"A Novel Approach Prediction of Heart Disease using Ordinary decision Tree, J48, Random Forest, RandomTree and Rep Tree in Weka Tool"

Er.Meha Seth, Er. Parminder Singh, Dr. Naveen Dhillon
M.Tech Scholar in R.I.E.T, Phagwara
Head of Computer Science Department in R.I.E.T, Phagwara
Principal in R.I.E.T, Phagwara

ABSTRACT

The Healthcare industry collects large amounts of Healthcare data, but unfortunately not all the data are mined which is required for discovering hidden patterns and effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining modeling techniques can help remedy this situation. Cardiovascular disease is the leading cause of death in many countries. 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. 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 and stress. 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. In our research we explore with different classification trees like ordinary decision tree (ODT), Hoeffding tree, J48 (C4.5), Random Forest (RF), Random tree (RT) and REP tree for rule induction in order to identify high quality data set. To execute of classification decision trees, we will use WEKA tool on the dataset. Finally results of all trees are compared using the performance measures like accuracy, correct and incorrect instances and execution times. Based on validation results on the data set, Random forest (RF) and Random tree (RT) are highest correct instances and in case of root mean square error (RMSE), Random tree (RT) is on top with lowest error rate as compare to others decision tree algorithms.

.

Keyword: - *Data mining, ODT, J48, RF, REP Tree and RT.*

---

1. **INTRODUCTION: -** Data mining is defined as the process in which useful information is extracted from the raw data. In order to acquire essential knowledge it is essential to extract large amount of data. This process of extraction is also known as misnomer [1]. Currently in every field, there is large amount of data is present and analyzing whole data is very difficult as well as it consumes a lot of time. This present data is in raw form that is of no use hence a proper data mining process is necessary to extract knowledge. The process of extracting raw material is characterized as mining. This is a world where having a lot of information leads to power and success and this is possible only because of sophisticated technologies such as satellites, computers [2].

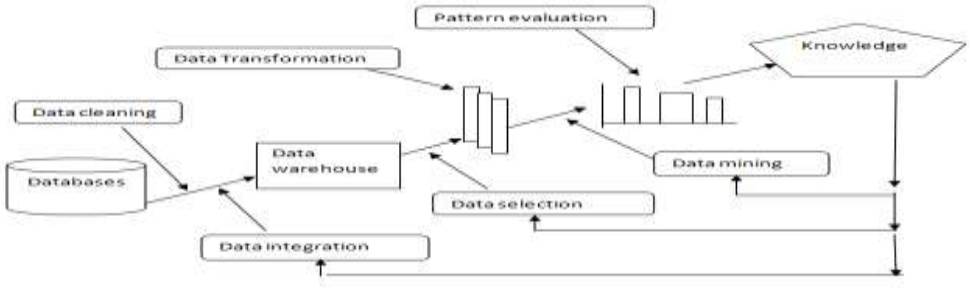

Figure: 1.1 Data Mining Process

1.2 Different functions of data mining

- **Association:** Association is classified as the best technique among others in the data mining technique. For the transaction of the similar data from one particular image to other, can be done with the help of association in which a pattern is discovered on the basis of relationship [3].
- **Clustering:** In the data mining technique, clustering is a technique in which clustering of objects are identified. An automatic technique has been utilized for this purpose as it has the similar characteristics. This clustering technique defined the classes and objects in order to define the process that how objects are assigned into a predefined classes.
- **Classification:** In the data mining technique, classification is a method that is based on machine learning. In the classification of the data, each item in the data set is classified into predefined set of groups. There are several mathematical techniques that have been utilized by the classification method.
- **Prediction:** It is possible to recognize a relationship among the variables which are dependent and independent using the data mining's another approach named as prediction. In the various fields this techniques can be utilized in order to predict profit for future. **Outlier detection:** In many cases, simply recognizing the overarching pattern can't give you a clear understanding of your data set. You also need to be able to identify anomalies, or outliers in your data [4].
- **Regression:** Regression, used primarily as a form of planning and modeling, is used to identify the likelihood of a certain variable, given the presence of other variables. For example, you could use it to project a certain price, based on other factors like availability, consumer demand, and competition.
- **Tracking patterns:** One of the most basic techniques in data mining is learning to recognize patterns in your data sets. This is usually recognition of some aberration in your data happening at regular intervals, or an ebb and flow of a certain variable over time. **Evolution and deviation analysis:** Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data.

1.3 Decision Tree Classifiers

Decision Tree Classifier is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receive an answer, a follow-up question is asked until a conclusion about the class label of the record is reached[4]. The classification technique is a systematic approach to build classification models from an input data set. For example, decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and naive Bayes classifiers are different technique to solve a classification problem.

- The decision tree classifiers organized a series of test questions and conditions in a tree structure. The following figure shows a example decision tree for predicting whether the person cheats. In the decision tree, the root and internal nodes contain attribute test conditions to separate records that have different characteristics. All the terminal node is assigned a class label Yes or No.

- Once the decision tree has been constructed, classifying a test record is straightforward. Starting from the root node, we apply the test condition to the record and follow the appropriate branch based on the outcome of the test. It then lead us either to another internal node, for which a new test condition is applied, or to a leaf node. When we reach the leaf node, the class label associated with the leaf node is then assigned to the record.

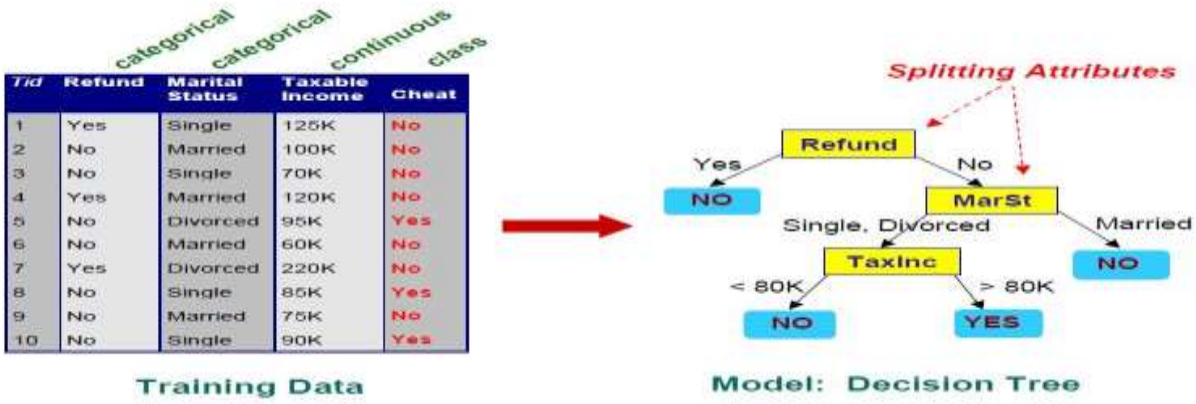

Figure 1.2: Decision tree methods.

2. REVIEW OF LITERATURE

[1] Ievgen Meniailov et. al in year 2020: The measurable qualities of patients impact deciding the probability of coronary illness. To decide the sickness in clinical diagnostics, factual techniques are frequently utilized Data Mining, which with a lot of data and complex connections can give more exact evaluations, particularly with an enormous number of comparative attributes. In this paper, we consider the assignment of grouping information to decide the probability of coronary illness for patients with comparable attributes.

[2] Haolin Wang Meniailov et. al in year 2020: Information driven methodologies can possibly recognize the high-hazard people by catching the intricate examples of certifiable information. To empower clinically relevant expectation of intravenous immunoglobulin opposition tending to the inadequacy of clinical information and the absence of interpretability of AI models, a multistage strategy is created by incorporating information missing example mining and understandable models.

[3] Jian Ping LI Meniailov et. al in year 2020: In this article, we proposed a proficient and precise framework to analysis coronary illness and the framework depends on AI strategies. The framework is created dependent on arrangement calculations incorporates Support vector machine, Logistic relapse, Artificial neural organization, K-closest neighbor, Naïve straights, and Decision tree while standard highlights determination calculations have been utilized.

[4] Luiz Antonio da Ponte Junior et. al in year 2020: The quantity of individuals determined to have uneasiness problems has been expanding as of late. The right conclusion of such issues isn't generally an insignificant assignment, at times compelling a person to talk with numerous clinicians and playing out a few clinical tests. Post-horrendous Stress Disorder (PTSD) is an issue identified with experienced occasions, which introduced a specific level of danger to a person.

[5] Anjan Nikhil Repaka et. al in year 2019: The exploration centers around coronary illness determination by thinking about past information and data. To accomplish this SHDP (Smart Heart Disease Prediction) is constructed through Navies Bayesian to foresee hazard factors concerning coronary illness. The expedient headway of innovation has prompted noteworthy ascent in versatile wellbeing innovation that being one of the web application. The necessary information is gathered in a normalized structure.

3.1 PROBLEM FORMULATION

The mechanism through which the future possible scenarios can be predicted based on previous outcomes is known as prediction analysis. The prediction analysis techniques are based on the clustering and classification. The machine learning algorithms are the most popular algorithms which are applied for the dataset prediction. Decision trees can support classification and regression problems. Decision trees are more recently referred to as Classification and Regression Trees (CART). They work by creating a tree to evaluate an instance of data, start at the root of the tree and moving town to the leaves (roots) until a prediction can be made. The process of creating a decision tree works by greedily selecting the best split point in order to make predictions and repeating the process until the tree is a fixed depth. After the tree is constructed, it is pruned in order to improve the model's ability to generalize to new data. The depth of the tree is defined automatically, but a depth can be specified in the maximum Depth attribute. We can also choose to turn of pruning by setting the no Pruning parameter to true, although this may result in worse performance.

3.2 Objectives

1. The study of different datasets prediction based algorithms of data mining.
2. To **implement** decision tree classifier for the dataset predication in data mining.
3. To propose voting approach for particular dataset predication in data mining.
4. The existing and proposed algorithm will be compared in terms of time and accuracy.

3.3. Research Methodology

This research work is related to decision tree classifier which has four phases. The first phase is of pre-process, second phase is of clustering with back propagation algorithm and last phase of classification for final predications.
`

Figure 3.1: Research Process

This research compares the effectiveness and evaluate the performance of decision tree classifiers on large scale data set of different algorithms as following:-

1. **Decision stump**: - A decision stump is a machine learning model consisting of a one-level decision tree. It is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature. Depending on the type of the input feature, several variations are possible.
2. **The Hoeffding tree**: - It is an incremental decision tree learner for large data streams that assumes that the data distribution is not changing over time. It grows incrementally a decision tree based on the theoretical guarantees of the Hoeffding bound. A node is expanded as soon as there is sufficient statistical evidence that an optimal splitting feature exists, a decision based on the distribution-independent Hoeffding bound.
3. **J48**:-This algorithm is one of the best machine learning algorithms to examine the data categorically and continuously. When it is used for instance purpose, it occupies more memory space and depletes the performance and accuracy in classifying medical data. J48 algorithm is seen to help in an effective detection of probable attacks which could jeopardize the network confidentiality.
4. **Random forest**: - It is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest.
5. **Reduced Error Pruning Tree**:-Rep Tree uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree.

4. RESULT

We compare the Decision Tree Classification algorithm for example Decision Stump, Hoeffding Tree, J48, Random forest Tree and REP Tree of Classification.

4.1 WEKA Implementation:-

We are using the WEKA tool for implementation thesis. To import the dataset click on open file and select the data set, which is must be a CSV or Arff file format.

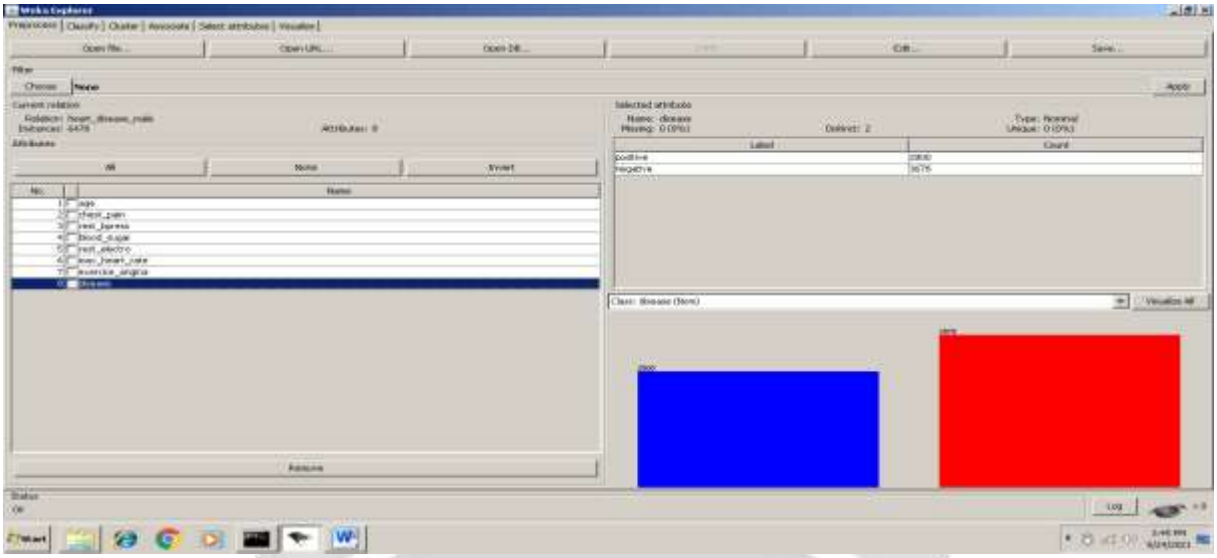

Fig. 4.1 Shows the last attribute "Disease" is taken as a class attribute by the WEKA . This attribute contains two categories i.e. positive and negative.

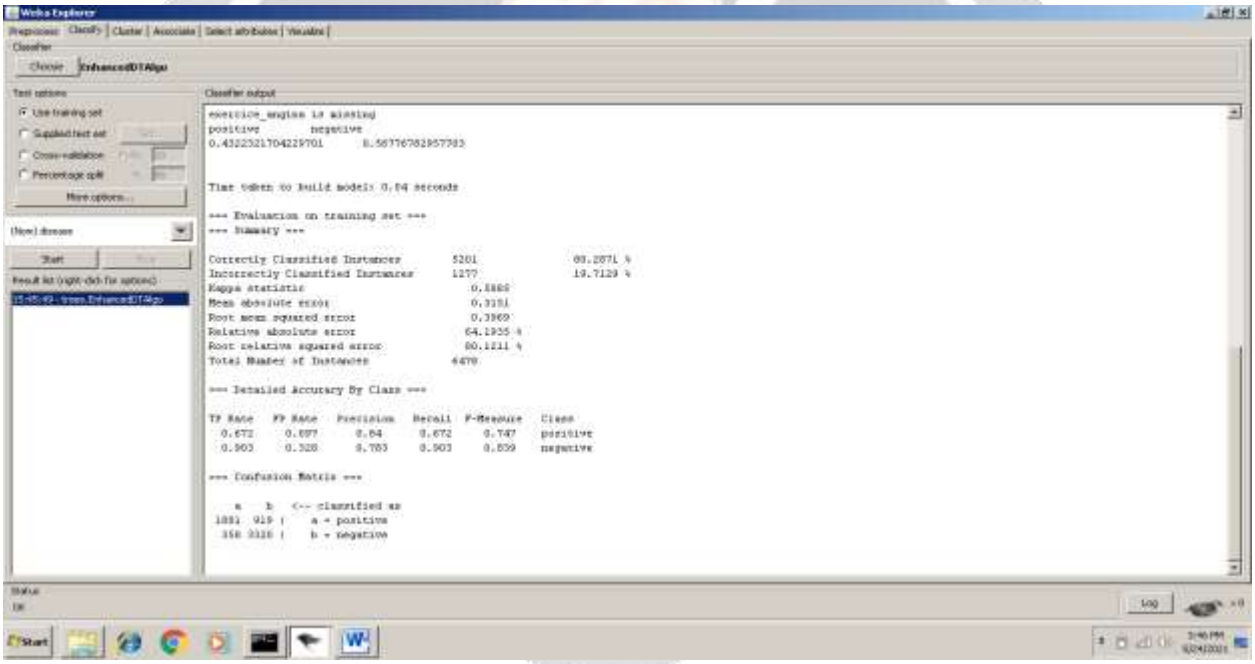

Fig. 4.2 Shows the values of parameters Correctly Classified Instances, Incorrectly Classified Instances and Error Rate using Enhanced Decision Tree Algorithm.

4.2 Result analysis

We compare the Decision Tree Classification algorithm for example Decision Stump, Hoeffding Tree, J48, Random forest Tree and REP Tree of Classification (Correct and Incorrect instances), Execution time and Root Related Square Error rate on heart disease data set.

Table 4.1: Correct and Incorrect instances

Classification	Correct instances	Incorrect instances
Decision Stump	**4350**	**2128**
Hoeffding Tree	**5425**	**1053**
J48 Tree	**5201**	**1277**
Random Forest tree	**3678**	**2800**
REP Tree	**4000**	**2478**

In the results for evaluating execution time of five Decision Tree based algorithms. Decision Stump has the best performance in execution time. Other hand Random forest Tree has worst performance in same parameter with 0.5 seconds.

Table 4.3: Execution time of five rule based algorithms in Seconds.

Classification	Execution Time(Seconds)
Decision Stump	**0.05**
Hoeffding Tree	**0.165**
J48 Tree	**0.11**
Random Forest tree	**0.5**
REP Tree	**0.06**

5. CONCLUSIONS

Recently data mining techniques have encompassed every field in our life. Data mining techniques are being used in the medical, banking, insurances, education, retail industry etc. Prior to working in the data mining models, it is very important to have the knowledge of the existing essential algorithms. In this research shows Five Decision Tree Classification algorithms Decision Stump, Hoeffding Tree, J48 tree, Random Forest and REP tree introduced and experimentally evaluated using Segment data sets. Decision Tree Classification algorithms are experimentally compared based on number of classified instances, accuracy and error rate using WEKA tool and the comparative results are presented in the form of table and graph We used cross validation testing options for our experiments . From the result it is evident that Random Forest Tree is Decision Tree algorithm when compared to the other studied Decision Tree algorithms. It is analyzed in terms of accuracy, precision-recall and execution time that voting based method high performance as compared to other classification methods.

6.REFERENCES

[1]Ievgen Meniailov, Dmytro Chumachenko, Ksenia Bazilevych, "Determination of Heart Disease Based on Analysis of Patient Statistics using the Fuzzy C-means Clustering Algorithm", August 21-25, 2020, Lviv, Ukraine.

[2]Haolin Wang , Zhilin Huang, Danfeng Zhang, Johan Arief, Tiewei Lyu, And Jie Tian, "Integrating Co-Clustering and Interpretable Machine Learning for the Prediction of Intravenous Immunoglobulin Resistance in Kawasaki Disease" https://creativecommons.org/licenses/by/4.0/ VOLUME 8, 2020, IEEE, 2020.

[3]Jian Ping Li, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, And Abdus Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare", IEEE ACCESS, 2020.

[4]Luiz Antonio da Ponte Junior, Rita de C´assia Alves3, Liana Catarina Lima Portuga, "Identifying Post-Traumatic Stress Symptoms Using Physiological Signals and Data Mining", 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), 2020.

[5]Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore, 2019.
[6]Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M, "Heart Disease Diagnosis Using Data Mining Technique", International Conference on Electronics, Communication and Aerospace Technology ICECA 2017.
[7]Cincy Raju, Philipsy E, Siji Chacko, L Padma Suresh, Deepa Rajan S, "A Survey on Predicting Heart Disease using Data Mining Techniques", Proc. IEEE Conference on Emerging Devices and Smart Systems (ICEDSS 2018).
[8]Chaitanya Suvarna, Abhishek Sali, Sakina Salmani, "Efficient Heart Disease Prediction system using Optimization Technique", Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC), 2017.
[9]Mafizur Rahman, Maryam Mehzabin Zahin, Linta Islam, "Effective Prediction On Heart Disease: Anticipating Heart Disease Using Data Mining Techniques", Second International Conference on Smart Systems and Inventive Technology (ICSSIT 2019) IEEE Xplore, 2019.
[10]Yukti Sharma, Rikku Veliyambara, Prof. Rajashree Shettar, "Hybrid Classifier for Identification of Heart Disease", IEEE, 2019.