Cricket Result Prediction using

Data Mining

**Team Members**

**G.Hariharan - 1705076**

**B.Naraen - 1705088**

**Algorithms Used**

1. **Step Forward Feature Extraction**
2. **Random Forest Classifier**

**Objective:**

To Predict the result of the cricket matches using attributes like Team name,Venue,Toss and Toss decision.

**Dataset:**

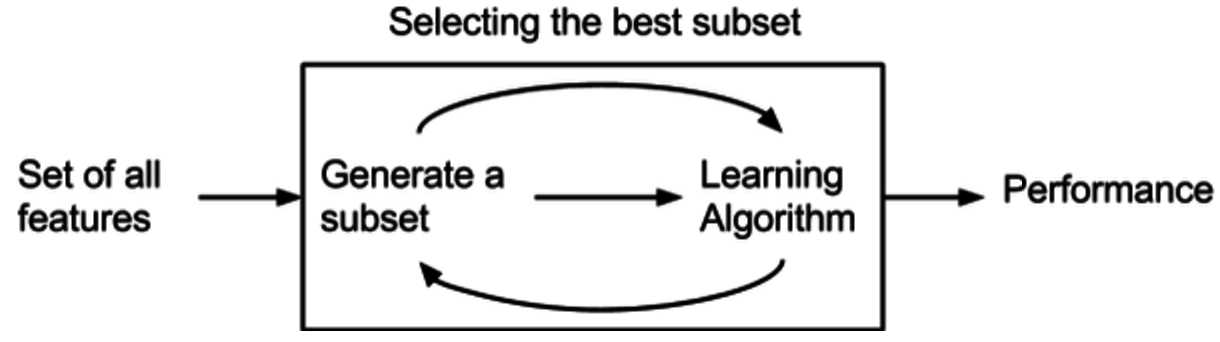
The Dataset I have used for this project contain the details of the 8 season IPL matches in which 8 teams are played.

Initially the Dataset have 16 attributes and the attributes is reduced to 6 for more accuracy.

The Dimensionality reduction is done using Step Forward Feature Extraction Algorithm,By using Step Forward Feature Extraction the attributes are reduced to 6, Which are Match date,Toss,Toss decision,venue,Is the match Result or not.

**Step Forward Feature Extraction:**

Step forward feature selection starts with the evaluation of each individual feature, and selects that which results in the best performing selected algorithm model. What's the "best?" That depends entirely on the defined evaluation criteria (AUC, prediction accuracy, RMSE, etc.).In my case Prediction Accuracy is the evaluation criteria. Next, all possible combinations of the that selected feature and a subsequent feature are evaluated, and a second feature is selected, and so on, until the required predefined number of features is selected.



By the Step Forward Feature Extraction the best set of attributes I have derived are

['Match\_Date',’Team\_Id’,'Opponent\_Team\_Id','Venue\_Name','Toss\_Winner\_Id','Toss\_Decision', 'IS\_Result'] with a accuracy score of 0.5357142857142857.

**Decision Tree:**

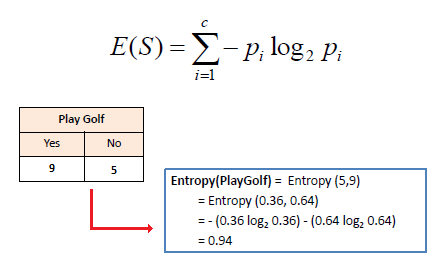
The Decision tree for the below mentioned dataset..

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Team\_Id | Opponent\_Team\_Id | Venue | Toss | Toss decision | Is\_Result | Result |
| CSK | MI | Chennai | MI | Field | yes | MI |
| CSK | MI | Mumbai | CSK | Field | yes | CSK |
| CSK | MI | Chennai | CSK | Field | yes | MI |
| CSK | MI | Mumbai | MI | Bat | yes | MI |
| CSK | MI | Mumbai | CSK | Field | yes | CSK |

The Decision tree is Constructed using Best Split and Entropy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Entropy** | | |  |  |
| A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar value.ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one. | | |  |  |
|  |  |
|  |  |

The Entropy is calculated using the formula



For the single attribute Result the Entropy is calculated as

1. >MI=3
2. >CSK=2

E(S)=Entropy(3/5,2/5)

E(S)=-(0.6log20.6)-(0.4log20.4)

E(S)=0.9709

|  |  |  |
| --- | --- | --- |
| **Information Gain** |  |  |
| The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches). |  |  |

**Step 1:**Calculate Entropy of the Target

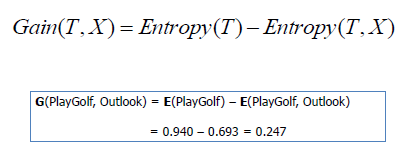
E(S)=Entropy(3/5,2/5)

E(S)=-(0.6log20.6)-(0.4log20.4)

E(S)=0.9709

**Step 2:** The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy.

The Formula for Information Gain is



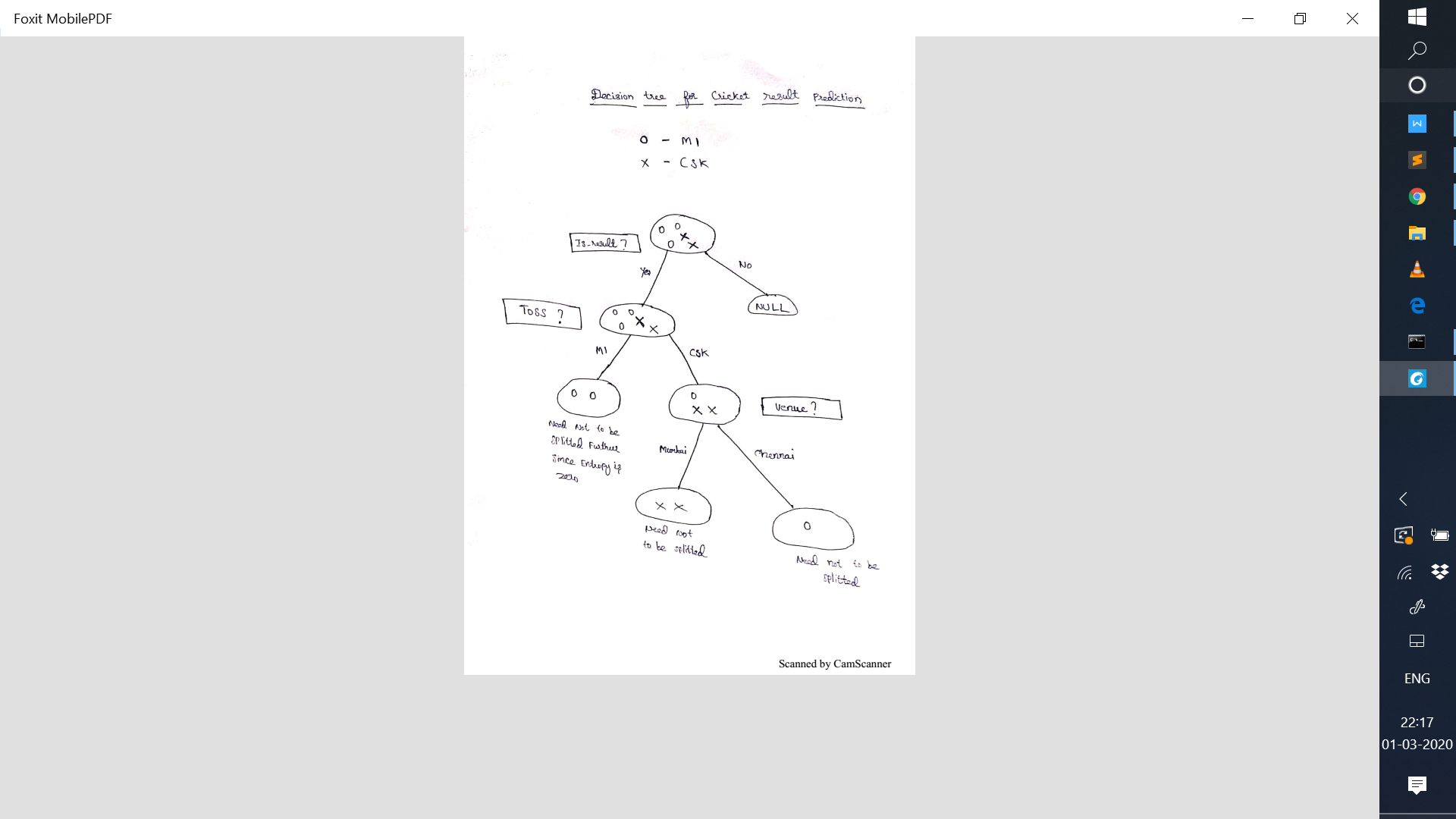
**Step 3:**

Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

**Step 4:**

* The Branch with Entropy 0 is the Leaf Node.
* The Branch that doesn’t reach the Entropy 0 needs to be Splitted.

**Decision Tree for the above Dataset:**

****

Entropy for Is Result= 0.97

Entropy for Toss=0.97

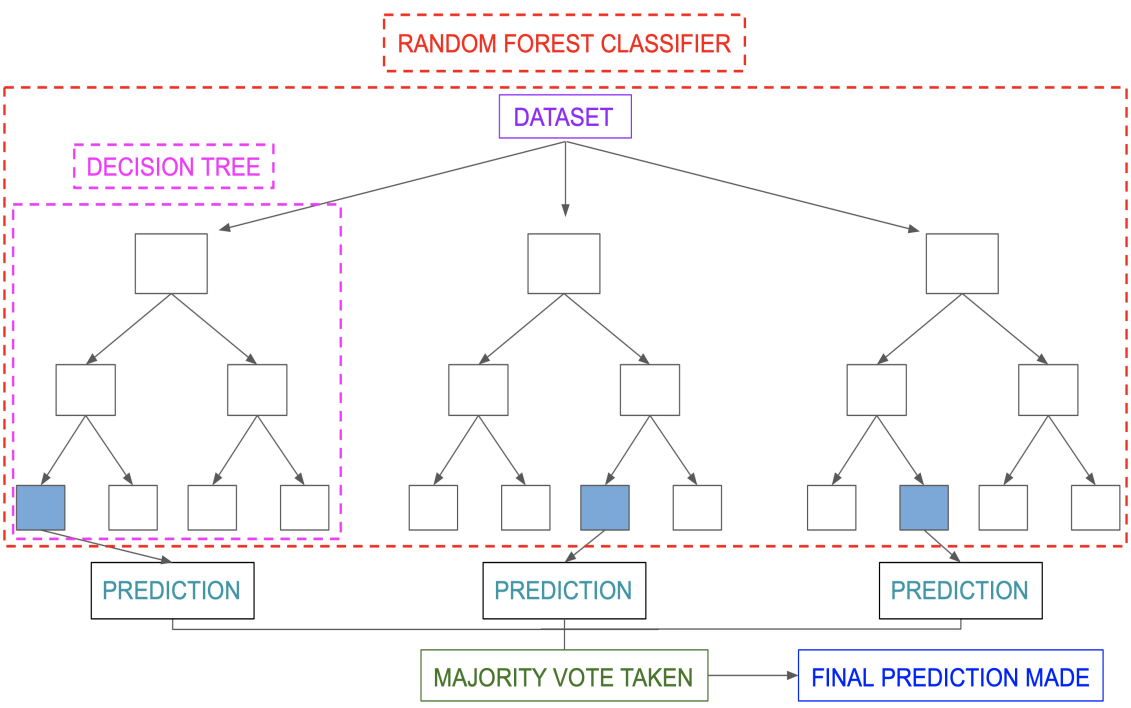
Entropy for Toss win by MI=0 (Since Entropy is zero need not to be splitted again)

Entropy for Toss win by CSK=0.92

Entropy for venue Mumbai and chennai= 0

**Random Forest:**

 A randomly sampled subset of the data on which classification is to be performed is passed through each of the decision trees in the forest in order to train that tree. Later, when we are performing classification for a datapoint, each tree constituting the forest gives a predicted class (label). The final prediction is then decided by measuring which prediction was made by the most number of trees in the forest.



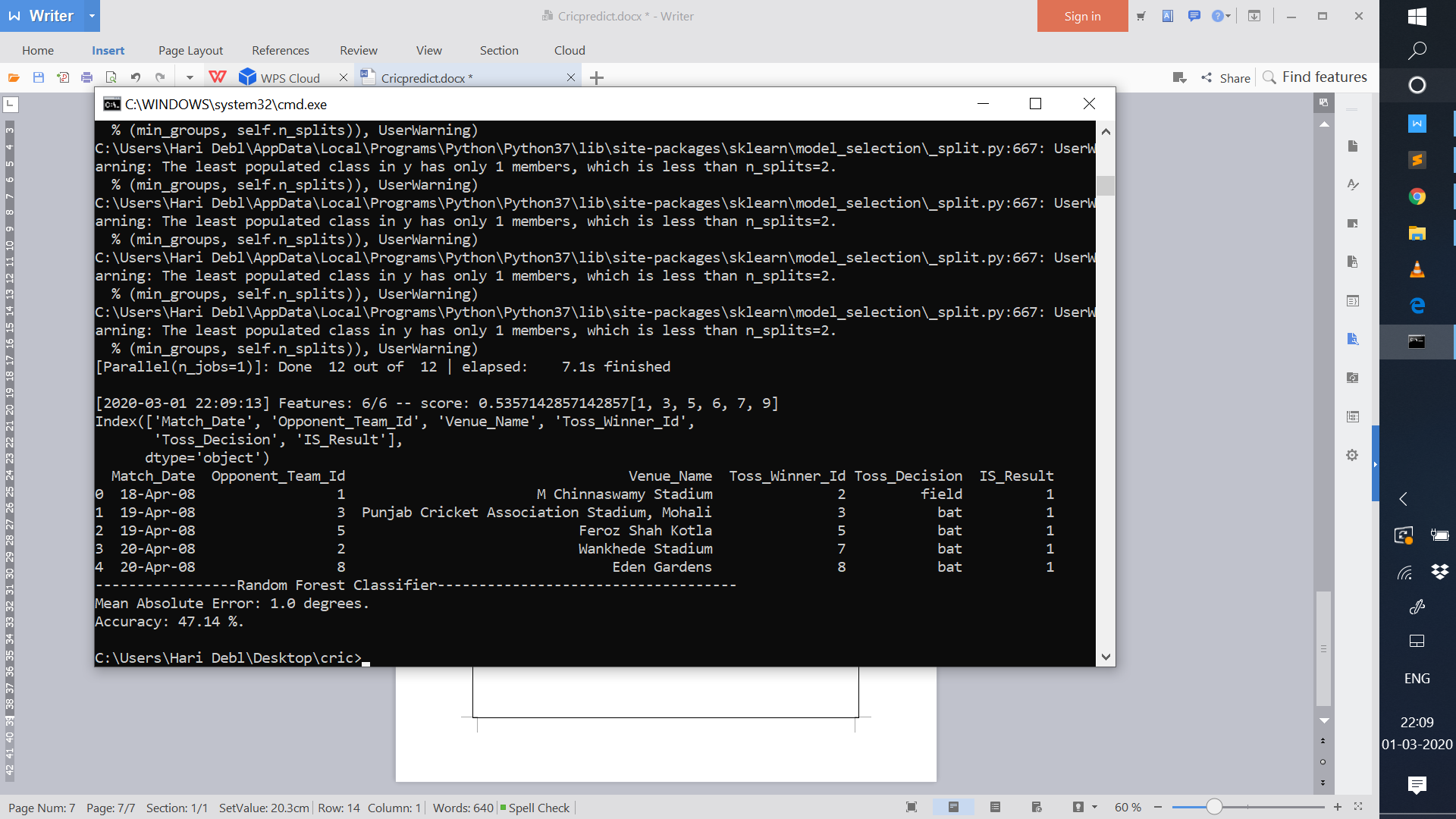
**Output:**

The Accuracy of the model is predicted using the below mentioned formula,

Error=| Predictions - Test Data |

%Error=100\*(mean(Error)/Test Data)

Accuracy=100- %Error



The Accuracy of the Cricket Match Result Prediction using Random Forest Algorithm is

47.14%.