****

**PES UNIVERSITY**



**DEPARTMENT OF ELECTRONICS AND COMMUNICATION**

1. **ENGINEERING**

***A mini project report***

**TOPIC: ODI CRICKET PREDICTION**

**Submitted in Partial fulfilment of the Requirements for VI Semester**

**Bachelor of Technology in Electronics and Communication**

**Jan. - Apr. 2019**

**Under the guidance of**

**Dr. Koshy George K**

1. **Prof, Dept of ECE**

**TEAM MEMBERS**

**CHINMAY D KUCHINAD – 01FB16EEC081**

**CHIRANJEEVI N – 01FB16EEC083**

**DHEERAJ P ANIKAR – 01FB16EEC090**

**A machine learning approach to analyze ODI cricket predictors**

Cricket has become one of the most popular sports in the world and craze for the game is massive in a country like India. In the recent years with improvement in technology the game in 21st century has become spectacular than ever before. The popularity of ODI cricket has been high for the past few decades. Broadcasters and cricket experts make best efforts to provide a great experience of the game to the viewers. Pre-match predictions keep the fans and public hooked to the game. In this project we are analyzing ODI cricket predictors using machine learning. For this the understanding of factors that affect the game is essential.

**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 over fitting, 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 ESPN cricinfo dataset with information for matches between the years 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 remove any ambiguities if present. We have removed cricket matches that have been interrupted by rain, abandoned or ended in a tie as they affect our final outcome. 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 as the 2 models discussed below. Once this is done the regression coefficients are calculated as described.

**The logistic regression approach**

Outcomes of bilateral games for a team ‘i’ can be modeled by the following logistic regression approach.

= β0+ β1\*HMi+ β2\*DNi+ β3\*TSi+ β4\*BFi

where the predictors in this model are defined as follows.

HMi = 1 if it is a home game for the team i

0 otherwise

DNi = 1 if the game is a day game

0 if the game is a day-night game

TSi= 1 if the team i won the Toss

0 otherwise

BFi = 1 if the team i batted first

1. otherwise

Home field advantage includes cricket pitch, ground conditions and regional factors. Even though this model accounts for home-field advantage, it does not quantify the opposition’s foreign factor or the continent effect. The new variable we suggest here intends to account the home-field advantage of the home team with respect to the continent of the opposition team.

We have divided the cricket playing nations into their respective continents. Africa (South Africa, Kenya, and Zimbabwe), America (Canada and West Indies), Asia (Bangladesh, India, Pakistan, and Sri Lanka), Europe (England, Ireland, and Scotland) and Oceania (Australia and New Zealand).

The model 2 is the extension of model 1 to have five different continent parameters for each team including its own continent in place of HM.

Model 2 is as follows

= β0 + β1\*DNi + β2\*TSi + β3\*BFi + β4\*AFRi + β5\*AMRi +

β6\*ASAi + β7\*ERPi + β8\*OCNi

where continent variables for away teams are defined as

AFRi = 1 if the away team i∈ Africa

0 otherwise

AMRi = 1 if the away team i∈ America

0 otherwise

ASAi= 1 if the away team i∈ Asia

0 otherwise

ERPi = 1 if the away team i∈ Europe

0 otherwise

OCNi = 1 if the away team i∈ Oceania

1. otherwise

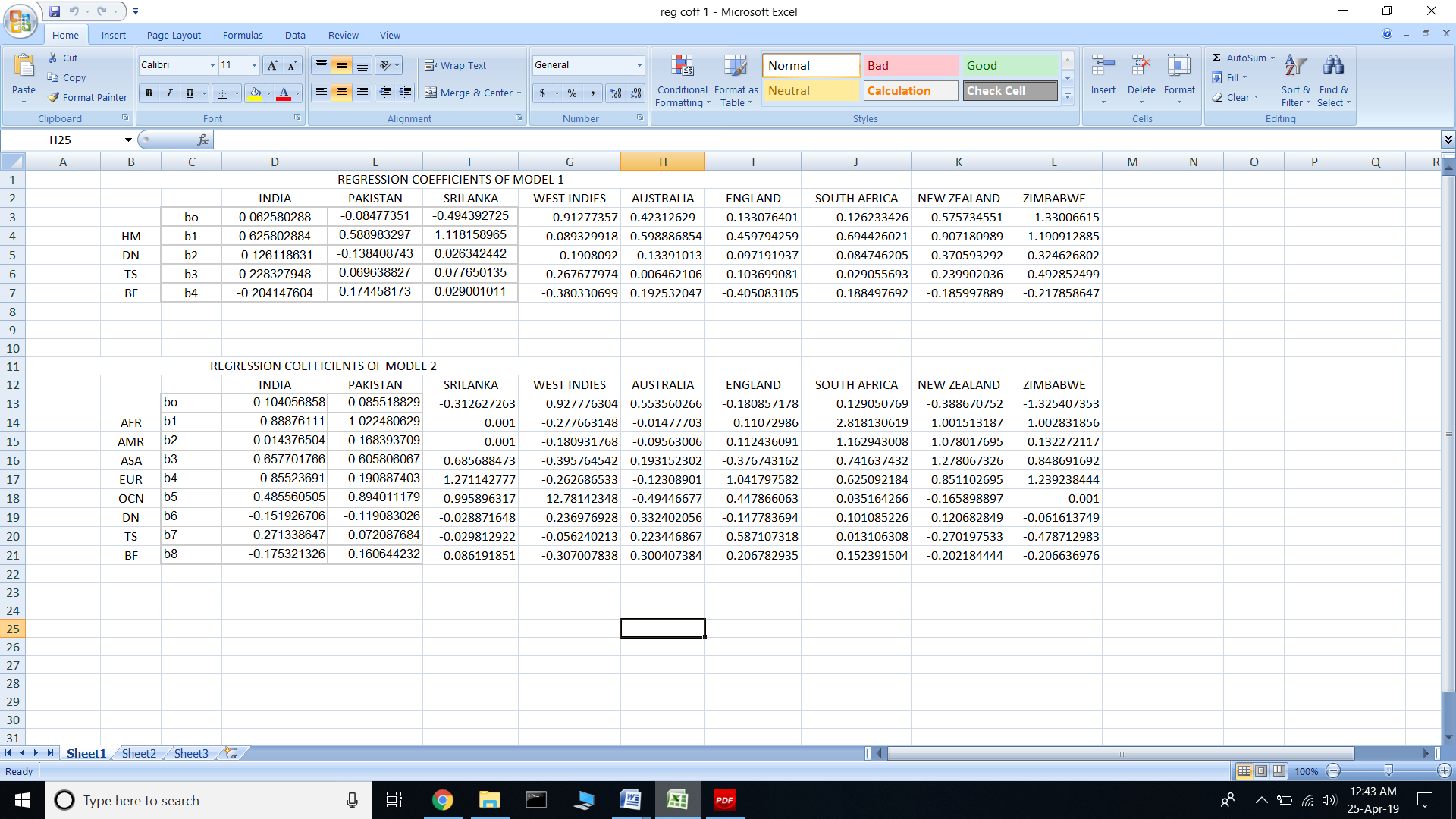
**Finding the regression coefficients**

We have found predictors using the following steps,

1. Find the logit (L) value using the above models
2. Find exponential of the logit and then calculate p(x) =
3. Using the logistic regression formula find log likelihood of p(x)

log likelihood = y\*log(p(x)) + (1 – y)\*(log(1 – p(x))

1. Using this, appropriate values of predictors/regression coefficients are calculated.



**Inferences that can be made from regression coefficients of model 1 and 2 are as follows**

1. Home field advantage is significant for teams like India, New Zealand, Australia and South Africa.
2. Australia and South Africa win many games when they bat first
3. Winning toss plays major role in the wins of India and England
4. West Indies win good number of games when they bat first
5. India has good chance of winning against most of the teams that tour India
6. Australia receive a good fight from England and South Africa while playing at home
7. South Africa wins many games against all the Asian and European nations

These are some of the many inferences that can be made.

In the next section we look at the CART – Classification and Regression Trees approach

**CART – CLASSIFICATION AND REGRESSION TREES**

CART is a popular ML technique that can be to predict outcomes and also to acknowledge the importance of predictors. The CART is a binary decision tree algorithm which recursively partitions data. Importantly, it can handle both continuous and categorical responses and predictors. In general, the trees are grown to the maximum possible size and then get rid of unimportant predictors (pruning) sequentially depending on a cost-complexity measure. Usually, this algorithm produces a set of pruned trees and finally selects the final tree as the best predictive model.

**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 class label of the record is reached.

**Gini impurity**

The order or the relative importance of these questions are kept track of using certain parameter 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.

Gini impurity function,

**i(t) = 1 − p(t)2 − (1 − p(t))2**

where p(t) is the conditional probability of class 1 provided the current node is t. The change or the gain of the impurity function due to a split of the parent node (tp) into left and right children nodes tL and tR,

respectively, is

**Δ i(t) = i(tp) − qi(tL) − (1 − q)i(tR)**

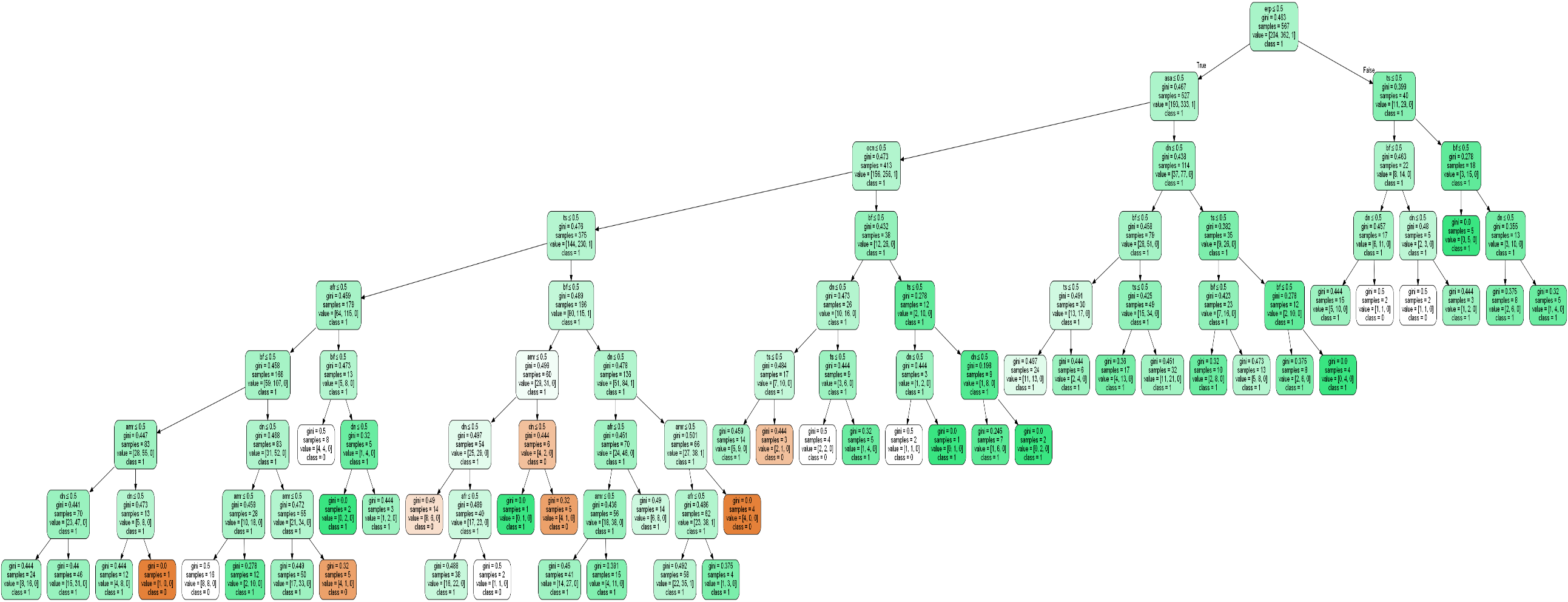
where q is the probability of the instances going to the left node. CART finds the best splitting criteria which maximizes the gain in impurity measure Δi(t).

**Implementing the decision tree**

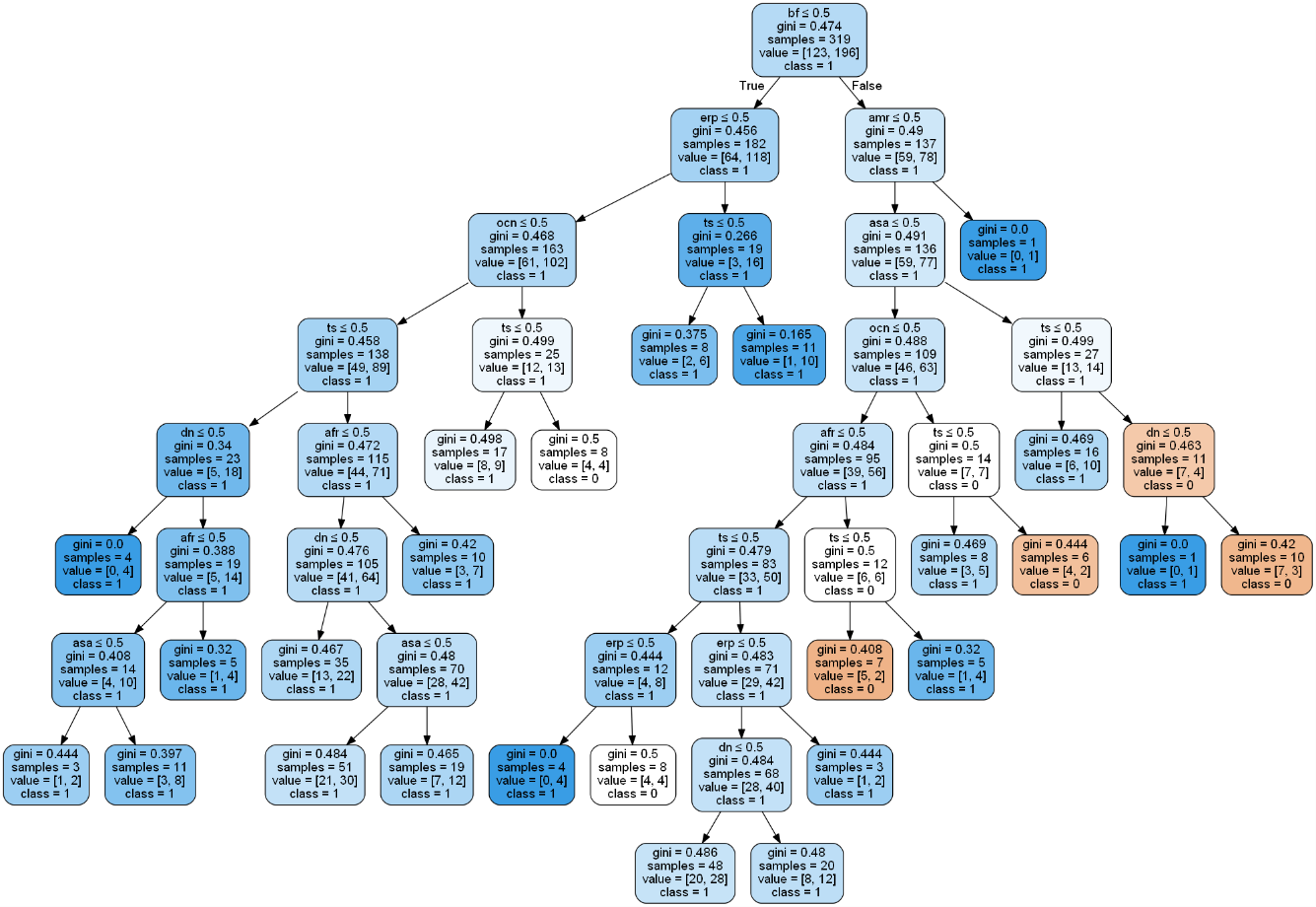
We use two approaches. One using the predefined function in scikit learn and the other involving writing the function on our own.

The scikit learn tree by default branches to the left child node for true cases and right child node for false cases. The parameters are placed hierarchically based on the gini impurity values and the relative importance of the parameters with respect to the final outcome is determined.

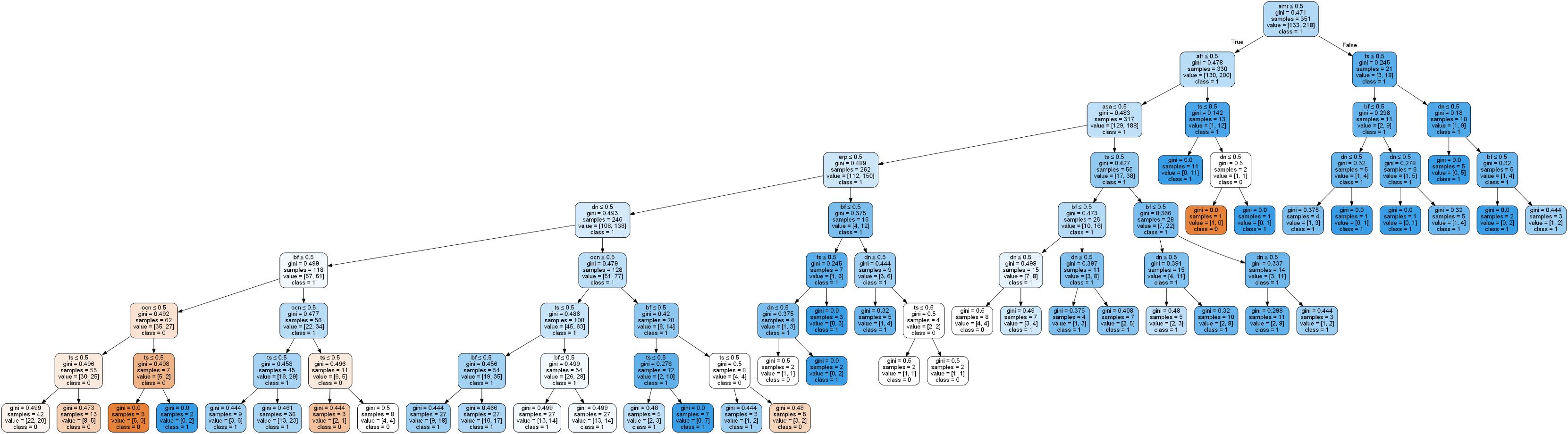
**Australia**

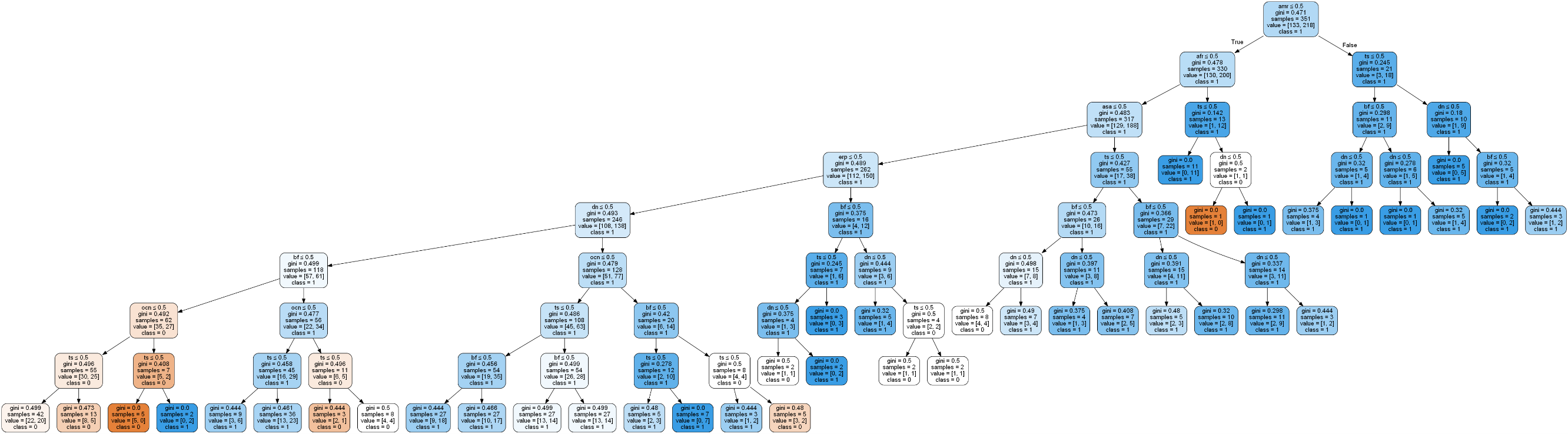


**West Indies**

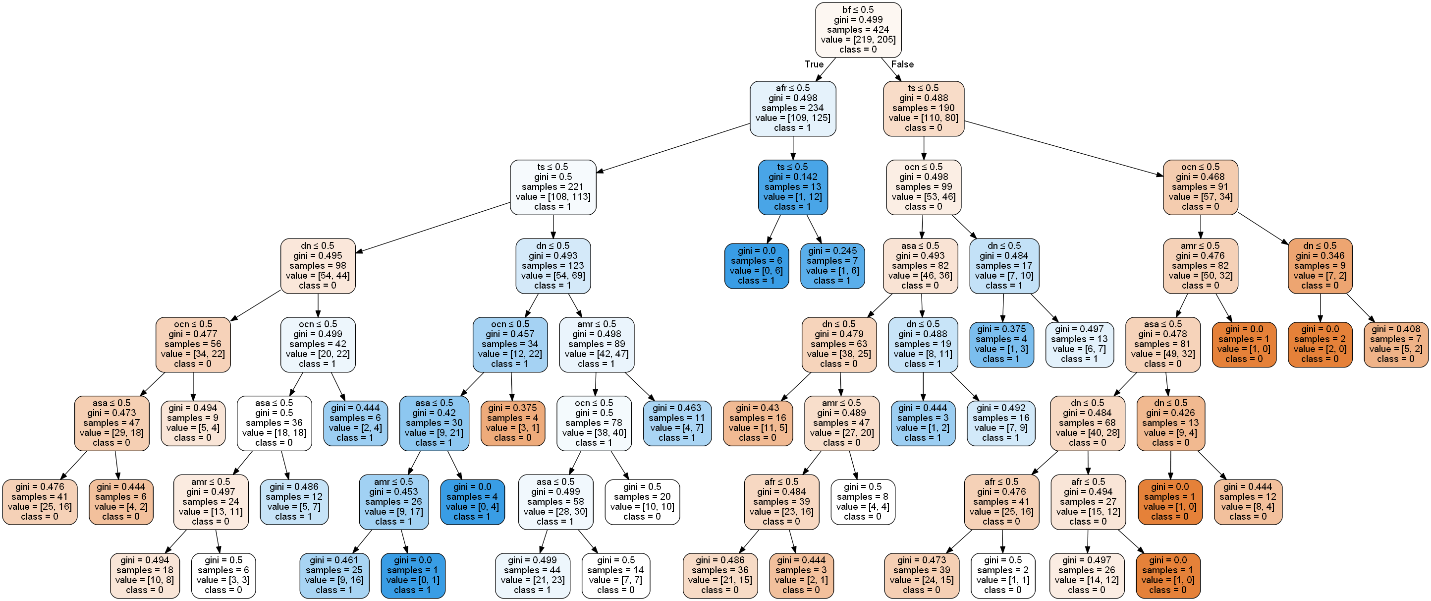


**South Africa**

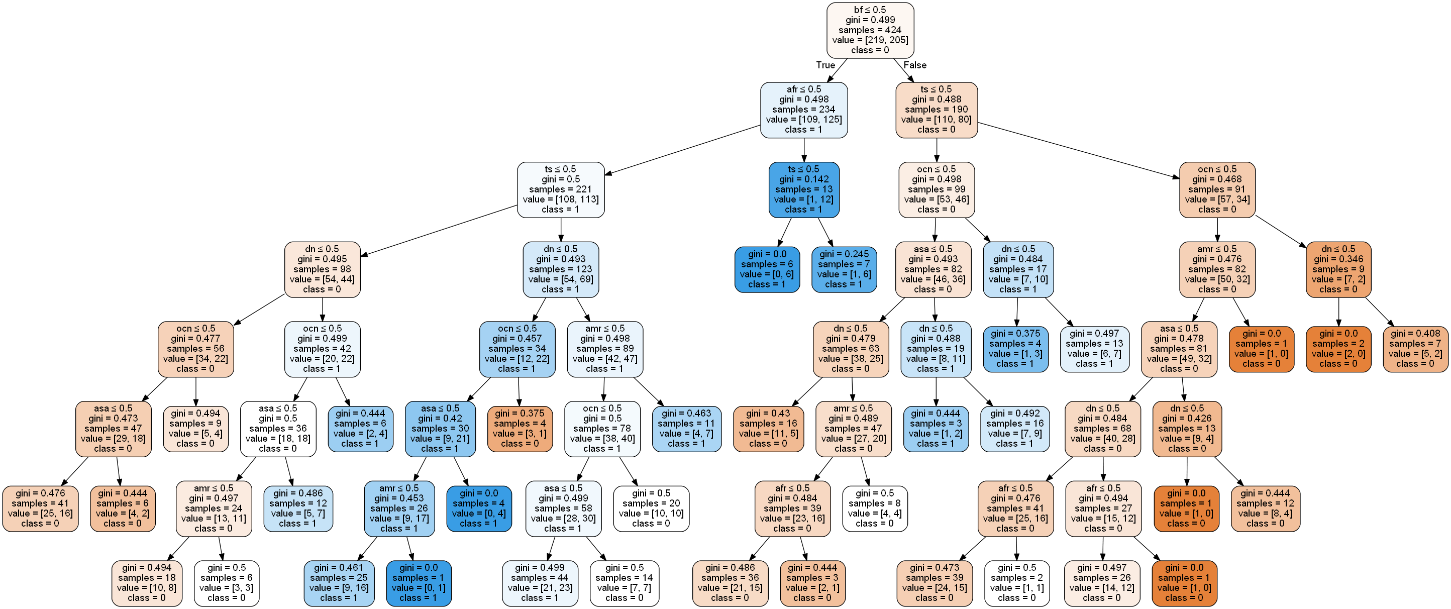
**New Zealand**



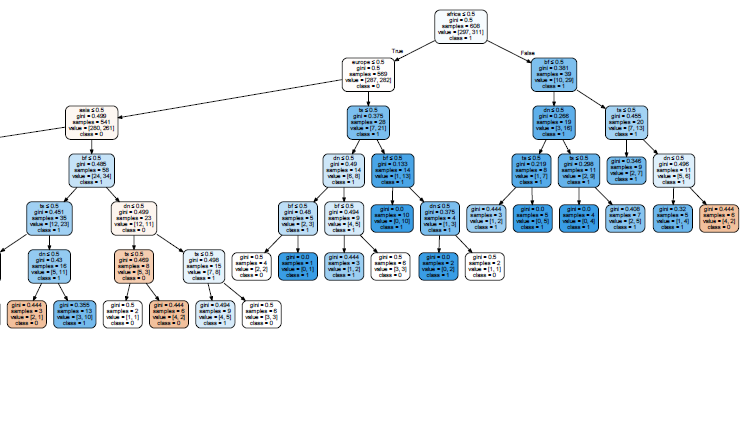
**England**



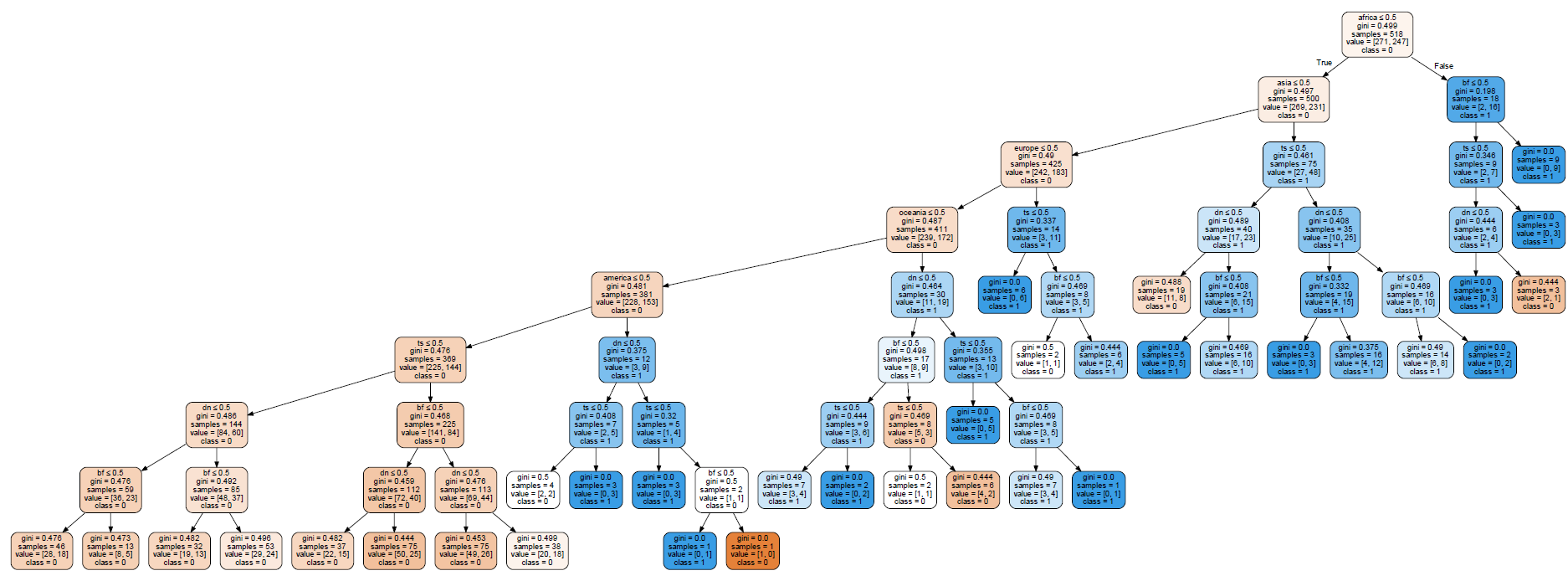
**Zimbabwe**

****

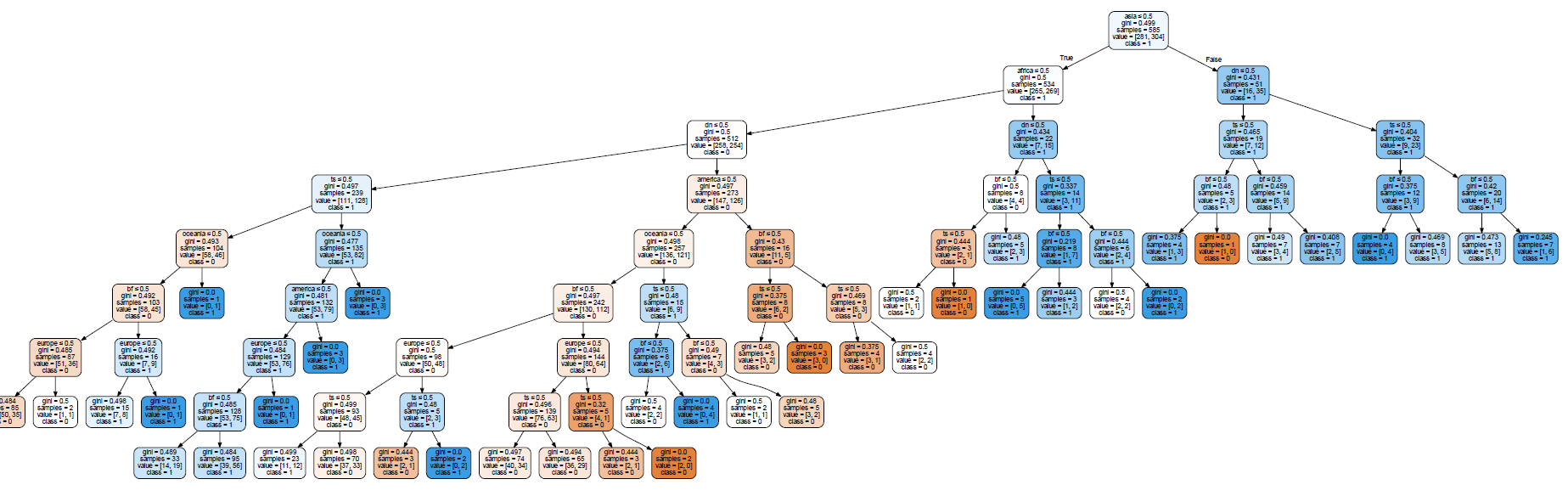
**India**

****

**Sri Lanka**

****

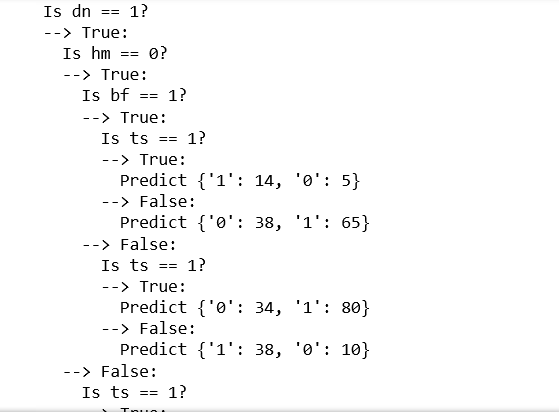
**Pakistan**

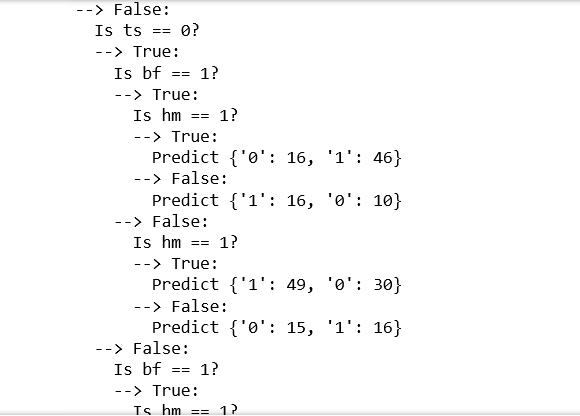
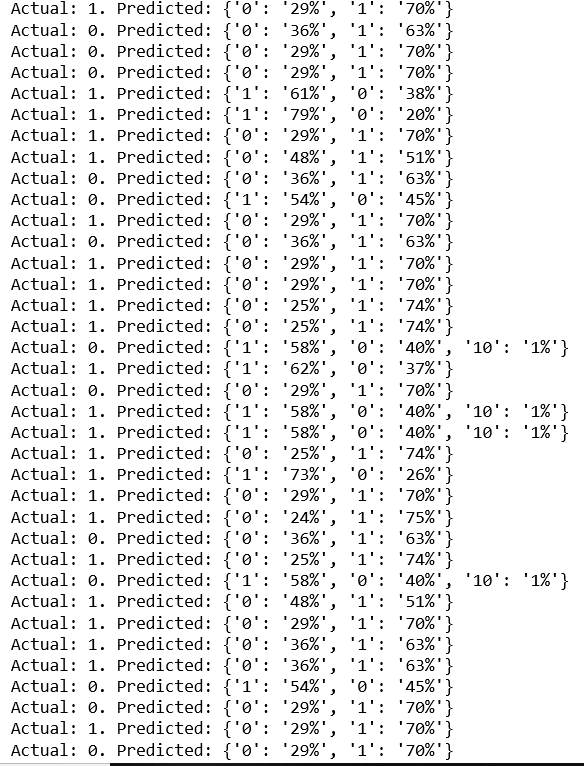
****

In the second approach we first design the questioning mechanism followed by the branching tree. It involves checking for valid numerical values followed by queries.

This is followed by partitioning the dataset into false rows and true rows depending on the query outcome. This leads to the formation of child nodes from the parent nodes. The gini impurity values for these rows are then calculated which further determines how the split will proceed.

Amongst all the queries, the hierarchical order of optimal queries is decided and the tree structure is obtained.





**CONCLUSION**

|  |  |  |
| --- | --- | --- |
| TEAM | ACCURACY SCORE | |
| Logistic regression | CART |
| India | 57.1269 | 55.6976 |
| Pakistan | 54.2874 | 54.3551 |
| Sri Lanka | 52.8412 | 51.2545 |
| Australia | 65.8421 | 62.2587 |
| South Africa | 63.2587 | 58.7946 |
| New Zealand | 66.9846 | 64.1634 |
| West Indies | 50.8842 | 50.2146 |
| England | 64.7663 | 61.9845 |
| Zimbabwe | 72.0671 | 69.7945 |

We conclude that logistic regression achieves a higher accuracy in contrast to CART approach but CART is easier for interpretation and explanation.

\*\*\*\*