Assignment-1 Report

Submitted By

Name : Shivam Singh
Roll No. : 2018201015
Subject : SMAI

Question 1

== Train decision tree only on categorical data. Report precision, recall, f1 score and accuracy.

Here the decision tree build is an n-ary tree and only categorical data has been used for training purpose. Dictionary has been used to save the Children nodes and isleaf attribute is there in each node to tell whether the node is leaf node or not. Entropy method has been used to select attribute.

Calculated Precision, Recall, F1 Score and Accuracy -

Precision : 0.9999999999999998
Recall : 0.001834862385321101
Accuracy : 0.7580071174377224
F1 Score : 0.0036630036630036626

Question 2

== Train the decision tree with categorical and numerical features. Report precision, recall, f1 score and accuracy.

Here the decision tree build is an binary tree and numerical as well as categorical data has been used for training purpose. A Class named Node is used to denote each node of tree which contains left and right attribute which points to left and right child and isleaf attribute is there in each node to tell whether node is leaf node or not where the result will be saved. Entropy method has been used to select attribute and well as for finding threshold value.

Calculated Precision, Recall, F1 Score and Accuracy -

Precision : 0.9293286219081273
Recall : 0.9651376146788991
Accuracy : 0.9737544483985765
F1 Score : 0.9468946894689468

Question 3

== Contrast the effectiveness of Misclassification rate, Gini, Entropy as impurity measures in terms of precision, recall and accuracy.

The work done in Question 2 has been repeated here but with including some extra functions regarding Gini Index and Missclassification rate and has been compiled in one to see the contrast between the three methods.

-- For only Categorical Data

```
Precision using Entropy : 0.9999999999999998
Recall using Entropy : 0.001834862385321101
Accuracy using Entropy : 0.7580071174377224
F1 Score using Entropy : 0.0036630036630036626

Precision using Gini Index : 0.0
Recall using Gini Index : 0.0
Accuracy using Gini Index : 0.7575622775800712
F1 Score using Gini Index : 0.0

Precision using Missclassification : 0.0
Recall using Missclassification : 0.0
Accuracy using Missclassification : 0.7575622775800712
F1 Score using Missclassification : 0.0
```

-- For Whole Given Dataset

```
Precision using Entropy : 0.9293286219081273
Recall using Entropy : 0.9651376146788991
Accuracy using Entropy : 0.9737544483985765
F1 Score using Entropy : 0.9468946894689468

Precision using Gini Index : 0.9353680430879713
Recall using Gini Index : 0.9559633027522936
Accuracy using Gini Index : 0.9733096085409253
F1 Score using Gini Index : 0.9455535390199635

Precision using Missclassification : 0.9850427350427351
Recall using Missclassification : 0.8458715596330275
Accuracy using Missclassification : 0.9595195729537367
F1 Score using Missclassification : 0.910167818361303
```

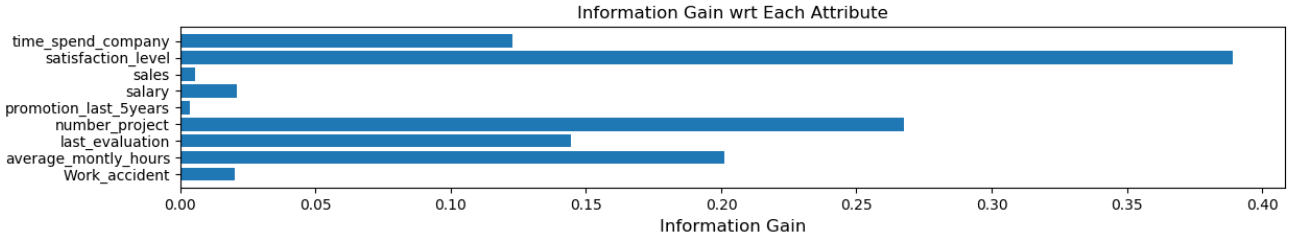
Question 4

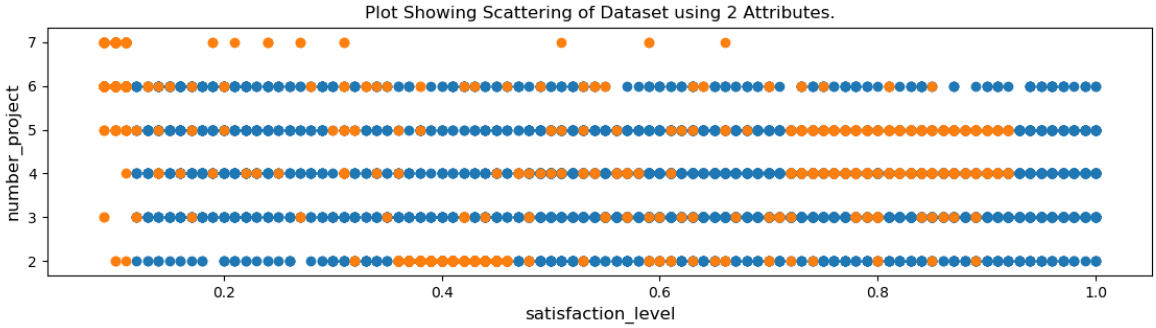
== Visualise training data on a 2-dimensional plot taking one feature (attribute) on one axis and other feature on another axis. Take two suitable features to visualise decision tree boundary

First Information Gain of every attribute is being Calculated and then those two attributes which have highest Information Gain have been selected to be plotted against each other to see better Decision Boundary.

The Two Attributes with Highest Information Gain were found to be satisfaction_level and number_project respectively. So, those two were selected and plotted on Graph.

Obtained Result and Graph was as follows :





Question 6

== Explain how decision tree is suitable handle missing values(few attributes missing in test samples) in data.

If in any test-case data is missing for any attribute then we will find all the possible values which that specific attribute can take from original dataset and turn-wise predict result for each of the possible value. Then from the obtained result whichever value has the highest probability among those that value will be as the predicted value.

For eg. in dataset if we have attributes like [Outlook, Temperature, Humidity, Wind, Play] and Play is the attribute for which value has to be predicted.

and we have a test-case like [overcast, hot, high,]

Here we can see that value for "Wind" Attribute is missing and we found that "Wind" can take following values ["weak", "normal", "strong"] from Given Dataset.

Then we'll find result for all the three possible cases i.e.

1) [overcast, hot, high, weak]
2) [overcast, hot, high, normal]
3) [overcast, hot, high, strong]

If it is found to be that predicted values are Yes, Yes and No for above cases then we'll return "Yes" as our predicted value or whichever has highest probability.