Requirements

Overview***:***

There are eight data tables embedded in the SQLite DB. Each table covers some aspect of football match data, such as players, match scores, etc. Details are below:

* Country –The country origin of the soccer league (11 countries)
* League –The name of the soccer league (11 leagues)
* Match – Game data such as opponents, players, in game events, score, etc. (25,979 matches)
* Player – A list of players included in the match data (11,060 players)
* Player Attributes – Player attributes across various seasons for all players in the database (183,978 player attribute records; each record represents a player’s attribute at a given point in time)
* Team – The name of the teams in each league (299 teams)
* Team Attributes – The team style of play and ratings (1,458 team attribute records; each record represents a team’s attributes at a given point in time)
* Stadium – Stadium data for a subset of the matches and teams, includes geolocation data and capacity (164 records for home team stadiums)

I need a data set that is in final form with roughly 138 columns as defined in ‘Final Data Set’ within the Excel file.

# Task 1 (Build the Data Set)

Leverage the Match table as the base table from which other data will be appended. In other words, we are going to analyze this data at the individual match level, and not any deeper (for example, we don’t care about individual events at a certain, such as a red card in the X minute, we just care about total red cards, etc.)

* Create the dependent variable for the analysis called Match\_Results. See metadata spreadsheet for calculation.
* Convert the values in the **season** column to represent the start year only (for example, when the data says 2008/2009 convert that to just say 2008, if it’s 2009/2010 convert the value to 2009).
* Calculate the Home and Away team’s starting lineup average age. See metadata spreadsheet for calculation.

## Task 1A (Parsing XML):

* On the Match table there are six fields (shoton, shotoff, foulcommit, card, red cards (derived from card), corner, and possession) which house relevant match data but in XML format. Need to parse these out into a new table where each record/row is related to the row match ID. See sample data below which is relevant to a Match between Real Madrid and Barcelona on November 21, 2015.
  + Note, I pulled the sample data manually by just reading the XML and figuring out the counts, the numbers below might be slightly off compared to what’s in the final data set (but not a lot) because I “fat fingered” the data or read it incorrectly.

|  |  |
| --- | --- |
| Match ID | 24210 |
| Home Team (Real Madrid) Shots On | 7 |
| Away Team (Barcelona) Shots On | 6 |
| Home Team (Real Madrid) Shots Off | 5 |
| Away Team (Barcelona) Shots Off | 9 |
| Home Team (Real Madrid) Fouls | 13 |
| Away Team (Barcelona) Fouls | 10 |
| Home Team (Real Madrid) Cards | 4 |
| Away Team (Barcelona) Cards | 2 |
| Home Team (Real Madrid) Red Cards | 1 |
| Away Team (Barcelona) Red Cards | 0 |
| Home Team (Real Madrid) Corner | 10 |
| Away Team (Barcelona) Corner | 3 |
| Home Team (Real Madrid) Possession | 48 |
| Away Team (Barcelona) Possession | 52 |

* Append the above newly parsed fields back onto the main data (Match) set as new columns

## Task 1B (Player Attributes):

* Calculate the player attribute average for each team’s starting line-up across the 29 non-goalkeeper related attributes (for example, overall rating, shot\_power, sliding\_tacklet, etc.)
* Append the results to the main data set
* Include the five goalkeeper ratings and append them back to the main data set

## Task 1C (Team Attributes):

* Incorporate the team attributes for the Home and Away team, will need to compare the match date with the team attribute date to get the most current, past, attribute data (that is, you will not look for the team attribute data that is closest in date to the match, you will need to identify the past match info).
* Append the results to the main data set

## Task 1D (Home Team Stadium Attributes):

* Include the Home team stadium name and stadium capacity from the Stadium table to the main data set

# Task 2 (Build the Algorithms)

Leveraging the data set built in task 1 (138 columns of data including the dependent variable [match\_results]) build three supervised learning multi-class (win, loss, draw) classification models. For each model, you will use a hold-out method and split the data 80/20 (train/test). Ensure you set the seed so that I can reproduce your results to validate the prediction outcome.

* **Note, there is no expectation on accuracy, I’m simply asking you to put the “Lego” pieces together so I can continue to build on what you’ve created for subsequent analysis.**

## Task 2A (Random Forest [scikit]):

* Default parameters are fine, please let me know if you recommend any modifications (e.g., 10 trees, auto max\_features, etc.)
* Generate misclassification rate
* Generate confusion matrix
* Generate F1 score

## Task 2B (AdaBoost [scikit]):

* Default parameters are fine, please let me know if you recommend any modifications (e.g., Decision Tree Classifier, number of estimators, etc.)
* Generate misclassification rate
* Generate confusion matrix
* Generate F1 score

## Task 2C (NeuralNetwork [TensorFlow]):

* Default parameters are fine, please let me know if you recommend any modifications
* Generate misclassification rate
* Generate confusion matrix
* Generate F1 score