Exploring a Supervised Learning Approach for Enhanced Clinical Diagnosis and Discovery

Kallagunta Bhavanapriya

Department of Computer Science Sri Venkateswara University, Tirupati

***Abstract*—** *This paper introduces a supervised learning approach for constructing decision trees in clinical diagnosis. The primary aim is to develop an efficient classification model with a balance of high recall and moderate precision to enhance the effectiveness of disease prediction. Utilizing the ID3 algorithm for decision tree construction, the final model is evaluated using standard assessment methods. This model offers valuable insights into leveraging clinical data, particularly aspects often overlooked by existing techniques focused solely on high precision. Experiments conducted on diabetes and coronary heart disease datasets from the UCI repository demonstrate the decision tree's effectiveness in classification tasks. Based on these findings, we conclude that decision trees are well-suited for addressing disease prediction challenges and recommend their adoption in similar classification problems.*

I. INTRODUCTION

With the rapid advancement of information technology and communication, various industries generate a substantial amount of data daily. However, raw data alone often fails to yield actionable insights, necessitating the extraction of hidden patterns from large datasets effectively. Data mining, the process of uncovering interesting patterns or knowledge from vast data, plays a crucial role in the data discovery process. It transforms a plethora of data into actionable insights, making it a fundamental step in data exploration. Data mining has emerged as a powerful tool for analyzing data from diverse perspectives and converting it into meaningful and actionable information [6].

Data mining finds wide application across various domains including clinical diagnosis, education, banking, and fraud detection. Classification, a supervised learning approach, involves prediction and categorization tasks in data mining, aimed at extracting patterns describing data classes or forecasting future data trends. The classification process comprises two stages: the learning phase, where training datasets are analyzed using classification algorithms to generate a model or classifier represented as classification rules or models, and the application phase, where the model is utilized for classification, and test datasets are used to evaluate the accuracy of classification rules [4].

In the realm of data mining and analysis, decision trees play a significant role. Decision tree learning entails utilizing a large set of training data to construct a decision tree that accurately categorizes the training data itself, with the expectation that it will also classify new data effectively. Decision trees vary across several dimensions such as splitting criteria, termination rules, branch condition (univariate, multivariate), branch growth style, and type of resulting tree. Recently, decision tree reasoning has gained popularity in clinical research, particularly in disease diagnosis. An example of clinical decision tree application involves diagnosing a disease based on observed symptoms, where the decision tree's classes may represent distinct clinical subtypes or different treatment options for patients with a particular condition.

II. CLASSIFICATION

Classification is the process of discovering a model or function that describes and distinguishes data classes and concepts, enabling the model to predict the classes of items whose class labels are unknown. Data classification involves a two-stage process: the learning phase, where a classifier is developed to describe a predetermined set of classes or concepts by analyzing the training set comprising data tuples and their associated labels [2]. In the subsequent phase, the model is applied for classification by initially assessing the predictive accuracy of the classifier established during the learning phase, using test data. Classifier accuracy on a given test set of tuples is the proportion of tuples accurately classified by the classifier. If the accuracy exceeds a certain acceptable threshold, the classifier can be utilized to predict future tuples whose class labels are unknown.

Classification, a form of data analysis, is utilized to construct models describing large data classes. It is a fundamental technique in data mining and finds applications in various domains such as pattern recognition, disease detection, customer relationship management, and targeted marketing. The objective of classification algorithms is to construct a model from a large set of training data whose target class labels are known, subsequently utilizing this model to categorize unseen instances [3].

Classification is one of the most common and widely used data mining techniques, involving the mapping of data into predefined groups or classes. It is often referred to as supervised learning since the classes are predetermined before analyzing the data. The process of classification aims to develop a model that distinguishes data classes, enabling the model to predict the class of items whose class labels are unknown. The efficacy of the learned model relies on the analysis of a substantial amount of training data. Datasets contain hidden information that can be leveraged for informed decision-making.

Developing accurate and efficient classifiers for large databases is a fundamental task in data mining and AI research. A wide range of classification methods have been proposed, including Decision Trees, Naive-Bayesian techniques, Neural Networks, Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbor, among others.

III. METHODOLOGY

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

3.1 Decision Tree Classifier

The decision tree methodology is a widely utilized data mining technique for constructing classification systems based on various covariates or for developing prediction algorithms for a target variable. This approach segments a population into branch-like segments that form an inverted tree structure with a root node, internal nodes, and leaf nodes. Being non-parametric, the algorithm can efficiently handle large, complex datasets without imposing a complicated parametric framework [1].

Decision trees represent their classification information in a tree structure, where each internal node performs a test on a feature, leading the instance down one branch if the test is satisfied and down another branch if it fails. To classify an instance, one starts at the root node and follows the path dictated by the feature tests until reaching a leaf node, which represents a decision or classification [4].

The process of constructing a decision tree involves two stages: tree building and pruning. During the tree building stage, a breadth-first recursive algorithm is employed to select portions of the training data and construct a decision tree until each leaf node corresponds to the same class [5][6]. Subsequently, in the pruning stage, the remaining data is utilized to evaluate and correct errors in the generated decision tree, ultimately refining the tree structure until an accurate decision tree is constructed. The decision tree construction algorithm is a recursive process, and pruning helps mitigate the impact of noisy data on classification accuracy.

The decision tree building process is driven by information gain, which aims to maximize the "purity improvement" gained by using features to partition the dataset. Consequently, information gain is utilized to select attributes for decision tree partitioning, identifying the feature with the greatest information gain.

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	93
Pima Diabetes	95

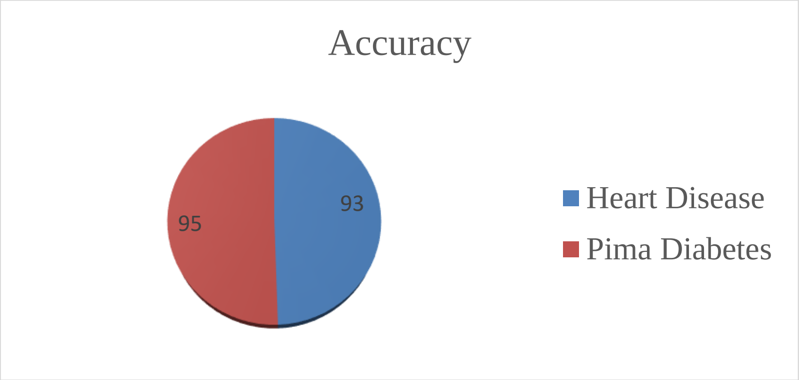
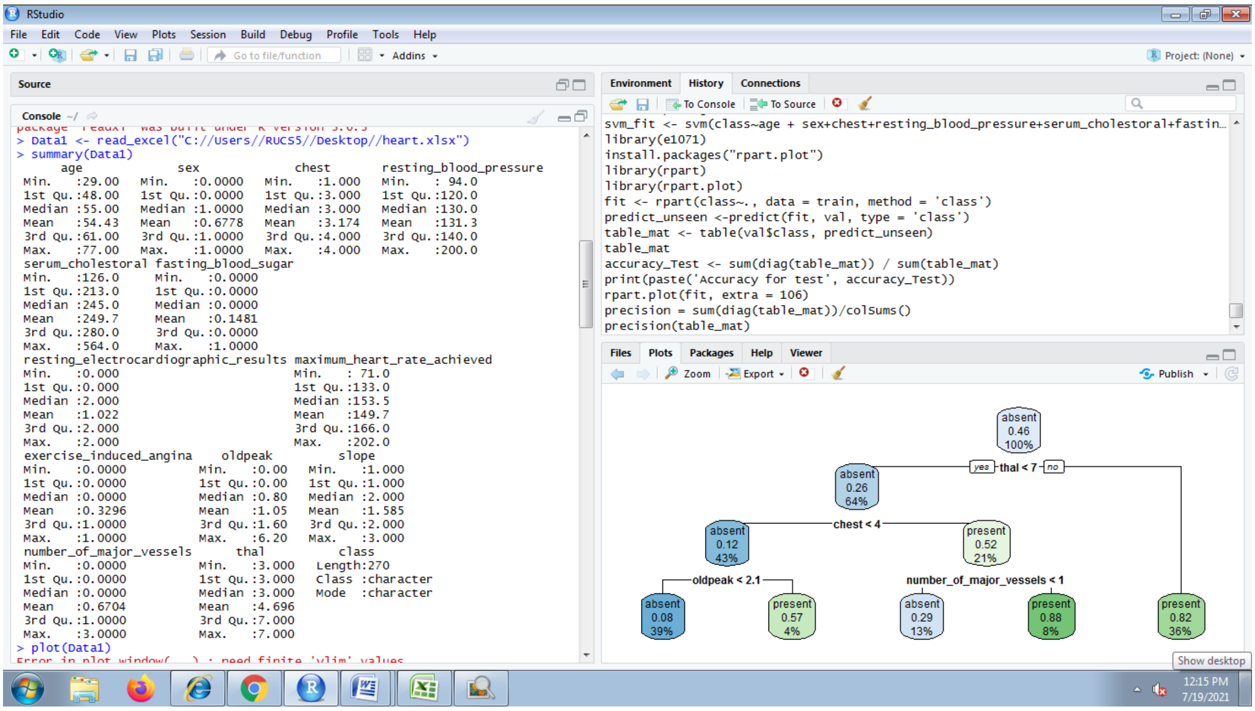


Figure-1: Performance of decision tree algorithm

From the figure-1, it tends to be seen that the decision tree calculation of precision on heart disease exactness is 93% and Pima Diabetes exactness is 95%.

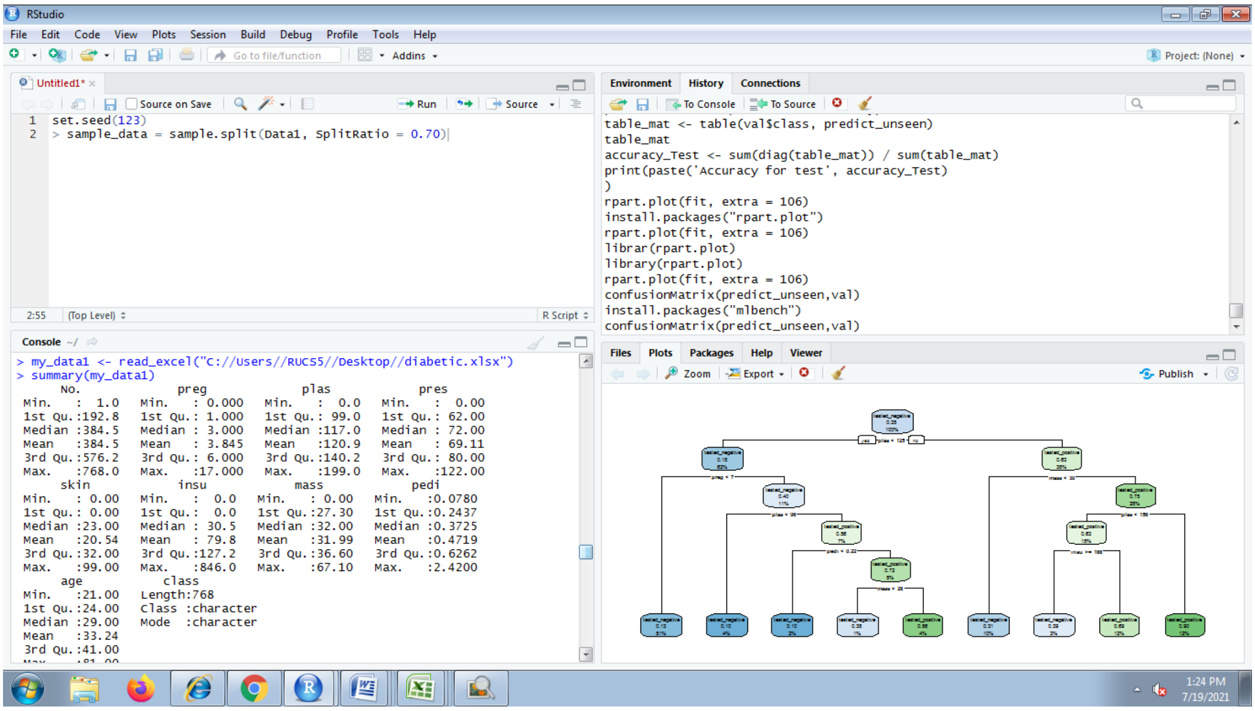
The experimental results of screen shots are shown in the figure-2 for heart disease and figure-3 for Pima Diabetes.



```
package 'readxl' was built under R version 3.6.3
> Data1 <- read_excel("C://Users//RUCS5//Desktop//heart.xlsx")
> summary(Data1)
      age             sex              chest       resting_blood_pressure
 Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0
 1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0
 Median :55.00   Median :1.0000   Median :3.000   Median :130.0
 Mean   :54.43   Mean   :0.6778   Mean   :3.174   Mean   :131.3
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0
 Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0
 serum_cholestoral fasting_blood_sugar
 Min.   :126.0     Min.   :0.0000
 1st Qu.:213.0     1st Qu.:0.0000
 Median :245.0     Median :0.0000
 Mean   :249.7     Mean   :0.1481
 3rd Qu.:280.0     3rd Qu.:0.0000
 Max.   :564.0     Max.   :1.0000
 resting_electrocardiographic_results maximum_heart_rate_achieved
 Min.   :0.000                        Min.   : 71.0
 1st Qu.:0.000                        1st Qu.:133.0
 Median :2.000                        Median :153.5
 Mean   :1.022                        Mean   :149.7
 3rd Qu.:2.000                        3rd Qu.:166.0
 Max.   :2.000                        Max.   :202.0
 exercise_induced_angina    oldpeak         slope
 Min.   :0.0000          Min.   :0.00   Min.   :1.000
 1st Qu.:0.0000          1st Qu.:0.00   1st Qu.:1.000
 Median :0.0000          Median :0.80   Median :2.000
 Mean   :0.3296          Mean   :1.05   Mean   :1.585
 3rd Qu.:1.0000          3rd Qu.:1.60   3rd Qu.:2.000
 Max.   :1.0000          Max.   :6.20   Max.   :3.000
 number_of_major_vessels      thal          class
 Min.   :0.0000          Min.   :3.000   Length:270
 1st Qu.:0.0000          1st Qu.:3.000   Class :character
 Median :0.0000          Median :3.000   Mode  :character
 Mean   :0.6704          Mean   :4.696
 3rd Qu.:1.0000          3rd Qu.:7.000
 Max.   :3.0000          Max.   :7.000
> plot(Data1)
Error in plot.window(...) : need finite 'ylim' values
```

```
svm_fit <- svm(class~age + sex+chest+resting_blood_pressure+serum_cholestoral+fastin…
library(e1071)
install.packages("rpart.plot")
library(rpart)
library(rpart.plot)
fit <- rpart(class~., data = train, method = 'class')
predict_unseen <-predict(fit, val, type = 'class')
table_mat <- table(val$class, predict_unseen)
table_mat
accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
print(paste('Accuracy for test', accuracy_Test))
rpart.plot(fit, extra = 106)
precision = sum(diag(table_mat))/colSums()
precision(table_mat)
```

Figure-2: Screen shot results of heart disease data



```
set.seed(123)
> sample_data = sample.split(Data1, SplitRatio = 0.70)
```

```
> my_data1 <- read_excel("C://Users//RUCS5//Desktop//diabetic.xlsx")
> summary(my_data1)
      No.             preg             plas             pres       
 Min.   :  1.0   Min.   : 0.000   Min.   :  0.0   Min.   :  0.00  
 1st Qu.:192.8   1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00  
 Median :384.5   Median : 3.000   Median :117.0   Median : 72.00  
 Mean   :384.5   Mean   : 3.845   Mean   :120.9   Mean   : 69.11  
 3rd Qu.:576.2   3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00  
 Max.   :768.0   Max.   :17.000   Max.   :199.0   Max.   :122.00  
      skin            insu            mass            pedi       
 Min.   : 0.00   Min.   :  0.0   Min.   : 0.00   Min.   :0.0780  
 1st Qu.: 0.00   1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437  
 Median :23.00   Median : 30.5   Median :32.00   Median :0.3725  
 Mean   :20.54   Mean   : 79.8   Mean   :31.99   Mean   :0.4719  
 3rd Qu.:32.00   3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262  
 Max.   :99.00   Max.   :846.0   Max.   :67.10   Max.   :2.4200  
      age            class          
 Min.   :21.00   Length:768         
 1st Qu.:24.00   Class :character   
 Median :29.00   Mode  :character   
 Mean   :33.24                      
 3rd Qu.:41.00                      
```

```
table_mat <- table(val$class, predict_unseen)
table_mat
accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
print(paste('Accuracy for test', accuracy_Test)
)
rpart.plot(fit, extra = 106)
install.packages("rpart.plot")
rpart.plot(fit, extra = 106)
librar(rpart.plot)
library(rpart.plot)
rpart.plot(fit, extra = 106)
confusionMatrix(predict_unseen,val)
install.packages("mlbench")
confusionMatrix(predict_unseen,val)
```

Figure-3: Screen shot results of Pima Diabetes data

V. CONCLUSION

In various data mining and AI techniques, clinical datasets are accessible, and a significant aspect of clinical data mining aims to enhance the accuracy and effectiveness of disease diagnosis. This study aims to demonstrate how classifying clinical data from publicly available raw datasets assists physicians in reaching accurate diagnoses. Results indicate classification accuracy of 95% for diabetes data and 93% for coronary heart disease data. Therefore, a decision tree classifier is recommended for clinical diagnosis prediction to improve accuracy and performance.

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, K.Nagamani and G.Anjan Babu, "A Framework of Dimensionality Reduction utilizing PCA for Neural Network Prediction", Lecture Notes on Data Engineering and Communications Technologies, Volume-37, Pages:173 – 180, Springer Nature Singapore Pte Ltd, 2020

[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/.