**FOOTBALL MATCH DATA ANALYSIS**



TABLE OF CONTENTS

[ INTRODUCTION: 3](#_Toc12141865)

[ PROBLEM STATEMENT: 4](#_Toc12141866)

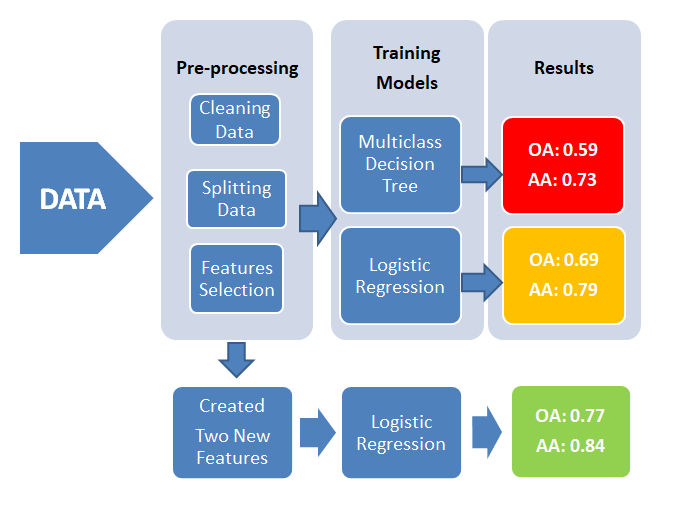
[ APPROACH: 4](#_Toc12141867)

[ PREPROCESSING: 5](#_Toc12141868)

[ BUILD AND INTERPRET 6](#_Toc12141869)

[ CONCLUSION 7](#_Toc12141870)

# Summary:



# INTRODUCTION:

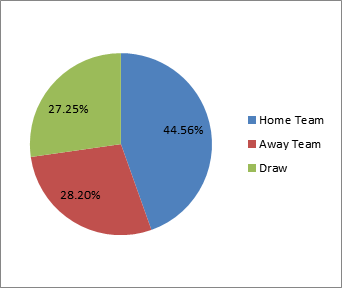
The objective of this project is to apply various machine learning concepts (models and algorithms) on the football dataset provided “football3.csv”and analyse the data accordingly. The dataset contains past match data of home team and away team for the season 2014-2015 (aug-may), and it has statistics like scores (half time, full time), shots by each side, shots on target by each side, fouls done by each side, corners received, cards (yellow, red) received by each side, odds (best, and average) and bookmakers data. Overall??? matches were played amongst ??? teams.

In this project, we have taken some of these features and used them to predict the target outcome FTR i.e. Full Time Result. We are trying to predict whether the home team won, or away team or it was a draw.

**Data Set:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Home Team Wins** | **Away Team Wins** | **Draw** | **Total** |
| **Overall games played** | **11793** | **7463** | **7212** | **26468** |

**Aggregate Win Percentage:**



# PROBLEM STATEMENT:

Topredict outcome for betting organization based on football match past data as– Home team, Away

Team and Draw

# APPROACH:

We followed following steps:

1. Clean/pre-process the data – The dataset was pre-processed using correlation between dependent and independent fields and following features were used.
2. Split the data. Training was done on 70% of data and testing was done on remaining 30% of match data. …why and how? Included the features like…..
3. Model was trained using following algorithms:

* **OBJECTIVE:**

After analysing the data, we have come up with an approach of splitting the prediction into two parts based on current and past data.

* This is to predict the outcome of winning team based on past data for Home team, away team or draw.
* Secondly, to estimate the number of goals in the match based on past data.
* **APPLICATIONS USED:**
* Technology used for number of classification methods is MICROSOFT AZURE which is used for developing the model for classification, clustering and deploying machine learning algorithm and data handling module.
* To estimate the number of goals, through regression method, ORANGE is used as component based software.
* Otherwise, WEKA is an optional tool which can be used for clustering as well as classification.

## Pre-processing

1. Use correlation between dependent and independent variables.

Based on correlation, following features have been selected-

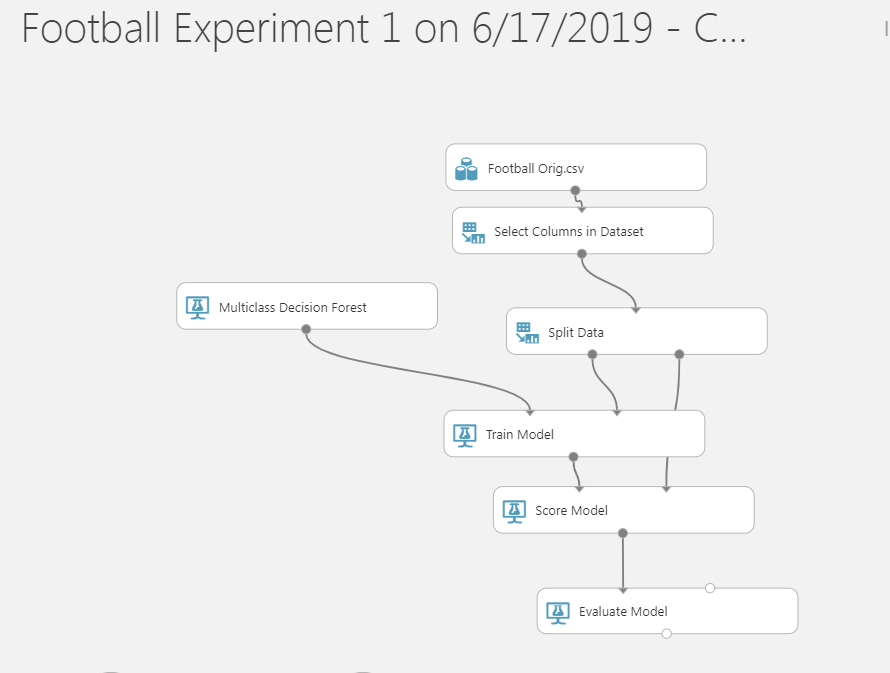
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| FTR | HST | BbAHh | AST | HTR | HF/AF | HY/AY | HR/AR | HC/AC | HS/AS |
| Full time result | Home team shots on target | Size of handicap for the home team | Away team shots on target | Half-time result | Fouls by each time | Yellow cards by each team | Red cards by each team | Corners by each team | Total shots by each team |

We have not taken book maker’s data into our features because the book maker’s data is already predicted data and there exists a collinearity with the match statistics data used to predict our model.

Moreover, the other pre-processing methods such as normalization and standardization are not considered because the features are not varying in scale and the range also doesn’t vary within the features.

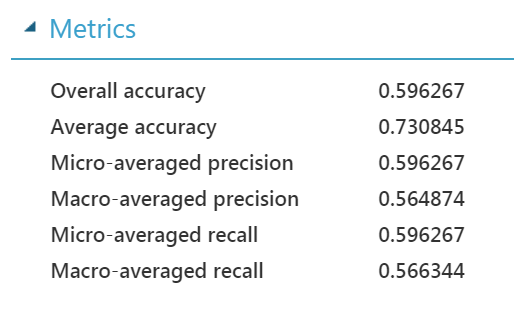
1. Next step is to split the data to create a training dataset and a test dataset. We have used 60-40 split of the data for training and test.
2. Then we used Train Model module to create a trained module using Multiclass Decision Forest algorithm. Here we tried different classification algorithms such as Baye’s, Decision jungle and logistic regression. However, on comparing the results of each, we finally decided to use Decision forest algorithm.
3. Then we used score model module to score the test dataset and finally evaluate model to see the accuracy of the model. Below is the snapshot of the experiment created –

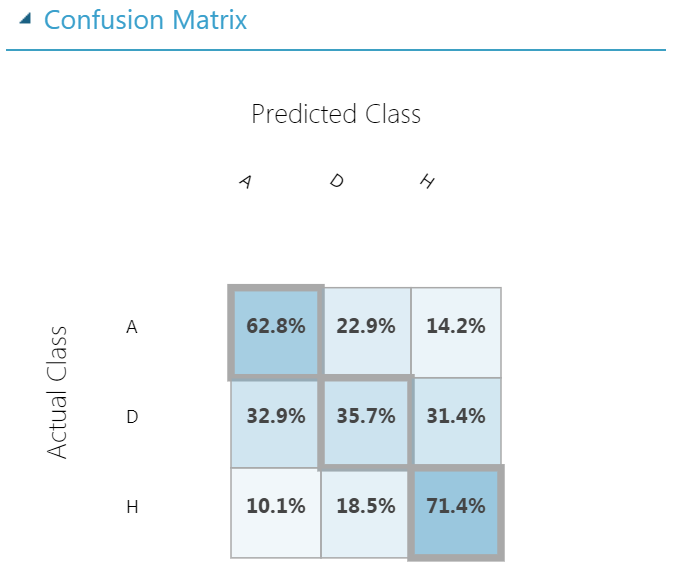
# **Multiclass Decision Tree**



# Evaluation

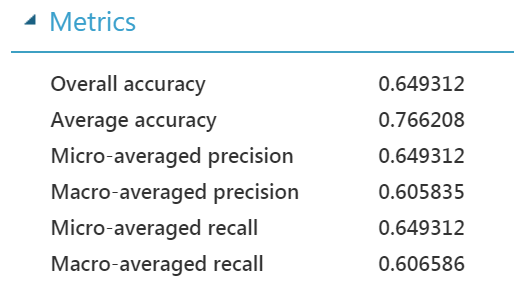
The key statistics from the evaluate model are –

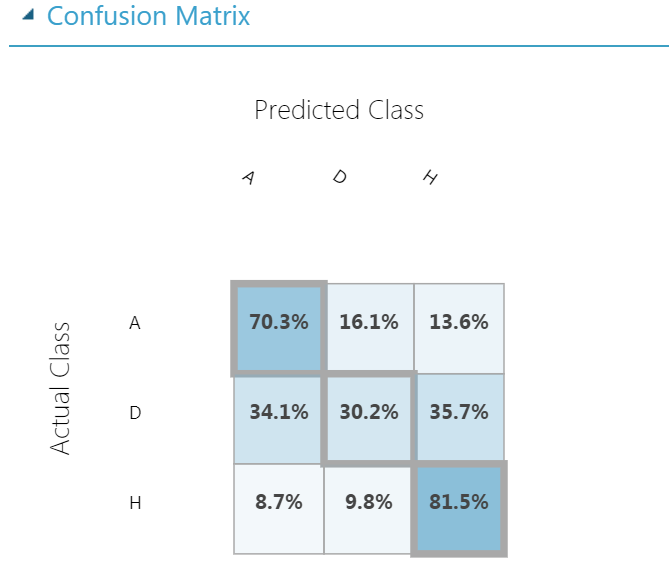




# Logistic Regression

We improved above statistics by using logistic regression approach –

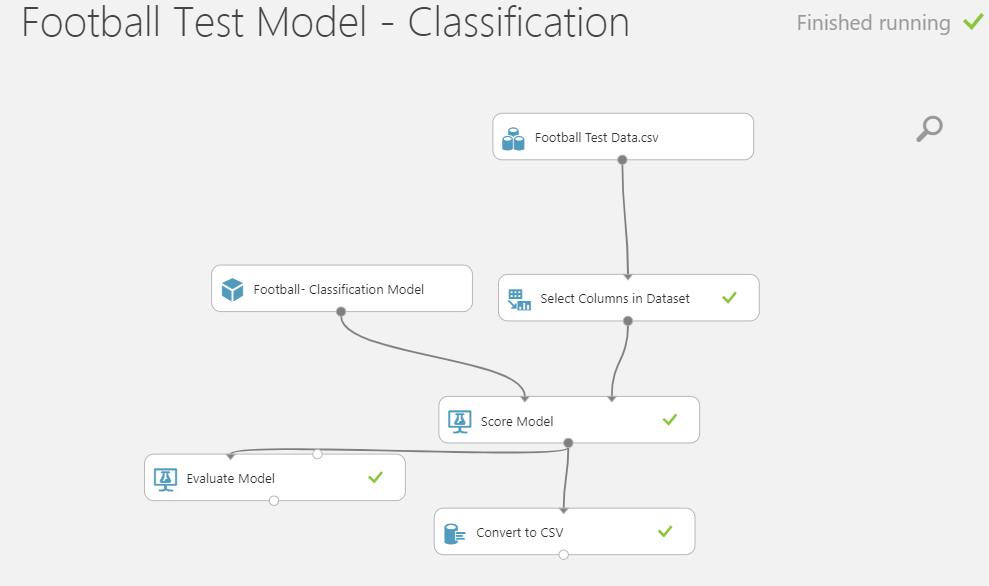




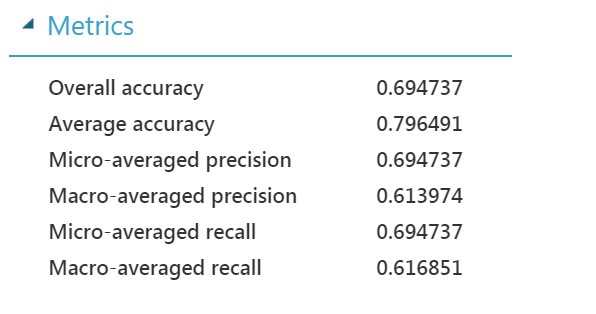
# Prediction

We used a test dataset for prediction; the dataset contains the E0 league matches played in 2018.





The output is as presented below –



# Logistic regression with two new features:

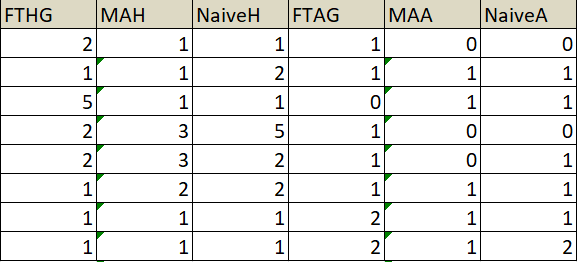
We have used two forecasting models to generate our new features. We have used FTHG(Full Time Home Goal) and FTAG(Full Time Away Goal) to generate two new features. Model used by us are:

1. **Naive Model:** Estimating technique in which the last period's actuals are used as this period's forecast, without adjusting them or attempting to establish causal factors. ) below Feature name NaiveH and NaiveA)

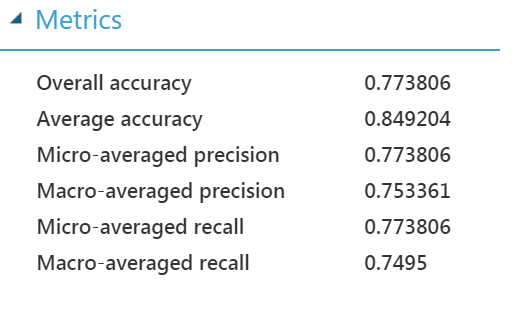
**Y(t) = Y(t-1)**

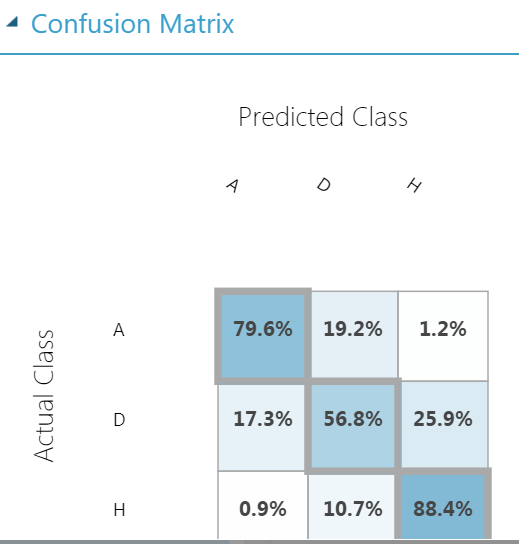
1. **Moving Average Method:** A technique to get an overall idea of the trends in a data set; it is an average of any subset of numbers. The moving average is extremely useful for forecasting long-term trends. In Excel we have used formula:

**Y(t) = Int(average(y(t-1), y(t-2)))**



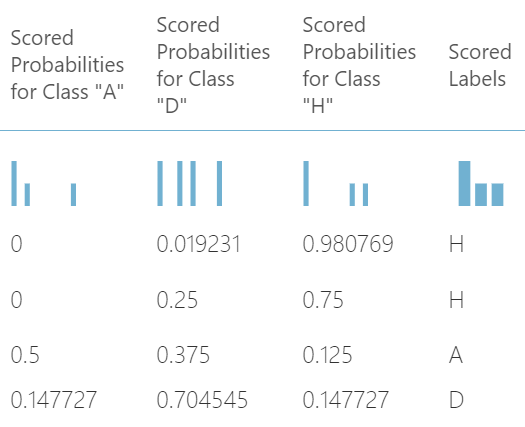
Below Output we got with these two new features:

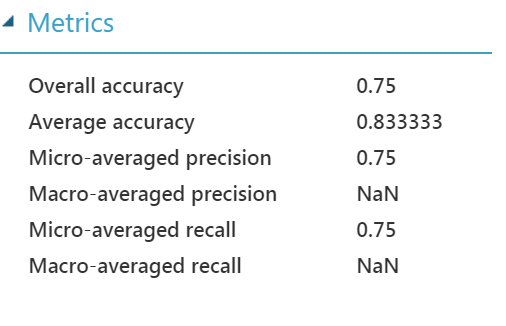




# Prediction

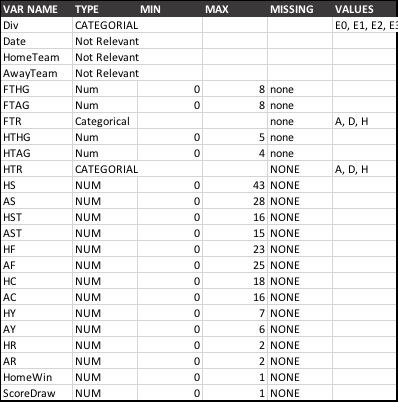
Our Prediction Score are given below:





# PREPROCESSING:

* MULTICLASS LOGISTIC REGRESSION

**Feature scaling**

For better results, algorithms need to have all features on same scale. This is for pre-processing process to visualize the data before being used for algorithm.

# BUILD AND INTERPRET

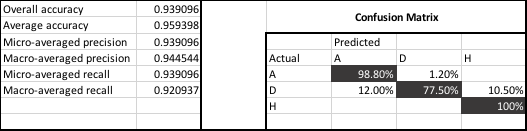
* CLASSIFICATION:

|  |  |  |  |
| --- | --- | --- | --- |
| **MULTI CLASS LOGISTIC REGRESSION** | | |  |
| FEATURE | **A** | **D** | **H** |
| FTHG | -6.38212 | -0.846134 | 8.18769 |
| FTAG | 7.67564 | 0 | -6.72663 |
| HTHG | -1.99604 | 0 | 1.51156 |
| HTAG | 1.30854 | 0 | -1.55115 |
| AST | 1.02221 | 0 | -0.587847 |
| HST | -0.89407 | 0 | 0.761049 |
| Bias | -0.198457 | 0.42206 | -0.223599 |
| HC | 0 | 0.0248297 | -0.364631 |
| AR | 0 | 0 | 0.304347 |
| HR | 0.248854 | 0 | 0 |
| AC | -0.109076 | 0 | 0.23143 |
| HY | -0.0119236 | 0.147611 | 0 |
| Div\_E1\_1 | 0 | 0.145818 | 0 |
| AS | 0 | -0.114452 | 0 |
| Div\_E3\_3 | -0.0973431 | 0 | 0.0334483 |
| AF | -0.0507512 | 0.009864 | 0 |
| Div\_E2\_2 | 0.036925 | 0 | 0 |
| Div\_E0\_0 | 0 | 0 | 0 |
| Div#unknown\_\_5 | 0 | 0 | 0 |
| HS | 0 | 0 | 0 |
| HF | 0 | 0 | 0 |
| AY | 0 | 0 | 0 |

**Interpretation:**

* The data shows intuitive relationship between factors like goals scored in full time, goals scored in half time, shots on target for both home and away games.
* The number of corners has counter intuitive signs which means style of play does not lead to goal scoring corners or there is collinearity with other variables.
* Being in division E0 has no impact on the result (Home/ Away/ Draw).
* E1 division teams are more likely to have Draws.
* E2 division teams are more likely to have Away wins whereas E3 division teams are more likely to have Home team wins

**Confusion Matrix:**



# CONCLUSION