Soccer Analytics

by

Devendra Deepkiran Adhikari

This thesis has been submitted in partial fulfillment for the degree of Bachelor of Science in Software Development

in the
Faculty of Engineering and Science
Department of Computer Science

December 2022

Declaration of Authorship

This report, Soccer Analytics, is submitted in partial fulfillment of the requirements of Bachelor of Science in Software Development at Munster Technological University Cork. I, Devendra Deepkiran Adhikari, declare that this thesis titled, Soccer Analytics and the work represents substantially the result of my own work except where explicitly indicated in the text. This report may be freely copied and distributed provided the source is explicitly acknowledged. I confirm that:

- This work was done wholly or mainly while in candidature Bachelor of Science in Software Development at Munster Technological University Cork.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Munster Technological University Cork or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this project report is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Devendra Deepkiran Adhikari

Date: 03/10/2022

MUNSTER TECHNOLOGICAL UNIVERSITY CORK

Abstract

Faculty of Engineering and Science
Department of Computer Science

Bachelor of Science

by Devendra Deepkiran Adhikari

Sports Prediction has been around since ancient times. From the time of Gladiators during the Roman empire to this day, people have tried to predict the winner of the game. It is in the human nature. Because of the betting apps, it has grown even more, which is now a multi-billion dollar industry. We can now bet any match outcome not only who will win or lose. Due to this, extensive research have been done using modern technologies like Machine learning (ML), deep learning (DL) and data analysis (DA) to predict any match outcome.

Soccer or football is one of the most popular sports in the world with millions of fans. According to the newspapers nearly a Billion people worldwide tuned the Germany vs. Argentina World Cup final in 2014. A soccer competition (league or tournament) consists of n teams playing against each other in a single or double round-robin schedule. With the availability of better match recording tools, intelligent video analytics technologies and growing interest for soccer in data science and analytics, it is expected that richer datasets will be available in the near future.

In this project, we aim at exploiting ML in the context of soccer analytics. This project will explore the use of DA mainly to identifying which performance attributes of the players will influence the outcome of a given match. We will use different machine learning algorithms like Support Vector Machine (SVM) and neural networks algorithm, Long Short Term Memory Network (LSTM) to predict the result and evaluate the result based on accuracy of the models.

Acknowledgements

I would like to genuinely thank Dr. Larkin Cunningham, lecturer at Munster Technological University in helping me with the research phase of the project. Without his direction in research phase, I would have been lost with all those information available. He pointed me to the right direction where to look for and what to look for...

Contents

List of Figures

List of Tables

Abbreviations

ML	**M**achine **L**earning
AI	**A**rtificial **I**ntelligence
DA	**D**ata **A**nalytics
DS	**D**ata **S**cience
MTU	**M**unster **T**echnological **U**niversity
IoT	**I**nternet **o**f **T**hings
LSTM	**L**ong **S**hort **T**erm **M**emory
SVM	**S**upport **V**ector **M**achine
NFR	**N**on **F**unctional **R**equirements
FR	**F**unctional **R**equirements
DNN	**D**eep **N**eural **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
IAAS	**I**nsfrastucture **A**s **A** **S**ervice

To my parents. Their contribution is my source of motivation…

Chapter 1

Introduction

Sports analytics has already become big. Clubs are investing heavily in learning about their opponent, bettors, betting agencies are investing hundreds of millions in knowing the match outcome. Big clubs are using the new technologies to analyze opponents.They know more about their opponents than the opponent know about themselves [1]. We can see its impact in National Basketball Association (NBA), major basketball league in the US. But the process is complex in case of football because there are infinite number of things that can affect match outcome: Player rating, player emotion, his physical health, ability to play with the team, referee decision, cards, fouls, penalties, free kicks, corners, team formation, substitutions,weather conditions, behavior of crowd and so on. [2] Another reason it is complex is because of less availability of quality football data.

In order to find patterns and solve these complex prediction problems, machine learning comes handy. Machine Learning (ML) is a part of Artificial Intelligence (AI). ML uses big data to make complex predictions by training a model from given dataset. There are different machine learning algorithms that are used to train a model. Once we decide which algorithm will be best fit for the dataset we have, we train the model fitting in the data. The model we created will be tested for the accuracy. One thing we must take care before we train the model is to make the data clean. Raw data can contain bias, noise,characters,unwanted special characters,unwanted columns etc. All these should be carefully removed before we train the model to get a model with high accuracy. If we are satisfied with the model we created, we will deploy it on the production for various use cases.

1.1 Motivation

I have always been fascinated about Soccer. We can agree to the fact that we have tried to predict the match outcome before watching a match by comparing players of opponent team,team formation,if the team is playing in home or away, etc.Watching a game, supporting a team and rooting for that team to win is why we enjoy watching sports.My motivation for this project came from the act of trying to make prediction. In addition to that, I am also interested in data science field. I believed this project would be great fit for me as I can combine both of my interests.

1.2 Contribution

Several factors made contribution for me to choose this project. The primary factor is the curriculum and module structure of Software Development(Hons) degree in MTU. Modules like Data Analytics, Machine Learning and 3rd Year group project module ignited my interest in Analytics domain of Computer Science. These modules helped me gain strong theoretical as well as practical knowledge in data cleaning, analysing data and using machine learning in making prediction from that data. In addition to these modules, I had also done some courses online, outside of college curriculum, majorly focusing on data science that helped me in choosing this project. I had done courses like Introduction to AWS Machine Learning and Python for Data Science and AI which gave me more confidence in this project. I analysed real world dataset to make house price prediction while doing this courses through which gained some experience similar to that of this project, even thought the project was beginner's level. I'll be using most of those tools and techniques learned from these courses and modules in my project.

1.3 Structure of This Document

Chapter 1 gives general introduction of this project. It digs deeper than the abstract but not as deeper as rest of the chapters. Chapter 2 explains different computer science areas this project is based on which are data analytics and machine learning. In addition to that, it provides literature review of those areas with respect to the project incorporated. Chapter 3 defines the problem we are attempting to solve. It also includes functional and non-functional requirements for the project. Chapter 4 highlights the implementation approach we are taking to solve the problem for the project. It includes proposed architecture, risk assessment, implementation methodology and evaluation. Chapter 5 concludes this project with our findings and how it can be made better in the future.

Chapter 2

Background

2.1 Thematic Area within Computer Science

The core project explores data science domain of computer science. Data Science in itself is a broad domain which incorporates programming, statistics, machine learning, predictive analytics, data analytics, business knowledge and more. However this project covers only data analytics and machine learning branch of data science. These will be discussed in more details in rest of the chapter.

2.1.1 Data Analytics

Data Analytics is the process of exploring and analysing business dataset to find unseen trends and patterns and convey valuable insights from the dataset. Data is being collected from everywhere be it be from social media,health tracking apps, embedded systems,Internet of Things(IoT), online shopping apps etc. Even the groceries store collects data of what you shop so that they can make business decision based on data of their customers. All these companies analyze their customer data in order to gain business insights in making fruitful decisions.Analytics process involves data collection, data cleaning, data exploration and analysis and result interpretation. Various tools are used in order to analyse and gain insights from data like python, R, Tableau, Power BI, Apache Spark and more. Python, R, excel and SQL are used to analyse data. Tableau, Power BI and python is used to visualise data into tables, charts, graphs etc. Apache Spark is used in in-memory caching so that the queries are optimized irrespective of data size. Spark is very handy for processing large data sets.

In this project, what we are doing is predictive analytics by using machine learning algorithms to find pattern in data set. Predictive analysis has a wide range of applications in fields like finance, healthcare, sports and law. It uses data, statistical algorithms and ML to make predictions look gaze into future. [3]

2.1.1.1 Data Analysis

Data Analysis is the act of inspecting, determining noises, cleaning and modelling the data to gain information from it.Raw data is of no value but when we analyse the data, we gain valuable insights from it.Raw data is just meaningless numbers. For example; 10 and 40 are just numbers that do not hold any information yet. But after we analyse the data and found out that 10 percent of total income comes from advertisement or if someone says 40 percent of individuals in our survey likes our product, then these hold valuable insights to the business. Data are, therefore, a form of wealth. In modern time, which ever company holds more data, the more valuable the company is because exploiting data results is a competitive advantage against the business's competitors [3]. Data Analysis not only involves cleaning or processing data but also collection of the data. All the tasks we do before using tools to find pattern in data is data analysis. Even though data analysis and data analytics may seem similar, they carry different theoretical definition. Data analysis is a component of data analytics which does not include the tools that we mentioned above to draw outcome or insights from data. However, we see these term being used together a lot also makes sense in a broader spectrum.

2.1.2 Machine Learning

Machine Learning is sub component of Artificial Intelligence (AI). It is a field of study, that trains computer to learn things without being programmed. Machine Learning draws patterns present in data to make prediction and eventually decision, see 2.2. We build machine learning models based on data we have which later can be used to make prediction on a new dataset. Machine Learning process relies on different set of algorithms.These algorithms are the by product of applied mathematics and computer science. These algorithms produce prediction and decision support tools based on data. [3]

ML is becoming more and more popular these days in most of the businesses because of availability of large chunk of raw data, decreasing cost of data storage and increase in computing power. Most commonly used algorithms can be classified into three types supervised,unsupervised and reinforcement machine learning algorithms.Depending upon data we have and the kind of problem we wish to solve, different algorithms are used accordingly. Even though machine learning is extensively used in data analytics, it is not a part of data science because analysts use them as pre-packaged tools without worrying about the internal workings. For most part, their concern is formulating the problems in first place and determining which algorithm best suits for the problem rather than developing an algorithm themselves.[4]

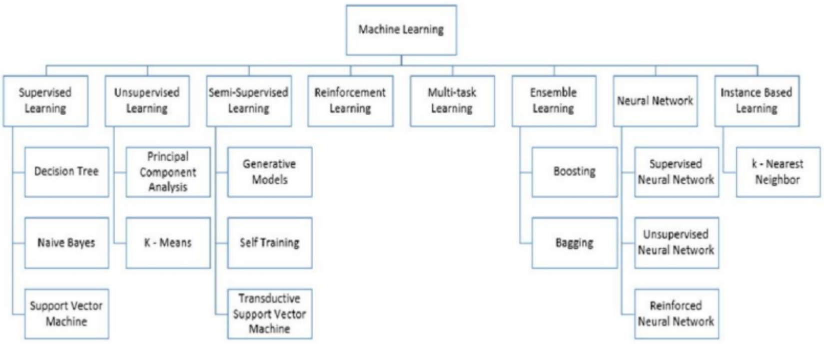

Figure 2.1: Classification of commonly used machine learning algorithms. [5]

2.1.2.1 Supervised Learning

In supervised learning, the algorithm is fed with input data and the algorithm predicts outcome based on past trends or instances. Data is divided into training and testing sets. First model is trained by passing the train set. Training data has input objects and output labels which needs to be predicted. The model finds pattern which can get the output from passed input from the train set. It uses function it learned from train set into test set to measure accuracy.If the data is clean and annotated well, these type of algorithm give high accuracy compared to other algorithms [6] Training data and testing data has to be splitted i.e. one should not overlap with other. Overlap will result in over-fitting of the data while testing. Overfitted models might perform well in test scenarios but might perform poorly while data outside of it's scope is passed. So, to avoid this condition, test and train data should be separated. Some of the examples of supervised learning is image classification in Google photos, price analysis etc. Some popular supervised algorithms are decision trees, support vector machine, Naive Bayes algorithm, K-nearest algorithm etc.[7]

2.1.2.2 Unsupervised Learning

Unlike supervised learning, unlabeled data are passed in unsupervised learning algorithms.Algorithm has to make prediction on its own but there is not correct output value. These algorithms are mostly used in clustering and feature reduction. K-Means Clustering algorithm is widely used for clustering which is an unsupervised algorithm.[7]

2.1.2.3 Reinforcement Learning

Reinforcement learning makes prediction based on train and error on a reward based system. Its main goal is to maximize rewards which in turn mean more accurate prediction. Even though there is reward for accurate prediction, the algorithm has to find it on its own based on series of decision. No hints are given to the algorithm but it must discover which actions yield most reward by trail and error method. This algorithm is useful in such cases where even the programmer does not know all the expected outcomes for example in autonomous vehicles, path finder in a maze game. In this type of learning there is trade off between exploration and exploitation. In order to gain more rewards, the algorithm has to exploit what it already knows to obtain reward, but has to explore to make better action selections in future. [8]

2.2 A Review of Machine Learning Algorithms and Data Analytics in Soccer

2.2.1 Prediction Algorithms

To build models for predictive analytics in this project, we use **supervised machine learning**.

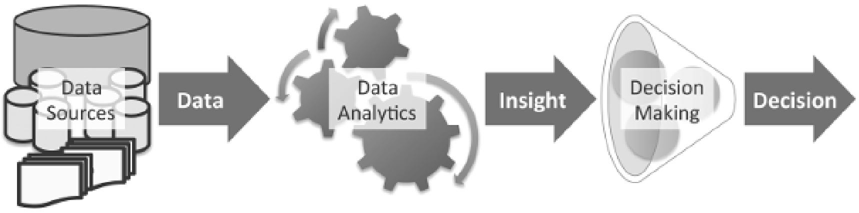

FIGURE 2.2: Predictive data analytics moving from data to insight to decision [9].

2.2.1.1 The Naive Bayes' Model

This model is probability based model. Since, we are dealing with probability of a team winning based on the team's players' features, this model comes handy. We can estimate the likelihood of an event happening based on the likelihood of that event happening on the past. It is given by the following formula,

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)} \tag{2.1}$$

It states that the posterior P[X | Y] is equal to the product of likelihood P[X | Y] and the prior P[Y] divided by the evidence P[X]. When we calculate the likelihood of a action directly from a dataset or using chain rule, we may end up having a probability of zero or an undefined probability because sometime there can be null values or no instances in the dataset where certain action has all features. It is almost impossible to collect a dataset big enough to cover all possible combination of feature values. So, to overcome this flaw, **conditional independence** and **factorization** concept came in practice [9]. If a knowledge of an effect happening is not dependent with the probability of another event happening, then we say those events are independent with one another. However, in real practice full independence is quite rare. A common phenomena in real word is what is called conditional independence. i.e. two or more events may be independent if we know a third event has happened. The Naive Bayes Model is based on the conditional independence. We call it naive, because the assumption of conditional independence between the features given the target feature is merely based on assumption weather or not it is correct. But despite the assumption, the approach seem to do surprisingly well across large number of domains[9]. In case of missing feature values, the model simply drops the conditional probabilities. It may have negative effect on the accuracy of posterior probabilities but may not directly impact the prediction errors. Another advantage of this model is simple to train and training is less as compared to other models.

2.2.1.2 Support Vector Machine

SVM is an error based learning model. In this type of model, we choose a parameter from a set of parameters that minimizes the total error across the predictions made by the model with respect to the training instances. The main goal of SVM is to find the decision boundary that leads to maximum margin and best separates the levels of target feature. Margin is the nearest distance of training instance to decision boundary. The instances in the training dataset that form the margin extents are called **support vectors**. These vectors define the boundaries hence, they are crucial in the dataset where as all other points are irrelevant for SVM. SVM requires a dot product calculation between each support vectors which is costly for computation as it is repeated multiple times while training. So, in order to make it less costly, **kernel trick** is used. Different kernel functions which give same results as of dot product as used. In general, Gaussian radial basis kernel is used which we won't be discussing in this paper.

2.2.1.3 Deep Learning Model

Deep learning is a subset of machine learning which focuses on neural networks and the algorithms associated with it.Deep learning is basically an attempt by the computer to depict human mind through algorithms[10].It consists of neural networks which attempt to simulate the human brain. Some of the real life examples of deep learning are image recognition, facial recognition, chatbots, autonomous cars etc. Artificial neural networks have input layer with several nodes, an output layer with single node and one or more hidden layer connecting input layer with output layer or other hidden layers, see figure 2.3. Deep learning has several hundreds of hidden layers which is called deep neural networks (DNNs), however using huge amount of hidden layer may cause overfitting. Neural networks is just an activation function with many input layers and one output layer.Activation layer is passed as one of the hyper-parameter while training.It decides weather the input from a particular neuron is important or not for making prediction using mathematical operations. Most commonly used activation layer is ReLU activation (Rectified Linear Unit) and Sigmoid activation.

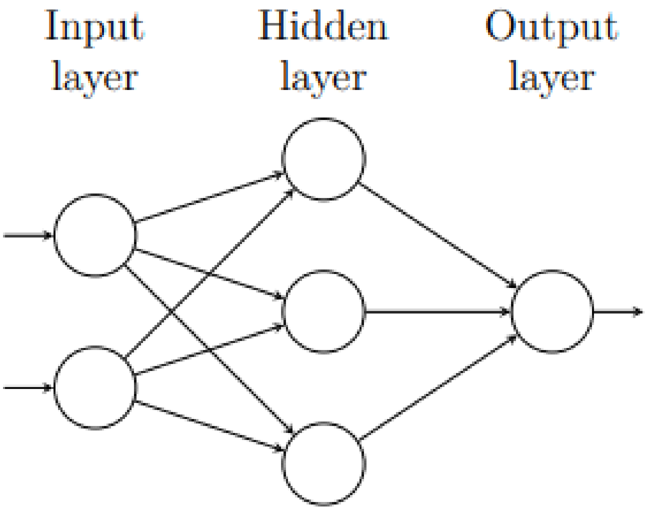

Figure 2.3: Neural Network with two input nodes, three hidden nodes and one output [11]

Long Short-term Memory(LSTM) Model is a type of RNN deep learning model that remembers values either for long or short duration of time. If the model finds a important feature in the beginning hidden layer, it carries it along many hidden layers, which acts as it is remembering the feature for a long period of time. This model is useful when information in one location is relevant in a more distant location. This model uses sigmoid activation function and tanh activation function as activation hyper-parameter

[10]. The important feature of LSTM is it has ability to control the weight of memorized input before activating certain neuron.

2.3 System Architectures for ML projects

2.3.1 Standalone Machine

In this architecture, entire ML process which includes data collection, data processing, model training and model serving is done in a single machine.External devices like sensors, trackers may be used for data collection.This architecture is best suited for students or those who are just introduced to ML. It is suitable for small projects that does not require big and complex data. It is cost effective but may sometimes consume a lot of processing and training time if data size is big enough for the machine.

2.3.2 Distributed Systems

In this architecture, the ML process is distributed across multiple machines. It is suitable while handling large data volume. Different approaches can be consider to build distributed system architecture for machine learning like cluster of computers where each machine is responsible for specific part of the the process like data processing, model training or model serving. Another approach is to use cloud based platforms like Amazon SageMaker, Google cloud ML. Using cloud based platforms is an efficient way to build the models because the ML engineer does not need to worry about the resources required and the complexity involved to distribute the tasks for data processing,model training and deploying [12]. It is handled by the cloud service provider. Containerization is another approach to implement distributed system architecture in ML projects [12]. Technologies like Docker can be used to make a package of ML application and its dependencies into self-contained unit which can be easily deployed and run on any machine. There is an advantage of easy scalability through additional containers if needed.

This process is more efficient for projects that require more processing power however, it is not as cost effective as standalone machine.

2.3.3 MicroServices

This architecture is useful in ML projects which have many moving parts and multiple teams working on the project. It allows the teams to work on their individual ML model or components of the system by breaking the large system into smaller modular components or services. This make it easier to add or update features. It allows each team to work on their service individually which allows faster development and deployment.The main use of micro-service architecture in ML is to provide Ml models outside of cloud platforms. A microservice can be made standalone and locally runnable, which is most useful in IOT devices [13]. However, implementing this architecture is complex and time consuming. It requires upfront planning and designing the architecture and needs robust APIs for communication between different services.

2.4 Related Works

Excessive studies and research have been done in order to predict game outcomes not only in just soccer, in almost all high crowd focus games like basketball, baseball and cricket. Because of use of analytics, small clubs with minimum budget are coming to limelight. **Moneyball**, a book which entails the story of such a club, *Oakland Athletics* baseball team and their team's analytical approach to assemble a competitive team with a limited budget. Similarly some research have been done in soccer as well to predict match outcome. One of the studies, that is similar to this project in some ways is the one done by [14]. The authors organized a **2017 Soccer Prediction Challenge** , where they provided a simple database, *Open International Soccer Database*, which contained goals scored by each team, teams involved, leagues, season and date on which the match was played. The challenge was to use that dataset to make prediction of 206 future match outcome. Different teams participated for the challenge and used different machine learning algorithms like Poisson model, Bayesian hierarchical model and some used variety of statistical and machine learning methods to predict. But the prediction were satisfactory with the highest being only **51.94%** accurate which was achieved by the authors of the article. The weakness of this finding can be blamed to the dataset they were using to make prediction. As they also mentioned, because of lack of proper dataset, they created a simple database and used that to find how accurate the prediction will be, keeping in mind all the possibilities, win, lose or draw. They were aware of the drawback of the data, i.e. it lacked more outcome-relevant information,such as fouls committed, cards, corners, players attributes, team attributes etc.

[15] used the match results based on player characteristics like age, weight etc and skills like ball control, dribbling, crossing etc to predict goal difference between home and away teams and to exploit the finding to construct a statistical betting strategy. It is a study done by department of Statistics and econometrics, University of Erlangen-Nürnberg, Germany. They used data of top leagues from England, Spain, Germany and Italy from

season 006/007 to 017/018. The authors used four different ML models: random forest regression, Boosting, SVM and linear regression and calculated the average prediction result for each model. Among them, random forest achieves the best result with the accuracy score of 81.26%. Even though this paper focuses on players' attributes the main goal is to find the best strategy for betting purpose and to maximize financial gain rather than finding the match outcome.

Another research similar this project is done in [16]. The goal of this paper was to build a predictive model that can be used to analyze the best player's combination among the squad based on player's attributes for the specific match. This was achieved by a model that has satisfactory accuracy. Using that model, they tried different player combination to find out the optimal player combination.In the paper, dataset was gained by crawling multiple sources. The paper used 139 datapoints at each step, which included 5 main attributes of players(shooting, passing, defensive skills, overall rating, pace control), home/away factor, result of previous match and obtained points percentage (OPP) from last five games. The paper used *many to one* architecture of LSTM model to train on seasons from 2011 to 2016 (except matches form 2013/2014 season, these matches were later used to validate the models). because the expected outcome is the result of the match where *many to many* architecture would not be of the best fit. The paper used different models to evaluate the outcome - classification model. regression model and dense model (the LSTM model with only Dense layers). Optimal result is given by the LSTM models that used Stochastic gradient descent (SGD) optimizer with accuracy around 54.5 - 55%.

The work done in this paper can be used by clubs in selecting the starting 11 of the the player against specific opponents and increase the probability of wining. It can also be used in betting odds if betting odds is included as a feature vector in the training model.

2.5 Application of Soccer Analytics

Soccer analytics is mainly used by clubs, prediction websites like **FiveThirtyEight** and for betting purposes. Major sports clubs have been hiring data scientists to evaluate their data and make meaningful predictions. The sport analytics industry is expected to reach $6.34 billion by 2030, according to a 2021 report from Research and Markets [17]. Many startup companies like Opta, Catapult are in sports analytics and had made analytics their main source of revenue. Some companies like Hudl, Stats Perform are making technologies and wearable devices that can capture more data from the players and can use those data to help clubs/coaches and athletes to review game, to pinpoint improvement areas and to develop tactics. We can not ignore betting where there is prediction involved. Bet365, Sportradar,Paddypower are leading betting companies which are also using analytics to calculate betting odds. There companies were among the first to apply data analytics in the practice to ensure the odd were not exploited against themselves. However, analytics can also be used by individual gamblers. Instead of solely believing in luck, some gamblers have made huge profit by using sport analytics and predictive analytics in gambling against the odds.

Chapter 3

Problem - Match outcome prediction

3.1 Problem Definition

This project aims at identifying which performance attribute of the players will impact the match outcome by using different machine learning algorithms and deep learning models like LSTM model to train a model and use that model to predict outcome given different players attributes. The project aims at solving following problems:

- Are player attributes helpful in making match predictions?
- Which model is optimal among the models selected for training and prediction?

3.2 Objectives

The main objective of this project is to make use of data analytics and create a machine learning model from already available dataset, that helps to identify which attributes of players have much impact in the game outcome using deep learning and machine learning models. The main objective of the thesis is to have a decent prediction accuracy. Match outcome not only depends on player's attributes. There are numerous factors that may impact the match outcome. However, in this project, our interest is to see how the ML models perform by just passing in players' attributes.

3.3 Functional Requirements

Functional requirements are the features that developers implement on the system as per the business needs. These features are requested to the developers. It describes how the system behave under certain conditions. Some functional requirements for the project are:

- To display the accuracy of the model once input details are passed.
- Users should be able to input players attributes themselves through console as a excel file of two different teams.
- Exploratory Analysis: Since, this thesis not only include ML but also DA, it is expected that we provide information of dataset, training time and prediction accuracy diagrammatically for each model. It should include plots that help in data exploration like mean, median, normal distribution etc.

3.4 Non-Functional Requirements

Like proper software systems, ML models also have some qualities which we define as non-functional requirements (NFR). However, in ML solutions, some of the knowledge of regular software may not apply because the way ML solutions are made, designed and run differs from software systems not including ML [18]. In general, NFR relates to how the system should act without the developer being told. These are expected qualities of the system like the system should be secure, it should maintain privacy, should run smoothly etc.

Some NFR, that relates to this projects are described in brief below.[18]

3.4.1 Run time

Run time is another important NFR that should be considered while making ML models. The model should not take forever to make prediction. However, it might take some time during training phase based on the size of training data. To make the model's prediction time low, we are going to use small number input features and reduce the model complexity.

3.4.2 Accuracy of the model

Accuracy of a model is calculated by first training the data and using another set of data to make the prediction of the model that is trained. Since, the chance of winning the game is 33% only (because there are 3 possibilities: win, draw or lose), we aim for the model in this thesis to be able to make prediction more than 50%. Since, we are making prediction with out considering factors like players' psychological state, substitutions, weather conditions, intensity of the match etc, we can not expect high accuracy from the model.

3.4.3 Transparency

It should be transparent what was done to achieve the results. Proper documentation and explanation are the key for transparent results. Transparency builds more trust in the model we prepare.

3.4.4 Reliable

Model should be robust to new input data. It should know what it does not know i.e. it should know the uncertainty.[19] It should be in the optimal range of over-fitting and under-fitting.

Chapter 4

Implementation Approach

This chapter includes my plan to achieve what we discussed in previous chapters which I will be doing in implementation phase of the project.

4.1 Architecture

For training a machine learning models, different architecture can be used based on several factors, like volume and complexity of data and scalability and reliability of the system, see 2.3.

This project will be coded extensively using Python programming language. Python is a powerful, flexible, simple and human readable programming language. It is dynamically-typed language which means the type of the variable is determined during runtime rather than explicitly declaring while writing code by programmer. The main reason why Python is used for most of Data Science projects is because of availability of various libraries that are extensively useful for the the projects. Libraries like Pandas, Numpy, Matplotlib and Scikit-learn are useful for this project. Pandas and numpy are used for data processing, Matplotlib is used for data visualization and Scikit-learn is used for ML side of the project. The code will be iterative meaning each model is trained multiple times to find optimal parameters useful for finding right accuracy of the model.

To do the data processing and ML model training and testing, I am thinking of using cloud based architecture, Google Colab. Google Colab is a cloud based platform which combines client-server architecture of Jupyter notebooks, a web-based interactive architecture for writing and running python code, with Infrastructure-as-a-service model of cloud computing. All the resources needed for small projects are provided for free until certain limits. Programmers do not have to worry about these resources and install libraries we discussed earlier in their personal local machine. It makes coding experience smooth and interactive. Since, the project we are doing does not require extensive resources we could be using standalone machine architecture. However, I chose the cloud based distributed system architecture for simplicity and efficiency. Microservice architecture is complex to implement and out of the scope of this project.

4.2 Risk Assessment

Table 4.1: Initial risk matrix

Frequency/ Consequence	**1-Rare**	**2-Remote**	**3-Occasional**	**4-Probable**	**5-Frequent**
4-Fatal	🟨	🟥	🟥	🟥	🟥
3-Critical	🟩	🟨	🟨	🟥	🟥
2-Major	🟩	🟩	🟨	🟨	🟥
1-Minor	🟩	🟩	🟩	🟨	🟨

4.2.1 Risk 1: Data Quality

Frequency = Probable
Consequence = Fatal
Quality of data that we use to train the ML models determines how much the model can extract insights from the raw data. If the data is biased, flawed, has missing bits, unreliable and the quality is poor overall, the ML model will not be able to make predictions with high accuracy. Hence, most of the time in DS is spent preparing a data with high quality. The only way to mitigate this risk is first, by finding a source of data that is trustworthy and then cleaning and carefully curating the data. Removing null bits, duplicate entries and visualizing the data helps in improving the quality of data.

4.2.2 Risk 2: Overfitting

Frequency = Occasional
Consequence = Critical
Overfitting occurs when a model is trained too closely to training data. Overfitted model is poorly generalized to new and unseen data but has high accuracy for training data. Hence a model with overfitted data performs poorly in real world. To mitigate this risk, data should be sufficient for training and techniques like cross-validation and regularization should be used to make the data less complex and more generalized.

4.2.3 Risk 3: Underfitting

Frequency = Remote
Consequence = Critical
Underfitting is opposite to overfitting. When model is not complex enough to capture underlying patterns of the data, underfitting happens which leads to poor performance in both real and training data. To mitigate underfitting, more data should be used and data should be transformed in such a way model can capture the underlying pattern. Optimal hyper parameters should be selected for the model by iterating the training process with different parameters.

4.2.4 Risk 5: Time Management

Frequency = Probable
Consequence = Critical
Time management would be a high risk factor for implementing the project. A lot has to be done in short time frame given strict deadline and other modules as well. Training a model may take a lot of time, making prediction form it as well because the process will be iterative. To mitigate this risk, I need a proper plan, see 4.4 and continuity from the beginning of the implementation phase.

4.3 Methodology

The methodology to train for every ML model is straight forward and most of the projects follow these steps.

- Formulating a question, which is the main theme of this project, predicting match outcome using players' attributes.

- Gathering data: In order to make the Ml models more accurate, more generalized and not overfitted or underfitted, I need data that is clean, most importantly not biased. The whole project depends on this bit, without this we can not move forward. Hence, I plan to give time it deserves.

- Cleaning data: The data in real world is raw. We need to clean data and make it more fit for out question. Majority of time is consumed in cleaning the raw data and making it suitable for the models we are training. It is important we do not miss important bits while cleaning and we include as less unnecessary data as possible. As Abraham Lincoln said "*If I only had one hour to chop down a tree, I* *would spend the first 45 minutes sharpening my axe*", I plan to do so in the project because if the axe is sharp, cutting down the tree should not take that much time.

- Explore and Visualize: Even-though data visualization does not have direct impact on the model training , it gives programmer a good understanding of the data s/he is working with. it is needed to find hidden patterns in the data so that programmer can choose the right model. This bit is called data analysis.
- Train Algorithm: This is where the actual training of the ML algorithm happens. As we have already have a list of algorithms we will be working with, see 2.2, we just need to tune in the algorithms, train models using those algorithms and we are ready for evaluation.
- Evaluation: Once the models are trained, we need to evaluate them based on testing data. Training and evaluation can be a iterative process if model does not make prediction to our satisfaction, we need to retain the model with different hyper-parameters and again re-evaluate. After we are satisfied with the model, we can deploy the model in out application or develop a REST API to send in real world test data as a request and receive prediction as a response.

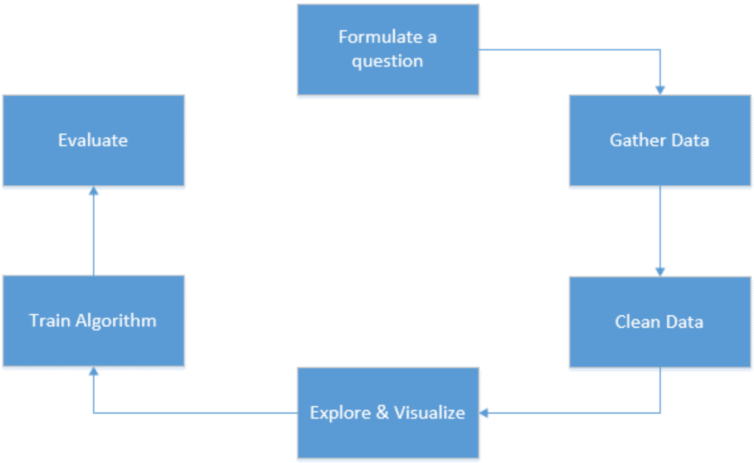

FIGURE 4.1: ML Model Building Steps [20]

4.4 Implementation Plan Schedule

We plan to make a schedule based on the methodology discussed above. The schedule will be 12 weeks long based on the final semester calendar weeks. I will be following waterfall model as it is most suitable for ML projects, even-though it has its own drawbacks.

- **Week 1 and 2** Look for good dataset that has good metric of both teams' players' features, match outcomes for matches they played. Features should include but is not limited to dribbling skills, header skills, overall rating, pass accuracy, shots on target, shots, age, number of goals and ball control.

- **Week 3** Data Cleaning. As said earlier, quality of data is most important factor that determines how our model perform. Cleaning data includes removing null values, normalization of data, data mapping, features selection,finding and removing outliers.

- **Week 4** Data Visualization. Visualizing dataset so that we get to know hidden anomaly in the dataset. It helps in finding the outliers, communicating data patterns and in analyzing the dataset statistically. We can use bar plots, pie chart, graphs, scatter plots and box diagrams.

- **Week 5 and 6** Further cleaning and visualization of data. After first visualization, we may know what anomalies present in the data. Hence, we may need to do further cleaning of data set.

- **Week 7, 8 and 9** Train models and evaluation. We have finalized candidate algorithms which we will be using to train the models. Training all those models may take a lot of time based on the size of data and more research may need to be dine on how to implement those models code-wise. This phase not only consists of training but also selecting and tuning hyper parameters best suited for the algorithms and using cross-validation method to optimize the model's performance. Once trained, we will evaluate the model's performance using test dataset and calculate accuracy of the model. If we are not satisfied with the model's performance, we will retrain the model.

- **Week 10** It depends on how our model performs in past weeks. If we are not satisfied with the model, we need to find if model is being overfitted, underfitted or what is the issue, solve it and retrain and re-evaluate the model. I hope one week will be enough for this. In case we are satisfied with the model, we will dive into deployment options.

- **Week 11** Deploy the model. How exactly will I be deploying is still to be decided. My plan was to use the console to print out match prediction. If I get more time, I may create a REST API endpoint where we can request with players' attributes of both teams and model responds with win prediction. Also, we need to monitor how the model performs in real world environment after it is deployed.

- **Week 12** Finalizing the bit and pieces. Revisiting the past weeks works and complete any unfinished works like documentation, code formatting.

4.5 Evaluation

Evaluation of ML models is done by using the model to predict test data set. Dataset is divided into train set and test set. Model is trained by using train set and tested using test set. The prediction made by the model is compared with the actual result of test set. That's how we know how accurate the model is. To find which algorithms gives the best accuracy, we compare the mean accuracy of each model. The one which has the highest accuracy among all is the model we want. If the prediction accuracy is somewhat similar, we measure the evaluation time of the models and which ever model has less evaluation time becomes the optimal model.

Chapter 5

Conclusions and Future Work

5.1 Discussion

The research phase had a slow start. Not knowing what to do, how to carry the project forward, I spent significant amount of time researching from unauthorised sources. Due to the reason, I had major problem of time management during this phase. As the semester went on, more assignments and project keep coming on and I could not give the time I wanted for this phase. However, with the right guidance of DR. Larkin, supervisor of this project and weekly meeting with him pushed me to do more each week.

Gaining a understanding of all the algorithms and how they are implemented was another problem that I encountered during this phase. Deciding which would be a right algorithm to make prediction with such a huge amount of input data was complex. However, with a short chat with ML lecturer Dr. Christian Beder, extensive research of related past works by other researchers gave me some direction on this project.

Finally, the major achievement of this phase is that we know what algorithms would make better prediction for the problem like ours.I gained in depth knowledge of these algorithm which will aid in the implementation phase.

5.2 Conclusion

When I started this project 12 weeks ago, I only had a little knowledge on what is being done in the world to predict the very game we watch everyday. Many organizations, researchers and students like me have done a lot in predicting sports, not only soccer. I gained a new way of thinking. Whenever I watch a new match, I have gained a different angle to view the game, more in statistical way, which I have been enjoying. I gained experience in doing research, which I know will be helpful in different walks of life. I got a chance to have a peek in a researcher's life through this project. While I enjoyed some of it, there were times where I felt overwhelmed as well. But overall, I enjoyed the research phase. I am coming out of this phase with a lot of knowledge in the domain I am interested in.

5.3 Future Work

If I had more time to work on this project, one thing I would do is integrate the model to an App or website, where people, sports betters and soccer enthusiasts could try and make prediction about two teams. They would be able to select teams and the app would automatically feed in the player's attributes from a database and display match prediction.

Another thing I would like to do is make the model incorporating bit that we did not cover in this project. We only used players attributes to make prediction. There are numerous other attributes that would impact match outcome. If I had time, I would try to incorporate as many attributes as I could and make the prediction more realistic.

Bibliography

[1] G. Fialho, A. Manhães, and J. P. Teixeira, "Predicting sports results with artificial intelligence – a proposal framework for soccer games," *Procedia Computer Science*, vol. 164, pp. 131–136, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050919322033

[2] W. Dubitzky, P. Lopes, J. Davis, and D. Berrar, "The open international soccer database for machine learning - machine learning," Jul 2018. [Online]. Available: https://link.springer.com/article/10.1007/s10994-018-5726-0

[3] S. Sedkaoui, *Data Analytics and big data.* John Wiley and Sons, Incorporated, 2018. [Online]. Available: https://ebookcentral.proquest.com/lib/cit-ebooks/detail.action?docID=5401178

[4] F. Cady, *Data science: The executive summary: A technical book for non-technical professionals.* Wiley, 2021.

[5] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.

[6] K. Ramasubramanian and J. Moolayil, *Applied Supervised Learning with R: Use Machine Learning Libraries of R to Build Models That Solve Business Problems and Predict Future Trends.* Birmingham: Packt Publishing, Limited, 2019.

[7] J. Bell, *Machine Learning: Hands-On for Developers and Technical Professionals.* Somerset: John Wiley and Sons, Incorporated, 2014;2015;.

[8] J. Hearty, *Advanced machine learning with python: solve challenging data science problems by mastering cutting-edge machine learning techniques in Python*, 1st ed. Birmingham: Packt Publishing Ltd, 2016.

[9] J. D. Kelleher, A. D'Arcy, and Brian, *Fundamentals of machine learning for Predictive Data Analytics: Algorithms, worked examples, and case studies.* MIT Press, 2020. [Online]. Available: https://ebookcentral.proquest.com/lib/cit-ebooks/detail.action?docID=6383434

[10] O. Campesato, *Artificial Intelligence, Machine Learning, and Deep Learning.* Mercury Learning &; Information, 2020.

[11] D. Pettersson and R. Nyquist, “Football match prediction using deep learning,” *CHALMERS UNIVERSITY OF TECHNOLOGY Gothenburg, Sweden*, pp. 23–29, 2017.

[12] H. Wang, D. Niu, and B. Li, “Distributed machine learning with a serverless architecture,” in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019, pp. 1288–1296.

[13] M.-O. Pahl and M. Loipfinger, “Machine learning as a reusable microservice,” in *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, 2018, pp. 1–7.

[14] W. Dubitzky, P. Lopes, J. Davis, and D. Berrar, “The open international soccer database for machine learning,” *Machine learning*, vol. 108, no. 1, pp. 9–28, 2019.

[15] J. Stübinger, B. Mangold, and J. Knoll, “Machine learning in football betting: Prediction of match results based on player characteristics,” *Applied Sciences*, vol. 10, no. 1, p. 46, 2019.

[16] N. Danisik, P. Lacko, and M. Farkas, “Football match prediction using players attributes,” in *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 2018, pp. 201–206.

[17] A. Schroer, “How sports analytics are used today, by teams and fans.” [Online]. Available: https://builtin.com/big-data/big-data-companies-sports

[18] J. Horkoff, “Non-functional requirements for machine learning: Challenges and new directions,” in *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 2019, pp. 386–391.

[19] D. Tran, J. Liu, M. W. Dusenberry, D. Phan, M. Collier, J. Ren, K. Han, Z. Wang, Z. Mariet, H. Hu *et al.*, “Plex: Towards reliability using pretrained large model extensions,” *arXiv preprint arXiv:2207.07411*, 2022.

[20] S. Pande, D. Torres, and K. Abhishek, “Building machine learning models to solve practical problems,” Apr 2022. [Online]. Available: https://www.red-gate.com/simple-talk/development/data-science-development/building-machine-learning-models-to-solve-practical-problems/