Performance Analysis of Various Classification Algorithms in Heart Diseases Using Orange Tool

[1] Sarangam Kodati, [2] Nara Sreekanth, [3]K.Pradeep Reddy, [4]Hari Kumar Bandaru
[1,2,3,4] Assistant Professor
[1]Department of Computer Science and Engineering, Brilliant Institute of Engineering & Technology - , Hyderabad, Telangana ,India.
[2.] Department of Computer Science and Engineering, BVRIT Hyderabad College of Engineering for Women, Hyderabad,Telangana,India.
[3] Department of Computer Science and Engineering, Tirumala engineering college,Bogaram, Hyderabad, Telangana ,India.
[4]Department of Computer Science and Engineering, Brilliant Institute of Engineering & Technology - , Hyderabad, Telangana ,India.

***Abstract*--** Data mining is one of the essential areas on research up to expectation is more popular within health organization. Data mining plays an effective role for discovery recent trends in healthcare organization as is helpful because of every the parties associated with this field. Heart disease is the leading cause regarding death within the world over the past 10 years. Heart disease is a term that assigns to a large number of medical conditions related to heart. Performance of a number of classification algorithms certain namely Decision trees, Support Vector Machine, KNN, Naive Bayes is compared after it are applied over two type discrete heart disease dataset. Comparison concerning performance into prediction n on benign and heart disease categories are found out..

Keywords: Data mining, Orange Tool, Heart disease, Classification Algorithms

I.INTRODUCTION

Among all fatal disease, heart attacks diseases are considered as the most prevalent [1]. Medical practitioners conduct different surveys on heart diseases and gather records on heart patients, theirs symptoms and disease progression. Increasingly are reported in relation to patients together with common diseases who have typical symptoms. Thus, in that place is valuable information hidden in their dataset to be extracted. After creating order in “multiple bags of words or a data set” the next aim is “mining” the data for knowledge discovery. Mining data in a structured format e.g. multiple databases. or text mining: how to deal with unstructured data e.g. natural language documents. An extra challenge in mining data is trying to find related data in other resources, and clustering data.

Data mining aims to find useful patterns in data. A problem for many enterprises is the large availability of rich data. More specific to extract useful information from these large amounts of data. Analyzing (often) large datasets and trying to find trivial (hidden) patterns / relationships is a challenge. With the growing amount of data it is harder to retrieve knowledge from several datasets Data mining may help to find unsuspected relations between data. Data mining is also recognized so Knowledge Discovery from Data (KDD). Data mining is used to replace or enhance human intelligence by scanning through massive storehouses on data according to discover meaningful new correlations[2]. Data mining consists of an iterative sequence of the following steps

1) Data cleaning (to remove noise and inconsistent data)

2) Data integration (where multiple data sources might also be combined)

3) Data selection (where data relevant to the analysis challenge are retrieved from the database)

4) Data mining (an imperative process where intelligent methods are applied in order to extract data patterns)

5) Pattern evaluation (to perceive the truly interesting patterns representing knowledge based about some interestingness measures)

6) Knowledge presentation (where visualization and knowledge representation techniques are used in conformity with present the mined knowledge according to the user)

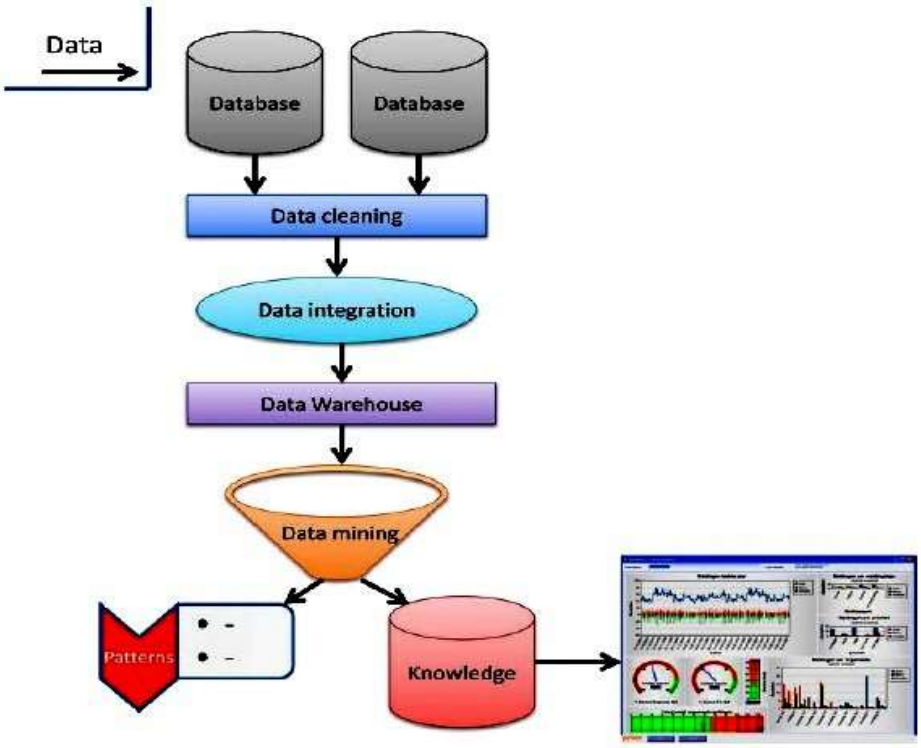


Fig No:1 Knowledge Discovery from Data

II.CLASSIFICATION ALGORITHMS

The system which does classification needs knowledge about the data and its associations and so the system has to be trained. The training dataset is needed to train the system and then the system will be ready for classifying the real time data. In this paper various classification algorithms such as Decision Tree, KNN,SVM, Naïve Bayes are used to classify the heart disease data.

II. CLASSIFICATION ALGORITHMS

2.1. DECISION TREE

Decision tree is similar according to the flowchart in which every non-leaf nodes denotes a check on a particular attribute and every branch denotes an result regarding that test and every leaf node have a class label. The node at the top most labels into the tree is referred to as root node. Using Decision Tree, decision makers can choose best alternative then traversal from root according to leaf indicates unique class split based regarding maximum information gain[6]. Decision trees are produced by algorithms that are used in imitation of identify various methods of splitting a data set in segments. These segments form an inverted decision tree. That decision tree originates including a root node at the top about the tree.

2.2. K- NEAREST NEIGHBOR CLASSIFIER

KNN is one of the most simple and straightforward lazy learning data mining technique. It is also called as memory-based classification namely the training samples need according to remain in the memory at runtime [8]. KNN will become popular due to its simplicity and relatively high convergence speed.KNN is known as lazy learning because it does not hold any training phase.

In the classification step, we will be given an instance S; whose attributes we wish refer to as much S.A i and we wish according to know instance class.KNN classification has two stages 1) Find the k instances in the data set so much are closest to S 2) These k instances then vote according to determine the category of S[3].

2.3. NAÏVE BAYES

A naive Bayes classifier is a simple and efficient probabilistic classifier based regarding applying Bayes' theorem including strong (naive) independence assumptions[4]. It considers as the presence/absence of a specific attribute about a class is another according to the presence/absence of any other attribute when the class variable is given [7]. It helps according to make the count process at all convenient and that has got better speed and accuracy because huge data. Also, it has the ability according to calculate the most viable output based over the input and according to add new raw data at runtime [7].

2.4. SUPPORT VECTOR MACHINE

A support vector machine is a type of model used according to analyze data and discover patters in classification or regression analysis. Support vector machine (SVM) is used when thy data has exactly two classes. An SVM classifies data through discovering the best hyper plane so much separates all data points regarding certain class beside those on the other class[1]. The larger margin in the two classes, the better the model is. A margin must have no points within its interior region. The support vectors are the data points so over the boundary on the margin. SVM is based totally regarding mathematical features or used according to model complex, and real world problems.

.

III. ORANGE TOOL

Orange is an open source machine learning technology and data mining software. Orange can remain used for exploration data analysis and visualization. It gives a platform because experiment selection, predictive modeling, and recommendation systems and to be used among gnomic research, bio medicine, bioinformatics, and teaching. It is intended because both experienced users and researchers about machine learning any want to prototype new algorithms while reusing as like much of the code as much possible, and because of those just entering the field who perform either write short Python scripts because of data analysis and enjoy into the robust while easy-to-use visible programming environment. Orange is always preferred so the component of innovation, quality, or reliability is involved. Orange includes a range about techniques, such and data management and preprocessing, supervised and unsupervised learning, performance analysis, and a range of data and model visualization techniques.

IV. HEART DISEASE DATASETS

We performed computer simulation regarding one dataset. Dataset is a Heart dataset. The dataset is available in UCI Machine Learning Repository. Dataset carries 303 samples and 14 input features as well as 1 outturn feature. The applications pencil a financial, personal, then convivial characteristic over mortgage applicants. The output feature is the decision class which has value 1 because of Good credit and 2 for Bad credit. The dataset-1 contains 700 instances shown as a Good credit while 300 instances as bad credit. The dataset contains features expressed on nominal, ordinal, or interval scales. [5].

V. PROPOSED MODEL

The proposed Model is implemented and tested using Orange data mining tool.
i) The data set is supplied to Classification Tree Algorithm which uses Entropy based attribute selection methodology.
ii) The dataset is supplied to SVM classifier which classifies based on finding linear separator.
iii) The dataset is supplied to logistic regression which classifies by creating regression equation
iv) The training data is supplied to predictions to use that as test data.
v) The results are compared using ConfusionMatrix.

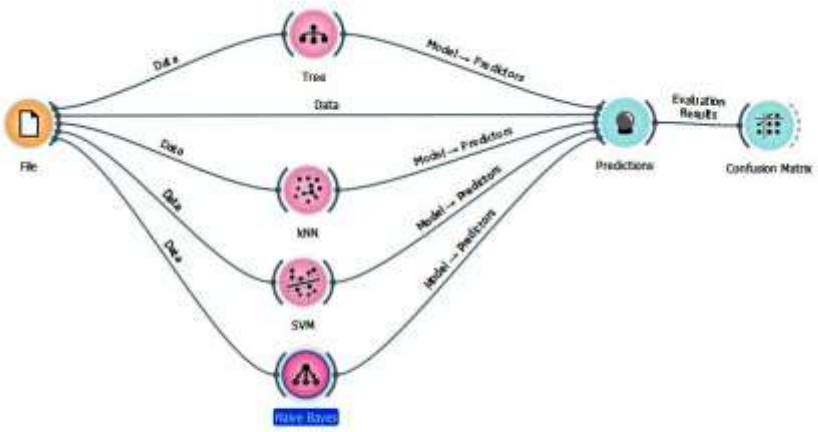


Fig No:2 Classification using in Orange tool Work flow Diagram

VI. RESULTS OF PREDICTIONS

Once the classification algorithms got trained, the entire training data set is given as test data, and the trained systems predicted the classes and the confusion matrix of those predictions are given below.

Table1: Decision Tree

		Predicted 0	Predicted 1	Σ
Actual	0	96.3 %	3.7 %	164
	1	7.9 %	92.1 %	139
	Σ	169	134	303

Table 2: K- Nearest Neighbor

		Predicted 0	Predicted 1	Σ
Actual	0	80.5 %	19.5 %	164
	1	28.1 %	71.9 %	139
	Σ	171	132	303

Table 3 Support Vector Machine

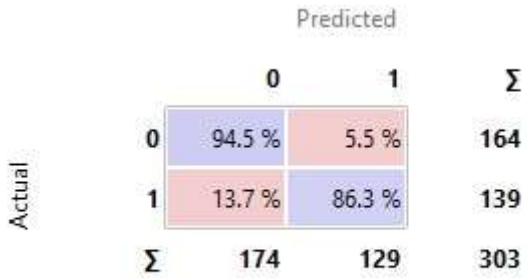
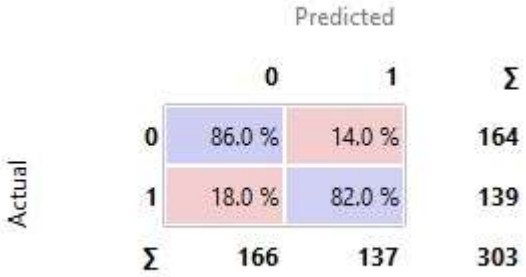


Table 4 : Naive Bayes



VII. CONCLUSION

Heart Disease is a fatal disease by its nature. This disease makes a life threatening complexities such as heart attack and death. The importance of Data Mining in the Medical Domain is realized and steps are taken to apply relevant techniques in the Disease Prediction. The various research works with some effective techniques done by different people were studied. The observations from the previous work have led to the deployment of the proposed system architecture for this work. Though, various classification techniques are widely used for Disease Prediction, The comparison of various classification algorithms here shows SVM classifiers is the best classifier for the Heart disease . The further study can be designed in such way to analyze very large data set using SVM.

VIII. REFERENCES

[1] Yuh-Jeng Lee. Smooth Support Vector Machines. Preliminary Thesis Proposal Computer Sciences Department University of Wisconsin. 2000

[2] R. Jyoti, G. Preeti, "Analysis of Data Mining Techniques for Diagnosing Heart Disease," International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 5, ISSUE. 7, July 2015

[3] . Prototype Selection for Composite Nearest Neighbor Classifiers. Department of Computer Science University of Massachusetts. 1997

[4] Chotirat Ann and DimitriosGunopulos. Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection. Computer Science Department University of California.

[5] Sarangam Kodati, Dr. R Vivekanandam and Dr. R P. Singh, Comparative Analysis In Diagnosis of Heart Disease With Data Mining Orange Tool, Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 09-Special Issue, 2018

[6] Apte & S.M. Weiss, Data Mining with Decision Tree and Decision Rules, T.J. Watson Research Center we issue_with_cover.pdf,(1997).

[7] A Comparative Study of Classification Techniques in Data Mining Algorithms. ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY. 8: p. 13-19

[8] E. Alpaydin - Voting over multiple condensed nearest neighbors, Artificial intelligence review, pp 115-132,1997