**Factors affecting the prediction**

By far the most crucial factor to be considered is the venue where the game is being played. Multiple studies have been carried out on home/away field advantages to the teams and have found considerable success in doing so. Also this approach has been taken further by incorporating the information about the continent to which the away team belongs .

But this may not always lead to desired results because the home factor may not be uniformly influential. Adopting this may lead to a high degree of overfitting, that too in sports prediction where the data available is not huge.

This can be overcome to a certain extent by taking into account additional factors, which may not necessarily be of uniform significance to the final outcome, but when considered together may lead to a closer approximation of actual results.

These additional factors affecting the predictions can broadly be classified as controllable, uncontrollable and partially controllable. While factors such as team combinations, on field strategies can be grouped under controllable , while venue , game type and coin toss fall under uncontrollable.

Coming to partially controllable factors, the choice of opting to bat or field can be considered one. Our approach considers these factors and a logistic regression model is to find the regression coefficients.

For our model we are considering home/away, toss result, choice after toss, game type as the parameters.

**Data Preprocessing**

We start out by considering the espncricinfo dataset with information for matches between 1972 to 2015. Information about teams playing, toss outcomes , game type, choice after winning the toss is obtained from it.

Data cleaning is performed to removed any ambiguities if present. The next step involves the team-wise partitioning of data . 9 such separate datasets are formed from the parent dataset.

This is followed by assigning discrete values of 0 or 1 to the above mentioned parameters. Once this is done the regression coeffiecients are calculated.

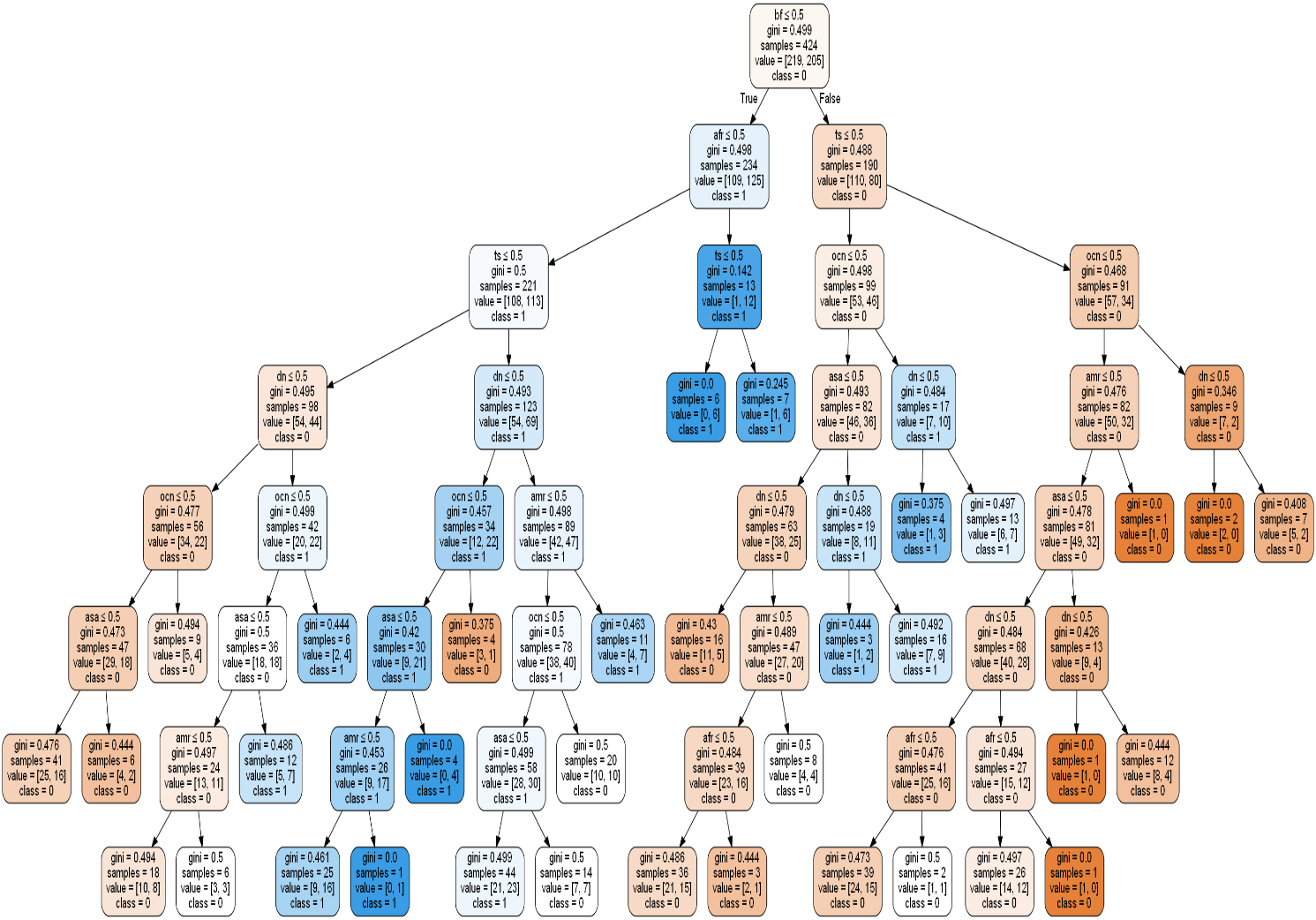
**Decision Tree Algorithm**

Decision Tree Classifier is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the calss label of the record is reached.

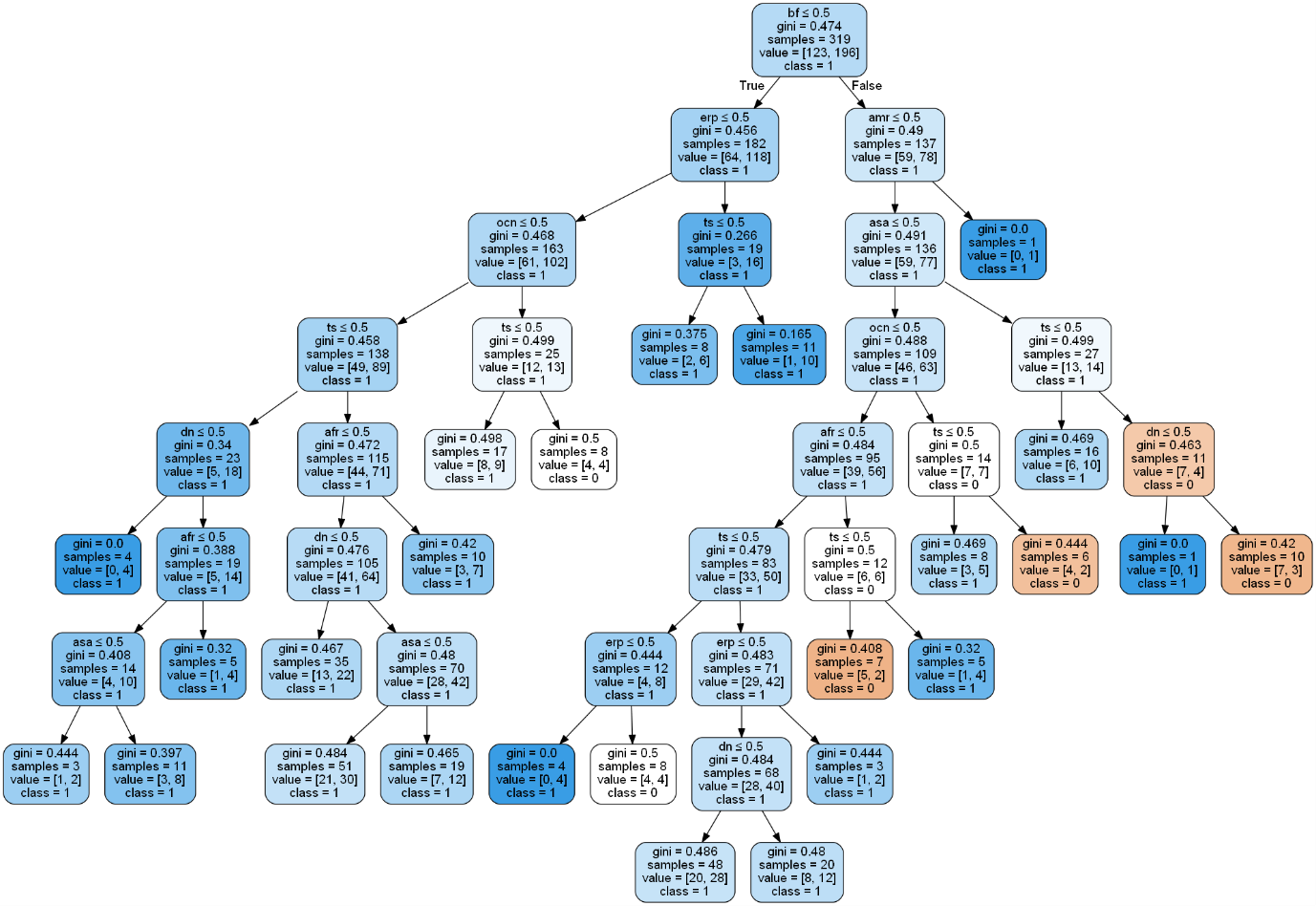
The order or the relative importance of these questions are kept track of using certain parameters which in our case is gini impurity. Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

**Decision Tree Outputs**

**England**



**West Indies**



**Australia**

