A Study on Decision Tree Learning in Data Mining for Medical Diagnosis

Gnaneswar B[1], Sreedevi M[2]

[1]Dept of Computer Science, S V University, Tirupati
[2]Assistant Professor, Dept of Computer Science, S V University, Tirupati

Abstract— *In this paper, we present a managed learning procedure of building a decision tree for clinical finding. The principal objective is to fabricate a proficient characterization model with high review under moderate accuracy to work on the productivity and viability of the illness expectation measure. We utilized ID3 calculation for choice tree development and the last model is assessed utilizing the normal assessment techniques. This model gives a logical view to utilize the important data in clinical information, particularly which is generally overlooked by the vast majority of the current strategies when they focus on high forecast exactness's. We have done the analyses on diabetes and coronary illness datasets from UCI storehouse. Test results show that decision tree incredibly works on the nature of arrangement. With these outcomes, we surmise that the decision tree is more appropriate in taking care of the grouping issue of illness expectation, and we suggest the utilization of these methodologies in comparative order issues.*

I. INTRODUCTION

With the fast improvement of data innovation and organization innovation, various exchanges produce a lot of information consistently. The actual information can't carry direct advantages so need to viably mine concealed data from enormous measure of information. Information digging manages looking for fascinating examples or information from huge information. It transforms a huge assortment of information into information. Information mining is a fundamental advance during the time spent information revelation. The information mining has become an interesting apparatus in examining information according to alternate point of view and changing over it into valuable and significant data [6]. Information mining has been generally applied in the space of medical finding, Intrusion identification framework, Education, Banking, Fraud discovery. Grouping is a regulated learning. Forecast and arrangement in information mining are two types of information investigation task that is utilized to separate models depicting information classes or to foresee future information patterns. Characterization measure has two stages; the first is the learning interaction where the preparation informational indexes are dissected by grouping calculation. The learned model or classifier is introduced as arrangement rules or examples. The subsequent stage is the utilization of model for grouping, and test informational collections are utilized to assess the exactness of characterization rules [4]. With the ascending of information mining, choice tree assumes a significant part during the time spent information mining and information investigation. Decision tree learning includes in utilizing a bunch of preparing information to produce a choice tree that accurately arranges the preparation information itself. Assuming the learning cycle works, this choice tree will effectively group new information too. Decision trees contrast along a few measurements like parting basis, halting standards, branch condition (univariate, multivariate), style of branch activity, sort of conclusive tree. All the more as of late, choice tree philosophy has gotten famous in clinical examination. An illustration of the clinical utilization of choice trees is in the determination of an ailment from the example of side effects, wherein the classes characterized by the choice tree could either be distinctive clinical subtypes or a condition, or patients with a condition who ought to get various treatments.

II. CLASSIFICATION PROCESS

Game plan is the way toward finding a model or a limit that depicts and perceives data classes and thoughts, to use the model to predict the classes of things whose class mark isn't known. Data request can be viewed as a two-stage measure: learning step in which a classifier is developed depicting a fated game plan of classes or thoughts by separating the readiness set contained informational index tuples and their connected names [2]. In the resulting advance model is used for request by first evaluating the judicious exactness of classifier worked during the underlying advance. It is done using the test data. The precision of classifier on a given test set tuples is level of tuples that are precisely requested by the classifier. If the exactness is over some satisfactory level, the classifier can be used to expect future tuples whose class mark isn't known.

Portrayal is a kind of data assessment that can be used to create models portraying huge data classes. Game plan is a data mining methodology used to predict pack interest for data models. It is one of the critical systems in data mining and is used

in various applications, for instance, plan affirmation, sickness assurance, customer relationship the leaders, and assigned displaying. The goal of the portrayal estimations is to assemble a model from a lot of getting ready data whose target class names are known and subsequently this model is used to bunch covered cases [3].

Plan is the most normal and most renowned data mining techniques. Game plan maps data into predefined social occasions or classes. It is typical suggested as managed learning considering the way that the classes are settled preceding taking a gander at the data. Course of action is the way toward finding a model that perceives data classes, to use the model to predict the class of things whose class name is dark. The decided model relies upon the assessment of a lot of getting ready data. Informational collections are rich with concealed information that can be used for watchful dynamic.

Building definite and useful classifiers for enormous data bases is one of the crucial tasks of data mining and AI research. Building fruitful request systems is one of the central tasks of data mining.

A wide extent of kinds of assortment structures have been proposed recorded as a printed copy that unite Decision Trees, Naive-Bayesian systems, Neural Networks, Logistic Regression, Support Vector Machines (SVM) and K-Nearest Neighbor, and so forth.

III. METHODOLOGY

At the present time, clarified about Decision Tree procedure structure model for clinical infection grouping issue.

3.1 Decision Tree Classifier

Decision tree philosophy is a usually utilized information digging technique for setting up characterization frameworks dependent on various covariates or for creating expectation calculations for an objective variable. This strategy characterizes a populace into branch-like portions that develop an upset tree with a root hub, inward hubs, and leaf hubs. The calculation is non-parametric and can proficiently manage huge, convoluted datasets without forcing a muddled parametric construction [1]. Decision trees are classifiers that address their characterization information in tree structure. Every inside hub of a decision tree is a test on a property. Fulfilling that test causes the case being characterized to remove one branch from that hub, bombing the test makes the example take the other branch. A decision tree is utilized to group an example by beginning at the root hub of the decision tree and following the way the property tests direct until a leaf hub is experienced [4]. Each leaf hub in a decision tree is a choice, i.e., addresses an order. An occasion that winds up at some specific leaf hub is arranged with the class allocated to that leaf hub. A second sort of tree is a class likelihood tree. This has a vector of class probabilities at each leaf rather than a choice. The fundamental calculation constructs a tree top down utilizing the standard voracious inquiry guideline, in light of recursive parceling. The parceling calculation incorporates halting, parting and pruning rules. At the point when the example size is sufficiently huge, study information can be separated into preparing and approval datasets. Utilizing the preparation dataset to assemble a choice tree model and an approval dataset to settle on the fitting tree size expected to accomplish the ideal last model.

The way toward developing a decision tree is separated into two stages: tree building and pruning. The initial step is the tree building stage, which chooses part of the preparation information and fabricates a choice tree by the expansiveness first recursive calculation until each leaf hub has a place with a similar class [5][6]. The subsequent advance is the pruning stage, which utilizes the leftover information to check the produced choice tree and right the blunders, and it at long last prunes the choice tree and adds hubs until a right choice tree is fabricated. The decision tree building calculation is a recursive interaction that eventually brings about a choice tree, and pruning lessens the effect of boisterous information on arrangement precision. As a general rule, the more noteworthy the data acquire, the more prominent the "immaculateness improvement" got by utilizing highlights to parcel the dataset. Subsequently, data gain can be utilized to choose credits for choice tree dividing, which is to choose the trait with the best data acquire.

IV. EXPERIMENTAL RESULTS

The analyses have been directed by utilizing R programming Language. R is a sophisticated statistical software package, which provides new approaches to data mining., it is an open-source tool for analysis of data mining algorithms. The R Language is a bundle for information characterization, grouping and representation. We have considered the Two UCI Machine Learning Repository datasets [7], including heart disease and Pima Diabetes for assessing the productivity and adequacy of decision tree calculation. The characteristic data information is consolidated in Table-1. The standard dataset is parceled into two sets one for training (70%) and another set for testing (30%).

TABLE 1
DATASET INFORMATION

S. No	Name of the Dataset	No. of Attributes	No. of Instances	No. of Classes
1	Heart Disease	13	270	2
2	Pima Diabetes	9	768	2

To approve the expectation consequences of the decision tree arrangement and the 10-overlap hybrid approval is utilized. The k-overlap hybrid approval is normally used to lessen the mistake came about because of irregular examining in the correlation of the exactness's of various forecast models. The current investigation partitioned the information into 10-folds where 1-crease was for trying and 9-folds were for preparing for the 10-overlap hybrid approval.

The performance of a chosen classifier is validated based on accuracy. The classification accuracy is noted for two datasets of decision tree classifier is taken in to account. The accuracy of two UCI data sets is presented in Table-2 and Accuracy of decision tree are shown in figure-1.

TABLE 2
PERFORMANCE OF DECISION TREE ALGORITHM

Name of the Dataset	Accuracy
Heart Disease	82
Pima Diabetes	84

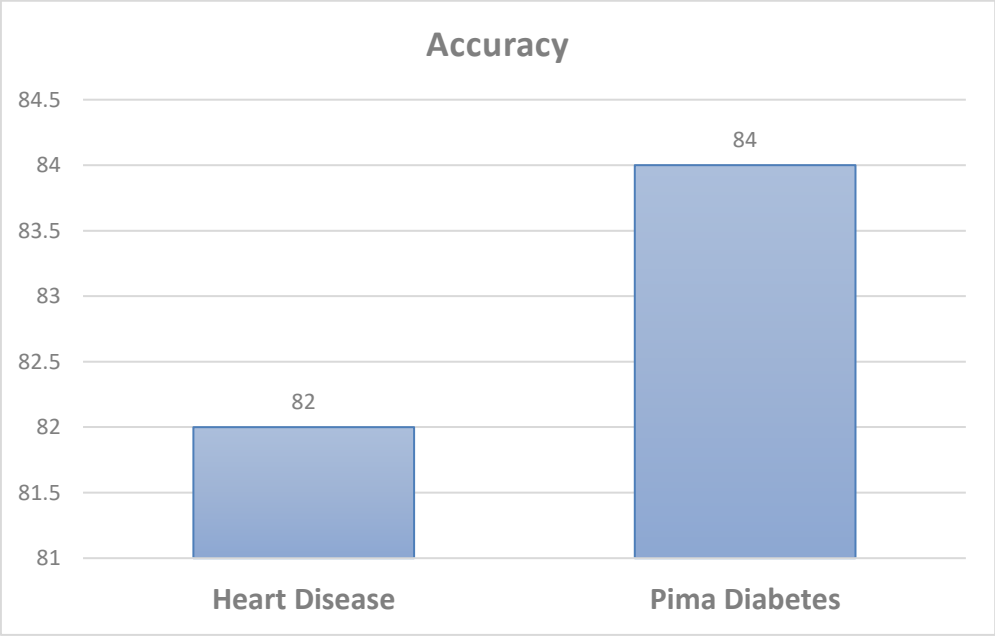

Figure 1: Performance of decision tree algorithm

From the figure-1, it tends to be seen that the decision tree calculation of precision on heart disease exactness is 82% and Pima Diabetes exactness is 84%.

The experimental results of screen shots are shown in the figure-2 for heart disease and figure-3 for Pima Diabetes.

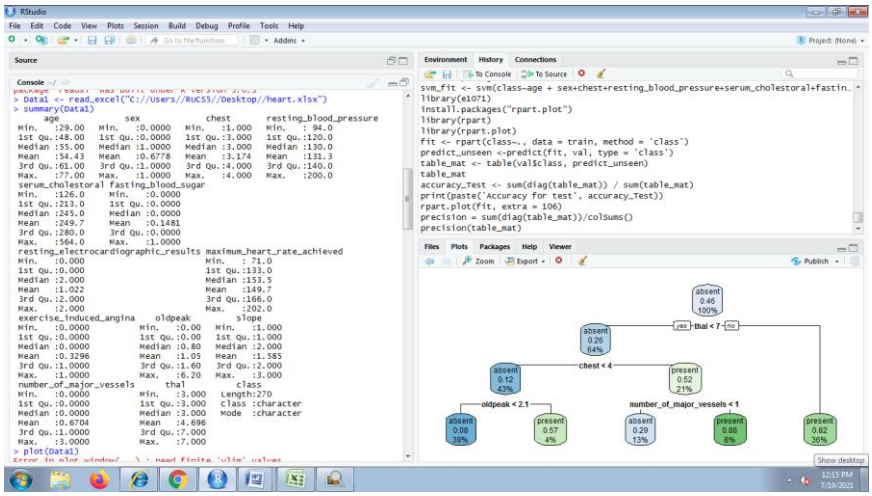

Figure-2: Screen shot results of heart disease data

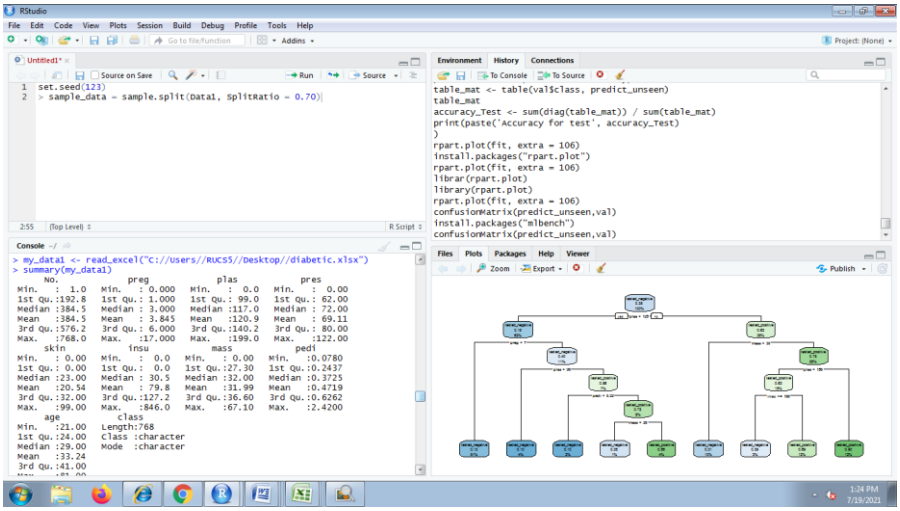

Figure-3: Screen shot results of Pima Diabetes data

V. CONCLUSION

The clinical dataset in the different information mining and the AI strategies are accessible and afterward the significant part of clinical information mining is to expand the exactness and effectiveness of sickness finding. The goal of this examination work is planned to show the classes of clinical information from the accessible crude clinical dataset assists the doctor with showing up at a precise finding. The outcomes are assessed dependent on the precision of arrangement is 94% for diabetes information and 82% for coronary illness information. Subsequently decision tree classifier is proposed for analysis of clinical determination expectation-based order to improve results with precision and execution.

REFERENCES

[1] Freund, Y., and Schapire, R. E., ―A decision-theoretic generalization of on-line learning and an application to Boosting‖, J. Comput. Syst. Sci. 55(1):119–139, 1997

[2] G. Ravi Kumar, Venkata Sheshanna Kongara & Dr. G. A. Ramachandra, "An Efficient Ensemble Based Classification Techniques for Medical Diagnosis", International Journal of Latest Technology in Engineering, Management and Applied Sciences, Volume II, Issue VIII, Pages: 5-9, ISSN-2278-2540, August-2013

[3] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[4] J Han, "Data Mining Concepts and Techniques", Second Edition. Morgan Kaufmann Publisher, 2006, pp.123-134.

[5] N. Michael, "Artificial Intelligence - A Guide to Intelligent Systems", 2nd edition, Addison Wesley, 2005.

[6] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[7] UCI machine learning repository.