A Review of Managed Learning Process for Constructing Decision Trees in Clinical Diagnosis

Badinehalu Mrutha

Department of Computer Science Sri Venkateswara University, Tirupati

Abstract— *In this paper, we introduce a managed learning approach for constructing a decision tree aimed at clinical diagnosis. Our primary goal is to develop an efficient classification model with high recall and moderate precision to enhance the efficiency and effectiveness of disease prediction processes. We employed the ID3 algorithm for decision tree construction, and the final model was assessed using standard evaluation methods. This model offers a systematic framework for leveraging relevant information in clinical data, particularly aspects often overlooked by existing methods overly focused on high predictive accuracy. Our analysis was conducted on datasets related to diabetes and coronary disease sourced from the UCI repository. Test results highlight the decision tree's significant contribution to classification quality. Based on these findings, we conclude that decision trees are particularly suitable for addressing disease prediction classification challenges and advocate for their adoption in similar classification tasks.*

I. INTRODUCTION

With the rapid advancement of both data technology and networking, a plethora of transactions generate vast amounts of data daily. While raw data alone may not yield direct benefits, effective mining is necessary to extract hidden insights from this abundance of information. Data mining involves the search for interesting patterns or knowledge within large datasets, transforming raw data into actionable insights. It serves as a crucial step in the data discovery process, enabling analysis from diverse perspectives and conversion into valuable information. Widely applied across various domains such as medical diagnosis, intrusion detection systems, education, banking, and fraud detection, data mining encompasses supervised learning techniques like classification, prediction, and grouping. Classification, in particular, involves a two-step process: learning, where training datasets are analyzed by classification algorithms to derive classification rules or patterns, and application, where the learned model is utilized to classify new data and evaluate accuracy. Decision trees play a pivotal role in data mining and analysis, utilizing a set of training data to generate a tree structure that accurately classifies the data. The decision tree methodology has gained popularity in clinical research, notably in disease diagnosis scenarios where it aids in identifying ailments based on symptom patterns, potentially guiding treatment decisions.[4,6]

II. CLASSIFICATION PROCESS

Classification is the process of discovering a model or boundary that describes and identifies data classes or concepts, enabling the model to predict the classes of items whose class label is unknown. Data classification can be seen as a two-step procedure: the learning step involves developing a classifier that describes a predefined arrangement of classes or concepts by analyzing the training set comprising data tuples and their associated labels. In the subsequent step, the model is employed for classification by first assessing the practical accuracy of the classifier constructed during the initial step using test data. Classifier accuracy on a given test set of tuples represents the proportion of tuples correctly classified by the classifier. If the accuracy surpasses a predefined threshold, the classifier can be utilized to predict the class of future tuples with unknown class labels.

Representation, a form of data analysis, can be utilized to construct models describing significant data classes. Classification, a data mining technique, is employed to predict group membership for data instances. It is a pivotal method in data mining and finds applications in various domains such as pattern recognition, disease diagnosis, customer relationship management, and targeted marketing. The objective of classification algorithms is to build a model from a set of training data with known class labels, which is subsequently used to classify unseen instances.

Classification is the most common and widely used data mining technique, mapping data into predefined groups or classes. It is often referred to as supervised learning since the classes are determined prior to examining the data. Classification involves finding a model that identifies data classes, enabling prediction of the class of items with unknown class labels. The constructed model relies on the analysis of a large set of training data. Datasets harbor latent information conducive to informed decision-making.

Developing accurate and efficient classifiers for large databases constitutes a fundamental task in data mining and AI research. Constructing effective classification systems is integral to data mining endeavors.

A wide array of classification structures have been proposed, including Decision Trees, Naive-Bayesian methods, Neural Networks, Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbor, among others.[2,3]

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 into 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

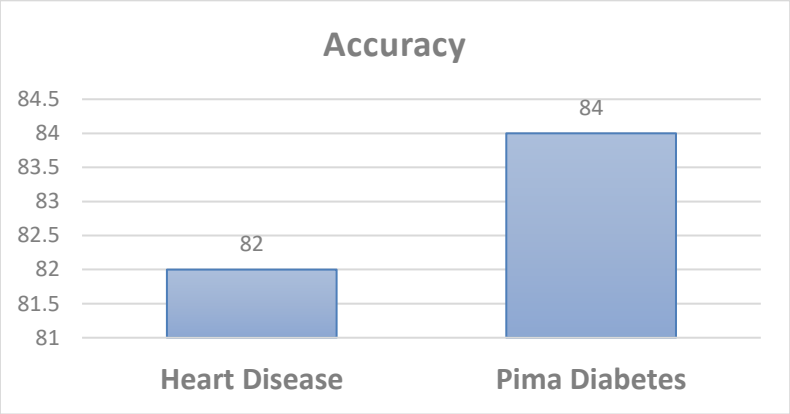

Figure-1: Performance of decision tree algorithm

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.

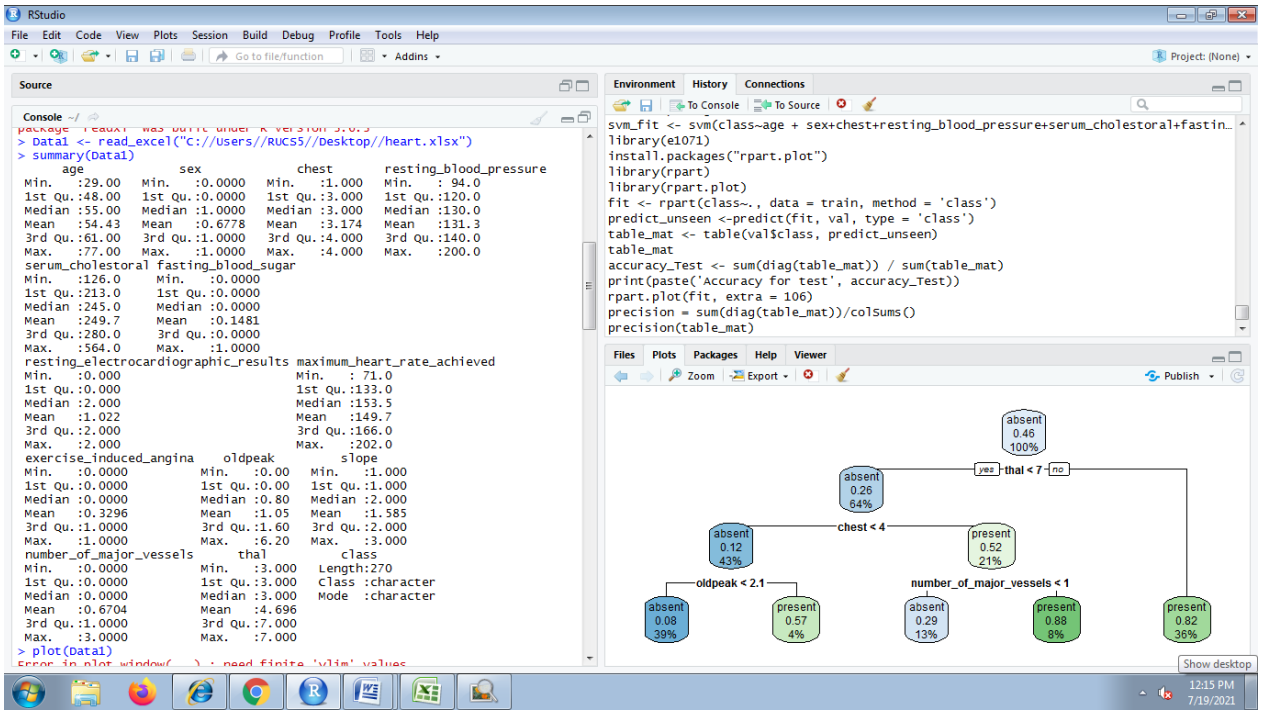

Figure-2: Screen shot results of heart disease data

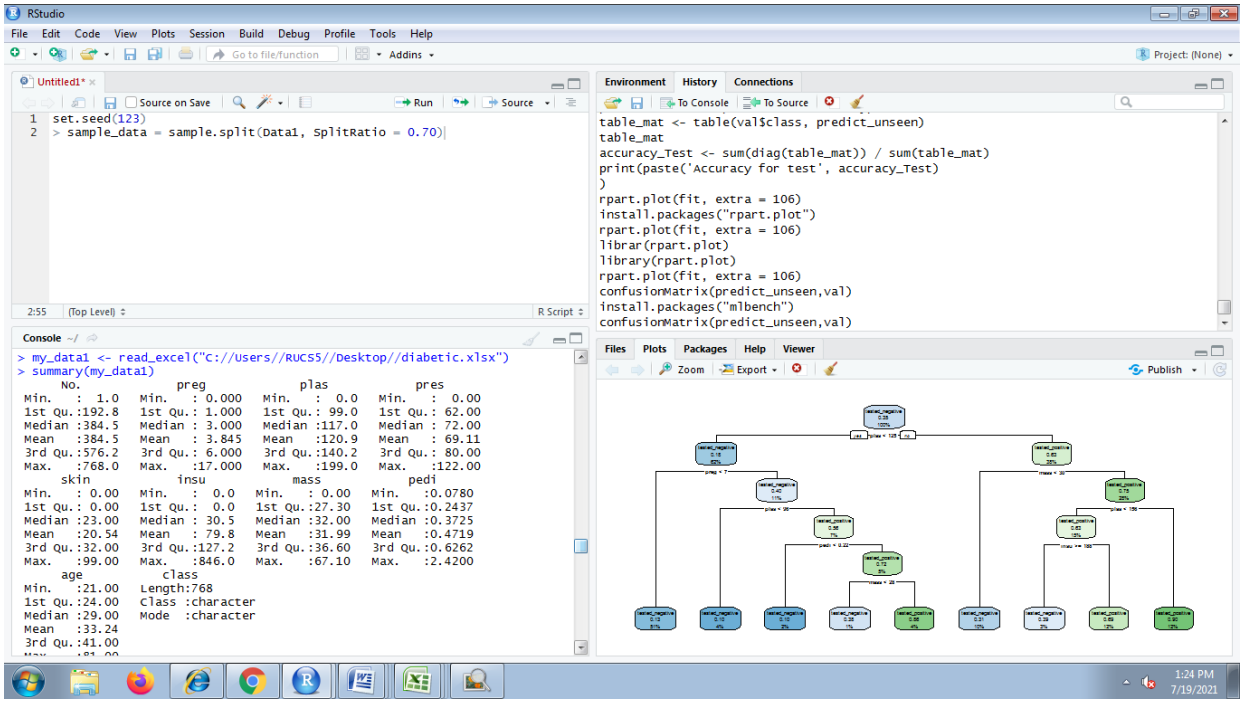

Figure-3: Screen shot results of Pima Diabetes data

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 Boostingǁ, 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. http://archive.ics.uci.edu/ml/