**BENG0095 Project Meeting Notes**

**(Team) Mon 21/11**

*Action points*

* Take a look through data and think of columns that might be relevant
* Play with data and find inconsistencies to clean up
* Understand if data is missing or irrelevant content wise (teams that might not be relevant because not participating etc)
* Think about data in the context of predicting world cup results (is the result of one match dependant on that of the previous etc)
* Look through models we might want to use

**(Team) Thu 24/11**

* Need to think of a way to split training data into training and validation sets
* Problem is multinomial classification (home win, lose, draw)
* We need to understand how to evaluate the data properly (train vs validate vs test)

*match\_history*

* Data from 2004 to 2021, should we give more weight to recent matches
* There is always a home and away team, no need to worry about this
* Neutral location should be interesting, giving a further meaning to who the home team is
* FIFA rank in never zero but before 2011-08-27 all FIFA total points are zero, maybe we can get rid of the total points columns
* Is there any reason to believe teams will perform differently in different tournaments? It seems like the model should already take the differences into account when checking the teams’ continents given the regional nature of the different tournaments
* Country and city are probably irrelevant and will add unnecessary noise to the model
* Shoot\_out relevant?
* Null values appear in all position scores, these seem to be concentrated within team and match so could be a good idea to just take average of the non-null values

*Action points*

* Read through jupyter notebook on classification (Pima Indians Diabetes Dataset)
* Replicate for our dataset
* Don’t think, just code

**(Team) Thu 1/12**

* How can we standardise qualitative data? (Countries, continents, etc)
* Created github repo (<https://github.com/rasensiod/ML_Qatar22>)

*Action points*

* Look over the other repo (<https://github.com/davidcamilo0710/QATAR_2022_Prediction>)
* Have some sort of model working, outputting stuff
* Get comfortable with python, github and jupyter

**(Informal) Wed 7/12**

* Look for mathematical representations or more rigorous explanations of models used
* Try to find some relevant papers we can reference in some small literature review section (using ML for predicting football matches / other sports / classification algos that work best / different methodologies)

**(Team) Thu 8/12**

* Discussed representation of string columns and whether to use dummies or change to number
* Maybe include more ways of representing data

*Action points*

* Include all relevant columns using dummies or change to number
* Find some way of doing hyperparameter optimisation
* (Optimistic case) run models on test data
* Start looking into mathematical background of models (think about report sections: intro, lit review, mathematical background)
* Understand 0.86 performance of LR in prediction 2022 notebook