Application of Decision Tree and Random Forest in Caesarean and Diabetes Prediction

Dr. D. Veeraiah[1], M.V.Komalatha, T.Navya Sree, P. Dhanyaka
[1]*Associate Professor*
Computer Science and Engineering, Lakireddy BaliReddy College of Engineering

Abstract - In health care operations, where a decision has to be taken instantly and accurately, there comes the need of data mining techniques. When we consider some healthcare scenarios like accouchement and diabetes which are the most happening cases in healthcare systems, a decision has to be taken at that instant or a prediction has to be done based on some symptoms and condition of a person etc..Choosing a best technique for the prediction or evaluation of the condition of a patient is necessary because it effects the lives. We can make these assumptions and predictions based on the past data that is available for us. But the huge data which is to be processed for results is difficult to handle. [1].So, we use any of the data mining technique which has higher accuracy and prediction rate. So, here we use the data mining techniques on two datasets one is related to problem of accouchement[2] and with the diabetes prediction. An effective data mining technique is thus found for achieving higher accuracy rate.

Keywords - Accouchement, Diabetes, Data mining, prediction, accuracy.

I. INTRODUCTION

Data Mining usually consists of seven steps .cleaning or pre-processing of the data, integration of data, selection of data, transformation, mining and pattern evaluation and presentation of knowledge. The data that is taken for knowledge discovery may contain missing values, unused attributes , inconsistent data values etc..These has to be removed by using pre processing techniques and the data has to be made fit for algorithm. We need to convert nominal attributes into numerical values in order to represent the data instance in vector form[3].

II. LITERATURE SURVEY

Inductive reasoning is that the method of moving from concrete examples to general models, wherever the goal is to find out the way to classify objects by analyzing a group of instances (already resolved cases) in their whose categories area unit proverbial. Instances area unit generally painted as attribute-value vectors. Learning input consists of a group of such vectors, every happiness to a proverbial category, and therefore the output consists of a mapping from attribute values to categories. This mapping ought to accurately classify each the given instances and alternative unseen instances. a call tree [Quinlan, 1993] could be a formalism for expressing such mappings and consists of tests or attribute nodes connected to 2 or a lot of sub-trees and leafs or call nodes tagged with a category which suggests the choice. A take a look at node computes some outcome supported the attribute values of associate instance, wherever every attainable outcome is related to one in every of the sub-trees. Associate instance is assessed by beginning at the foundation node of the tree. If this node could be a take a look at, the end result for the instance is set and therefore the method continues victimisation the suitable sub-tree. Once a leaf is eventually encountered, its label provides the anticipated category of the instance. The finding of an answer with the assistance of call trees starts by making ready a group of resolved cases. the full set is then divided into 1) a coaching set, that is employed for the induction of a call tree, and 2) a testing set, that is employed to ascertain the accuracy of associate obtained resolution. First, all attributes shaping every case area unit delineate (input data) and among them one attribute is chosen that represents a call for the given drawback(output data). For all input attributes specific worth categories area unit outlined. If associate attribute will take just one of a number of separate worth's then every value takes its own class; if associate attribute will take varied numeric values then some characteristic intervals should be outlined, that represent totally different categories. Every attribute will represent one internal node during a generated call tree, conjointly referred to as associate attribute node or a take a look at node Such associate attribute node has precisely as several branches as its range of various worth categories. The leaves of {a call|a choice|a call} tree area unit choices and represent the worth categories of the choice attribute – decision categories once a call should be created for associate unresolved case, we tend to begin with the foundation node of {the call|the choice} tree and moving on attribute nodes choose branches wherever values of the suitable attributes within the unresolved case matches the attribute values within the decision tree till the leaf node is reached representing the choice

III. METHODOLIGIES

Decision tree - A Decision Tree is a combination of a number of decisions or rules that are formed with one or more combinations of attributes that are present in the given dataset. In this research, we use c4.5 decision tree algorithm which is developed by Ross Quinlan[4]. The algorithm is basically an extension of ID3 algorithm and works in a better way than ID3 . In principle, Decision Trees are used to predict the membership of objects to different categories named (classes), taking into account the values that correspond to their attributes[5].

The Decision Tree algorithm is a classification as well as regression algorithm provided by Microsoft SQL Server Analysis Services (SSAS) specially for use in predictive modeling of both discrete and continuous attributes. You can see a simple DT in fig. 1 which is a Univariate Tree.

Decision Trees can be constructed using a variety of methods. For example, C4.5 uses information-theoretic measures and Classification and Regression Trees (CART) uses statistical methods[6].

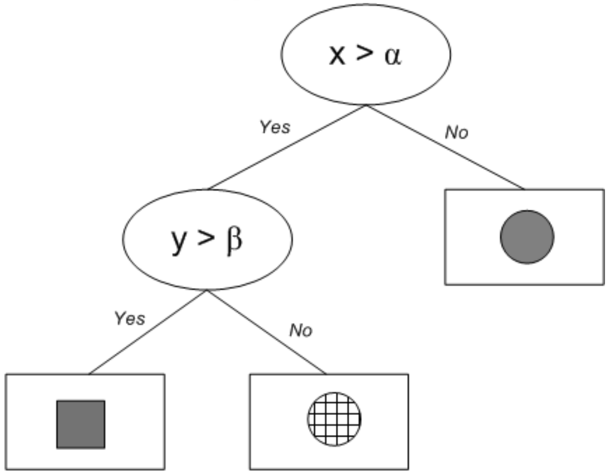


Figure 1: Example of UniVariate tree

Random forest algorithm - Random forests or **random decision forests** are combination of learning method techniques for classification, regression and other tasks that operates by constructing a multiple of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

The training algorithm for random forests algorithm applies the general technique of bootstrap aggregating, or bagging, to tree learners.

In a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$, occurring repeatedly (*B* times) selects a random sample with replacement of the training set and fits trees to these samples.

After training, predictions for unseen samples x' (x_test) can be made by averaging the predictions from all the individual regression trees on x'(x_test).

IV. PROPOSED METHOD

In the proposed method ,the two datasets we considered for the research i.e., caesarean dataset and diabetes prediction dataset are given for prediction analysis to both C4.5 and Random forest algorithm .The results for then checked and compared in terms of accuracy and other parameters of data analysis.

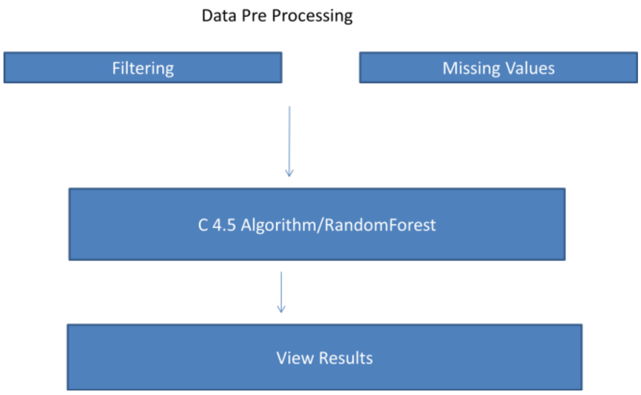


Figure 2: Architecture of the System

V. IMPLEMENTATION

Weka, the tool for data analysis and machine learning contains different classifiers and clustering techniques which can be applied on different datasets in .arff format and the output is given in terms of many parameters like precision,accuracy etc..

Weka tool has the pre-processing capability [8],visualization of outputs like decision trees etc..

The Random forest algorithm is then applied on the same two datasets using python as the platform and the results include the accuracy of the classification set.

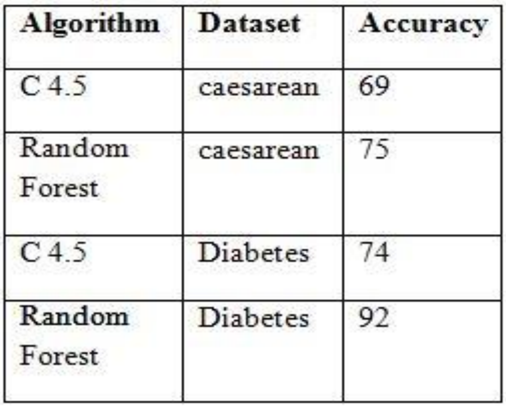
C 4.5 in Weka is available under J48 classifier and is choosen because of abilities to apply negotiation strategies, diagnose delivery method in pregnant women and high accuracy in medical applications.

Weka has processing options like[8] cross validation, cross folding, training sets etc.

VI. RESULTS

The results of C4.5 include accuracy, precision, f1-score where as the random forest gives sensitivity, specificity and accuracy of the algorithm. The following table conveys the result of the research Weka tool [7] has the pre-processing capability, visualization of outputs like decision trees etc..Python being a Scripting language and a big repository of in built libraries and functions, helps in easy implementation and processing of an algorithm.

Table 1: Result of the research Weka tool



Algorithm	Dataset	Accuracy
C 4.5	caesarean	69
Random Forest	caesarean	75
C 4.5	Diabetes	74
Random Forest	Diabetes	92

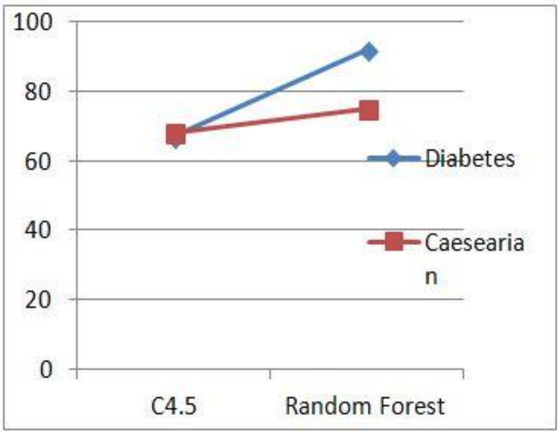


Figure 3: Comparing results through graph and table

VII. CONCLUSION

Development of technology and computer techniques made analysis and evaluation of a huge data very easy and accessible to normal people. This easy methodology can be made more accurate and useful to future scenarios as required with furtive analysis of data using data mining techniques like decision trees and random forests. Both, being the most accurate techniques of data mining can be compared and analysed for finding the best among two. The algorithms when applied in health care gives results that could save a life in future.

VIII. REFERENCES

[1]. F.S. Gharehchtopogh, Z.A. Khalifelu, "Application Data Mining Methods for Detection Useful Knowledge in Health Center: A Case Study Using Decision Tree", 2011 International Conference on Computer Applications and Network Security.

[2]. A.A. Walter, "Data Mining Industry: Emerging Trrends and New Opportunities", Massachusetts Institute of Technology, pp. 13-15, 2000.

[3]. Kantardzic, Mehmed. Data Mining: Concepts, Models, Methods, and Algorithms. John Willey & Sons, 2003.

[4]. H. Jiawei and K. Michelline, Data Mining: Concepts and Techniques, vol. 2, Morgan Kaufmann Publishers, 2006.

[5]. Christy, T. (1997). Analytical tools help health firms fight fraud. Insurance & Technology, Vol .22(31), pp 2- 26.

[6]. Biaforre, S. (1999). Predictive solutions bring more power to decision makers. Health Management Technology, Vol.20 (10), pp 12- 14.

[7]. Indranil Bose, Radhaa K. Mahapatra, Business data mining — a machine learning perspective, Information &amp; Management, Volume 39, Issue 3, 20 December 2001, Pages 211-225, ISSN 0378-7206, 10.1016/S0378- 7206(01)00091-X.

[8]. H. Witten and F. Eibe, Data Mining: Practical Machine Learning Tools and Techniques, vol. 2, Diane Cherra Publishers, 2005.

[9]. Chia-Ming Wang, Yin-Fu Huang, Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data, Expert Systems with Applications, Volume 36, Issue 3, Part 2, April 2009, Pages 5900-5908, ISSN 0957-4174, 10.1016/j.eswa.2008.07.026.