AMAT 502 FINAL PROJECT PROPOSAL

Although combining sports and statistics is not a modern phenomenon, Movies like Moneyball are a prime example of popularizing the subject and garnering attention today more than ever. Sports analytics is an emerging field that has gained significant traction in the past decade with the rise of available data and the necessary processing power to derive pertinent insights from it [2] [3]. FIFA 2017/2018 are football simulation video games developed as a part of Electronic Arts' FIFA series. In a sport like football (Soccer), each player adds significant value to the team's success. Hence, it is essential to understand player's and team's skills. In this project, we will utilize the complete data science toolbox on a dataset of players statistics from the FIFA 17 [1] and FIFA 18 [4] game.

In this project, two datasets will be used for comparative analysis. The first dataset contains information regarding over seventeen thousand players and over fifty features for the year 2017. The second dataset contains all the features contained in the previous dataset with an additional twenty-two. Some of the highlighting features contributing to both data sets are Age, Club, Nationality, Ball Control, Speed, Stamina, Interceptions, Shot Power, etc. The 2017 data set contains 17341 samples, while the 2018 one, on the other hand, has 17929 samples. Therefore, preprocessing and data transformation must be applied to ensure that the appropriate features are present to meet the goal of the analysis and redundant, unneeded data is removed from the dataset.

In terms of analysis to be performed on the data, we will be employing **Principle Component Analysis (PCA)** to reduce the dimensionality of the dataset. This method will be applied to the various player attributes included in the dataset. As another analysis, we will be grouping the players by their respective teams (Clubs), a ranking system will be created and assigned to each team. This new ranking feature will be created by incorporating all the necessary and essential features for each player in that team. The project's goal is to answer the following question: Based on the ranking created, can we **predict**, within a reasonable degree of certainty, if a team is going to win or not? Since this prediction task involves a prediction of win or lose, we will be using a classification machine learning model such as **logistic regression**. Furthermore, a **clustering method** will be employed based on the ranking system that outputs clusters by the Club or nationality. Moreover, hypothesis testing will be employed with an alpha value of 0.05 to test a few hypotheses regarding the dataset like age impacting potential, speed of a player affecting the stamina and fitness of the individual, etc. Along with this, we will also be employing some exciting data science mechanisms to dig deep into the data to get some compelling trends and insights about the dataset.

References

[1] Soumitra Agarwal. *Complete FIFA 2017 Player dataset (Global)*. https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global/data?select=FullData.csv. [Online]. 2017.

[2] Luke Bornn, Dan Cervone, and Javier Fernandez. "Soccer analytics: Unravelling the complexity of "the beautiful game"". In: *Significance* 15.3 (2018), pp. 26–29.

[3] Luca Pappalardo et al. "A public data set of spatio-temporal match events in soccer competitions". In: *Scientific data* 6.1 (2019), pp. 1–15.

[4] Aman Srivastava. *FIFA 18 Complete Player Dataset*. https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset. [Online]. 2018.