**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)