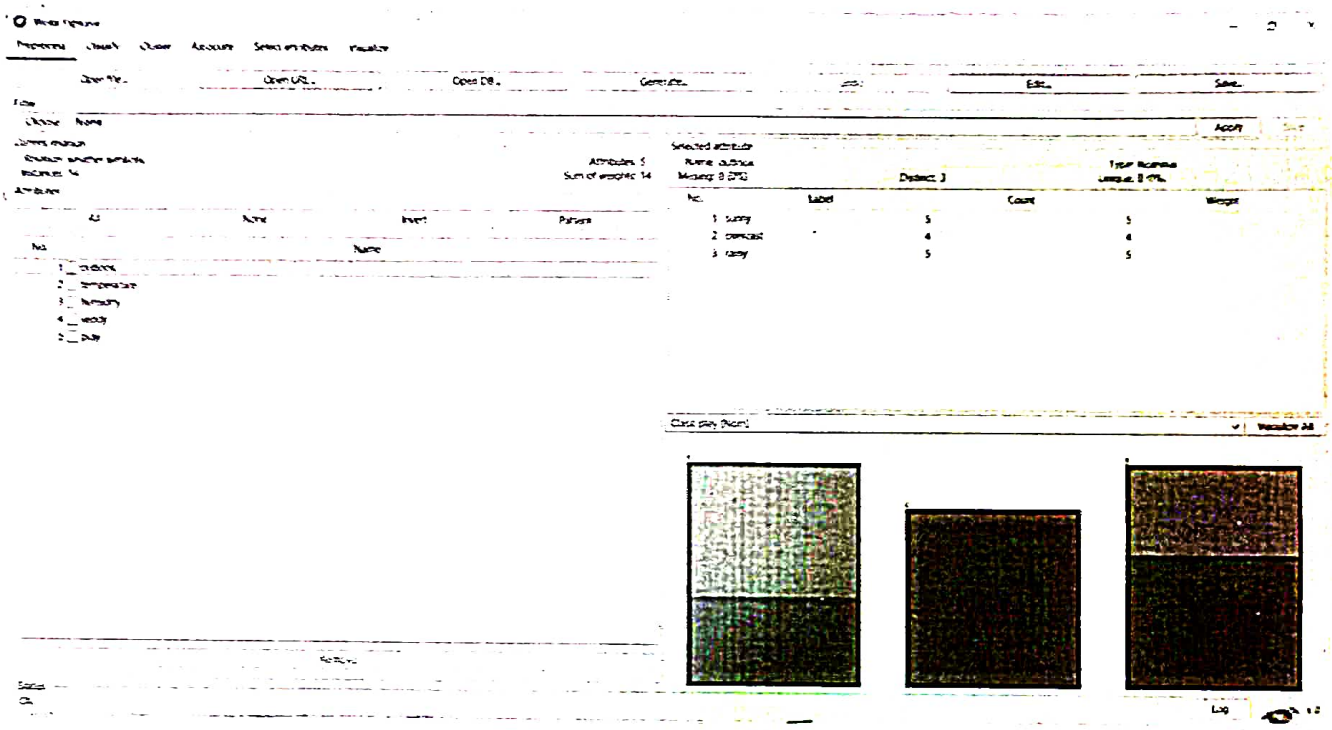
Practical No: 03

Aim: To demonstrate performing classification on data set



Date : 15/02/2023

Practical No. 03

Aim : To demonstrate performing classification on data sets.

Theory :

Classification :

Classification is the process for finding a model that describe the data values and concept for the purpose of prediction.

Classification in data mining is a common technique that separates data points into different classes.

It allows you to organize data set of all sorts, including complex and large datasets as well as small and simples ones. It primarily involves using algorithms that you can easily modify to improve the data quality. The algorithm establishes the link between the variables for predictions. The algorithm you use for classification in data mining is called the classifier, and observations you make through the same are called instances.

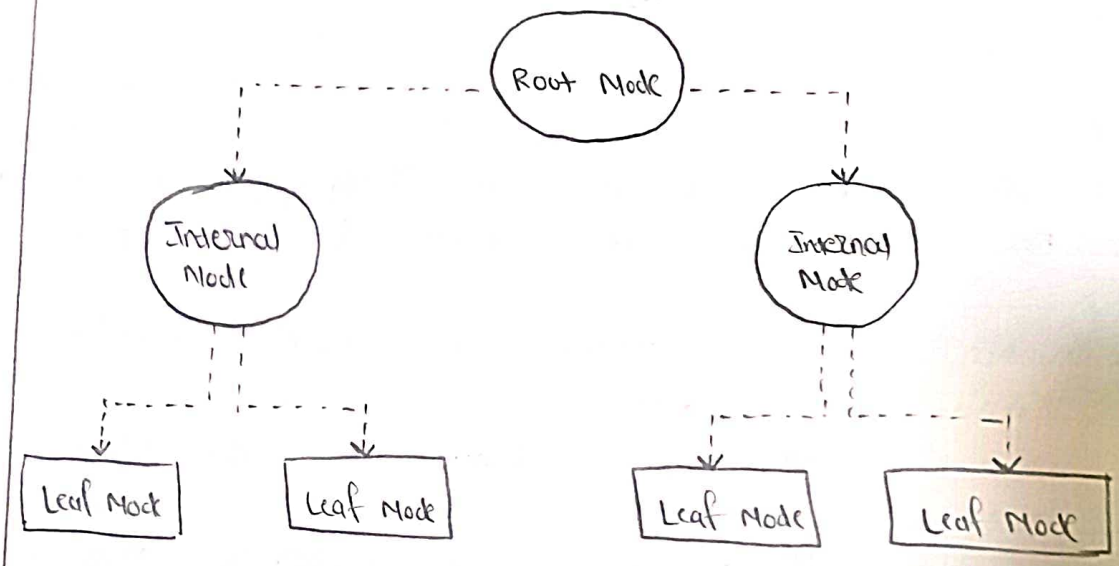
There are multiple type of classification algorithm, each with its unique functionality and application. All of those algorithm are used to extract data from a dataset.

Data Mining algorithm for classification :

- Decision Trees
- Logistic Regression
- Naive Bayes classification
- k - nearest neighbors
- Support vector machine.

Decision Tree :

A decision tree is a structure that includes a root node, branches and leaf nodes. Each internal node denotes a test on an


Root Node
Internal Node
Internal Node
Leaf Node
Leaf Node
Leaf Node
Leaf Node

Date :

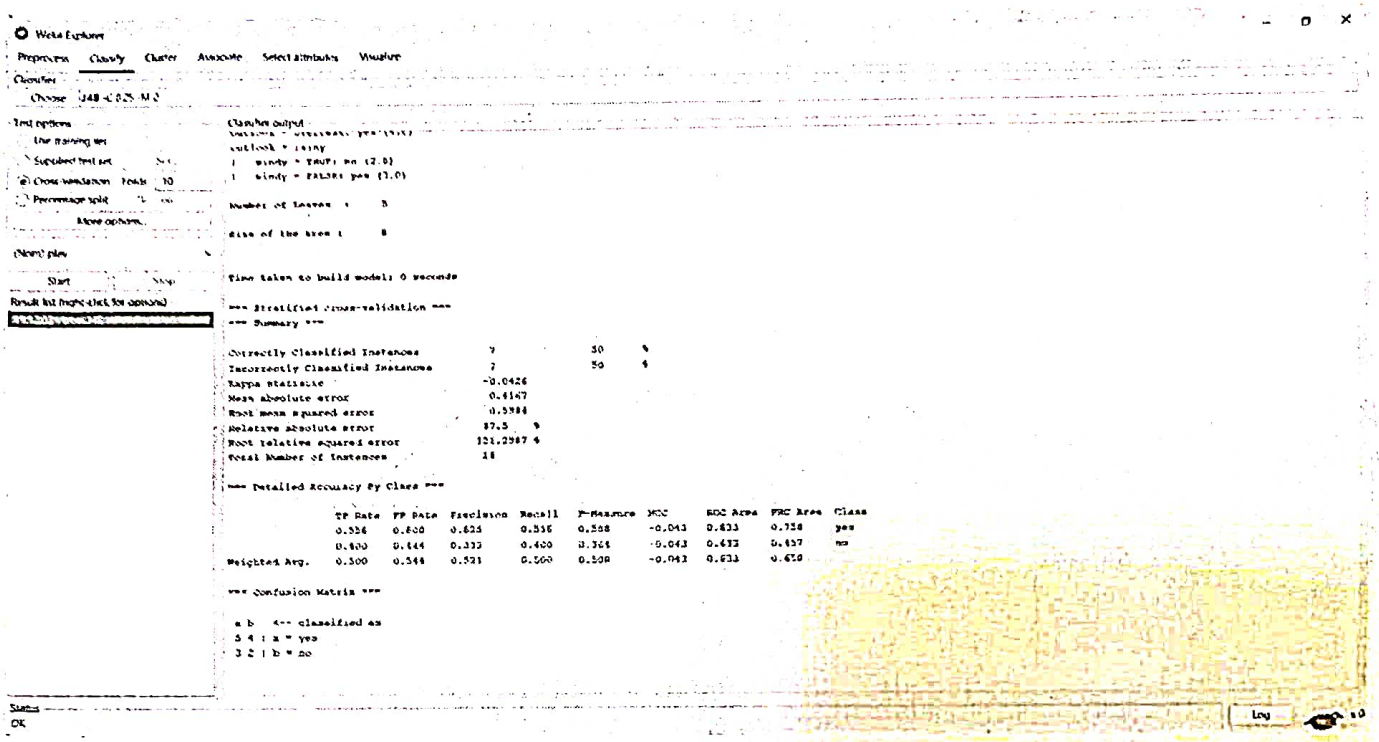
attribute, each branch denotes the outcome of a test and each leaf node holds a class label. The topmost node in the tree is the root node.

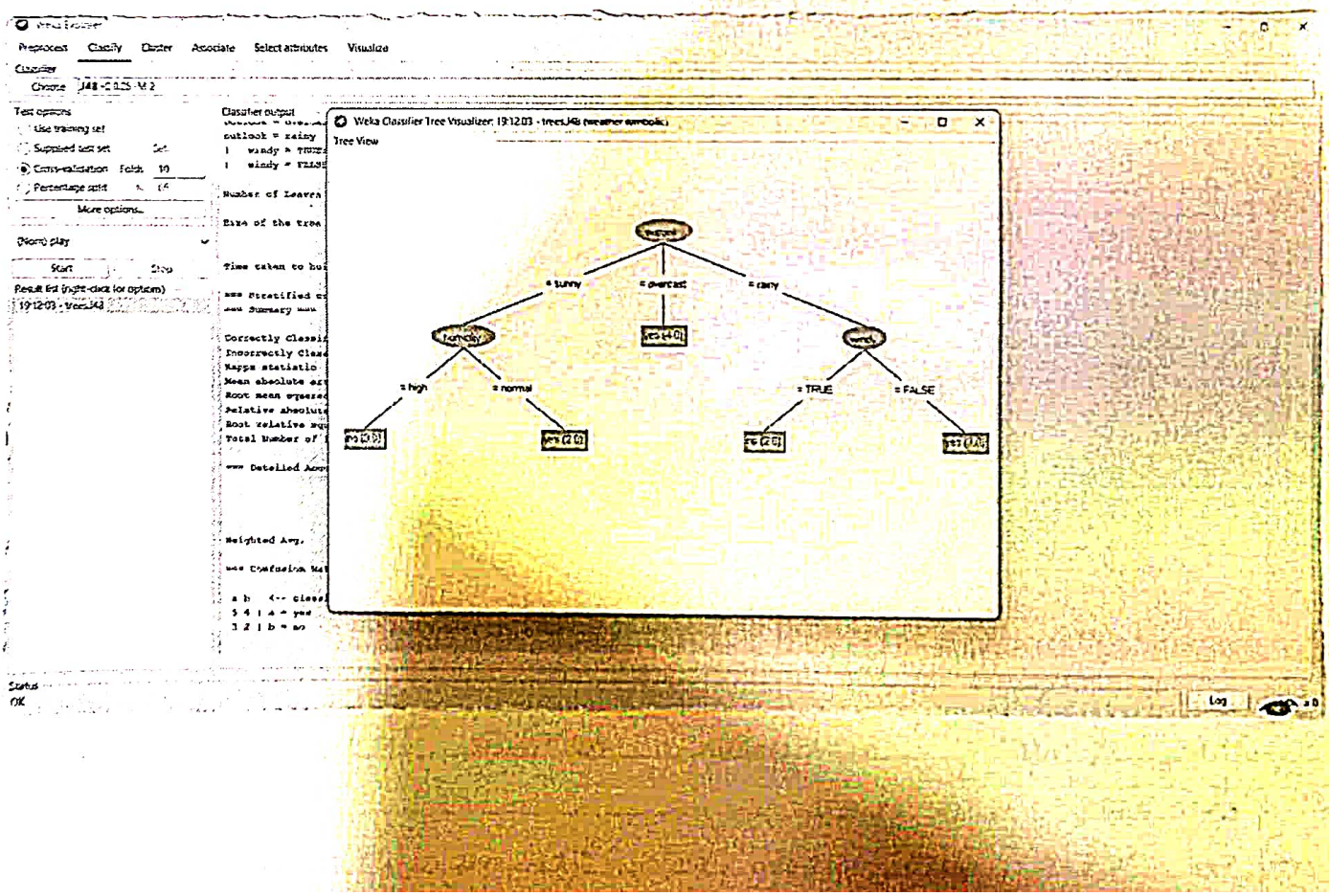
A decision tree is a classification scheme to generate a tree consisting of root node, internal nodes and external nodes. Root nodes representing the attributes. Internal nodes are also the attributes. External nodes are the class and each branch represent the values of the attributes.

Decision tree also contains set of rules for a given data set; there are two subsets in decision tree, one is a training data set and second one is a testing data set. Training data set is previously classified data. Testing data set is newly generated data.

The benefits of having a decision tree are as follows :-

- It does not required any domain knowledge.
- It is easy to comprehend
- The learning and classification steps of a decision tree are simple and fast.
- J48 classification and its decision Tree
- C4.5 algorithm / J48
- The C4.5 algo. is a classification algo. which provides decision trees based on information theory. It is an extension of Ross Quinlan's earlier ID3 algorithm also known as weka as J48, J standing for Java. The decision trees generated by C4.5 are used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.
- The J48 implementation of the C4.5 algo. has many additional features including accounting for missing values, decision trees pruning, continuous attribute value range, derivation of rule and etc..
- In WEKA, data mining tool, J48 is an open Page No. Java


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier output
outlook = rainy
| windy = TRUE: no (2.0)
| windy = FALSE: yes (3.0)
Number of Leaves : 3
Size of the tree : 5
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 7 50 %
Incorrectly Classified Instances 7 50 %
Kappa statistic -0.0426
Mean absolute error 0.4147
Root mean squared error 0.5984
Relative absolute error 97.5 %
Root relative squared error 121.2967 %
Total Number of Instances 14
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.556 0.600 0.625 0.556 0.588 -0.043 0.433 0.738 yes
0.400 0.444 0.333 0.400 0.364 -0.043 0.433 0.457 no
Weighted Avg. 0.500 0.544 0.521 0.500 0.508 -0.043 0.433 0.630
=== Confusion Matrix ===
a b <-- classified as
5 4 | a = yes
3 2 | b = no
Status
OK
Log


Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Weka Classifier Tree Visualizer: 19:12:03 - trees.J48 (weather.symbolic)
Tree View
outlook
= sunny
= overcast
= rainy
humidity
windy
= high
= normal
= TRUE
= FALSE
Status
OK
Log

Date :

implementation of the C4.5 algorithm. It allows classification via either decision trees or rules generated from them.

Generating a decision tree from training tuples of data partition D

Algorithm : Generate_decision_tree

Input :

Data partition, D, which is a set of training tuples and their associated class labels.

attribute_list, the set of candidate attributes.

Attribute selection method, a procedure to determine the splitting criterion that best partitions that the data tuples into individual classes. This criterion includes a splitting_attribute and either a splitting point or splitting subset.

Output :

A decision tree

Method :

create a node N;

if tuples in D are all of the same class, C then
    return N as leaf node labeled with class C;

if attribute_list is empty then
    return N as leaf node with labeled
    with majority class in D; // majority voting

Apply attribute_selection_method(D, attribute_list)
to find the best splitting_criterion;
label node N with splitting_criterion;

if splitting_attribute is discrete-valued and
    multiway split allowed then // no restriction to binary trees.

Recall :

Then, the classification on Data set is performed by decision tree (J48) method.

Date :

attribute_list = splitting attribute // remove splitting attribute
for each outcome j of splitting criterion

// partition the tuples and grow subtrees for each partitions
let Dj be the set of data tuples in D satisfying outcome j ;
// a partition

if Dj is empty then
    attach a leaf labeled with the majority
    class in D to node N;
else
    attach the node returned by Generate
    decision tree (Dj, attribute list) to node N;
end for
return N;

Result:
Thus the classification on data set is performed by decision tree (J48) method.

Viva Question :

① what is classification ?
→ Classification is the process for finding a model that describe the data values and concept for the purpose of prediction.

② what is the need of classification ?
→ The need of classification is to accurately predict the target class for each case in the data, it allows you to organize data set of all sorts, including complex and large datasets as well as small and simples ones.

Page No.

Date :

Q.3 What are the different methods of classification?

→ Logistic regression, Naive Bayes, Decision tree, k-Near Neighbours, Support vector machine, Bayes classifiers, function classifier, lazy classifier, meta classifier and so on.

Q.4 What are the advantages of a decision tree classifier?

→ ① It is easy to comprehend

② It does not require any domain knowledge.

③ The learning and classification steps of decision tree are simple and fast

④ less data preparation, Non-parametric versatility, Non-Linearity.