Predicting the Winner of a Tennis Match Using Machine Learning Techniques

MSc Research Project
Data Analytics

Akshaya Sekar
Student ID: x18138977

School of Computing
National College of Ireland

Supervisor: Dr Dondio Pierpaolo

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Akshaya Sekar
Student ID:	x18138977
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Supervisor:	Dr Dondio Pierpaolo
Submission Due Date:	20/12/2018
Project Title:	Predicting the Winner of a Tennis Match Using Machine Learning Techniques
Word Count:	5620
Page Count:	LastPage

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th December 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting the Winner of a Tennis Match Using Machine Learning Techniques

Akshaya Sekar
x18138977

Abstract

Winning is the primary goal of any sport. Predicting the winner of the match in advance has gained a lot of attention by sports organizations and potential bidders as it involves a lot of time and money invested in it. Now-a-days, sports organizations realize the value of data and the science in the data which can be used as an advantage to players, coaches also the potential bidders using machine learning techniques. Tennis is a challenging and unpredictable sport, yet the most exciting sport which is enjoyed by fans from all over the world. Machine learning techniques have helped in predicting the outcomes of tennis matches using various attributes. Removing the irrelevant attribute plays a very important role in getting high accuracy hence in this research we have used the Principle component analysis (PCA) for dimensioanlity reduction and the machine learning classifiers such as SVM, Naive Bayes, Logistic regression and Random forest are used to forecast the winner of the match. The hyper-parameter tuning is performed to highlight the most significant parameters which help in increasing the accuracy of the models used. These models are finally evaluated in terms of accuracy, F1 score, Cohens kappa and AUC. The results show that after hyper parameter tuning Random Forest with F1 score has outperformed all other models with 78% accuracy.

—

1 Introduction

1.1 Background

Tennis is very popular sport which is enjoyed and worshiped by fans from all over the world. Tennis has four major tournaments known as the grand slam tournaments namely the Wimbledon, Australian Open, US Open and French Open Somboonphokkaphan et al. (2008). It is usually played by players on three different types of surfaces (Clay, Hard, Grass) Gorgi et al. (2019). Tennis is an extremely unpredictable sport which is played by professional tennis players from diverse backgrounds and different styles Cornman et al. (n.d.). Each player has a unique style and technique which makes the game even more interesting and challenging to predict the winner.

Today, machine learning is used in many sports such as soccer, cricket, baseball, tennis etc Haghighat et al. (2013). As we know data is everywhere and tennis is defined by data, and machine learning techniques are already making waves in the field of tennis not only for professional players but also for coaches, fans and potential bidders. Online sports betting involves huge amount of money and time involved by potential bidders which can be saved since statistical analysis has helped to remodel the game of tennis by diving deep in to the insights of the game and predicting the results with great accuracy Easton and Uylangco (2010). This has not just increased the efficiency of the betting markets but also helped players and coaches to get better understanding about the game.

1.2 Motivation

Tennis scoring system is in a hierarchical manner. A singles tennis match which occurs between two opponents is composed of individual points, games and sets. A set can be best of 3 sets or best of 5 sets depending upon the tournament. The sets are in turn composed into games, which are composed of individual points Barnett and Clarke (2005). Tennis players put in years of practise, hard work, time and effort into the game but still in certain tournaments, they are not able to play up to their full potential Sipko and Knottenbelt (2015). This is mainly because of the lack of understanding of the details in the game and the areas which need improvement must be focused better for superior results Cornman et al. (n.d.). One of the most important factors to win a game is the right technique and good anticipation of the next shot. With the right technique, there is better accuracy of strokes, more power to shots and high level of consistency in the game Barnett and Clarke (2005). It is mandatory for the coaches to understand the strength and weakness of each player and model them accordingly. Gorgi et al. (2019)

This research paper is completely focused on various attributes related to tennis serve which contribute in predicting the winner of the match. A serve is a big stroke in tennis. It can "make or break a game". The main objective of this project is to predict the winner of the tennis match using individual player statistics and with the help of various parameters of tennis serve and the individual set score of each and every player, the winner of each match is predicted. The point-by-point analysis gives an insight and better understanding of the game. Analysing the various metrics in the game of tennis has helped to answer many unanswered questions Sipko and Knottenbelt (2015). Previous research has focused on other factors such as court type, players technique, shot by shot analysis etc. This research has used completely new set of attributes focusing only on the tennis serve and compared the average and maximum betting odds to predict the winner of each match using machine learning techniques and hyper-parameter tuning.

1.3 Project specification

Research Question: "*How can machine learning algorithms be employed in predicting the winner of a tennis match based on player statistics?*"

There are different stakeholders who are benefited from this research. Firstly, it helps professional tennis players and coaches to understand their strength and weakness in the game and it helps to anticipate the next shot. Secondly, the results give the potential bidders a clarity on who to bet on, as the money invested in sports betting is too high, it is important to give good accurate results.

In section 2 the Related work is discussed. In section 3, the Methodology used, process flow diagram, deployment and steps in architecture are seen. Followed by section 4 and 5 with Design specification and Implementation of models respectively. In section 7, the Evaluation and result is discussed and the last section has the conclusion and future work followed by references.

2 Related Work

2.1 Introduction

Even though sports prediction and online sports gambling is a vast topic with extensive research work done previously, there is always space for some more research with much more efficient models and better accuracy as the money and time involved by investors in gambling is too high [Philpott et al. (2004)].

The literature work of this paper is discussed below. Each paper used here is carefully analyzed to find the attributes used, the machine learning techniques applied and the limitations and drawbacks from previous work.

2.2 Machine learning techniques for Tennis

Using Machine learning techniques to predict the winner of a tennis match is a subject in various research papers. Cornman et al. (n.d.) in his paper has used various top performing machine learning algorithms such as SVM, Random forest, Logistic regression and neural network to make sure that it gives accurate and reliable results. The two main goals achieved in this paper is to forecast the winner of the match and based on the results obtained see the probability of betting odds. Here the tennis match statistics data set was merged with the betting data set. The features are selected based only on the domain knowledge of the author, there is no clear explanation why the features were selected. The results show that SVM has outperformed all other models in predicting the winner with 69.6% accuracy and 3.3% profit per match.

Similarly, In Easton and Uylangco (2010) point by point data of the 2007 Australian open tennis match was collected and analysed to find the probability if a player will win the match as implied by betting odds. This paper has collected data set of 49 matches of 2007 Australian open and the betting data before and after each point on a continuous basis. It was identified that there was a great correlation between the model and the betting market. This project lacks the information regarding the data set collection and the feature extraction. By performing simple statistical analysis the author implies the importance of serve and service points in tennis as it might help to increase the winning probability. The results show that the efficiency of betting market is good in both men's and women's tennis.

Learning (2017)s has used two different datasets and combined both the historical data and real-time tennis match dataset to predict the winner using machine learning models such as SVM with Liner, RBF and polynomial kernals, Naive Bayes and logistic regression. The feature selection was done with Principle component analysis and recursive feature elimination. The project lacks the accuracy with the extracted features due to high bias. It is seen that the TTL(Total points won) is the most contributed feature in forecasting the winner of the tennis match.

Unlike other research papers, here Sipko and Knottenbelt (2015) has predicted the winner of the tennis match based on the probability of serve points won which in turn gives the probability of a player winning the match. The author has used three types of betting data and a deep insight into the player statistics of ATP matches in the year 2013 and 2014 has given 22 important features such as player fatigue and condition of injury of the players has helped to outperform the Knottenbelt's - Common-Opponent model using two supervised machine learning models such as the logistic regression and artificial neural networking. The author believes that machine learning can bring new innovation in tennis betting as the neural networking generates a 4.35% return on investment which is an improvement of about 75% in the betting market.

Another major attribute in predicting the winner of tennis match is the court type Gorgi et al. (2019). Tennis is played in three different types of court (Clay, Hard, Grass). Every player has a unique style of playing and their abilities might differ depending on the type of court. This paper shows that the court type has a correlation with the result of a match which helps to predict the winner of the match. The author has analyzed 17 years of tennis matches with approximately 500 players using a basic- high dimensional dynamic model. The result shows that the model has overcome the limitations of Bradley-Terry model and is better in predicting results compared to all the existing models which has used the same attributes, such as time-varying strength, Court type and performance of players in Grand Slam tournaments.

Somboonphokkaphan et al. (2008) has used the Multi-layer perception and time series in order predict the winner of the tennis match using the statistical data and environmental data combined together. The author has used the statistical and environmental data to build a prediction model which is accurate in results. Based on multi-Layer perception, using appropriate input features the author has predicted the winner of the tennis match. In this research paper, apart from the statistical data of players, the environmental data set used consist of the different court surface. Each player is comfortable playing at certain surface which can be used to predict the winner. Other external factors used by the author is considering the previous injury of players not considering the players rankings, this problem is over come by incorporating Time series for statistical data.

Barnett and Clarke (2005) uses the combined player statistics to forecast the outcome of the match. This project is done by setting up a Markov chain model in Excel. The dataset used for the analysis was a long match played at the 2003 Australian Open which was manipulated to combine player A and player B statistics. The author has used only simple mathematical models to find the result of the match and the duration on the match.

2.3 Machine learning techniques for other sports

Here, the authors Pathak and Wadhwa (2016) has used some of the modern machine learning classification techniques such as Support Vector Machine, Naive Bayes, and Random Forest to forecast the outcome of an ODI (One Day International) cricket match. It is seen that SVM has given better accuracy than all other models. The author has selected various attributes to do this research and based on the result obtained the probability to win or lose a match is predicted. The factors used to predict the outcome are which team will win the toss, Day/Night, Home game advantage, First and second innings, Fitness level of players and various strategies used by different teams .

Turning to football, Prasetio et al. (2016) has used the logistic regression classification method in determining the result of a match with a prediction accuracy of 69.5%. This research helps the managers and various clubs who invest on professional football players a clear shot idea in selecting the right players for the team. The author has selected the most important attributes to get accurate results. Also he has included the FIFA video game analysis to get better results. The parameters such as "Home defence" and "Away defence" help best to predict the win/loss record of the match in this research.

Similarly, in this paper Gevaria et al. (2015) some of the most significant features are used to predict the winner of a football match using machine learning classifiers such as SVM, logistic regression and Bayesian network. The feature selection is done with WEKA software program. It is shown that logistic regression has outperformed other models with 90% accuracy. The author explains the importance of choosing the right features and selecting the best classifiers to get high accuracy.

Today, Sports betting is a big business and the game of soccer has been of great interest to most of the potential bidders. Here, Hassanniakalager and Newall (n.d.) explored the mixed logistic regression machine learning technique for the data set which consist of eight seasons of English Premier League soccer and four types of betting data. The author compared the results based on three types of predictions which are i) most-skilled prediction ii) random strategy iii) least skilled-prediction. The results were compared with the loss of gamblers with their betting odds for all the bets. This helps the gamblers to avoid risk which are similar and helps in endorsing responsible gambling.

Even though online sports betting is challenging, today it has gained high popularity among fans. Also, basketball is the most predicted game in online sports gambling. The author Cao (2012) has used various data mining techniques such as (Simple Logistics Classifier, Artificial Neural Networks, SVM and Naive Bayes) to find out who will be the winner of a basketball match. Since the result of a game is a classification problem, the author uses the Naive Bayes machine learning technique. Besides which the multivariate linear regression method is applied to predict the winner of the match.

In this paper Mantovani et al. (2015) shows the effectiveness of Random search hyper parameter tuning. The results show that Random search hyper parameter tuning gives the models better performance and less computation time. The models with random search work the same as the models with Grid search and meta heuristic models. Since we know that Svm is great for classification problems, the author experiments various methods to see how the Random search model will differ in performance using SVM model.

Similarly, In Bergstra and Bengio (2012) has shown that the random search is better than the grid search hyper parameter tuning. The random search has a lot of advantages over the grid search as, it is more effective at searching better models with less computational space.

2.4 Research Gap

All the papers reviewed so far in this research has not used hyper parameter tuning. Mantovani et al. (2015) explains the importance of hyper parameter tuning in selecting the accurate parameters which in turn helps in increasing the accuracy of the models. This research will perform the hyper parameter tuning with the machine learning models such as SVM, Random Forest, Logistic Regression and Naive Bayes to get better accuracy than the previous models used.

3 Methodology

The most important and common data mining models used are KDD (Knowledge Discovery Database), CRISP-DM (Cross-industry standard process for Data Mining) and SEMMA (Sample, Explore, Modify, Model, Assess) Azevedo and Santos (2008). This research was guided with the CRISP-DM method. The CRSIP-DM is widely used to solve data mining problems. [1] It is a structured approach which follows a cyclical process. There are six phases in a CRISP-DM process. A detailed step by step implementation is explained below.

- Business Understanding
- Data Understanding
- Data Preparations
- Data Modelling
- Evaluation
- Deployment

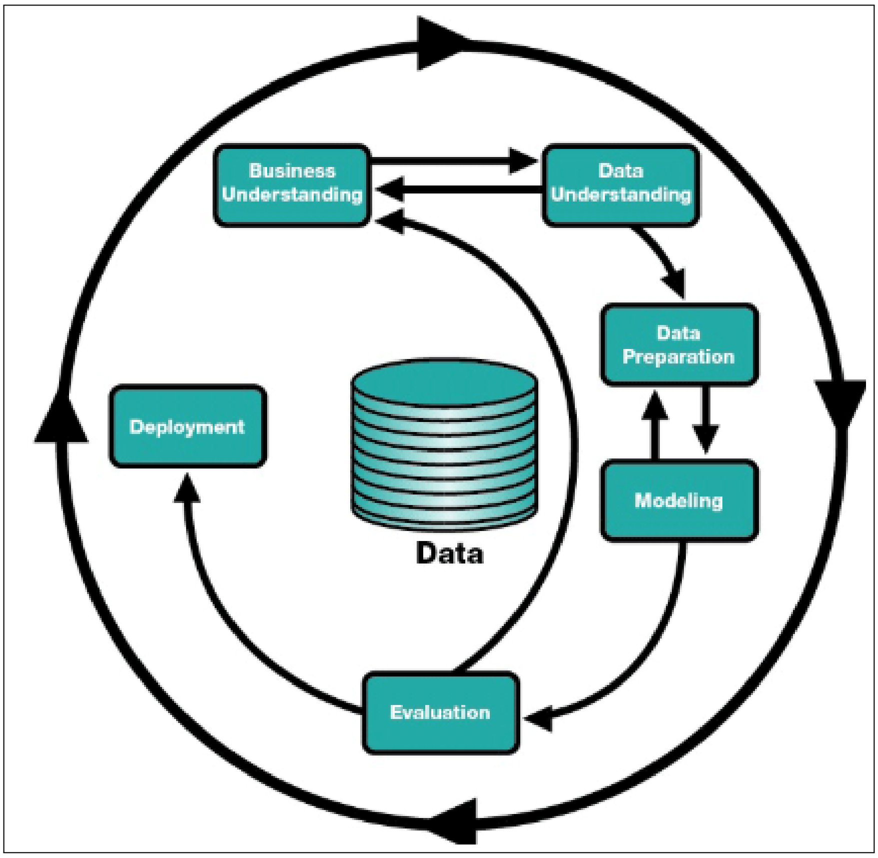


Figure 1: CRISP-DM
[2]

3.1 Business Understanding

Any data mining project should begin with a clear understanding and ground knowledge of the domain. Once we have a clear idea about the project objectives, it is important to convert this knowledge in to finding solutions using data mining techniques. In this

[1] https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome

research, the first and foremost step was to identify the factors that contribute the most in predicting the winner of a tennis match. Previously few researches were carried out to predict the winner of the tennis match and find the probability of betting odds based on the results obtained, this research is carried out to increase the accuracy of the models using a whole new set of attributes to predict the winner of the match. Since the amount of money involved in betting markets are too high it is important to have good accuracy in results and another important factor that this project was focused on was to have low latency. This is sub-sectioned into two categories such as

3.1.1 Setting objectives

It is mandatory to initiate with a primary objective from business point of view. This process involves understanding the needs of different stakeholders and working towards achieving it. It is vital that the objectives are clearly understood as this is the first step to be successful in this project. The main objective of this research project is to predict the winner of a tennis match using different parameters and models with better accuracy.

3.1.2 Produce Project Plan

In this stage, a proper sketch of the project objectives and business goals to be achieved should be collected. This process involves referring to previous research papers and having a clear idea and awareness about the models used previously and having the knowledge on how it can be implemented on our project.

This research project involves understanding the importance of serve in predicting the winner of a tennis match. It focuses on the various attributes of serve and shows how choosing the most relevant feature helps in increasing the accuracy of the model

3.2 Data Understanding

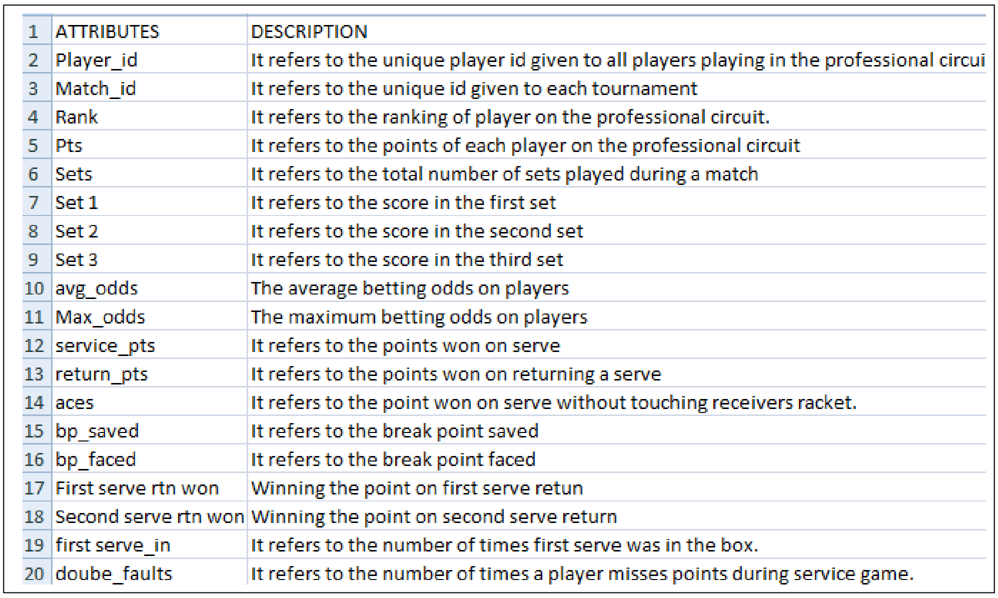
This process involves having a clear understanding about the data set, the sources from which the data is obtained, size of the data set, the details about each row and column is discussed in this section which helps in performing the project efficiently. The data understanding process is mainly focused on checking the quality of data. Verifying the data quality involves making sure that the data which is selected for this project is clean without any errors and if the data is from a trustful source. The data set used for this project was downloaded from a public data repository called Kaggle. This data set was gathered from various sources and combined together and uploaded on Kaggle by Taylor.

3.2.1 Data set

The stats csv file used for this project has the match statistics which has 20241 rows and 20 columns. The attributes contained in are match id, player id, individual set scores, break points won and saved, serve point won, fist and second serve points, aces, winner and loser of the match and finally it has the average odds and the maximum odds.

3.3 Data Preparations

This process involves preparing the data for the next stage by cleaning the data and selecting the attributes which are essential for analysis to obtain effective results. This



1	ATTRIBUTES	DESCRIPTION
2	Player_id	It refers to the unique player id given to all players playing in the professional circui
3	Match_id	It refers to the unique id given to each tournament
4	Rank	It refers to the ranking of player on the professional circuit.
5	Pts	It refers to the points of each player on the professional circuit
6	Sets	It refers to the total number of sets played during a match
7	Set 1	It refers to the score in the first set
8	Set 2	It refers to the score in the second set
9	Set 3	It refers to the score in the third set
10	avg_odds	The average betting odds on players
11	Max_odds	The maximum betting odds on players
12	service_pts	It refers to the points won on serve
13	return_pts	It refers to the points won on returning a serve
14	aces	It refers to the point won on serve without touching receivers racket.
15	bp_saved	It refers to the break point saved
16	bp_faced	It refers to the break point faced
17	First serve rtn won	Winning the point on first serve retun
18	Second serve rtn won	Winning the point on second serve return
19	first serve_in	It refers to the number of times first serve was in the box.
20	doube_faults	It refers to the number of times a player misses points during service game.

Figure 2: Data Description

is a very critical part of the project as this involves selecting attributes based on the business requirements, quality of the data and the technical constraints.

3.3.1 Data cleaning

Cleaning the data helps to increase the accuracy of the result and also improves the quality of the data set by replacing/removing the missing values, checking class imbalance, removing unwanted rows and columns.

3.4 Data Modelling

This part of the project is very important as it engages in understanding the various aspects of the project by doing a lot of research by referring to previous work and finding the best way to produce efficient results. Once the data pre-processing is done, the implementation process takes place. In this project, the first step was to do dimensionality reduction which helps to remove the irrelevant features for this project using PCA (Principle Component Analysis). Since the data set has continuous variables we are using PCA for dimensioanlity reduction. After this process, the machine learning models were applied. The machine learning models used for this project are SVM, Naive Bayes, Random Forest and Logistic Regression. The Machine learning models were hyper tuned in the end.

3.5 Evaluation

The various metrics used for evaluation in this project are accuracy, F1 score, precision, Kappa and AUC. These are used in machine learning models to interpret the performance and extract the best performing algorithm.

4 Design Specification

Initially this research study was done with the help of the Anaconda navigator. The Anaconda navigator serves as a platform for various web applications and has a collection of packages. Spyder which is a well known user interactive coding environment which is used for data exploration,execution and visualization representations.[3] To increase the execution time, this project was later executed on a cloud platform from Google Colab. This is a very well known platform with pre-installed packages widely used for performing machine learning projects.The GUP was selected as runtime for faster execution.

[3] https://anaconda.org/anaconda/spyder

5 Implementation

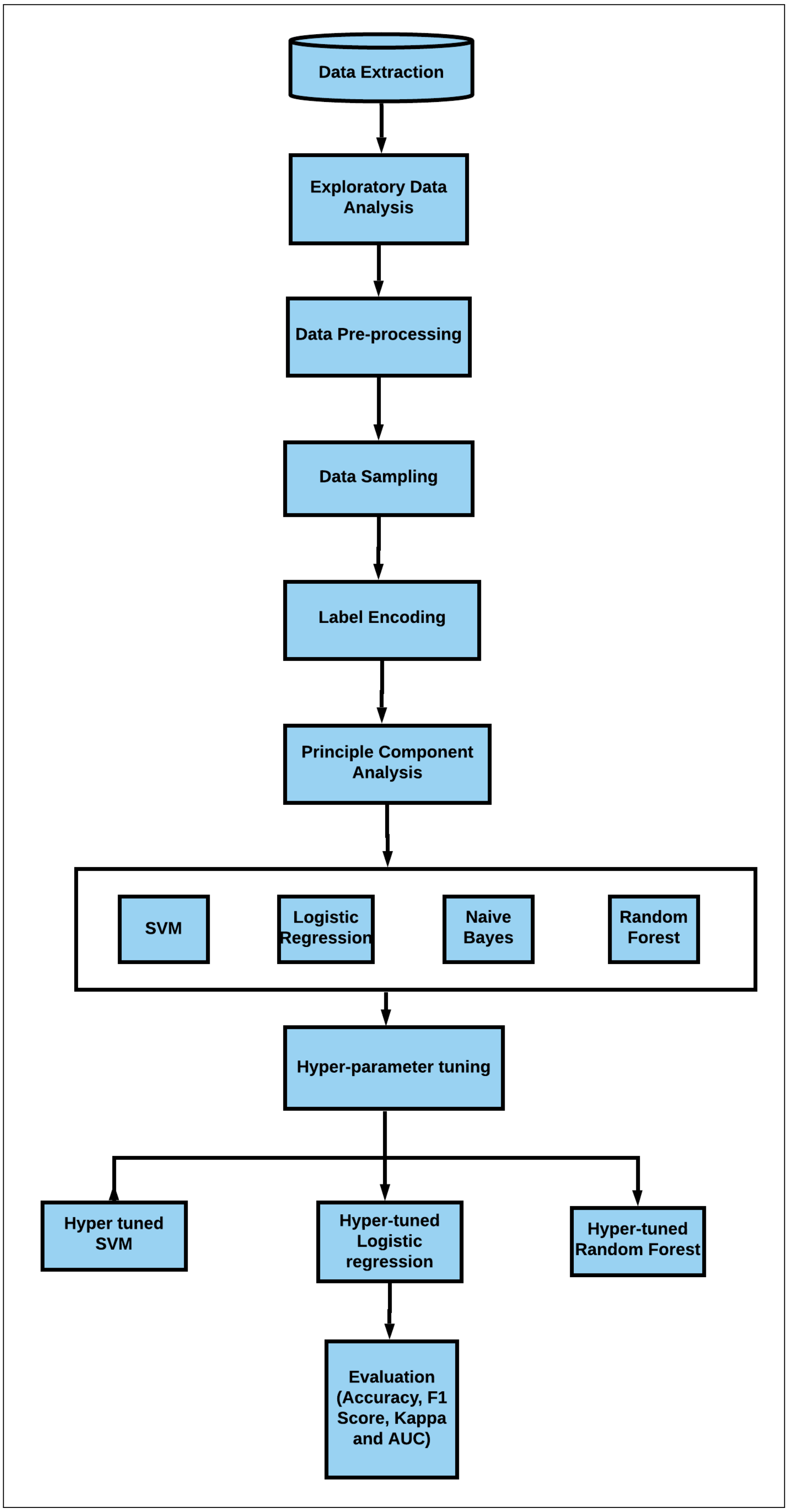

Data Extraction
Exploratory Data Analysis
Data Pre-processing
Data Sampling
Label Encoding
Principle Component Analysis
SVM
Logistic Regression
Naive Bayes
Random Forest
Hyper-parameter tuning
Hyper tuned SVM
Hyper-tuned Logistic regression
Hyper-tuned Random Forest
Evaluation (Accuracy, F1 Score, Kappa and AUC)

Figure 3: Work Flow Diagram

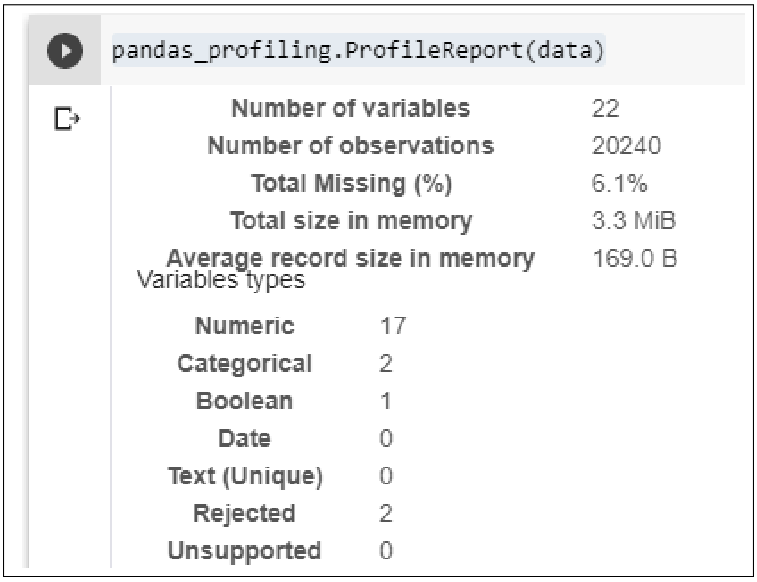
5.1 Data Selection

The data set used for this project was downloaded from a public repository called Kaggle. It was uploaded by gathering data from three different sources.Data was combined from Jeff Sackmann, Tennis-Data.co.uk and Ultimate Tennis statistics. The data was in 3 csv files (player csv, matches csv and stats csv) For this research paper, only the stats csv file was used which has all the service related statistics of each player in the match and the betting odds such as average and maximum odds on every player, the serve points, aces, break points won, service return points, first and second serve won etc.The raw data set consist of 20241 rows and 20 columns. It was seen that class was fairly balanced. The missing values were replaced and the error values were removed for better accuracy.

5.2 Data Description

5.2.1 Exploratory data analysis

After extracting the data set, it is necessary to understand the various details in the data such as number of columns, if the column contains numerical or categorical values, the number of null values, check outliers, the ratio of missing value, anomalies present in the data set. In this project, the pandas profiling is used which makes the EDA process easy and quick. The pandas is pre-defined in python, which helps to represent the variables in the form of visualizations. In this research the EDA was performed to do remove the null values, check class imbalance and remove unwanted columns. The Pearson correlation and the Spearman correlation was plotted to show the relationship between variables.



```
pandas_profiling.ProfileReport(data)
```

```
      Number of variables             22
   Number of observations             20240
        Total Missing (%)             6.1%
     Total size in memory             3.3 MiB
Average record size in memory         169.0 B
Variables types

       Numeric        17
   Categorical        2
       Boolean        1
          Date        0
  Text (Unique)       0
      Rejected        2
   Unsupported        0
```

Figure 4: Pandas Overview

5.3 Data Sampling

One of the most important problems that occur in classification problems is class imbalance. This class imbalance is resolved using two methods the down-sampling and up-sampling method. This data set which is used for the project was evenly balanced.

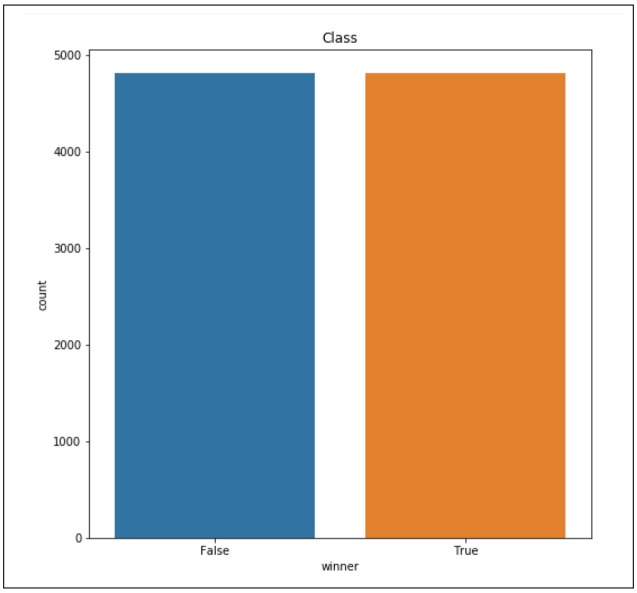


Figure 5: Class Imbalance

5.4 Data pre-processing

The Raw, unprocessed dataset usually is cleaned and pre-process before applying machine learning models . The data pre processing is done to remove the null values, inconsistent data and unwanted rows from the dataset. The pre-processing done in this project is included in the steps given below.n

5.4.1 Removing unwanted columns

The unwanted columns which do not contribute much to the project are removed to reduce the noise in the data. In this project, after the EDA (Exploratory Data Analysis) process we come to know that the set 4 and set 5 give less than 0.5 which can be neglected as they do not contribute much to the project.

5.4.2 Removing null values

The null values should be removed before applying the machine learning models as they do not give efficient results and reduce the quality of the data.

5.5 Dimensionality Reduction

The dimensionality reduction is one of the most important step in this research as it can make or break a research project. Dimensionality reduction helps in creating artificial data from the data. It is the process of reducing the amount of random variables. High dimensional data will affect the performance of the machine learning model by creating noise, takes a lot of space and computation time and also has offer fitting problems.It makes sure that the extracted features are highly informative ,useful and should have some relevance to the dependent variable. This process helps in achieving better accuracy. The PCA dimensionality reduction technique is used in this project .

5.5.1 Principle Component Analysiss

The PCA is the Principle component analysis. It is widely used as a solution to solve dimensionality reduction.The primary reason in choosing the PCA for dimensioanlity reduction is, it works well with categorical values. PCA is popularly known as linear dimension reduction as it takes the original variables and finds the best linear combination so that the variance in the new data set is maximum [4].

5.6 Encoding categorical variables

The process of encoding is a mandatory step , as the machine learning cannot be directly applied to special characters. In this research the Label Encoding is used after de-noising the data.

5.6.1 Label encoding

The Label encoder is used to encode the binary data and the nominal data. Since the data set used has categorical values, the label encoding is applied as it converts the categorical values in to o's and 1's. The two categorical values in the data set used are Winner and loser which is classified as true for Winner and False for loser. This is Encoded using the Label encoding as 0 for Winner and 1 for loser.

5.7 Classification using Machine learning algorithms

5.7.1 Support Vector Machine

The support vector machine commonly known as SVM is a supervised machine learning technique which is preferred by many data experts as it gives better accuracy with less computation power. [5] Even though SVM is mostly used for classification problems, it can be used for both regression and classification problems. As Svm has given high accuracy by outperfforming other models in Cornman et al. (n.d.), we have used SVM model in this research.

5.7.2 Naive Bayes

Naive Bayes is most popularly used for classification problems.There are three types of Naive Bayes classifiers namely, Multi nominal Naive Bayes, Gaussian Naive Bayes and Bernoulli Naive Bayes. [6] In this research project we have used the Gaussian Naive Bayes since the data set contains continuous values. Gaussian Naive Bayes works best with continuous values. Even though Naive Bayes is well known for its speed and easy implementation, the major drawback is the performance is highly disturbed if the predictors are independent. As Learning (2017) used various ML models such as Naive Bayes, Svm, Logistic Regression which has give efficient results, we have used Naive Bayes in this research.

[4] https://blog.paperspace.com/dimension-reduction-with-principal-component-analysis/

[5] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[6] https://towardsdatascience.com/introduction-to-naive-bayes-classification-4cffabb1ae54

5.7.3 Logistic Regression

Logistic Regression is mainly used for classification problem which is a predictive analysis algorithm and it is based on the concept of probability [7]. Since this project involves predicting the winner of a match, we have used logistic regression The two types of logistic regression are Binary and Multi-liner.Logistic regression is well known for sports predicting Gevaria et al. (2015) , Cornman et al. (n.d.), Learning (2017) are examples for successful results with logistic regression.

5.7.4 Random Forest

The Random Forest model is a collection of multiple decision trees. If the number of trees is larger, the accuracy is better in this model. It can be used for both classification and regression problems and it also avoids the problem of overfitting [8]. The Random forest model has two key concepts rather than just averaging the prediction trees. The first key concept is while building trees there should be random sampling of training data. The second concept is while splitting the nodes, there should be random subset of features selected [9]. In this reserach Random Forest has outperformed all other models with the F1 score gives 78% accuracy.ss

5.8 Hyper-parameter tuning

Machine learning algorithms consist of various parameters with specific values. The process of tuning these parameters to obtain the best parameter which can help the models to learn efficiently is known as Hyper-parameter tuning. In Hyper-parameter tuning, the values are set before starting the learning process [10]. There are two types of tuning strategies known as Grid search and Random search. Grid search follows the traditional strategy of searching through an optimal set of parameters but it is extremely time consuming. On the other hand, the random search saves a lot of processing time since it searches only a specific set of parameters. Mantovani et al. (2015) In this research we have used the random search hyper-parameter tuning which has increased the accuracy of the models.

6 Evaluation

6.1 Accuracy

Accuracy is the efficiency of the machine learning models which has classified the data. The performance of the ML model can be determined based on the accuracy. Accuracy works well only with symmetric data set i.e the class should be balanced properly. In this research, after performing the hyper tuning, we can see that SVM has outperformed all other models with accuracy of 76%. Accuracy can be calculated in confusion matrix using the formula given below.

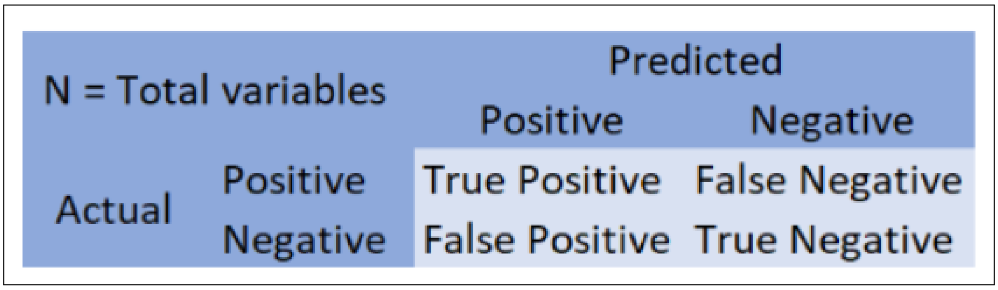
Accuracy = (TP+TN) / (TP+FP+TN+FN)

[7] https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

[8] https://syncedreview.com/2017/10/24/how-random-forest-algorithm-works-in-machine-learning/

[9] https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76

[10] https://towardsdatascience.com/hyperparameter-tuning-c5619e7e6624



N = Total variables		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Figure 6: Confusion Matrix

6.2 AUC

AUC is known as area under curve. It represents degree or measure of separability [11]. If the AUC is higher, the model works better. It predicts 0s as 0s and 1s as 1s. For a good model, the AUC is near 1 and for a poor model, the AUC is near 0. For a poor model, the prediction might be different such as 0s as 1s and 1s as 0s.

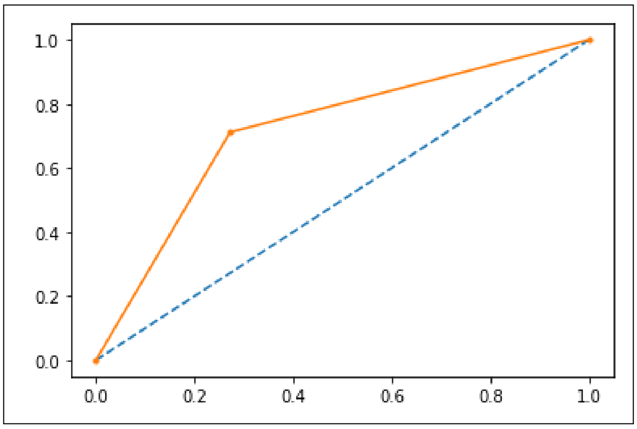


Figure 7: SVM

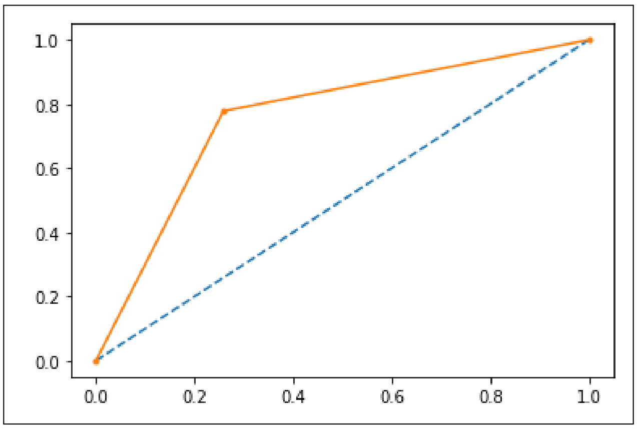


Figure 8: Random Forest

[11] https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

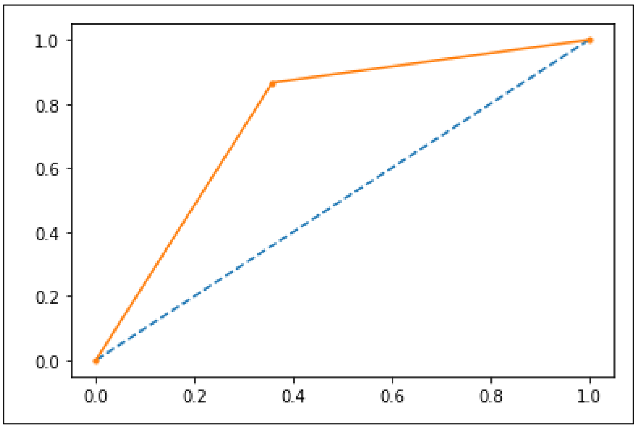


Figure 9: Logistic Regression

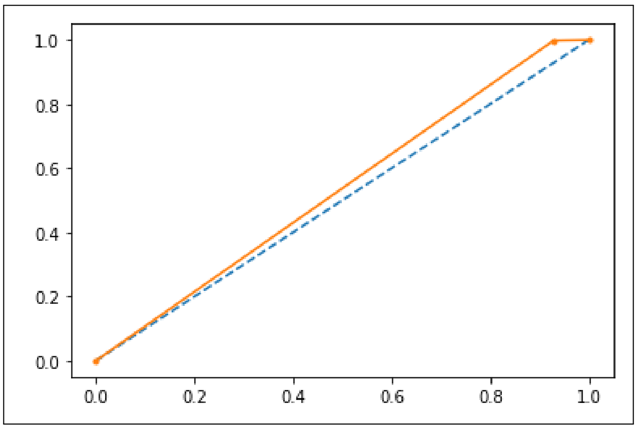


Figure 10: Naive Bayes

6.3 F1-Score

F1 score is defined as the weighed average of Precision and Recall. The main advantage of F1 score over accuracy is that it works well even with uneven class distribution. F1 score takes both positive and negative values. [12] In this project F1 score has got 78% accuracy with random forest. The formula for F1 score is given below.

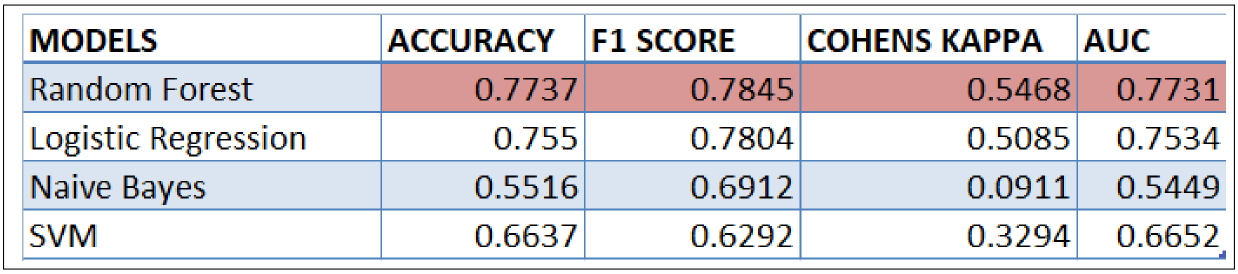
F1 = 2*(Precision * Recall) / (Precision + Recall)

6.4 Kappa Statistics

Kappa Statistics is usually used when machine learning faces multi-class classification problems.It is a qualitative measure, which is defined as the ability to compare different raters to classify subjects.

The comparison table shows the accuracy comparison of all four models evaluated. It is seen that Random Forest has the highest accuracy of 78% with F1 score.

[12] https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/



MODELS	ACCURACY	F1 SCORE	COHENS KAPPA	AUC
Random Forest	0.7737	0.7845	0.5468	0.7731
Logistic Regression	0.755	0.7804	0.5085	0.7534
Naive Bayes	0.5516	0.6912	0.0911	0.5449
SVM	0.6637	0.6292	0.3294	0.6652

Figure 11: Result table comparing all models

6.5 Discussion

In this section, there are few discussions related to the project which was encountered while performing this project.

Initially, the data pre-processing was a major challenge, as understanding the details and key terms in the raw data set was difficult and time consuming.

The Pandas profiling helped to perform the exploratory data analysis in a very efficient way by making the process quick and easy by reducing the time and effort to analyse the data effectively and present an overall report.

Since the data set had more than 10 variables, the data set was considered to be highly dimensional. Dimension reduction was very important step to avoid over fitting of the model accuracy and to get effective results. Hence the dimensionality reduction was done using the Principle Component Analysis (PCA).

Since the field of sports betting involves huge amount of money invested in it, obtaining high accuracy is important. For this purpose, the hyper parameter tuning was performed to tune the models to get better accuracy. A detailed research must be done to define the parameters and find the right combination of models to obtain high accuracy.

7 Conclusion and Future Work

This research study has shown that the winner of tennis match can be predicted with great accuracy by selecting the most significant features. Dimensionality reduction is the key in obtaining efficient results as it helps to eliminate the irrelevant features which do not contribute much to the project. Hence in this research we have successfully used the Principle Component Analysis (PCA) for dimensionality reduction which helps in attaining better accuracy and accurate results. Some of the top machine learning models such as SVM, Naive Bayes, Logistic Regression and Random forest was used and the results show that SVM has better accuracy compared to the other models. The Hyper-parameter tuning was done to select and highlight the most significant parameters that works the best with models to increase the accuracy of the models. The main objective of this research is to predict the winner of a tennis match by using the different attributes related to service points in tennis and to get high accuracy which has been achieved. This research has used the Hyper parameter tuning which has helped to increase the accuracy of the models. The randomized search hyper parameter approach was used in this project. All the parameters related to service points are discussed and analysed in this paper successfully.

This research could be further enhanced by selecting the data which helps to predict the winner of the match based on:-

- **Weather Condition:** The weather condition is one important aspect to consider in predicting the winner of a tennis match. Getting the weather data on the same day of the match will help to predict if it has any influence on the match results.

- **In Play bets:** Various In play bets can be conducted with the appropriate data, such as bets on predicting what the next shot will be or how many aces will the player hit before winning the match will be interesting and challenging to predict in the future.

- **Intensity of previous matches:** The intensity of the matches played before, the schedule of players, the time before previous injury are all factors which could be used to forecast the winner of the match.

8 Acknowledgement

- I would sincerely like to thank my supervisor Dr Dondio Pierpaul for his support and constant encouragement towards this research project. He was very helpful and guided me throughout my research project.

- I would like to thank my friends and family for being extremely supportive and helping me complete this reseach project successfully

- I would also like to thank all the staffs and professors of National college of Ireland for their support throughout the dissertation period.

References

Azevedo, A. I. R. L. and Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview, *IADS-DM* .

Barnett, T. and Clarke, S. R. (2005). Combining player statistics to predict outcomes of tennis matches, *IMA Journal of Management Mathematics* **16**(2): 113–120.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization, *Journal of Machine Learning Research* **13**(Feb): 281–305.

Cao, C. (2012). Sports data mining technology used in basketball outcome prediction.

Cornman, A., Spellman, G. and Wright, D. (n.d.). Machine learning for professional tennis match prediction and betting.

Easton, S. and Uylangco, K. (2010). Forecasting outcomes in tennis matches using within-match betting markets, *International Journal of Forecasting* **26**(3): 564–575.

Gevaria, K., Sanghavi, H., Vaidya, S. and Deulkar, K. (2015). Football match winner prediction, *International Journal of Emerging Technology and Advanced Engineering* **10**(5): 364–368.

Gorgi, P., Koopman, S. J. and Lit, R. (2019). The analysis and forecasting of tennis matches by using a high dimensional dynamic model, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* .

Haghighat, M., Rastegari, H. and Nourafza, N. (2013). A review of data mining techniques for result prediction in sports, *Advances in Computer Science: an International Journal* **2**(5): 7–12.

Hassanniakalager, A. and Newall, P. (n.d.). A machine learning perspective on responsible gambling. 2018.

Learning, M. (2017). Final project report: Real time tennis match prediction using machine learning.

Mantovani, R. G., Rossi, A. L., Vanschoren, J., Bischl, B. and De Carvalho, A. C. (2015). Effectiveness of random search in svm hyper-parameter tuning, *2015 International Joint Conference on Neural Networks (IJCNN)*, Ieee, pp. 1–8.

Pathak, N. and Wadhwa, H. (2016). Applications of modern classification techniques to predict the outcome of odi cricket, *Procedia Computer Science* **87**: 55–60.

Philpott, A. B., Henderson, S. G. and Teirney, D. (2004). A simulation model for predicting yacht match race outcomes, *Operations Research* **52**(1): 1–16.

Prasetio, D. et al. (2016). Predicting football match results with logistic regression, *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, IEEE, pp. 1–5.

Sipko, M. and Knottenbelt, W. (2015). Machine learning for the prediction of professional tennis matches, *MEng computing-final year project, Imperial College London* .

Somboonphokkaphan, A., Phimoltares, S. and Lursinsap, C. (2008). *Tennis winner prediction based on time-series history with neural modeling*, PhD thesis, Chulalongkorn University.