Assignment Description

This assignment asks you to predict international football results

based on a database of features from previous games.

You are provided with a subset of the data from the following source - https://github.com/davidcamilo0710/QATAR_2022_Prediction

This provides a set of international match results. The subset of data you have has been further cleaned and split into a training and a test data set. Your task is to use the training data to build a machine learning model that can predict the outcome of international games on the test data (from 2022). In addition to the data, you are provided a simple Python script to read the data and provide visualisation of the results of a very simple method.

You will need to build a model that predicts the outcome of each game (home win, draw, away win). You can use the training data along with a suitable evaluation method (e.g. splitting the training data into training and validation sets) to train and validate your model. You will be assessed primarily not so much on the final predictive accuracy which you obtain, but rather on your approach when attempting this task, and on the level of understanding that you demonstrate.

Be creative and ask questions of the data. Think about methods to generate features missing in the test data. Most of all, please reflect upon what you have learned during the term and seek to apply machine learning in a way that is appropriate for this task.

Be sure that all members of the group participate. Your contributions will be peer-assessed.

The assignment submission will take the form of:

1. A PDF file containing a 10-page report of your results on the task.

2. A Jupyter notebook containing the **Python** source code of your approach as well as (brief) in-line documentation.

Data Description

The data comprises three files: match_history.csv, test_set.csv, worldcup_2022.csv.

match_history.csv contains the data required to train your model. It consists of historical match information of all the teams worldwide. The features include:

- date : the date of the match
- home_team : the team playing at home
- away_team : the team playing away
- home_team_continent : the continent where the home team is from
- away_team_continent : the continent where the away team is from
- home_team_fifa_rank : the fifa rank of home team before the match
- away_team_fifa_rank : the fifa rank of away team before the match
- home_team_total_fifa_points : the fifa points of home team before the match
- away_team_total_fifa_points : the fifa points of away team before the match
- home_team_score : score of home team in the match
- away_team_score : score of away team in the match
- tournament : the type of match (i.e. World Cup Qualification)
- city : the city where the match took place
- country : the country where the match took place
- neutral_location : True if the match location is not home for both teams
- shoot_out : whether there is a shoot-out
- home_team_result : Win, Draw or Lose for home team
- home_team_goalkeeper_score : score of the goalkeeper of home team during the match
- away_team_goalkeeper_score : score of the goalkeeper of away team during the match
- home_team_mean_defense_score : score of defence players of home team during the match
- home_team_mean_offense_score : score of offense players of home team during the match
- home_team_mean_midfield_score : score of midfield players of home team during the match
- away_team_mean_defense_score : score of defense players of away team during the match
- away_team_mean_offense_score : score of offense players of away team during the match
- away_team_mean_midfield_score : score of midfiled players of away team during the match

test_set.csv contains the data of all the game that happened in 2022 with only the teamnames and the match result. This should be used as a test set to see how your model performs. This does not have all the features provided in the training files (because you would not have the in-match information before the match) and you should engineer your own features.

worldcup_2022.csv contains the match information during the coming world cup. This is provided for your interest and you can use your methods (maybe with extra data) to predict the Fifa World Cup Final results if you wish. No extra marks will be given.

Submission Format & Structure

The assignment submission will take the form of a zip file containing:

1. A **PDF file** containing a **report** of the task (this should be at most 10 A4 sides in length, including references). This PDF should be prepared using the LATEX files included on the module Moodle page under the ‘Group Assessment’ link, which provide a format similar in style to \preprint" publications such as arXiv.

2. A **Jupyter notebook .ipynb file** containing the **Python source code** of your approach as well as (brief) in-line documentation. The notebook should include an analysis of the performance of your classifier on the data from the test set file. The PDF report should adopt the following structure:

1. **Introduction**
A brief description of your approach to the problem and the results that you have obtained on the training data.

2. **Data Transformation & Exploration**
Any transformations that you apply to the data prior to training. Also, any exploration of the data that you performed such as visualization, feature selection, etc.

3. **Methodology Overview**
Start by describing in broad terms your methodology. Include any background reading you may have done and a step by step description of how you have trained and evaluated your model. Describe any feature engineering that you have applied. If you had attempted different approaches prior to landing on your final methodology, then describe those approaches here.

4. **Model Training & Validation**
This contains a breakdown of how your model was trained and evaluated.

5. **Results**
Here you show the results that you obtain using your model on the training data. If you have multiple variations or approaches, this is where you compare them.

6. **Final Predictions on Test Set**
This is the section where you perform your final predictions on the test set using the model that you have trained in the previous section.

7. **Conclusion**
This is the section where you consider your findings and suggest avenues for future research.

The Notebook should adopt the following structure:

1. **Introduction**
A brief precis of the equivalent section in your report.

2. **Data Import**
This section is how you import the data into the notebook. It should be written in such a way that I can modify it to run on my own machine by simply changing the location of the training data and any additional data sources that you have used.

3. **Data Transformation & Exploration**
Code for the equivalent section in your report, together with in-line documentation of that code.

4. **Methodology Overview**
Code for the equivalent section in your report, together with in-line documentation of that code.

5. **Model Training & Validation**
Code for the equivalent section in your report, together with in-line documentation of that code.

6. **Results**
Code for the equivalent section in your report, together with in-line documentation of that code.

7. **Final Predictions on Test Set**
Code for the equivalent section in your report, together with in-line documentation of that code.

Note:
- Your notebook need only contain brief in-line documentation, while the PDF should contain a more detailed description.
- You will be assessed primarily on the contents of your PDF report. The notebook is required so that we can check that your results are replicable.
- Keep in mind that your notebook should be written in such a way that we can modify the location of the data and then step through your notebook to obtain the same results as you have submitted.