MSc. in Computing
Practicum Approval Form

Section 1: Student Details

Project Title:	**Examining the use of machine learning to predict the match outcomes in English Premier League**
Student ID:	21260470, 21261256
Student name:	Aneeta Charly, Binit George
Student email	aneeta.vazhappillycharly2@mail.dcu.ie binit.george2@mail.dcu.ie
Chosen major:	Data Analytics
Supervisor	Andrew McCarren
Date of Submission	31-01-2022

Section 2: About your Practicum

Please answer all questions below. Please pay special attention to the word counts in all cases.

What is the topic of your proposed practicum? (100 words)

With our practicum we aim to examine the application of machine learning to predict match outcomes in English Premier League football. Using historic match statistics and player-related data, we aim to use a machine learning model to predict the match outcome in the EPL. The important features which contribute to the match outcome will be identified from the data collected. Consequently, using those features, a highly accurate predictive model will be determined and experimented with as part of our practicum.

Please provide details of the papers you have read on this topic (details of 5 papers expected).

1. Guan, S. and Wang, X., 2021. Optimization analysis of football match prediction model based on neural network. Neural Computing and Applications.
   Rank: Q1, H-index: 80
2. Baboota, Rahul & Kaur, Harleen. (2018). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting. 35. 10.1016/j.ijforecast.2018.01.003.
   Rank: Q1, H-index: 96
3. Constantinou, A., 2018. Dolores: a model that predicts football match outcomes from all over the world. Machine Learning, 108(1), pp.49-75.
   Rank: Q1, H-index:152

4. Joseph, A., Fenton, N. and Neil, M., 2006. Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems, 19(7), pp.544-553.
    Rank: Q1, H-index: 121
5. Razali, Nazim et al. "Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)." (2017).
    IOP Conference Series: Materials Science and Engineering, Volume 226, International Research and Innovation Summit (IRIS2017) 6–7 May 2017, Melaka, Malaysia
    Rank: H-index: 44

How does your proposal relate to existing work on this topic described in these papers? (200 words)

Generally, you find in the literature that the variable selection and features are related to how they influence the match outcomes. There can of course be variables that are obvious and other themes that you may want to engineer as they have been mentioned in the previous literature.

[1] uses three Back Propagation (BP) neural network models to predict and analyse the football league data. The authors have used a combination model combining a gray fuzzy prediction model and a gray extreme machine learning prediction model, which is used to improve the overall prediction accuracy of the model and performance of the system.

[2] uses feature engineering and exploratory analysis to identify the appropriate variables that aid the prediction of a match.

In [3], the author introduced a hybrid model: Dolores that predicted the football match results of one country or team by observing football match data in multiple countries, by using dynamic ratings and a Hybrid Bayesian Network. The author argues that this model can make a good prediction for a match result between two teams even when there is no historical data of participation between the teams.

In [4], the authors have used an expert Bayesian Network model which predicts match outcomes (win, draw or lose). They have also used 4 alternate ML techniques, such as MC4 decision tree, Naïve Bayes learner, Data-Driven Bayesian learner, and K-nearest neighbour, and compared the comparative prediction accuracy of the expert constructed BN network with the other ML models.

[5] uses Bayesian Networks (BN) to predict the outcome of EPL matches. The author claims that the average predictive accuracy across three seasons of EPL is 75.09% using BN which is considered as a high benchmark in terms of the related work using this method.

Similar to the existing work, we aim to identify those features or variables that affect the outcome of a match and evaluate how the selected variables contribute to the accuracy of the prediction of a match outcome.

What are the research questions that you will attempt to answer? (200 words)

Our practicum aims to achieve the following:

- What are the appropriate factors that influence the outcome of a match in the EPL? In particular, does a players' FIFA ranking have any predictive influence on the actual match outcome?

- Identify a machine learning model that can be highly accurate in predicting the match outcome.

How will you explore these questions? (Please address the following points. Note that three or four sentences on each will suffice.)

- What software and programming environment will you use?

Python, Google Colab

- What coding/development will you do?

Use python to scrape the website to collect player line up data for each match.

Further coding/development is yet to be decided.

- What data will be used for your investigations?

- https://www.football-data.co.uk/data.php

This site has historic data on various home and away matches in EPL.

- https://understat.com/

This website has the match line ups historic data of EPL matches.

- https://sofifa.com/

In this website, year-wise player skills ranking, and FIFA ranking is available.

- Is this data currently available, if not, where will it come from?

- Historic match statistics of EPL is available as downloadable csv files on the website(https://www.football-data.co.uk).
- FIFA rankings and match line-ups need to be web scraped from the above-mentioned websites (sofifa.com, understat.com).

- What experiments do you expect to run?

- Examining the data to find out which variables are appropriate to provide a highly accurate predictive model.
- Identify the machine learning models which are efficient and accurate in predicting the match outcome.

- What output do you expect to gather?

Appropriate factors that help to predict EPL match results and a predictive ML model which provides accurate results with less margin of error.

- How will the results be evaluated?

Cross-validate the prediction result to the actual match results to analyse the accuracy of the prediction models and the relevancy of the variables selected.