INTRODUCTION

Soccer produces considerable amount of statistics that could prove to be very useful in data mining. There is keen interest to predict what the results of future matches could be. Predicting results with the help of data mining clears one of the weaknesses with prediction – the emotional aspect. Many teams rely on historical data and prediction results in order to understand their own teams better. Data mining as a field has become very important but not a widely used platform to guide teams. The project aims to predict soccer results using data from previous matches. Certain prediction algorithms have been applied in order to get the best predictions of some soccer matches. Multiple approaches have been made use of to make sure the best features are selected since selection of features greatly impacts the accuracy of the results.

Dataset
• Data Collection

Feature Selection
• Elimination Approach
• Calculated Elimination
• Single Team Prediction

Algorithms
• Logistic Regression
• Decision Tree

The data was collected from football-data.co.uk. There was no real method of data cleaning that had to be followed since most of the data was taken from official sources. Data for a single season (2010/11) was collected for use. Two of the approaches mentioned in the feature selection approaches made use of this dataset.

FEATURE SELECTION APPROACH

1. Elimination Approach (EL)
2. Calculated Elimination (CEL)
3. A single team’s performance over a couple of seasons.

ALGORITHMS EXPLAINED

• We use results of an entire season to predict the results of the next game or a bunch of games.
• The prediction result is either a Home win (H), Away win (A) or Draw (D)

• LOGISTIC REGRESSION
  • A method to fit a regression curve $y = f(x)$, i.e. to predict the value of $y$ when given a set of $x$.
  • The glm() function in R fits the data

• DECISION TREE
  • rpart() function in R allows us to create decision. The decision is made in the form of $x \sim a + b + c$, which is then fed to the predict() function
  • The predict() function, part of the rpart() package then makes the prediction and the output is then written to a file.

RESULTS

Accuracy was the metric used to evaluate the performances of both the classifier algorithms.

<table>
<thead>
<tr>
<th></th>
<th>LOGISTIC REGRESSION</th>
<th>DECISION TREE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elimination Approach</td>
<td>56.34 %</td>
<td>57.89 %</td>
</tr>
<tr>
<td>Calculated elimination</td>
<td>65.47 %</td>
<td>62.36 %</td>
</tr>
<tr>
<td>Single team prediction</td>
<td>69.12 %</td>
<td>68.42 %</td>
</tr>
</tbody>
</table>