Explorer User Guide for version 3.* s.l. : University of Weikato, 2002. pp. 3-4.
33. *Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools.* **Kumar, M. Nikhil, Koushik, K. V. S. and Deepak, K.** 3, 2018, International Journal of Scientific Research in Computer Science, Engineering and Information Technology ©, Vol. 3, pp. 887-898.
34. *HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES.* **Benjamin, H., Fredrick, David and S. Antony, Belcy.** 1, 2018, ICTACT JOURNAL ON SOFT COMPUTING, Vol. 9.
35. *Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford Exercise Testing (FIT) Project.* **S. Sakr, R. Elshawi, A. Ahmed, W.T. Qureshi, C. Brawner, S. Keteyian, M.J. Blaha, M.H. Al-mallah, H. Ford.** 4, 2018, PLoS ONE, Vol. 13, pp. 1-18.
36. *Comparative study of Data Mining Approaches for prediction Heart Diseases.* **S, Hari Ganesh and M, Gajenthiran.** 7, 2014, IOSR Journal of Engineering (IOSRJEN), Vol. 4.
37. *Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods.* **Dbritto, Rovina, Anuradha and Joseph, Vincy.** 2, 2016, International Journal of Applied Information Systems (IJ AIS), Vol. 11.
38. *Intelligent heart disease prediction in cloud environment through ensembling.* **Gupta, Nishant, et al., et al.** 2017, Expert Systems.
39. *Cardiac Catheterization Procedure Prediction Using Machine Learning and Data Mining Techniques.* **Huda, Kutrani and Saria, Eltalhi.** 1, 2019, IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 21.
40. *Comparative Analysis of Data Mining Classification Techniques for Heart Disease Prediction.* **Dhara, Mehta and Nirali, Varnagar.** 12, 2018, International Research Journal of Engineering and Technology (IRJET), Vol. 5.
41. *A Study on WEKA Tool for Data Preprocessing, Classification and Clustering.* **Swasti, Singhal and Monika, Jena.** 6, 2013, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 2.
42. **al., Frank E. et.** *Weka-A Machine Learning Workbench for Data Mining.* [book auth.] Rokach L. (eds) Maimon O. *ata Mining and Knowledge Discovery Handbook.* 2nd. Boston : Springer Science+Business Media, 2010, pp. 1269-1277.

