An Investigation into the Statistical Properties of Ranking in the Sport of Tennis

Submitted By

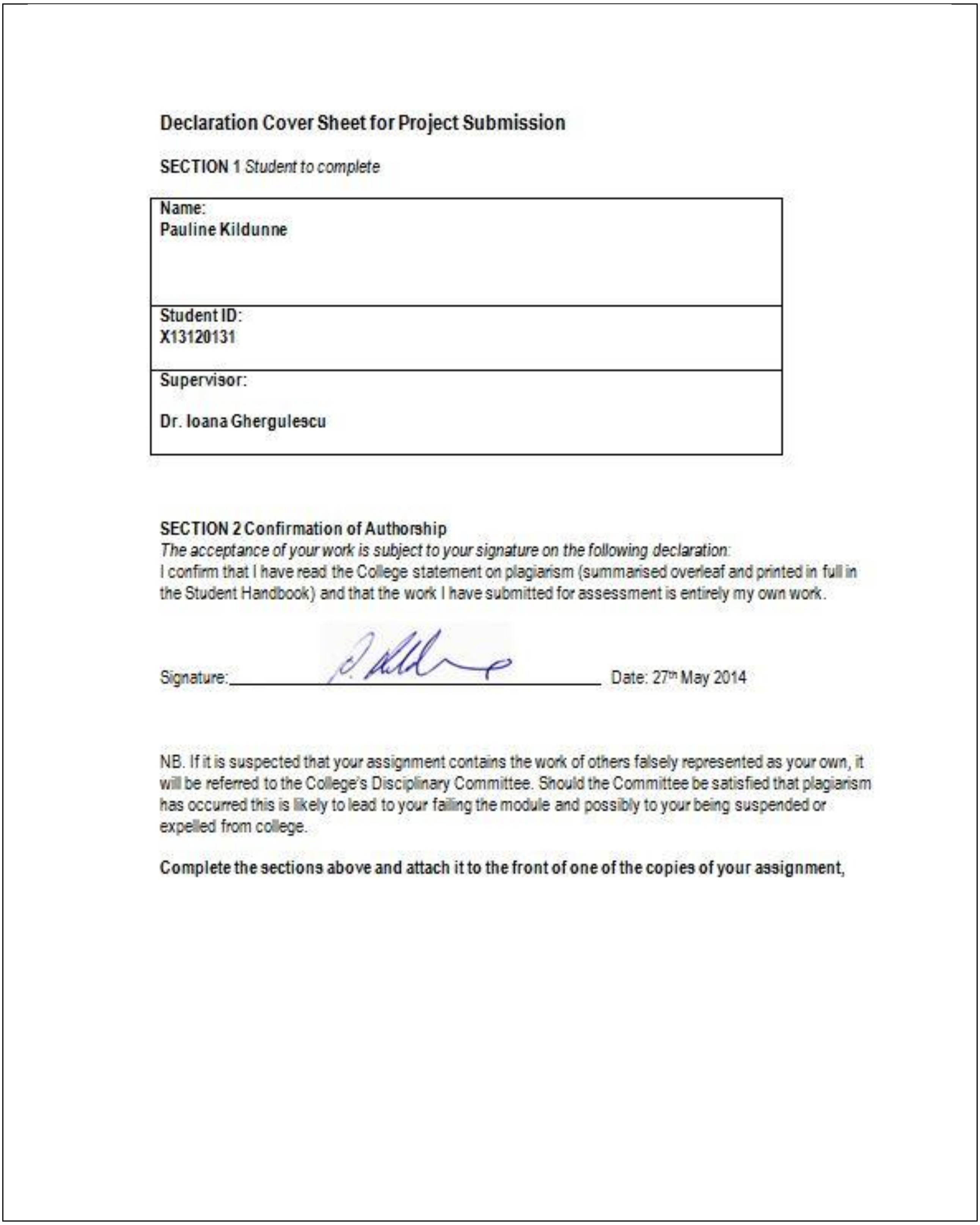
Pauline Kildunne (Student Number: x13120131)
Pauline.Kildunne@student.ncirl.ie

Higher Diploma in Science in Data Analytics

Module Project

Due Date: 28th May 2014

Submitted to: Dr. Ioana Ghergulescu



Declaration Cover Sheet for Project Submission

SECTION 1 *Student to complete*

Name: **Pauline Kildunne**
Student ID: **X13120131**
Supervisor: **Dr. Ioana Ghergulescu**

SECTION 2 Confirmation of Authorship

The acceptance of your work is subject to your signature on the following declaration:

I confirm that I have read the College statement on plagiarism (summarised overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature:______________________________ Date: 27th May 2014

NB. If it is suspected that your assignment contains the work of others falsely represented as your own, it will be referred to the College's Disciplinary Committee. Should the Committee be satisfied that plagiarism has occurred this is likely to lead to your failing the module and possibly to your being suspended or expelled from college.

Complete the sections above and attach it to the front of one of the copies of your assignment,

Tables of Contents

Figures Table

1. Executive Summary

This project is set out to give an insight into the ranking system in tennis. Through statistical analysis the requirements was to produce evidence that would present satisfactory answers to the research question outlined in 2.3. Research Questions. The findings for the Grand Slam semi-finals and finals were;

- The most predicable variable was MaxW (greedy algorithm). If the odds were greater than 1.68 the players with a lower rank were most likely to win the match, the least favourites.
- Djokovic N., - rank 3 won Federer R., - rank 2
- Murray A., - rank 8 won Djokovic N., - rank 2
- Kvitova P., - rank 8 won Sharapova M., - rank 6
- In the semi-finals the favourites won the match 17 out of 18 times.
- In the finals round it looked at the odds again (MaxW), 5 out of 6 times the higher ranked player won.
- 2011, Djokovic N., - rank 1 won Nadal R., - rank 2
- 2012, Nadal R., - rank 2 (least favourite) won Djokovic N., - rank 1.
- Federer R., - rank 3 (favourite) won Murray A., - rank 4.
- When the betting odds were less than or equal to 1.23 in the final round the less favourite won the match but of course on both occasions it was Serena Williams in 2012 at the US open and Wimbledon. She ranked 4 against her opponent Azarenka V, who was ranked 1 and she was ranked number 6 against Radwanska A. ranked 3. Serena Williams 2014 is ranked at number 1.

2. Introduction

The scope of the project is to develop an output of statistical analysis of ranking using the datasets of the tennis tournaments for the last three years, 2011, 2012, 2013 for the ATP and WTA.

The field of statistical analysis is an important mathematical tool that informs sponsors and decision makers on their choice of player to invest in. A higher performing player will attract lucrative sponsorship deal, media coverage and spectators. In the competitive world of sport bookies rely on decisions from the outputs of sound mathematical models for placing betting odds. The underlying premise to this paper is to add value to the official statistics, and produce a reasonable forecasting model.

2.1. Domain Description

Both associations, ATP and WTA rank professional tennis players and use their rankings to decide both the participation of players in tournaments, as well as the ultimate champion of the year.
The players ranking is dependent on how many points they accumulate. Each tournament is assigned points and the amount awarded is dependent on the level a player achieves in each tournament.

- Points are calculated over a 12 month period.
- How each player earns a particular ranking number like World No.1 is based purely on whether his total ranking points tally is greater than other players.

- Rafael Nadal as of May 2014 is ranked number one at 12,500 points where as Novak Djokovic is ranked number two with 11,850 points and Andy Murray is ranked number eight with 4,120 points[1].
- Each tournament has a grading system for how many points you can earn this is dependent on how far the player progresses in the tournament i.e. the number rounds he advances through[2].

  - ➢ Gram Slam Ranking Points:
  - ➢ Winner: 2,000 points
  - ➢ Runner-up: 1,200 points
  - ➢ Semi-finalist: 720 points
  - ➢ Quarter-finalist: 360 points
  - ➢ Round of 16: 180 points
  - ➢ Round of 32: 90 points
  - ➢ Round of 64: 45 points
  - ➢ Round of 128: 10 points
  - ➢ Win qualifying: 25 points

- It is probable that a higher-ranked player will win the tournament.
- A Grand Slam singles champion earns 2,000 ATP and WTA ranking points.
- The next highest point single champions earns is 1,500 in the BNP Paribas WTA Championships for the women[3] and Barclays ATP World Tour Finals for the men.

2.2. Motivation / Aims

The main motivation/aim is to give the decision maker(s) a better understanding of the drivers that effect the ranking system in tennis, this paper is complete with comprehensive research and statistics evidence to back up these views. Furthermore to supply them with information that will allow them to make better decisions, invest in players with high ROI (return of investment), forecast predictions and utilise the results to gain a competitive advantage.

2.3. Research Questions

The research questions have changed since the outset of this project this is mainly due to the time constraint and the feasibility to investigate all the possibilities initially set out.

a. Do the higher ranking players typically win the match?
b. Do higher ranking players typically outperform the lower ranking players on the different surface type, clay, grass, hard?
c. Analysing three years of data, 2011, 2012, 2013 and predict the probable variable that will determine the ‘winner rank’.

2.4 Solution Overview

The initial analysis was produced in RStudio[4], which uses the R programming language. Microsoft Excel was used in the pre-process stage of the data; it was also used to produce graphs due to its ease of use and pivot table function. The Comprehensive R Archive Network[5]

[1] Official Emirates *ATP Rankings (2014). ATP World Tour.* Available at: http://www.atpworldtour.com/Rankings/Singles.aspx [Accessed 20 May 2014].

[2] Samson, M. (2012) *ATP Men’s Tennis Rankings Explained.* Available at: http://grandslamgal.com/atp-mens-tennis-rankings-explained/ [Accessed at 20 May 2014].

[3] WTA Tennis (2014). Women’s Tennis Association. Available at: http://www.wtatennis.com/all-about-rankings [Accessed 20 May 2014].

[4] R (1993). The R Project for Statistical Computing. Available at: http://www.r-project.org/ [Accessed 30 April 2014].

[5] R (1993). The Comprehensive R Archive Network. Available at: http://cran.r-project.org/ [Accessed 30 April 2014].

library was used to test various programs to help in the analysis. Weka[6] was used for the machine learning algorithm. Google Chrome was used as the main browser to access the internet and carry out research on the sport, access different authors who have completed studies in the analysis of tennis and it was used to download the dataset.

2.5. Structure

The following is a brief overview of each section in the document:

- Section one contains the executive summary and details the results of the analysis in the Grand Slam semi-finals and finals.
- Section two contains the introduction and the background information for the project, along with the solution overview.
- Section three discusses related work in the field (literary research).
- Section four includes the system design including functional and non-functional requirements.
- Section five details the datasets used and the extraction of these datasets.
- Section six details the data cleansing.
- Section seven describes the data analysis, methodology and reults.
- Section eight gives the report's conclusions.
- Section nine details further development/research in the field.
- Section ten contains the report's references.
- Section eleven contains the appendix which includes Weka output from model, the initial project proposal, requirements specification and the management progress reports.

3. Related Work (Background)

Tennis is one of the most popular individual sports in the world. It engages millions of spectators and attracts substantial TV viewers who follow numerous tournaments which take place throughout year. Professional single tennis matches comprise of two players and only two possible outcomes of a match. The search for a model using techniques including statistical analysis, machine learning and data mining is constantly explored to predict the outcome of a match.

In psychological literature there exists a phenomenon known as the 'effect of psychological momentum' which was published by Jackson and Mosurski,[7] they state that this phenomenon is a major factor which can dictate the outcome of a match. A player could under perform in the first set but go on and win the next set and the whole match and likewise you have other players who start off playing well but gradually their play deteriorates as play progresses, due to fatigue or psychological strain. This can have an adverse effect to the accuracy of the predicted winning-odds.

A professional singles tennis match is played between two players. The objective is to score points in rallies throughout the duration of the event. The beginning of any rally is called a serve, and each player has at most two attempts to serve without a fail.

[6] Weka (1993). Machine Learning Group at the University of Waikato. Available at: http://www.cs.waikato.ac.nz/ml/weka/ [Accessed 25 April 2014].

[7] Jackson, D. and Mosurski, K. (1997) "Heavy Defeats in Tennis: Psychological Momentum or Random Effect?" *Springer International,* Volume 10(2), pp. 27-34. Available at: http://www.tandfonline.com/doi/abs/10.1080/09332480.1997.10542019#.U3njKyhzQ68 [Accessed 10 May 2014].

3.1. Game Set Match

Points in tennis are counted as follows: 0; 15; 30; 40. The first player to win four points scores a game. A score of deuce occurs when the players have three points each and the score is 40 – 40. The winner is the first player to score two points in a row.

Depending on any given tournament, a player is deemed a winner, in the case of the women's tournaments, when they beat their opponent in 2 out of 3 sets and 3 out of 5 in the case of the men's tournaments. Each set is composed of at least six games, and the first player to win six of those games is considered the champion. However, if each of the players has won five games, any one of them must be two games ahead to win the set. If, in turn, the score reaches six-games-all, a so-called tiebreaker game is played to decide on the winner (depending on the rules of the tournament a tiebreaker may or may not be played in the final set).

A tiebreaker is won by the first player to reach seven points and in the case of sets the players have to win two points in a row, to beat their opponent. In game and set, the players continue until one secures a two-point lead.

For this project I have researched several papers and their study in the area of tennis. The common theme to them all is their criticism of the ranking system set out by both the Association of Tennis Professionals (ATP) and Women's Tennis Association (WTA) tours and how this ranking system is not used to predict a player's chance of winning, however the official ranking of a player is used to seed them in tournaments, this is used also to determine the betting odds for a player and of course the rank of a player is considered a better bet for attracting prestige, publicity and sponsorship. If you take for example the current number ones in tennis ranking ATP, Rafael Nadal and WTA Serena Williams, they have been consistent contenders for the past five years and favourites at the bookies. In this paper I will look at the analyses that are intuitively used to predict the probability of winning a match.

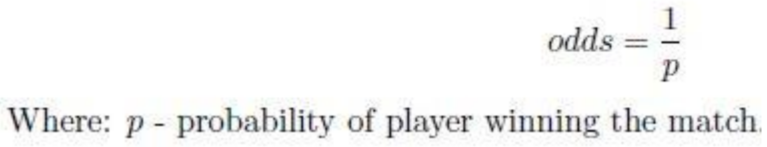
The four Grand Slam tournaments (the Australian Open, the French Open, Wimbledon, and the US Open) are the most important and prestigious tournaments on the professional tennis circuit. Each Grand Slam tournament has 128 entrants per gender, organised in a predetermined draw of 64 matches: the winner of a match advances to the next round, while the loser exits the tournament.

Some tournaments are played on different surfaces (grass - Wimbledon, clay – French Open, hard – US and Australian Open), and may be indoors or outdoors this can pose some difficulty for ranking tennis players as most players have a favourite surface, and their performance level changes with different surfaces. So therefore the official rating prior to a tournament can change depending on the type of play, speed of the surface and whether it is a home or away match.

3.2. Capturing the Probability and Confidence

In Madurska's[8] paper it looks at predicting the outcome of a tennis match by capturing the probability and confidence of the binary outcome of winning odds. It is fair to say the higher the odds the lower the probability of winning the match but you can expect higher financial gain if the player with higher odds wins. In the following diagram you can see the formula which Madurska used.

[8] Madurska M. (2012) "A Set-by-Set Analysis Method for Predicting the Outcome of Professional Singles Tennis Matches" *Imperial College London,* pp.13. Available at: http://www.doc.ic.ac.uk/teaching/distinguished-projects/2012/a.madurska%20.pdf [Accessed 11 May 2014].



$$odds = \frac{1}{p}$$

Where: p - probability of player winning the match.

Figure 1: Probability of winning or losing the match

Also in Madurska's[9] paper he describes the quantitative models of a tennis match through a ranking Markov model. He mentions that they rely on estimating the probability of winning a point on serve or return, against a certain opponent. The values are subsequently fed into a mathematical equation based on a Markov chain, to produce the probability of a given player winning the match. This model has been published in a number of papers, and they assume that the point-winning probabilities, once calculated, do not change throughout the match. However, the dynamics of the match can change on the day and are not taken into account and although mathematically very attractive because of its simplicity, does not describe tennis matches accurately.

3.3. Gender Differences in Performance

In Paserma[10] paper data was used from nine Grand Slam tournaments to evaluate whether men and women respond differently to competitive pressure in a real-world setting with large monetary rewards. The results reveal that the performance of both men and women can deteriorate in the final and decisive set. Women's decline in their performance is more pronounced than that of men, but the difference is not statistically significant.

However there was indication that there were significant differences between men's and women's performances at crucial stages of the match: there was a tendency for women to commit more unforced errors while they remain constant for men. Some of this difference can be explained by gender differences in type of play as points become more important: the evidence on rally length and on the speed and accuracy of first serves strongly suggests that women tend to adopt a safer and less aggressive strategy on important points.

Tennis fans are aware of the different styles in play between the men and women's tournaments as in the four grand slams where men play best-of-five-sets matches and women play best-of-three sets. As a result it has been much reported that fans prefer to watch men playing, because of rally length where they demand greater fitness, their play are less predictable, more interesting, sheer endurance and competitive.

Their matches are on average 75 percent longer than women's matches. In other words, male players spend on average 75 percent more time out on the court entertaining paying customers[11]. Virginia Wade, the last British woman to win Wimbledon, thinks women's tennis

[9] Madurska M. (2012) "A Set-by-Set Analysis Method for Predicting the Outcome of Professional Singles Tennis Matches" *Imperial College London,* pp.13. Available at: http://www.doc.ic.ac.uk/teaching/distinguished-projects/2012/a.madurska%20.pdf [Accessed 11 May 2014].

[10] Paserma, D.M. (2010) "Gender Differences in Performance in Competitive Environments? Evidence from Professional Tennis Players" *Boston University and Hebrew University.* Available at: http://people.bu.edu/paserman/papers/Paserman_Tennis_January2010.pdf [Accessed 12 May 2014].

[11] Johnson V.D. (2012) "Men's Tennis Is More Interesting Than Women's", *Boston Review,* 6 July. Available at: http://www.bostonreview.net/us/men%E2%80%99s-tennis-more-interesting-women%E2%80%99s-david-johnson [Accessed 19 May 2014].

has become boring and predictable due to the robotic personalities of the players and the lack of meaningful rivalries in the game[12]

3.4. SortRank and LadderRank

Tournaments are designed such that top players face the lower ranked players in the earlier rounds, this provides the top players with an unfair advantage as seeded tournaments make it increasingly difficult for lower ranked players to climb the ranking. The tournament draw is set up in a way such that the seeded players have a greater opportunity to reach the final and also gain maximum points which is a contributory factor to their ranking. This type of selection creates a bias towards the top 32 players. Players ranked lower than this find themselves facing top ranked players in the early rounds of the tournament but also if they lose they have less of an opportunity to earn substantial points.

In Spanias and Knottenbelt[13] paper they discuss two algorithms, SortRank and LadderRank, which rank professional tennis players. Both ideas make use of a quantitative tennis model to assess the performance of individual players and then compare them with each other

SortRank uses traditional sorting algorithms to rank the players using the result of a likeness match between the two players as the comparison criterion. LadderRank ranks players using a "sports-ladder" style repeat algorithm, which also compares players based on the result of a likeness match between them. Their findings resulted in comparing the LadderRank-Combined system's performance against the ATP rankings in terms of how well the rankings represent the set of matches used to generate them, the LadderRank algorithm outperformed the ATP rankings although it does not perform as well as the PageRank ranking system it still outperforms the ATP Official Rankings.

3.5. Independent and Identically Distributed

In another study a model was used to determine the probability of winning a point on service. In most work on tennis, points are assumed to be independent and identically distributed (iid), which implies that the key probability is constant for a player throughout a match. In Klaassen and Magnus[14] paper they test the independent and identically distributed hypothesis and rejected it. They challenged the independence assumption and concluded that there is dependence, possibly caused by 'psychological momentum'. There is study known as back-to-the-wall effect in which the player that is behind performs better, thus challenging the independence assumption.

Winning the previous point has a positive effect on winning the current point, the 'psychological momentum' continues and this can be seen even when a player is playing with an injury, sheer will and determination drives their ability to win the match. They also show that the deviations

[12] Panahi R. (2014)" For equal pay, women must play best of five", Herald Sun, 21 January. Available at: http://www.heraldsun.com.au/news/opinion/for-equal-pay-women-must-play-best-of-five/story-fni0fhh1-1226806188706 [Accessed 19 May 2014].

[13] Spanias D. A. and Knottenbelt W. B. (2013) "Tennis Player Ranking using Quantitative Models" *Department of Computing, Imperial College London.* Available at: http://www.doc.ic.ac.uk/~wjk/publications/spanias-knottenbelt-mis-2013.pdf [Accessed 12 May 2014].

[14] Klaassen F.J.G.M. and Magnus J.R. (2001) "Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model" *Journal of the American Statistical Association* Volume 96, Issue 454, pp. 500-509. Available at: http://amstat.tandfonline.com/doi/abs/10.1198/016214501753168217#.U3yPbyhzQ68 [Accessed 14 May 2014].

from iid depend on the quality of the players; the stronger a player, the smaller the deviation from the iid hypothesis. These results are the same for men and women. They suggest that players should be trained to “play every point as it comes.”

3.6. Conclusion

As shown above ranking and points accumulated in each tournament is a major factor to a player and what most papers based their analysis on and interesting most papers are concerned with only the top ranking players and the major tournaments. In some way or another, the algorithms used are to predict the outcome of the match and this is essentially is the tool which the betting agents used to wage their bets.

It is also shown that they is a different style of play for both men and women especially when they are under pressure and there is a phenomenon known as the ‘effect of psychological momentum’ which can affect the player psychologically and rejects the assumption that points played are independent and identically distributed.

This research has been beneficial as it has given a better understanding of the requirements, what variables might be relevant to use and the types of algorithms, in this case logistic regression and Decision Trees classifier.

4. System

4.1. System Design / Implementation

The analysis used the following programs/software:

- Internet Browser, Google Chrome to download the dataset (a precondition is an internet connection)
- Microsoft Excel to pre-process, review data and produce visuals for example histograms.
- RSudio to analyse data, visualise, carry out analysis and the Comprehensive R Archive Network library was used to test other libraries to see if they were conductive for the analyisis.
- Weka for data mining algorithms; the classification tree.
- All files were stored in the National College of Ireland depository.

Below is the overall data analysis process:

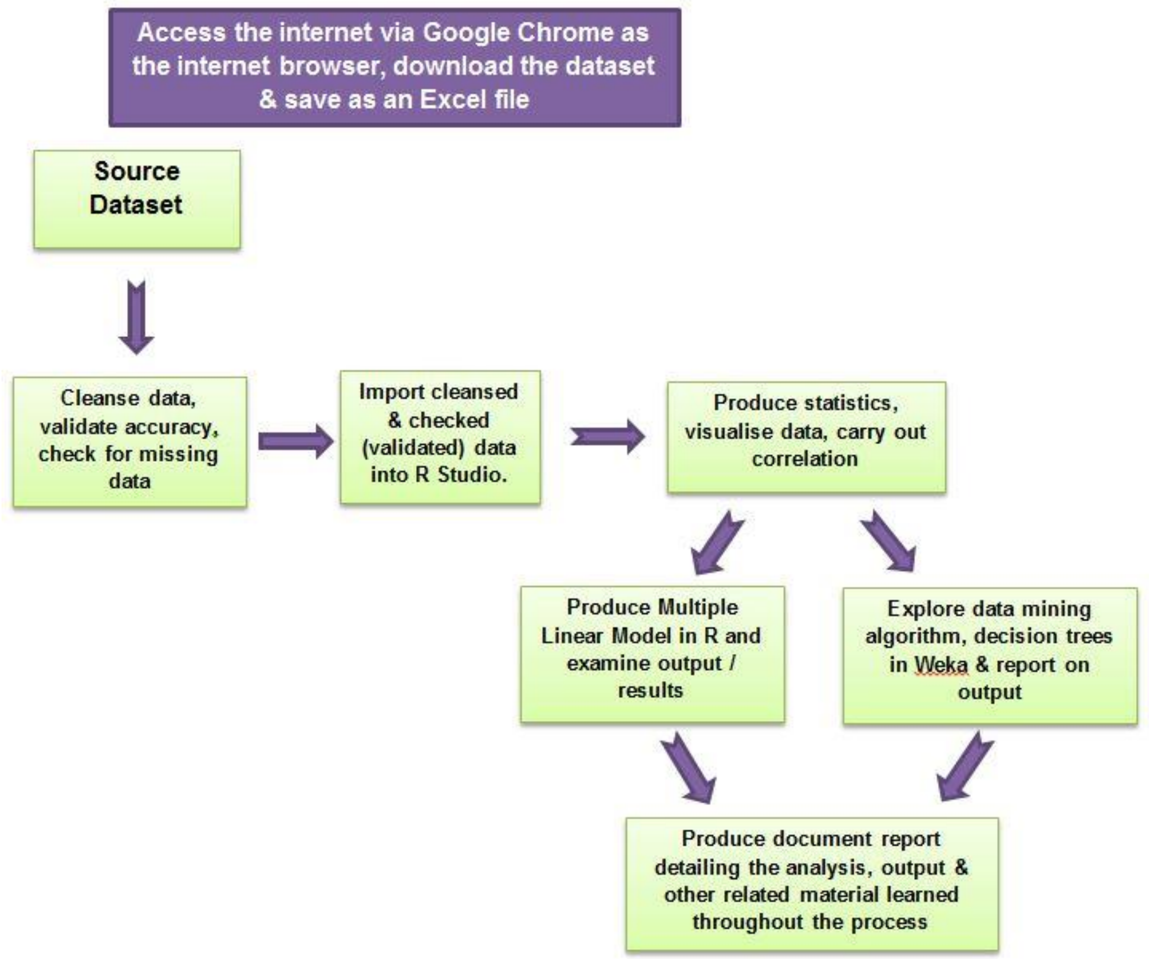

Access the internet via Google Chrome as the internet browser, download the dataset & save as an Excel file
Source Dataset
Cleanse data, validate accuracy, check for missing data
Import cleansed & checked (validated) data into R Studio.
Produce statistics, visualise data, carry out correlation
Produce Multiple Linear Model in R and examine output / results
Explore data mining algorithm, decision trees in Weka & report on output
Produce document report detailing the analysis, output & other related material learned throughout the process

Figure 2: Use Case Diagram

4.2. Requirements

4.2.1. Functional Requirements

The overall functional requirements describe how the system should conduct or operate, in this case there are four functional requirements.

- Extract data
- Prepare and clean data
- Analyse data
- Report data analysis

In Figure 4 the overall use case diagram of the project is illustrated. The diagram demonstrates each step of the process and how dependent each process is on the other. All requirements are documented in the following sections and the report aspect of the use case diagram is the entire drafted document.

Further information on the functional requirements can be found in appendices section under the heading, Initial Requirements Specification these requirements have remained the same since the initial requirements specification.

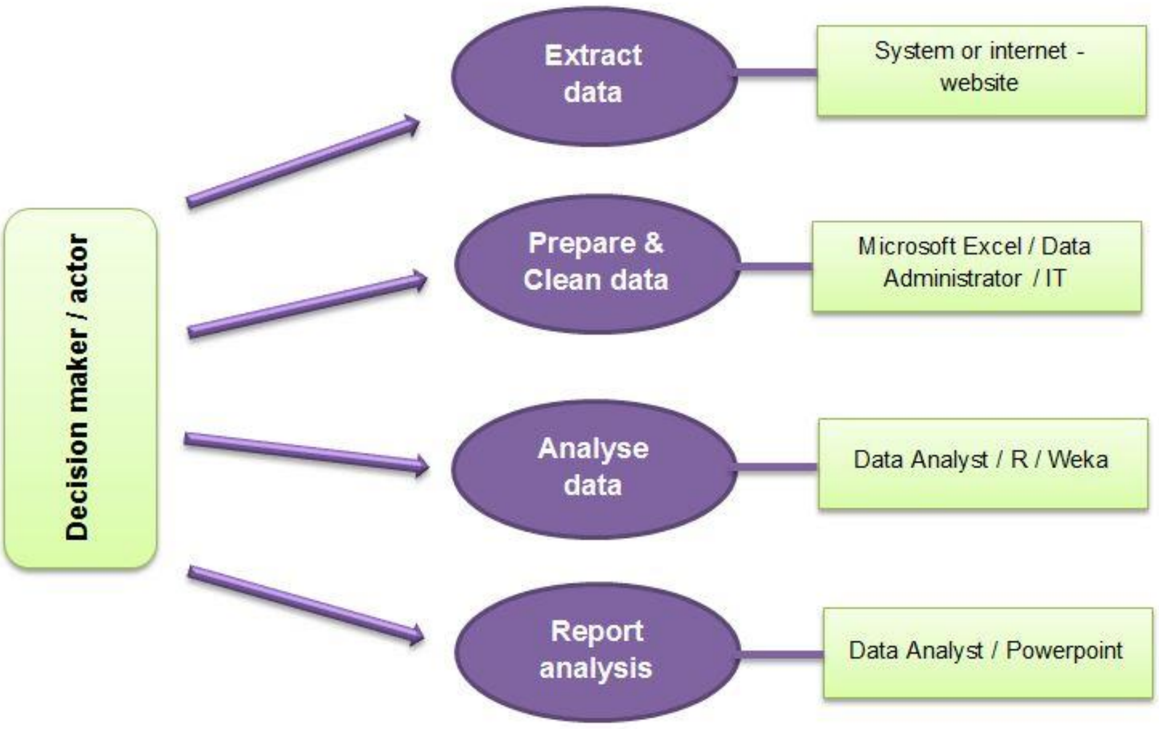


Figure 3: Overall Use Case Diagram

4.2.2. Other Requirements

Further information on the functional and non-functional requirements is documented and can be found in appendices section, Initial Requirement Specification.

5. Data Set and Data Extraction

5.1. Description

The data was collected from the web site, the Tennis-Data Betting, Results and Livescores Portal (2014)[15]. It was downloaded in three separate files in Microsoft excel format (zipped) and contained all matches played by professional tennis players during the period from 2011 to 2013 for both the Men's Association of Tennis Professionals (ATP) and the Women's Tennis Association (WTA).

The dataset consists of 38 variables for the past three years of all the tournaments from the men's ATP and women's WTA tours. There are 365 tournaments in total and 14,220 matches for 2011, 2012 and 2013.

- For the analysis carried out in R the three files were amalgamated. This consisted of 15,230 records.
- There were 21 variables used for the R programming analysis.
  - ATP = Tournament number (men) and Tournament number (women)
  - Gender = male =1 and female = 2
  - Location = Venue of tournament
  - Tournament = Name of tounament (including sponsor if relevant)
  - Data = Date of match (note: prior to 2003 the date shown for all matches played in a single tournament is the start date)
  - Series = Name of ATP tennis series and WTA tennis series
  - Court = Type of court (outdoors or indoors)

[15] Tennis Betting, Results & Livescores Portal (2014). The Tennis-Data Betting, Results and Livescores Portal. Available at: http://www.tennis-data.co.uk/alldata.php [Accessed 10th February 2014].

  - » Surface = Type of surface (clay, hard, carpet or grass)
  - » Round = Round of match
  - » Best of = Maximum number of sets playable in match
  - » Winner = Match winner
  - » Loser = Match loser
  - » WRank = ATP Entry ranking of the match winner as of the start of the tournament
  - » LRank = ATP Entry ranking of the match loser as of the start of the tournament
  - » WPts = ATP Entry points of the match winner as of the start of the tournament
  - » LPts = ATP Entry points of the match loser as of the start of the tournament
  - » Wsets = Number of sets won by match winner
  - » Lsets = Number of sets won by match loser
  - » Comment = Comment on the match (Completed, won through retirement of loser, or via Walkover)
  - » MaxW= Maximum odds of match winner (as shown by Oddsportal.com)
  - » MaxL= Maximum odds of match loser (as shown by Oddsportal.com)

- For the analysis carried out in Weka there was a new variable added called 'Winner Rank' and the dataset was reduced to include 10 variables.
  - » Gender = male =1 and female = 2
  - » Data = Date of match
  - » Series = Name of ATP tennis series and WTA tennis series
  - » Court = Type of court (outdoors or indoors)
  - » Surface = Type of surface (clay, hard or grass)
  - » Round = Round of match
  - » Comment = Comment on the match (Completed, won through retirement of loser, or via Walkover)
  - » MaxW = Maximum odds of match winner (as shown by Oddsportal.com)
  - » MaxL= Maximum odds of match loser (as shown by Oddsportal.com)
  - » Winner Rank = comparison between WRank = ATP Entry ranking of the match winner as of the start of the tournament and LRank = ATP Entry ranking of the match loser as of the start of the tournament
  - » Winner Point = comparison between WPts = ATP Entry points of the match winner as of the start of the tournament and LPts = ATP Entry points of the match loser as of the start of the tournament

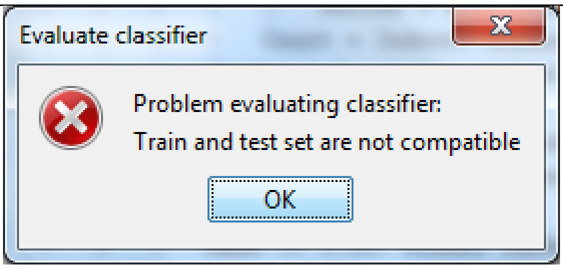
6. Data Cleansing

There were three files each for the men's ATP and the women's WTA ranging from 2011 to 2013 respectively. Prior to amalgamating both the men's and women's datasets and deleting any variable the files had a number of missing data points. The following issues where encountered:

6.1. Issues

- The women's tennis dataset was missing 15 data points in LRank and WRank variable out of 7,317 records. These values were compensated by checking the corresponding year on the Women's Tennis Association website for the correct ranking for the players.
- There were also 15 instances of missing data points for the LWPts and WPts variable. The true values was found in their archive section on the Women's Tennis Association website
- In the MaxW and MaxL variable there were 14 missing data points the values were replaced by estimating the average mean value, 2.02 and 4.23 respectively.
- The men's tennis dataset had 2 missing data points in WRank, 13 in LRank, 1 in WPts and 12 in LPts out of 7,913 records. These values were compensated by checking the corresponding year and tournament on the Men's Official Emirates *ATP Rankings* for the correct ranking for the players and winner or loser points.
- In the MaxW and MaxL variable there were 15 missing data points the values were replaced by estimating the average mean value, 2.01 and 5.27 respectively.

- There were missing data points in the following variables for both the men and women's datasets, W1 & L1 for men 48 and women 49, W2 & L2 for men 136 for women 132, W3 & L3 for men 4,282 for women 5,034. Because in some tournaments the men play best-of-five-sets matches and women take on best-of-three sets, there was no relevance for the analysis to have these variables in the dataset so they were removed.
- In Weka there was a problem when loading the test dataset it gave the following error message "training and test set are not compatible", it was shown 1,013 records instead of 5,054.


Evaluate classifier

Problem evaluating classifier:
Train and test set are not compatible

Figure 4: Error in Decision Tree Classifier

- This was due to how Weka converts csv files into ARFF files for the analysis. The structure of the training and test sets has to be exactly the same (same number of attributes, same type and in the case of categorical attributes, same number of labels and order).

Pre change attribute list:
@attribute Comment {Completed,Retired,Walkover,retired}
@attribute MaxW numeric
@attribute MaxL numeric
@attribute 'Winner Rank' {'winner ranked lower','winner ranked higher'}

Corrected attribute list -Test Set:
@attribute Comment {Completed,Retired,Walkover,Disqualified,Fin,R_P,R_Mo}
@attribute MaxW numeric
@attribute MaxL numeric
@attribute 'Winner Rank' {'winner ranked higher','winner ranked lower'}

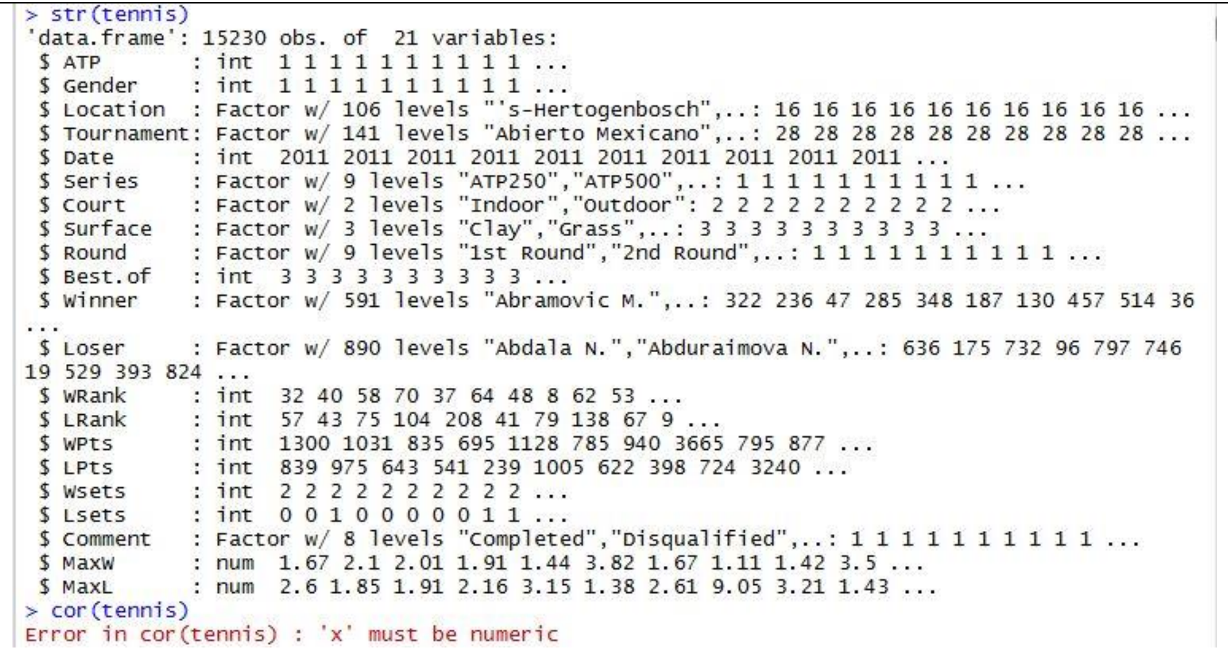
The reason for the issue was that the categorical attributes, Comment - had an extra label 'retired' with a lower case letter in the test set and the attribute 'Winner Rank' the ordering of the labels was different. This was solved by saving down the ARFF files and manually changing the attribute structure so that they matched.

7. Data Analysis Methodology and Results

7.1. Visualisation

In order to get an visual of the elements in the data, the dataset was imported into R studio, using the **str()** function it showed that the tennis data was a data frame with 21 variables and 15,230 observations. From this function it gave a visual of the internal structure of the datasets that is; the type for each of the 21 variables. From the 21 variables a test was used to see if there was a correlation between the variables.

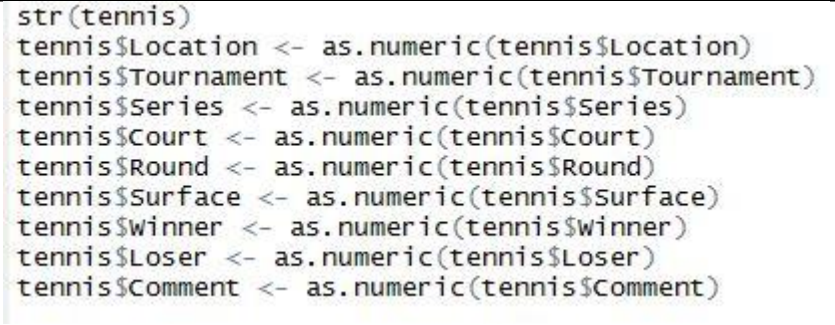
Upon using the **cor(tennis)** function it failed to give an output as can be seen in Figure 5 an error occurred because a number of the independent variables were not numeric.



```
> str(tennis)
'data.frame':	15230 obs. of  21 variables:
 $ ATP       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Gender    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Location  : Factor w/ 106 levels "'s-Hertogenbosch",..: 16 16 16 16 16 16 16 16 16 16 ...
 $ Tournament: Factor w/ 141 levels "Abierto Mexicano",..: 28 28 28 28 28 28 28 28 28 28 ...
 $ Date      : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
 $ Series    : Factor w/ 9 levels "ATP250","ATP500",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Court     : Factor w/ 2 levels "Indoor","Outdoor": 2 2 2 2 2 2 2 2 2 2 ...
 $ Surface   : Factor w/ 3 levels "Clay","Grass",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Round     : Factor w/ 9 levels "1st Round","2nd Round",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Best.of   : int  3 3 3 3 3 3 3 3 3 3 ...
 $ Winner    : Factor w/ 591 levels "Abramovic M.",..: 322 236 47 285 348 187 130 457 514 36
...
 $ Loser     : Factor w/ 890 levels "Abdala N.","Abduraimova N.",..: 636 175 732 96 797 746
19 529 393 824 ...
 $ WRank     : int  32 40 58 70 37 64 48 8 62 53 ...
 $ LRank     : int  57 43 75 104 208 41 79 138 67 9 ...
 $ WPts      : int  1300 1031 835 695 1128 785 940 3665 795 877 ...
 $ LPts      : int  839 975 643 541 239 1005 622 398 724 3240 ...
 $ Wsets     : int  2 2 2 2 2 2 2 2 2 2 ...
 $ Lsets     : int  0 0 1 0 0 0 0 0 1 1 ...
 $ Comment   : Factor w/ 8 levels "Completed","Disqualified",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ MaxW      : num  1.67 2.1 2.01 1.91 1.44 3.82 1.67 1.11 1.42 3.5 ...
 $ MaxL      : num  2.6 1.85 1.91 2.16 3.15 1.38 2.61 9.05 3.21 1.43 ...
> cor(tennis)
Error in cor(tennis) : 'x' must be numeric
```

Figure 5: Str() Function to Visualise

To rectify this error the factor variables were changed to numeric as shown in Figure 6.



```
str(tennis)
tennis$Location <- as.numeric(tennis$Location)
tennis$Tournament <- as.numeric(tennis$Tournament)
tennis$Series <- as.numeric(tennis$Series)
tennis$Court <- as.numeric(tennis$Court)
tennis$Round <- as.numeric(tennis$Round)
tennis$Surface <- as.numeric(tennis$Surface)
tennis$Winner <- as.numeric(tennis$Winner)
tennis$Loser <- as.numeric(tennis$Loser)
tennis$Comment <- as.numeric(tennis$Comment)
```

Figure 6: Transformed to Numeric

The function was ran again **cor(tennis).** In order to build a model this function is used to look at the relationships between the x variables (independent) and check for any redundant relationships. The results are too large to display in this document so a brief outline will be given to show where there was intermediate to strong correlation.

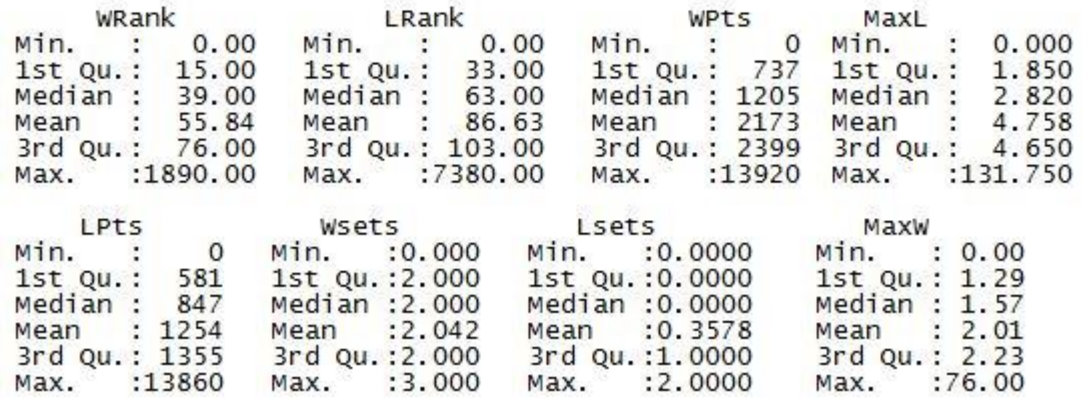
Correlation Coefficient measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the ‘Pearson product moment correlation coefficient’. The correlation coefficient will vary from -1 to +1. A -1 indicates perfect negative correlation, and +1 indicates perfect positive correlation.

From examining the entire dataset the following variables have a measure of association;

- Series (name of the ATP and WTA tennis series) and Gender (male = 1 and female = 2) have a moderate correlation of 0.546369.
- WSets (number of sets won be match winner and Best of (maximum number of sets playable in a match) have a strong correlation of 0.69599.
- WRank (the entry ranking of a match winner as of the start of the tournament) and WPts (the entry points of a match winner as of the start of the tournament) have an intermediate to weak negative correlation of -0.464629.
- MaxL (Maximum odds of match loser - as shown by Oddsportal.com) and WPts which have a moderate correlation of 0.55324.

- Comment (Comment on the match - completed, won through retirement of loser, or via Walkover) and WSets which have a strong negative correlation of -0.663198.

Upon using the function **summary(tennis)** it gave a summary output of the variables. As you can see in Figure 7 there are only 8 variables selected.

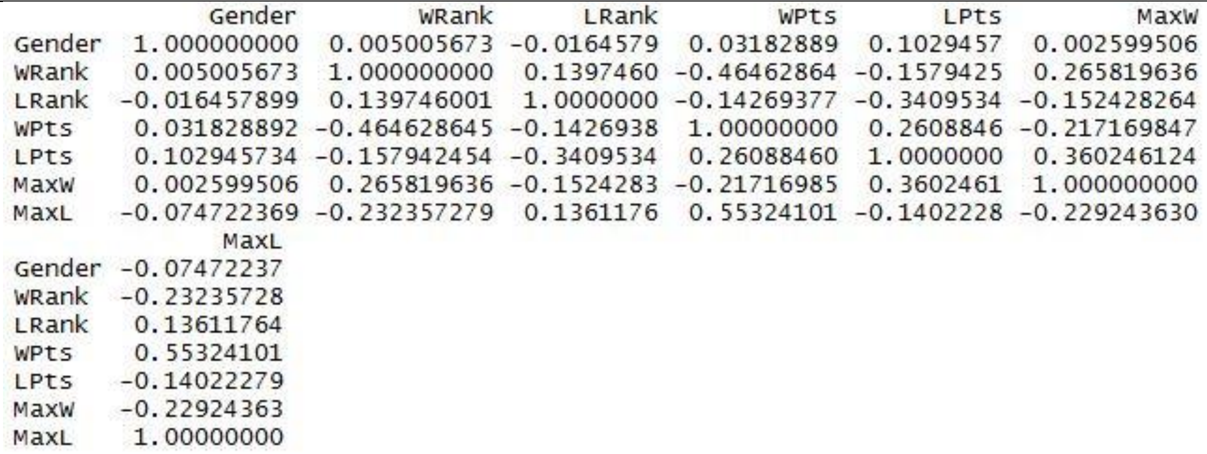


```
     WRank             LRank              WPts            MaxL        
 Min.   :   0.00   Min.   :   0.00   Min.   :    0   Min.   :  0.000  
 1st Qu.:  15.00   1st Qu.:  33.00   1st Qu.:  737   1st Qu.:  1.850  
 Median :  39.00   Median :  63.00   Median : 1205   Median :  2.820  
 Mean   :  55.84   Mean   :  86.63   Mean   : 2173   Mean   :  4.758  
 3rd Qu.:  76.00   3rd Qu.: 103.00   3rd Qu.: 2399   3rd Qu.:  4.650  
 Max.   :1890.00   Max.   :7380.00   Max.   :13920   Max.   :131.750  

      LPts            Wsets           Lsets              MaxW      
 Min.   :    0   Min.   :0.000   Min.   :0.0000   Min.   : 0.00  
 1st Qu.:  581   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.: 1.29  
 Median :  847   Median :2.000   Median :0.0000   Median : 1.57  
 Mean   : 1254   Mean   :2.042   Mean   :0.3578   Mean   : 2.01  
 3rd Qu.: 1355   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.: 2.23  
 Max.   :13860   Max.   :3.000   Max.   :2.0000   Max.   :76.00  
```

Figure 7: Summary data related to the individual objects

As it is not necessary to use all the variables the function **cor()** was used again and included the following variables as shown in Figure 8.

R Codes - cor(tennis[c("Gender","WRank","LRank","WPts","LPts","MaxW","MaxL")])



```
             Gender        WRank      LRank        WPts       LPts         MaxW
Gender  1.000000000  0.005005673 -0.0164579  0.03182889  0.1029457  0.002599506
WRank   0.005005673  1.000000000  0.1397460 -0.46462864 -0.1579425  0.265819636
LRank  -0.016457899  0.139746001  1.0000000 -0.14269377 -0.3409534 -0.152428264
WPts    0.031828892 -0.464628645 -0.1426938  1.00000000  0.2608846 -0.217169847
LPts    0.102945734 -0.157942454 -0.3409534  0.26088460  1.0000000  0.360246124
MaxW    0.002599506  0.265819636 -0.1524283 -0.21716985  0.3602461  1.000000000
MaxL   -0.074722369 -0.232357279  0.1361176  0.55324101 -0.1402228 -0.229243630
              MaxL
Gender -0.07472237
WRank  -0.23235728
LRank   0.13611764
WPts    0.55324101
LPts   -0.14022279
MaxW   -0.22924363
MaxL    1.00000000
```

Figure 8: Correction Model

- The correlation between the ranking of the match winner at the start of the tournament and the entry points of the match winner at the start of the tournament is -0.4646629, (WPts and WRank). This result indicates a fairly moderate negative linear relationship between these two variables, as the value of one variable increases the value of the other variable decreases. This would make sense as you earn more points you ranking decreases.
- The correlation between the Maximum odds of match loser (as shown by Oddsportal.com) and the entry points of the match winner as of the start of the tournament is 0.55324, (MaxL and WPts). This result indicates a fairly strong positive linear relationship between these two variables, as the value of one variable increases the value of the other variable increases

An example of this is in the Australian Open in 2011; Roger Federer entered the tournament with 9,245 points the maximum odds for his opponent Lucas Lacko was 43 (MaxL), the match loser.

Similarly Andy Roddick in the same tournament entered with 3,565 points and the maximum odds for his opponent Jan Hajek was 39.75 (MaxL), the match loser. The more points a player has the less likely the opponent is going to win therefore the betting odds are higher.

Figure 9 visual depicts the relationship between these 7 variables in particular WPts and WRank, MaxL and WPts.

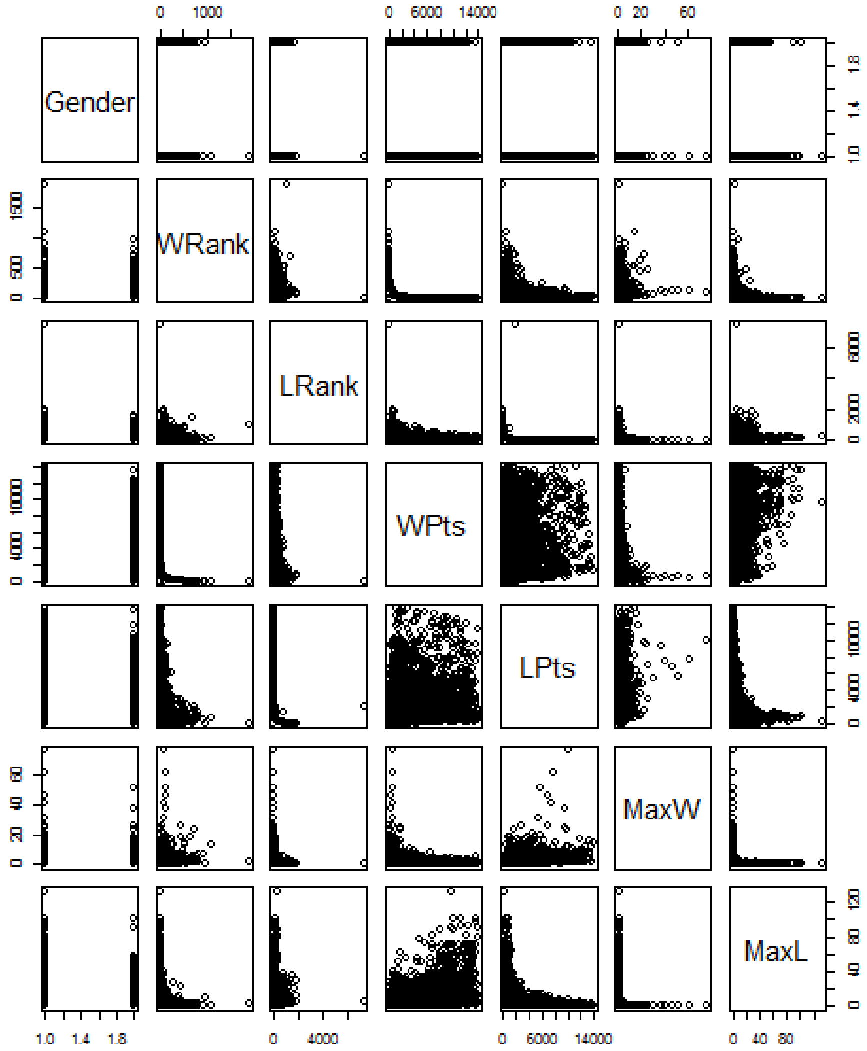


Figure 9: Visual of Data in R

Exploring the dataset further a test was carried out using a multiple regression model. The ‘Winner’ variables was used as the dependent and WRank, LRank, WPts, LPts, MaxW and MaxL as the independent variables.

As shown in Figure 10, most (95 percent) of the standardised residuals fall within two standard deviations of the mean, which in this case is –2 to +2.There should be more residuals hovering around zero and there should be fewer and fewer of the residuals as they go away from zero. If the residuals fall in a straight line that means the normality condition is met as it is shown in Figure 11, the conditions has not quite met.

In Figure 12 the residuals output provides a summary statistic for the errors in our prediction. The median value should be close to 0 (in this case it is 7.72)

- Since a residual is equal to the true value minus the predicted value, the maximum error of 417.83 suggests that the model is under-predicting.

- On the other hand, 50 percent of errors fall within the 1Q and 3Q values (the first and third quartile), so the majority of predictions were between -149.70 over the true value and 151.82 under the true value

- The stars (for example, ***) indicate the predictive power of each feature in the model. The significance level (as listed by the ‘Signif. Codes’ in the footer) provides a measure of how likely the true coefficient is zero given the value of the estimate.

- The presence of three stars indicates a significance level of 0, which means that the feature is extremely unlikely to be unrelated to the dependent variable.

- The common practice is to use a significance level of 0.05 to denote a statistically significant variable. The model in this case show WRank as the only variable with statistically significant, it indicates that our features are not very predictive of the outcome.

- The Multiple R-squared value (also called the coefficient of determination) provides a measure of how well our model as a whole explains the values of the dependent variable.

- It is similar to the correlation coefficient in that the closer the value is to 1.0, the better the model perfectly explains the data. In this case R-squared value is 0.002175, which tells us that the model is not performing well.

- There is also evidence of outliers present these are considered bad data points.

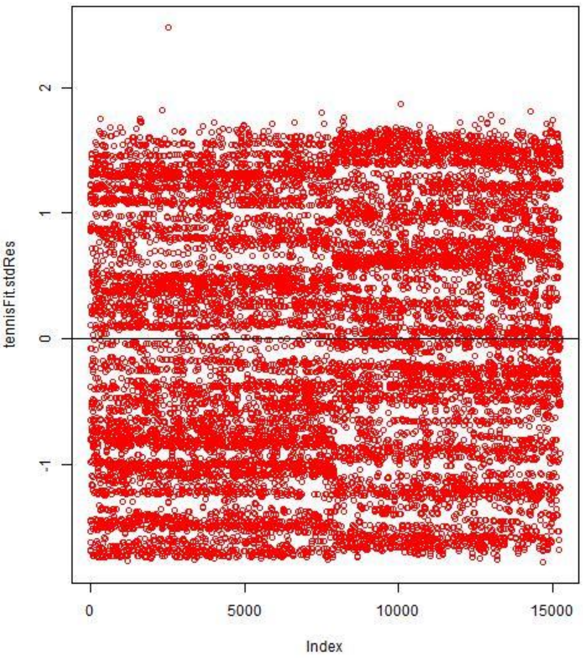


Figure 10: Model Fit Residuals

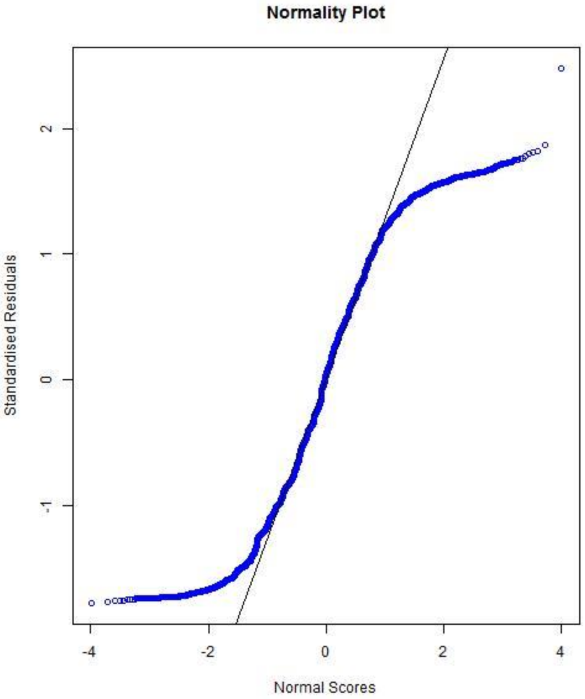
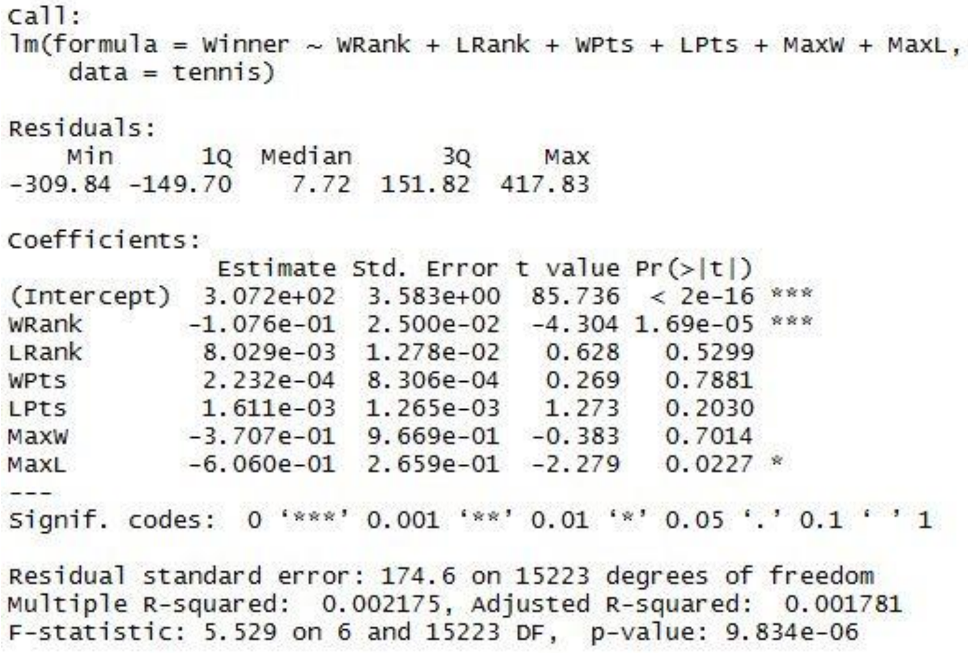


Figure 11: Normality plot in R



```
Call:
lm(formula = Winner ~ WRank + LRank + WPts + LPts + MaxW + MaxL,
    data = tennis)

Residuals:
    Min      1Q  Median      3Q     Max 
-309.84 -149.70    7.72  151.82  417.83 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.072e+02  3.583e+00  85.736  < 2e-16 ***
WRank       -1.076e-01  2.500e-02  -4.304 1.69e-05 ***
LRank        8.029e-03  1.278e-02   0.628   0.5299    
WPts         2.232e-04  8.306e-04   0.269   0.7881    
LPts         1.611e-03  1.265e-03   1.273   0.2030    
MaxW        -3.707e-01  9.669e-01  -0.383   0.7014    
MaxL        -6.060e-01  2.659e-01  -2.279   0.0227 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 174.6 on 15223 degrees of freedom
Multiple R-squared:  0.002175, Adjusted R-squared:  0.001781 
F-statistic: 5.529 on 6 and 15223 DF,  p-value: 9.834e-06
```

Figure 12: Multiple Regression Model (Dependent variable: Winner)

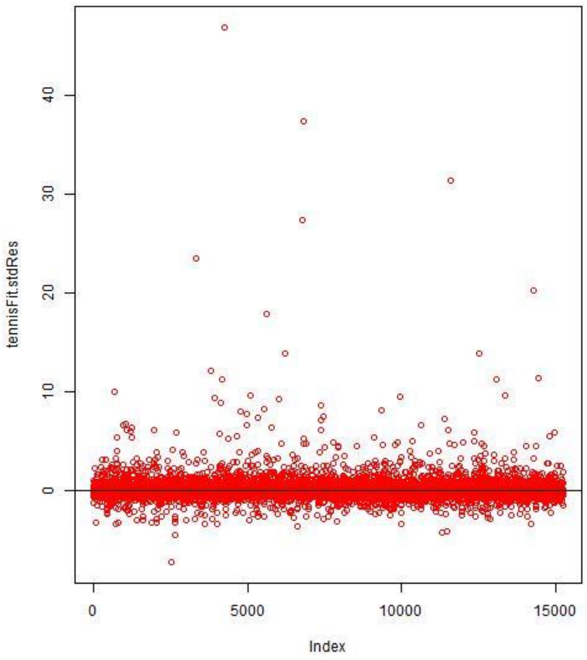


Figure 13: Model Fit Residuals

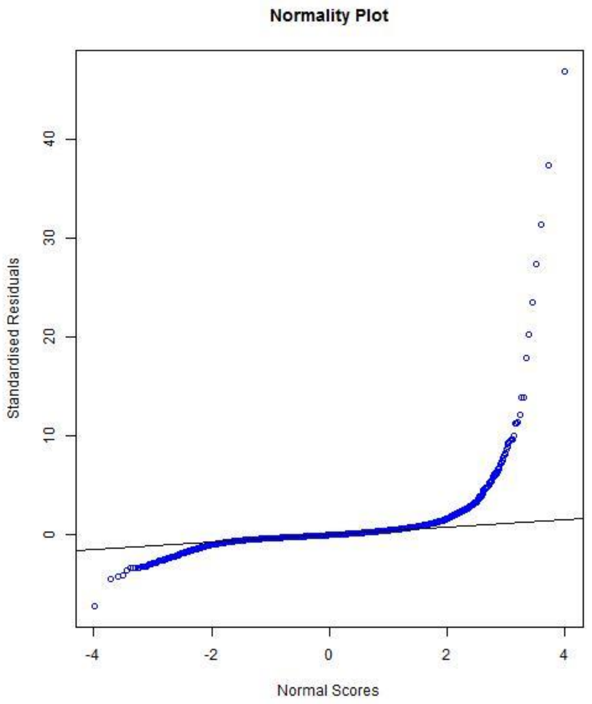
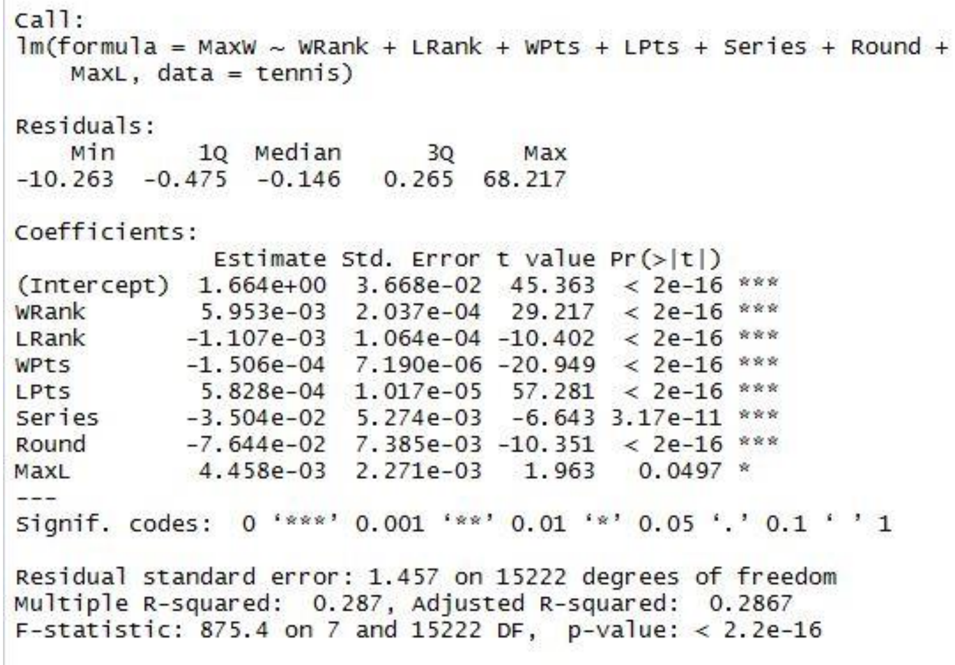


Figure 14: Normality plot in R



```
Call:
lm(formula = MaxW ~ WRank + LRank + WPts + LPts + Series + Round +
    MaxL, data = tennis)

Residuals:
    Min      1Q  Median      3Q     Max
-10.263  -0.475  -0.146   0.265  68.217

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.664e+00  3.668e-02  45.363  < 2e-16 ***
WRank        5.953e-03  2.037e-04  29.217  < 2e-16 ***
LRank       -1.107e-03  1.064e-04 -10.402  < 2e-16 ***
WPts        -1.506e-04  7.190e-06 -20.949  < 2e-16 ***
LPts         5.828e-04  1.017e-05  57.281  < 2e-16 ***
Series      -3.504e-02  5.274e-03  -6.643 3.17e-11 ***
Round       -7.644e-02  7.385e-03 -10.351  < 2e-16 ***
MaxL         4.458e-03  2.271e-03   1.963   0.0497 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.457 on 15222 degrees of freedom
Multiple R-squared:  0.287,	Adjusted R-squared:  0.2867
F-statistic: 875.4 on 7 and 15222 DF,  p-value: < 2.2e-16
```

Figure 15: Multiple Regression Model (Dependent variable: MaxL)

A multiple regression model was carried out using MaxW as the dependent variable and WRank, LRank, WPts, LPts, Series, Round and MaxL as the independent variables. As shown in Figure 13, most (95 percent) of the standardised residuals fall within two standard deviations of the mean, which in this case is –10 to +10. In this case we see more residuals hovering around zero and we there is fewer and fewer of the residuals as they go away from zero. In Figure 14, the residuals do not fall in a straight line which means the normality condition is not met.

In Figure 10 the residuals output provides a summary statistic for the errors in our prediction. The median value should be close to 0 (in this case it is -0.146 which is acceptable)

- Since a residual is equal to the true value minus the predicted value, the maximum error of 68.217 suggests that the model is under-predicting.

- 50 percent of errors fall within the 1Q and 3Q values (the first and third quartile), so the majority of predictions are between -0.475 over the true value and 0.265 under the true value

- The presence of three stars indicates a significance level of 0, which means that the feature is extremely unlikely to be unrelated to the dependent variable.

- The common practice is to use a significance level of 0.05 to denote a statistically significant variable. The model in this case show all variables, WRank, LRank, WPts, LPts, Series, Round and MaxL with statistically significant, it indicates that our features are very predictive of the outcome.

- However the Multiple R-squared value (also called the coefficient of determination) In this case R-squared value is 0.287, which tells us that the model is not performing well (the closer the value is to 1.0, the better the model perfectly explains the data).

- There is also evidence of outliers present which is a data point that does not fit the general trend of the data, outliers are considered bad data points.

7.1. Machine Learning Algorithm

For the purpose of this project there was one datasets to carry out statistical analysis using a machine learning algorithm, the tennis dataset is **2011_2012_2013_men_women.csv** saved in Microsoft Excel as Comma Separated Values File (.csv). For the years 2011, 2012, and 2013, the total number of instances captured is **15,230** with **10** classes used for this project.

In this section it sets out the algorithm which was tested, the decision trees classifier. The paper will present the performance of this classifier, how effective WEKA is at classifying a result and try to come up with a model that can predict the probable ‘Winner Rank’ using the selected variables in the dataset.

7.2. Decision Trees

The statistical tool used for the machine learning classification process is Weka. The dataset was examined **2011_2012_2013_men_women.csv** and a new categorical variable was created, ‘Winner Rank’ to use for the prediction of the model. This variable compared the two existing variables; winner rank (WRank) and loser rank (LRank), if WRank is less than WRank the winner is ranked higher. A tennis player performance is ranked according to a numerical system whereby the lower the number is the higher prestige that player has and the lower the betting odds assigned to that player. By comparing these two variables WRank and LRank it was clear to identify in the new variable ‘Winner Rank’ whether the winner was rank higher or lower.

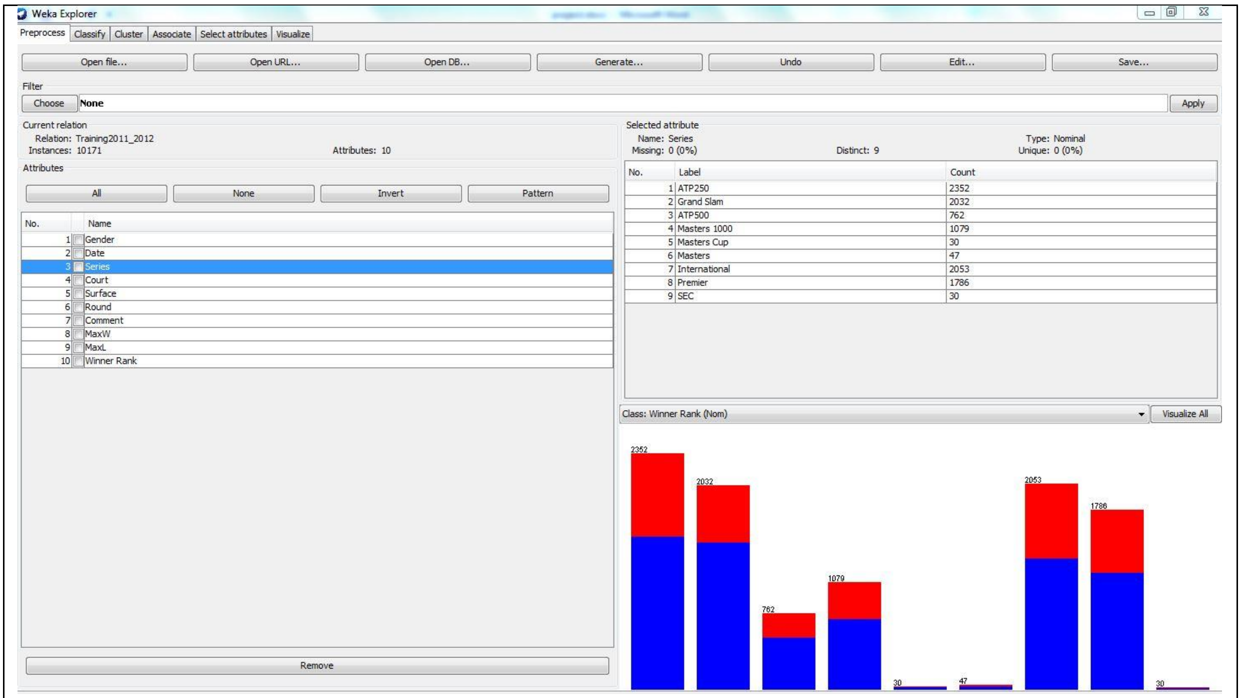
One important concept of the classification tree is the concept of using a "training set" to produce the model. The entire training set was taken and divided it into two parts:

Training2011_2012.csv and **Test2013.csv** the latter consisting of 50 percent of the data.

The training set was tested first, **Training2011_2012.csv** with **10,171** instances and used to create the model and then the remaining data **Test2013.csv** with **5,059** instances was put it into a test set, which was used immediately after creating the model to test the accuracy of the model, this is done to check and ensure that the accuracy of the model built doesn't decrease with the test set

To begin processing the data, the training set was imported into the pre-process panel. In Figure 16 the diagram shows the name of the file, the number of instances and the number of attributes (descriptors + class). It is shown also in this diagram on the left side of the frame the number of instances which is **10,171**, whereas the number of descriptors is 10.

The **Attributes** frame allows user to modify the set of attributes using select and remove options. Information about the selected attribute is given in the **selected attribute** frame in which a histogram depicts the attribute distribution. One can see that the value of the currently selected descriptor ‘Series’ shows the distribution of the attribute values in the dataset. Also you may see that the number of missing, unique and distinct values in the **selected attribute** frame.



Filter: None

Current relation
Relation: Training2011_2012
Instances: 10171
Attributes: 10

Attributes

No.		Name
1	□	Gender
2	□	Date
3	□	Series
4	□	Court
5	□	Surface
6	□	Round
7	□	Comment
8	□	MaxW
9	□	MaxL
10	□	Winner Rank

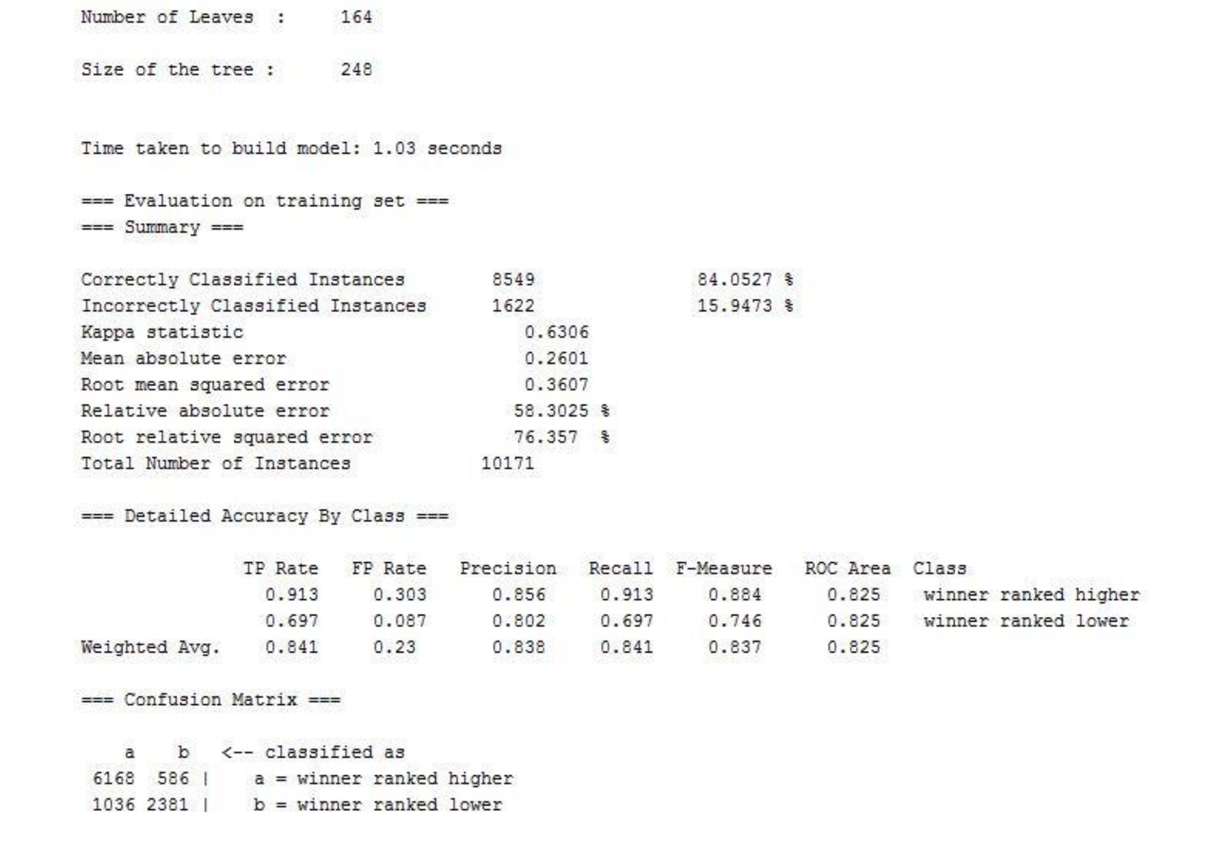
Selected attribute
Name: Series
Missing: 0 (0%)
Distinct: 9
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	ATP250	2352
2	Grand Slam	2032
3	ATP500	762
4	Masters 1000	1079
5	Masters Cup	30
6	Masters	47
7	International	2053
8	Premier	1786
9	SEC	30

Class: Winner Rank (Nom)

Figure 16: Training dataset visual in pre-process panel

In order to use the data in the experiment the next step of the process is in the ‘Classify’ panel where there is an option to select the procedure for the analysis. In this case the **trees** node was select, followed by the **J48** leaf and finally the **Training Set** in the test option section. The output of the model is shown in Figure 17.



```
Number of Leaves  :     164

Size of the tree :      248


Time taken to build model: 1.03 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        8549               84.0527 %
Incorrectly Classified Instances      1622               15.9473 %
Kappa statistic                          0.6306
Mean absolute error                      0.2601
Root mean squared error                  0.3607
Relative absolute error                 58.3025 %
Root relative squared error             76.357  %
Total Number of Instances            10171

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.913     0.303      0.856     0.913     0.884      0.825    winner ranked higher
                 0.697     0.087      0.802     0.697     0.746      0.825    winner ranked lower
Weighted Avg.    0.841     0.23       0.838     0.841     0.837      0.825

=== Confusion Matrix ===

    a    b   <-- classified as
 6168  586 |    a = winner ranked higher
 1036 2381 |    b = winner ranked lower
```

Figure 17: Results of Training set 2011 & 2012

The correctly classified instances show 84.0527% of test instances that were correctly classified (Accuracy) and the incorrectly classified instances show 15.9473% of test instances that were incorrectly classified (Error Rate). This is a good result.

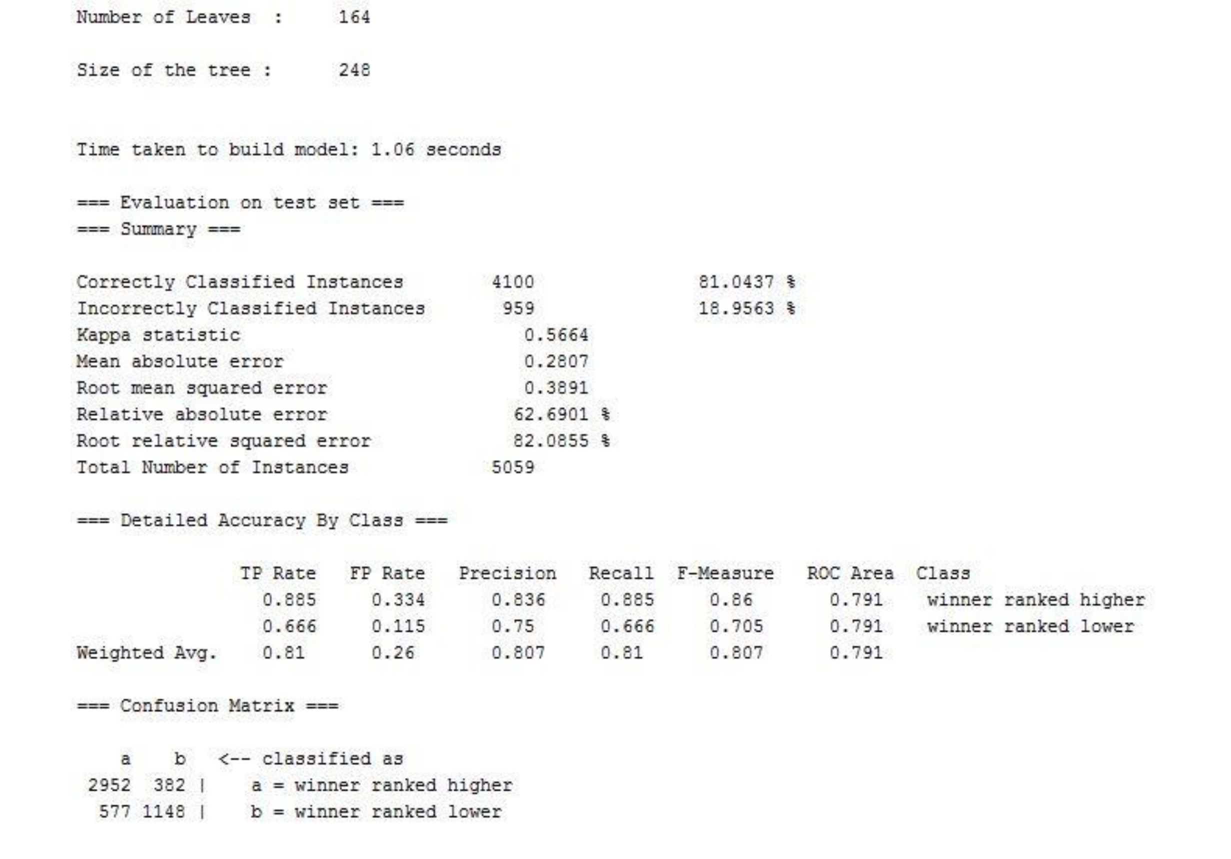
The Contingency table or confusion matrix, where class ‘a’ and ‘b’ represent the number of instances correctly or incorrectly classified. For class ‘a’, 6168 instances were correctly classified but 586 were put into class ‘b’ (false positives). For class ‘b’ 2381 instances were correctly classified but 1036 were put into class ‘a (false negatives)’.

The Confusion Matrix illustrates an imbalance to how the instances are classified and there is some bias towards the ‘winner ranked higher’ class. The "balance" of the dataset needs to be taken into account when interpreting results. Unbalanced datasets in which a disproportionately large amount of instances belong to a certain class may lead to high accuracy rates even though the classifier may not necessarily be particularly good. This fact clearly indicates that the accuracy cannot be used for assessing the usefulness of classification models built using unbalanced datasets. So to satisfy the output further some of the other numbers will be looked at, firstly the ROC Area, or area under the ROC curve. An ‘optimal’ classifier will have ROC area values approaching 1, with 0.5 being comparable to "random guessing" (similar to a Kappa statistic of 0), in this case it is 0.825 which tells us it is not random guessing.

Secondly it was important to look at the Kappa statistic which is a chance-corrected measure of agreement between the classifications and the true classes. It's calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. A value greater than zero means that your classifier is doing better than chance and it is shown from the output the Kappa Statistic is 0.6306, which indicates there is existence of statistical dependence.

This means that this model can assess the value of the probability that the winner rank with particular characteristic can be predicted as ‘winner ranked higher’ and ‘winner ranked lower’, in the nodes and leaves output it shows that ‘MaxW’ as the first predictable variable and ‘Round’ as the second predicable variable.

To validate the classification tree, the next step was to run the test set through the model. In the **Test options**, the user selects the **Supplied test set** radio button and clicks **Set**. The file is then loaded into Weka, **Test2013.arff** which contained **5,059** records. In Figure 18, the output of the test model is shown.



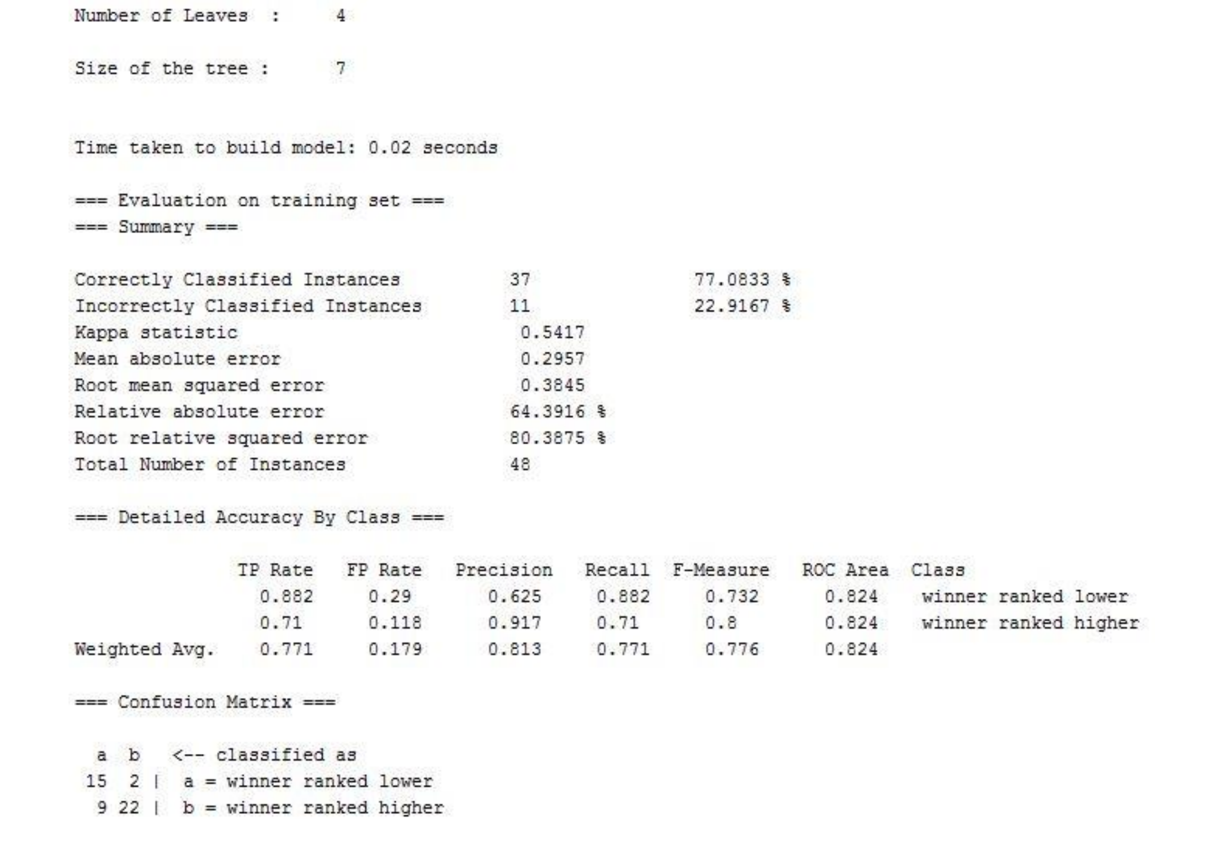
```
Number of Leaves  :     164

Size of the tree :      248


Time taken to build model: 1.06 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances        4100               81.0437 %
Incorrectly Classified Instances       959               18.9563 %
Kappa statistic                          0.5664
Mean absolute error                      0.2807
Root mean squared error                  0.3891
Relative absolute error                 62.6901 %
Root relative squared error             82.0855 %
Total Number of Instances             5059

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.885     0.334      0.836     0.885     0.86       0.791    winner ranked higher
                 0.666     0.115      0.75      0.666     0.705      0.791    winner ranked lower
Weighted Avg.    0.81      0.26       0.807     0.81      0.807      0.791

=== Confusion Matrix ===

    a    b   <-- classified as
 2952  382 |    a = winner ranked higher
  577 1148 |    b = winner ranked lower
```

Figure 18: Results of Test set 2013

In this case the correctly classified instances show 81.0437% of test instances that were correctly classified (Accuracy) and the incorrectly classified instances show 18.9563% of test instances that were incorrectly classified (Error Rate). This also shows a good result.

The Confusion Matrix, where class ‘a’ and ‘b’ represent the number of instances correctly or incorrectly classified. For class ‘a’, 2952 instances were correctly classified but 382 were put into class ‘b’ (false positive). For class ‘b’ 1148 instances were correctly classified but 577 were put into class ‘a’ (false negative). The ROC Area is 0.791 which is a good result and the Kappa Statistic is 0.5664. Comparing the ‘Correctly Classified Instances’ from this test set (81.0437 percent) with the ‘Correctly Classified Instances’ from the training set (84.0527 percent), we see that the accuracy of the model is pretty close.

In order to get a deeper insight into the dataset, the dataset was filtered to include Grand Slam Series only for the semi-finals and finals of 2011, 2012 and 2013. These datasets were saved in Microsoft excel and divided into two parts as **Training2011_2012 Grand Slam Finals Only.csv,** for the training set and **Test2013 Grand Slam Finals Only.csv** for the test set. The training set was used to create the model, and the test set was used to verify that the model was accurate and not overfitting. The output of the training set is shown in Figure 19.



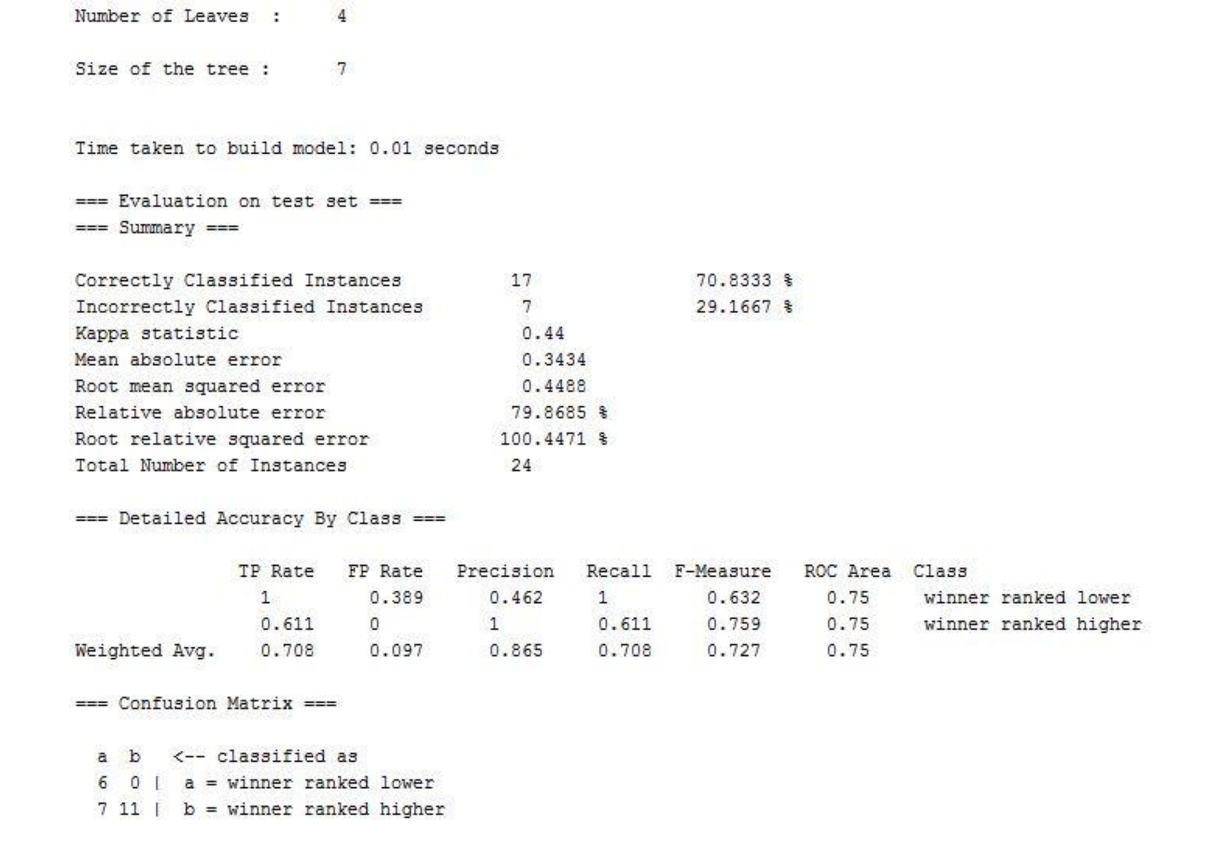
```
Number of Leaves  :     4

Size of the tree :      7


Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          37               77.0833 %
Incorrectly Classified Instances        11               22.9167 %
Kappa statistic                          0.5417
Mean absolute error                      0.2957
Root mean squared error                  0.3845
Relative absolute error                 64.3916 %
Root relative squared error             80.3875 %
Total Number of Instances               48

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.882     0.29      0.625     0.882     0.732      0.824    winner ranked lower
                 0.71      0.118     0.917     0.71      0.8        0.824    winner ranked higher
Weighted Avg.    0.771     0.179     0.813     0.771     0.776      0.824

=== Confusion Matrix ===

  a  b   <-- classified as
 15  2 |  a = winner ranked lower
  9 22 |  b = winner ranked higher
```

Figure 19: Results of Training set for Grand Slam semi-finals and finals 2011 & 2012

In the training set there were 48 records used. The correctly classified instances show 77.0833% of test instances that were correctly classified (Accuracy) and the incorrectly classified instances show 22.9167% of test instances that were incorrectly classified (Error Rate). This also shows a good result.

The Confusion Matrix, where class ‘a’ and ‘b’ represent the number of instances correctly or incorrectly classified. For class ‘a’, 15 instances were correctly classified and only 2 were put into class ‘b’ (false positive). For class ‘b’ 22 instances were correctly classified but 9 were put into class ‘a’ (false negative). The ROC Area is closer to 1 at 0.824 and the Kappa Statistic is 0.5417, which indicates there is existence of statistical dependence. Overall this indicates a good model.

To validate the classification tree, the next step was to run the test set through the model. The file **Test2013 Grand Slam Finals Only.arff** was used with contained 24 records. The output is shown in Figure 20.



```
Number of Leaves  :      4

Size of the tree :       7


Time taken to build model: 0.01 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          17               70.8333 %
Incorrectly Classified Instances         7               29.1667 %
Kappa statistic                          0.44
Mean absolute error                      0.3434
Root mean squared error                  0.4488
Relative absolute error                 79.8685 %
Root relative squared error            100.4471 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.389     0.462     1         0.632      0.75     winner ranked lower
                 0.611     0         1         0.611     0.759      0.75     winner ranked higher
Weighted Avg.    0.708     0.097     0.865     0.708     0.727      0.75

=== Confusion Matrix ===

  a  b   <-- classified as
  6  0 |  a = winner ranked lower
  7 11 |  b = winner ranked higher
```

Figure 20: Results of Test set for Grand Slam semi-finals and finals 2013

In this case the correctly classified instances show 70.8333% of test instances that were correctly classified (Accuracy) and the incorrectly classified instances show 29.1667% of test instances that were incorrectly classified (Error Rate). This also shows a good result.

The Confusion Matrix, where class ‘a’ and ‘b’ represent the number of instances correctly or incorrectly classified. For class ‘a’ all instances were correctly classified. For class ‘b’ 11 instances were correctly classified but 7 were put into class ‘a’ (false negative). The ROC Area is close to 1 at 0.75 and the Kappa Statistic is 0.5664. Comparing the ‘Correctly Classified Instances’ from this test set (70.8333 percent) with the ‘Correctly Classified Instances’ from the training set (77.0833 percent), we see that the accuracy of the model is pretty close.

In Figure 21 the two predicable values are MaxW and Round. Using the variable MaxW is a good predictor when classifying a higher ranking player when the betting odds are less than or equal to 1.68 for the semi-finals and when the betting odds are greater than 1.68 it is probable that the lower ranked player will lose the match.

```
MaxW <= 1.68
|   Round = Semifinals: winner ranked higher (18.0/1.0)
|   Round = The Final
|   |   MaxW <= 1.23: winner ranked lower (2.0)
|   |   MaxW > 1.23: winner ranked higher (6.0/1.0)
MaxW > 1.68: winner ranked lower (22.0/9.0)
```

It has been proven that the decision tree is valid and has given a good result and that there is satisfactory result to the research question whereby there statistical evidence to show that by analysing three years of data, 2011, 2012, 2013 that the probable variables, MaxW and Round will predict and determine the ‘winner rank’.

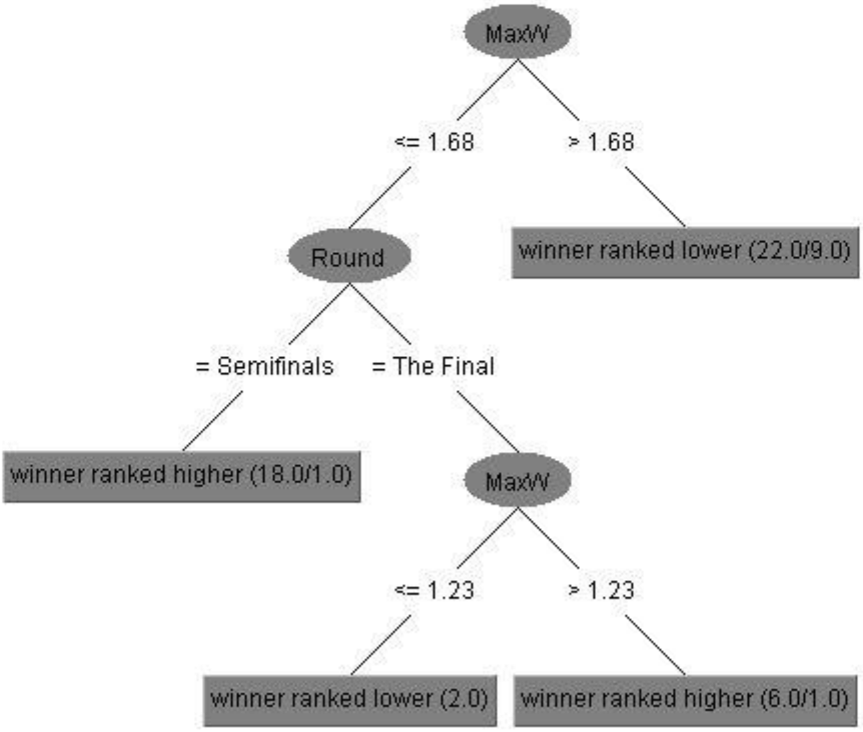

MaxW
<= 1.68
> 1.68
Round
winner ranked lower (22.0/9.0)
= Semifinals
= The Final
winner ranked higher (18.0/1.0)
MaxW
<= 1.23
> 1.23
winner ranked lower (2.0)
winner ranked higher (6.0/1.0)

Figure 21: Decision Tree: Training & Test set for Grand Slam semi-finals and finals 2011, 2012, 2013

Do the higher ranking players typically win the match?

Also from the above analysis using the variable 'Winner Rank', it is shown in Figure 22 that it is 66 per cent (10,088) probable that the higher ranking player typically wins the match. There is only a 34 per cent (5,142) probability that the lower ranked player will lose the match.

Do higher ranking players typically outperform the lower ranking players on the different surface type, clay, grass, hard?

Upon using the same variable again, 'Winner Rank' in Figure 23, it is shown that the higher ranking player outperform the lower ranking player on the different surface types.

- On clay surface 65 per cent (2,992) of higher ranked player compared to 35 per cent (1,584) of lower ranking players
- On grass surface 65 per cent (1,059) of higher ranking players outperform the lower rank players to 35 per cent (575).
- On hard surface there was 67 per cent (6,037) of higher ranking players compared to 33 per cent (2,983) lower ranking players.

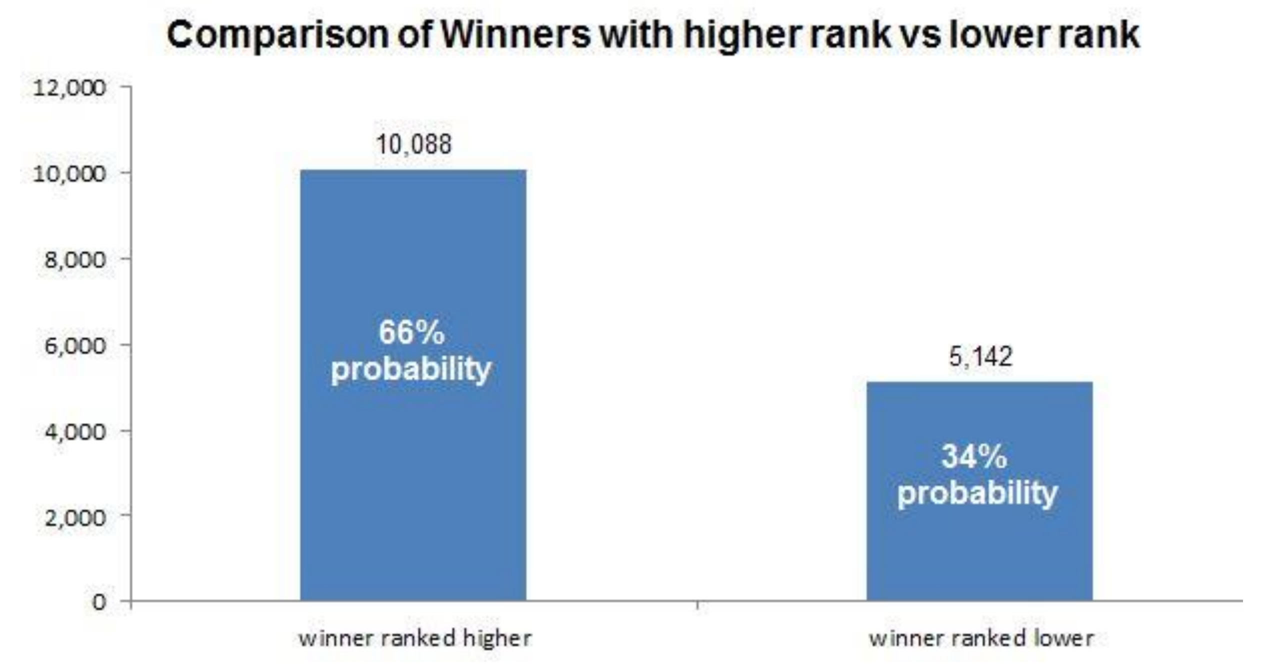


Figure 22: Comparison of Winners with higher rank vs lower rank

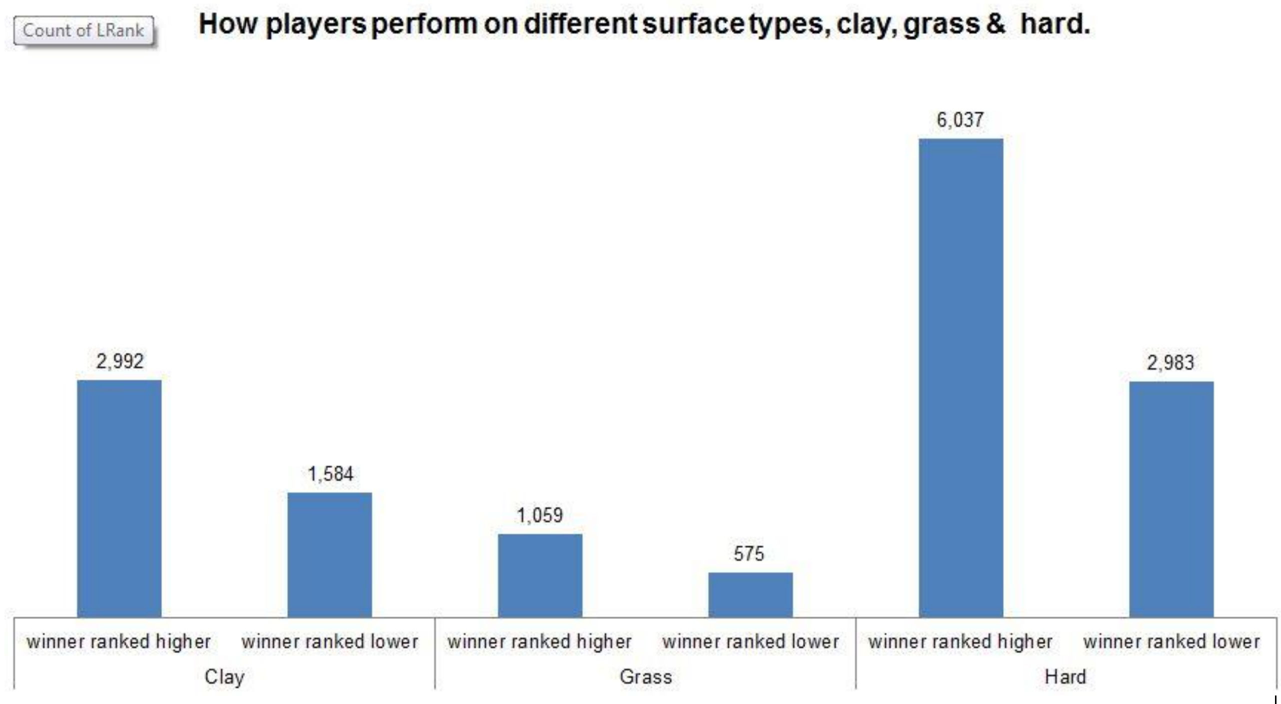


Figure 23: How players perform on different surface types, clay, grass & hard

8. Conclusion

The output for the dataset with 15,230 records showed a good result. The ROC Area is close to close which means that it was not random guessing and the Kappa Statistic showed strong statically evidence which indicates there is existence of statistical dependence. Comparing the 'Correctly Classified Instances' from this test set (81.0437 percent) with the 'Correctly Classified Instances' from the training set (84.0527 percent), illustrates that the accuracy of the model is pretty close.

This means that this model can assess the value of the probability that the winner rank with particular characteristic can be predicted as ‘winner ranked higher” and ‘winner ranked lower’ and in the nodes and leaves output it shows that ‘MaxW’ as the first predictable variable and ‘Round’ as the second predicable variable. This answers one of the research questions whereby could there be a model that could to predict the probable variable that would determine the ‘winner rank’.

The first question asks if the higher ranked players typically win the match. Figure 22 diagram shows the percentage output, 66% for the higher ranked players compared to 34% for the lower players.

The second question address the surface types and do the higher ranked players typically outperform the lower ranked players on the different surfaces. It is shown in Figure 23 overall it the results are favourable to the higher ranked players particularly strong evident is shown on the hard surface.

To summarise the results show a success to the data analysis conducted in for this paper, all the research questions were answered and sound statistical evidence was produced to support the findings.

9. Further Development

Although the project has achieved the goals and objects set out, further development using other attributes such as the percentage of unforced error during the four grand slams in the quarter, semi and final rounds and see if there is a comparison between the men and women. This could draw on the independence assumption as in Klaassen and Magnus[16] paper where they tested the independent and identically distributed hypothesis and rejected it.

It would be interesting to look at some common factors that revolt this assumption, does the dependence come from the ‘psychological momentum’ or is a player on top of his/her game all the time or are there other factors such as the prestige of the tournament? Where Paserma[17] stated in his paper men and women respond differently to competitive pressure in a real-world setting with large monetary rewards.

10. References

Official Emirates *ATP Rankings (2014). ATP World Tour.* Available at: http://www.atpworldtour.com/Rankings/Singles.aspx [Accessed 20 May 2014].

Samson, M. (2012) *ATP Men’s Tennis Rankings Explained.* Available at: http://grandslamgal.com/atp-mens-tennis-rankings-explained/ [Accessed at 20 May 2014].

WTA Tennis (2014). Women’s Tennis Association. Available at: http://www.wtatennis.com/all-about-rankings [Accessed 20 May 2014].

[16] Klaassen F.J.G.M. and Magnus J.R. (2001) “Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model” *Journal of the American Statistical Association* Volume 96, Issue 454, pp. 500-509. Available at: http://amstat.tandfonline.com/doi/abs/10.1198/016214501753168217#.U3yPbyhzQ68 [Accessed 14 May 2014].

[17] Paserma, D.M. (2010) “Gender Differences in Performance in Competitive Environments? Evidence from Professional Tennis Players” *Boston University and Hebrew University.* Available at: http://people.bu.edu/paserman/papers/Paserman_Tennis_January2010.pdf [Accessed 12 May 2014].

R (1993). The R Project for Statistical Computing. Available at: http://www.r-project.org/ [Accessed 30 April 2014].

R (1993). The Comprehensive R Archive Network. Available at: http://cran.r-project.org/ [Accessed 30 April 2014].

Weka (1993). Machine Learning Group at the University of Waikato. Available at: http://www.cs.waikato.ac.nz/ml/weka/ [Accessed 25 April 2014].

Jackson, D. and Mosurski, K. (1997) “Heavy Defeats in Tennis: Psychological Momentum or Random Effect?” *Springer International,* Volume 10(2), pp. 27-34. Available at: http://www.tandfonline.com/doi/abs/10.1080/09332480.1997.10542019#.U3njKyhzQ68 [Accessed 10 May 2014].

Madurska M. (2012) “A Set-by-Set Analysis Method for Predicting the Outcome of Professional Singles Tennis Matches” *Imperial College London,* pp.13. Available at: http://www.doc.ic.ac.uk/teaching/distinguished-projects/2012/a.madurska%20.pdf [Accessed 11 May 2014].

Madurska M. (2012) “A Set-by-Set Analysis Method for Predicting the Outcome of Professional Singles Tennis Matches” *Imperial College London,* pp.13. Available at: http://www.doc.ic.ac.uk/teaching/distinguished-projects/2012/a.madurska%20.pdf [Accessed 11 May 2014].

Paserma, D.M. (2010) “Gender Differences in Performance in Competitive Environments? Evidence from Professional Tennis Players” *Boston University and Hebrew University.* Available at: http://people.bu.edu/paserman/papers/Paserman_Tennis_January2010.pdf [Accessed 12 May 2014].

Johnson V.D. (2012) “Men’s Tennis Is More Interesting Than Women’s”, *Boston Review,* 6 July. Available at: http://www.bostonreview.net/us/men%E2%80%99s-tennis-more-interesting-women%E2%80%99s-david-johnson [Accessed 19 May 2014].

Panahi R. (2014)” For equal pay, women must play best of five”, Herald Sun, 21 January. Available at: http://www.heraldsun.com.au/news/opinion/for-equal-pay-women-must-play-best-of-five/story-fni0fhh1-1226806188706 [Accessed 19 May 2014].

Spanias D. A. and Knottenbelt W. B. (2013) “Tennis Player Ranking using Quantitative Models” *Department of Computing, Imperial College London.* Available at: http://www.doc.ic.ac.uk/~wjk/publications/spanias-knottenbelt-mis-2013.pdf [Accessed 12 May 2014].

Klaassen F.J.G.M. and Magnus J.R. (2001) “Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model” *Journal of the American Statistical Association* Volume 96, Issue 454, pp. 500-509. Available at: http://amstat.tandfonline.com/doi/abs/10.1198/016214501753168217#.U3yPbyhzQ68 [Accessed 14 May 2014].

Tennis Betting, Results & Livescores Portal (2014). The Tennis-Data Betting, Results and Livescores Portal. Available at: http://www.tennis-data.co.uk/alldata.php [Accessed 10th February 2014].

Klaassen F.J.G.M. and Magnus J.R. (2001) “Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model” *Journal of the American*

Statistical Association Volume 96, Issue 454, pp. 500-509. Available at:http://amstat.tandfonline.com/doi/abs/10.1198/016214501753168217#.U3yPbyhzQ68 [Accessed 14 May 2014].

Paserma, D.M. (2010) “Gender Differences in Performance in Competitive Environments? Evidence from Professional Tennis Players” *Boston University and Hebrew University.* Available at: http://people.bu.edu/paserman/papers/Paserman_Tennis_January2010.pdf [Accessed 12 May 2014].

11. Appendices

11.1. Weka Model

11.1.1. Training & Test for Grand Slam semi-final and finals output

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Training2011_2012 Grand Slam Finals Only
Instances:    48
Attributes:   10
              Gender
              Date
              Series
              Court
              Surface
              Round
              Comment
              MaxW
              MaxL
              Winner Rank
Test mode:evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------

MaxW <= 1.68
|   Round = Semifinals: winner ranked higher (18.0/1.0)
|   Round = The Final
|   |   MaxW <= 1.23: winner ranked lower (2.0)
|   |   MaxW > 1.23: winner ranked higher (6.0/1.0)
MaxW > 1.68: winner ranked lower (22.0/9.0)

Number of Leaves  :     4

Size of the tree :      7


Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         37               77.0833 %
Incorrectly Classified Instances       11               22.9167 %
Kappa statistic                      0.5417
Mean absolute error                     0.2957
Root mean squared error                    0.3845
Relative absolute error               64.3916 %
Root relative squared error             80.3875 %
Total Number of Instances              48

=== Detailed Accuracy By Class ===

              TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                0.882     0.29       0.625     0.882     0.732      0.824    winner ranked lower
                0.71      0.118      0.917     0.71      0.8        0.824    winner ranked higher
Weighted Avg.    0.771     0.179      0.813     0.771     0.776      0.824

=== Confusion Matrix ===

  a  b   <-- classified as
 15  2 |  a = winner ranked lower
  9 22 |  b = winner ranked higher


=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Training2011_2012 Grand Slam Finals Only
Instances:    48
Attributes:   10
              Gender
              Date
              Series
              Court
              Surface
              Round
              Comment
              MaxW
              MaxL
              Winner Rank
Test mode:user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

J48 pruned tree
------------------

MaxW <= 1.68
|   Round = Semifinals: winner ranked higher (18.0/1.0)
|   Round = The Final
|   |   MaxW <= 1.23: winner ranked lower (2.0)
|   |   MaxW > 1.23: winner ranked higher (6.0/1.0)
MaxW > 1.68: winner ranked lower (22.0/9.0)

Number of Leaves  :    4

Size of the tree :         7
```

```
--------------------------------------------------------------------------------

Time taken to build model: 0 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          17               70.8333 %
Incorrectly Classified Instances         7               29.1667 %
Kappa statistic                          0.44
Mean absolute error                      0.3434
Root mean squared error                  0.4488
Relative absolute error                 79.8685 %
Root relative squared error            100.4471 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.389      0.462     1         0.632      0.75     winner ranked lower
                 0.611     0          1         0.611     0.759      0.75     winner ranked higher
Weighted Avg.    0.708     0.097      0.865     0.708     0.727      0.75

=== Confusion Matrix ===

  a  b   <-- classified as
  6  0 |  a = winner ranked lower
  7 11 |  b = winner ranked higher
```

11.2. Initial Project Proposal

An Investigation into the Statistical Properties of Ranking in the Sport of Tennis

Pauline Kildunne

Computing Department

Higher Diploma in Science in Data Analytics

Submitted to:

Dr. Ioana Ghergulescu

Higher Diploma in Science in Data Analytics

Table of Contents

Objectives & Contribution to the Knowledge

In this paper I will consider the data sets of all tennis matches played by professional players during all tournaments for the last three years, 2011, 2012, 2013. Each player can attain points in each tournament and this is reflected on the level they reach, therefore how a player performs in each tournament will influence their ranking position

The field of statistical analysis is an important mathematical tool that informs sponsors and decision makers on their choice of player to invest in, also, a higher performing player will attract lucrative sponsorship and media coverage. Betting odds are also based on decisions from the outputs of sound mathematical models. The underlying premise to this paper is to define a hypothesis and let the data disprove or not, add value to the official statistics, and produce a reasonable forecasting model.

The paper will utilise statistical hypothesis tests to identify the following questions;

- Is it possible for a player to become rank number 1 by entering into more less ranking tournaments?
- Do the higher ranking players typically win the match?

- Does court type and surface type influence their performance and in turn influence their ranking?
- Analysing the previous three years of data predict the winning players for each Grand Slam in 2014.

Background

The data was collected from the web site, the Tennis-Data Betting, Results and Livescores Portal (2014).[18] I downloaded, a complete excel file (zipped), all matches played by professional tennis players from 2011 to 2013 for both the Association of Tennis Professionals (ATP) and the Women's Tennis Association (WTA). In total there are 365 tournaments.

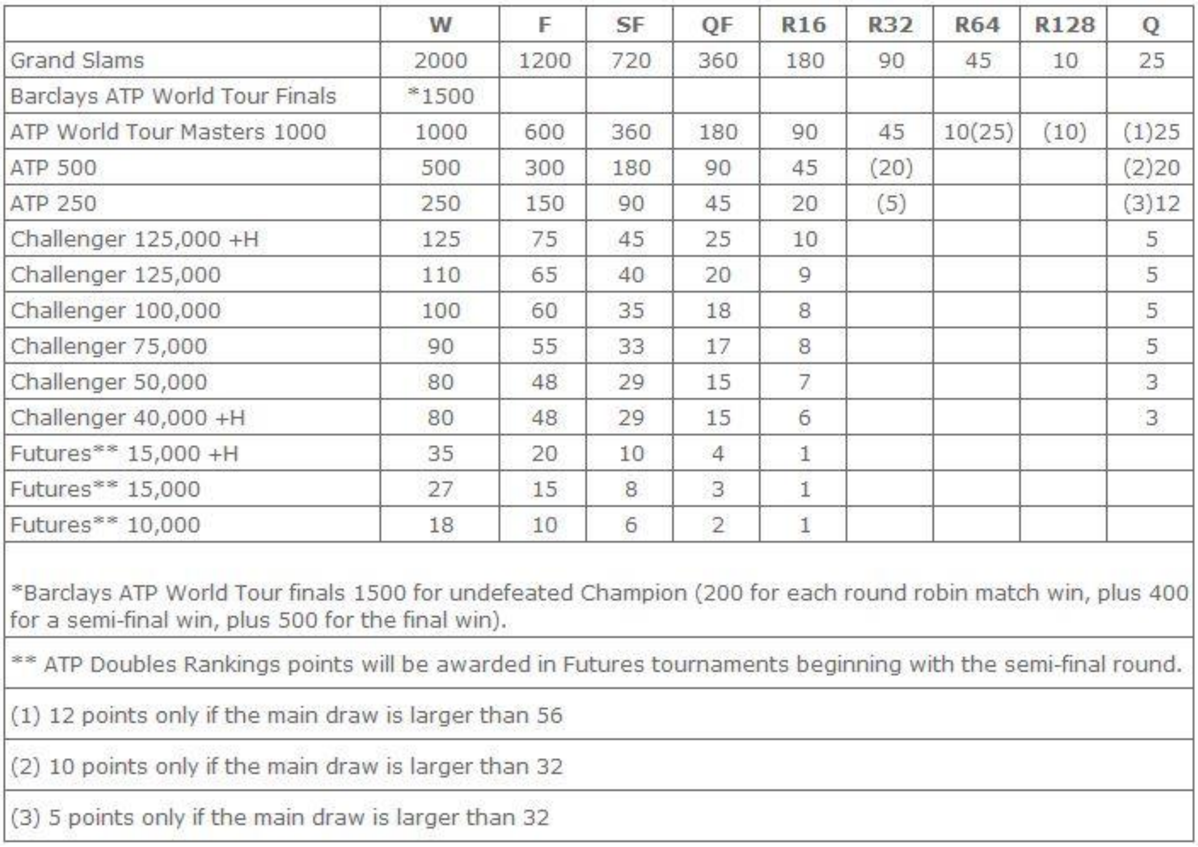
It states on the web site that all data is free to use.

The tennis Grand Slam is composed of four tournaments: the Australian Open, the French Open, Wimbledon and the US Open. Even though these tournaments are similar in prestige and prize money, they differ in terms of court surfaces. The Australian Open and the US Open are played on hard courts, the French Open is played on clay, and Wimbledon is played on grass. Each draw is composed of 128 players. Thus, 127 matches are played in each tournament for both men and women.

Both associations, ATP and WTA rank professional tennis players and use their rankings to decide both the participation of players in tournaments, as well as the ultimate champion of the year. Therefore, it is probable that a higher-ranked player will win the tournament, however both Williams sister won a Gram Slam tournament, even though at the time they were ranked outside the top ten (Serena Williams won the 2007 Australian Open when she was ranked 81st; Venus won the 2007 Wimbledon when she was ranked 31st).

The players ranking is dependent on how many points they accumulate. Each tournament is assigned points and the amount awarded is dependent on the level a player achieves in each tournament. In Figure 1, we can see the breakdown to the point's structure, again this is pertinent to their ranking at the start and end of a tournament.

[18] Tennis Betting, Results & Livescores Portal (2014). The Tennis-Data Betting, Results and Livescores Portal. Available at: http://www.tennis-data.co.uk/alldata.php [Accessed 10th February 2014].



	W	F	SF	QF	R16	R32	R64	R128	Q
Grand Slams	2000	1200	720	360	180	90	45	10	25
Barclays ATP World Tour Finals	*1500								
ATP World Tour Masters 1000	1000	600	360	180	90	45	10(25)	(10)	(1)25
ATP 500	500	300	180	90	45	(20)			(2)20
ATP 250	250	150	90	45	20	(5)			(3)12
Challenger 125,000 +H	125	75	45	25	10				5
Challenger 125,000	110	65	40	20	9				5
Challenger 100,000	100	60	35	18	8				5
Challenger 75,000	90	55	33	17	8				5
Challenger 50,000	80	48	29	15	7				3
Challenger 40,000 +H	80	48	29	15	6				3
Futures** 15,000 +H	35	20	10	4	1				
Futures** 15,000	27	15	8	3	1				
Futures** 10,000	18	10	6	2	1				

*Barclays ATP World Tour finals 1500 for undefeated Champion (200 for each round robin match win, plus 400 for a semi-final win, plus 500 for the final win).

** ATP Doubles Rankings points will be awarded in Futures tournaments beginning with the semi-final round.

(1) 12 points only if the main draw is larger than 56

(2) 10 points only if the main draw is larger than 32

(3) 5 points only if the main draw is larger than 32

Figure 1. Ranking Points Structure

There has been previous analysis undertaken in this area before by Boulier and Stekler (1999)[19] who found that the ranking difference between contestants is a good predictor for victory in professional tennis and Klaassen and Magnus (2003)[20] proposed a method of forecasting the winner of a match at the beginning of the match, as well as during it, they used a measure based on nonlinear differences in rankings.

Technical Approach

Data Preparation

Data mining is finding patterns in data using statistics. For the purpose of data processing the data sets in their current state will need to be cleaned up, reshaped, transformed and aggregated. I will need to deal with the missing values as missing value often causes no solution. Some of the data is not pertinent to the data mining exercise, and can be ignored. Time initially will be spent in excel.

[19] Boulier, B., & Stekler, H. (1999). Are sports seedings good predictors? An evaluation. *International Journal of Forecasting*, *15*, 83–91. [Internet]. Available at: http://mres.gmu.edu/pmwiki/uploads/Main/Boulier%20Stekler%201999.pdf [Accessed 11th February 2014].

[20] Klaassen, F., & Magnus, J. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*, *148*,257–267. [Internet]. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.5640&rep=rep1&type=pdf [Accessed 12th February 2014.

Data Mining Techniques

A number of hypothesis tests will be carried out beginning with the assumption that the hypothesis is true. There are two hypotheses, the null hypothesis assuming there is no difference and the alternative hypothesis which if proven asserts the discovery of new knowledge.

Prediction Analysis and Multi Regression

The key idea of prediction analysis is to discover the relationship between the dependent and independent variables. By using historical data from both either linear or nonlinear regression techniques can produce a fitted regression curve that can be used for predictions in the future.

Evaluation

Which tool to should be used to visualise the data mining results? In order to properly interpret knowledge patterns it is important that an appropriate visualisation tool is used, I may use pie charts, histograms, box plots, scatter plots, and distributions, as the analysis begins this will become more evident.

Building and testing models – how do you know you are right? And have I achieve my objective?

Special Resources Required

I intend to use Python[21] and R[22] for this project; I will be using their online site for resources. Other books which I may find useful, these can be found in the National College of Irelands' library, is the study of analytical evidence which is examined in Lewis (2003) book where he explores the importance of rigorous statistical analysis for sport.[23] Another helpful resource is in Defusco (2007) book where he explores the application of quantitative analysis.[24]

Project Plan

I have created a Gantt chart outlining the project schedule. In Figure 2, I have illustrated the start date and duration for each task of the project. There are 6 tasks, Project Proposal, Requirements Specification, Management Progress Report 1, Preliminary Presentation, Management Progress Report 2 and Showcase, Presentation & Dissertation.

The final presentation for the project is due 3rd May 2014. During the all tasks will be carried out my Project Leader (who is the author of this proposal). The schedule will be reviewed and updated continuously throughout the project.

[21] Python (2014). The official home of the *Python* Programming Language. Available at: http://www.python.org/ [Accessed on 9th February 2014].

[22] R (2014). The R Project for Statistical Computing. Available at: http://www.r-project.org/ [Accessed on 10th February 2014].

[23] Lewis, M. (2003) Moneyball: *The Art of Winning an Unfair Game*. United States. W.W. Norton & Company Inc

[24] Defusco, R.A. (2nd Ed.). (2007). *Quantitative investment analysis.* New Jersey. John Wiley & Sons

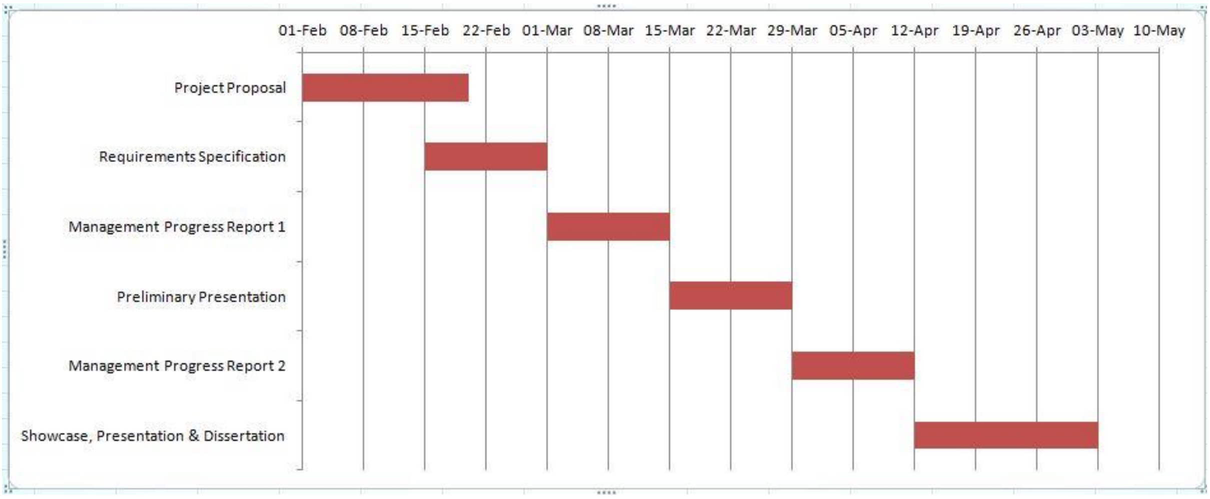


Figure 2. Gantt chart

Technical Details

As mentioned previously I will be using Python and R for this project. There are a number of modules available for statistical analysis which I will need to research and investigate and if suitable download. These modules are NumPy (Numeric Python) and SciPy (Scientific Python) which works with NumPy with a greater collection of applied mathematical techniques.

Systems / Datasets

The data consists of 38 variables for the past three years of all the tournaments from the men's ATP and women's WTA tours. There are 365 tournaments in total and 14,220 matches for 2011, 2012 and 2013. For the purpose of this project I will consider the 15 variable and they are as follows;

ATP = Tournament number (men)

WTA = Tournament number (women)

Location = Venue of tournament

Tournament = Name of tournament (including sponsor if relevant)

Data = Date of match (note: prior to 2003 the date shown for all matches played in a single tournament is the start date)

Series = Name of ATP tennis series (Grand Slam, Masters, International or International Gold)

Tier = Tier (tournament ranking) of WTA tennis series.

Court = Type of court (outdoors or indoors)

Surface = Type of surface (clay, hard, carpet or grass)

Round = Round of match

Best of = Maximum number of sets playable in match

Winner = Match winner

Loser = Match loser

WRank = ATP Entry ranking of the match winner as of the start of the tournament

LRank = ATP Entry ranking of the match loser as of the start of the tournament

WPts = ATP Entry points of the match winner as of the start of the tournament

LPts = ATP Entry points of the match loser as of the start of the tournament

Evaluation, Tests and Analysis

For the purpose of this proposal I will be carrying out a paired t test, an independent test, correlation and multiple regression tests. All hypothesis tests will be under taken at α = 0.05 level of significance. As the project progresses there may be additional testing and analysis also as I get a deeper knowledge of the modules there may be opportunities to experiment with other software and statistical techniques.

Consultation with Specialisation Person(s)

At the moment there has been no direct consultation with any lecture, however from completing this proposal I understand the challenges ahead and I do intend to consult with the supervisor, Dr. Ioana Ghergulescu.

11.3. Initial Requirements Specification

Requirements Specification (RS)

Document Control

Revision History

Date	Version	Scope of Activity	Prepared	Reviewed	Approved
27/02/2014	1	Create	PK	Dr.Ioana Ghergulescu	Dr.Ioana Ghergulescu
13/02/2014	2	Update	PK		

Distribution List

Name	Title	Version
Dr.Ioana Ghergulescu	Lecturer	1
Pauline Kildunne	Author	1

Related Documents

Title	Comments
Title of Use Case Model	
Title of Use Case Description	

Table of Contents

Introduction

In today's competitive world of sport it is not just whether you win or lose but also how you play the game, the difference is between Tennis player and Tennis champion. When trying to enhance or improve performance it is important that all performances are analysed, especially when so much time, dedication and investment is put into each player. Sport performers, coaches/trainers need to understand why the player was good or if not, where are the areas for improvement. Analysing and evaluating performances creates a competitive advantage over their rivals and enable them to optimize their strategies.

Purpose

The purpose of this document is to set out the requirements needed in order to carry out statistical hypothesis tests to identify the following questions;

- Is it possible for a player to become rank number 1 by entering into more less ranking tournaments?
- Do the higher ranking players typically win the match?
- Does court type and surface type influence their performance and in turn influence their ranking?
- Explore the possibility of analysing the previous three years of data and predict the winning players for each Grand Slam in 2014.

When tennis data is captured and analysed it can highlight certain tendencies and patterns which emerge from a match point, set and game. This information can provide players and coaches with another tool to use in training and in developing match strategies.

The other intended stakeholders which can utilise the data effectively are the sponsors (tennis players typically earn money through sponsorship, sponsors want to invest in the most popular player, high ranking and in order to manage their future budgets they need analysis of past performance), sport journalists (they want to make sure they make front page news with high ranking players) and bookies / gamblers (Betting odds are also based on decisions from the outputs of sound mathematical models)..

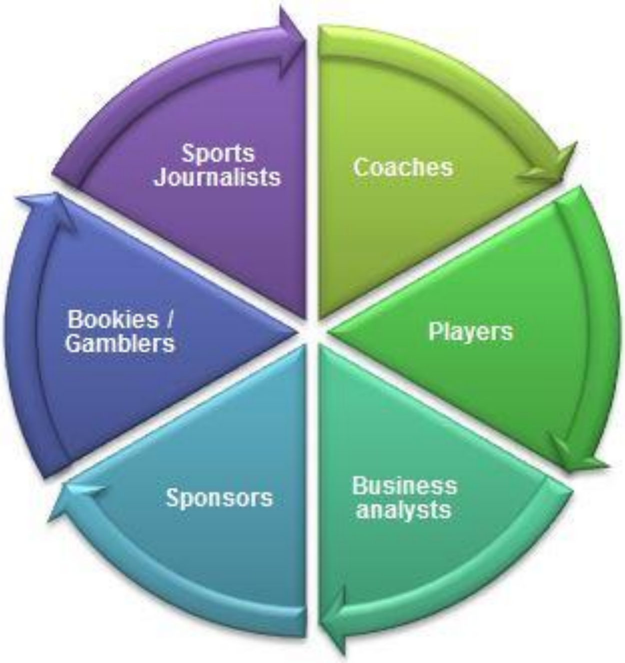


Figure 1

Project Scope

The scope of the project is to develop an output of statistical analysis of ranking using the data sets of the tennis tournaments for the last three years, 2011, 2012, 2013 for the ATP and WTA.

Illustrate how factors such as points, court and surface type influence their ranking and explore the possibility of carrying out predictive analysis to determine the winning players for each Grand Slam in 2014.

The work will be carried out in the National College of Ireland Campus using their licensed software on their personal computers. Following on from the submission of the project proposal there will be a meeting with the faculty supervisor to discuss and sign off the project scope.

The shareholders are identified in Figure1, they have been informed that the delivery of the project is scheduled in 12 weeks' time, 3rd May 2014.

In order to meet this deadline and ensure the project is successful it is the responsibility of the Business Analyst to recognise and take on board the stakeholder's ideas, views and inputs.

This Requirement Specification is version 1. The first task of the requirement specification is to elicit requirements from the stakeholders. During the course of the project there will be revised versions to accommodate additional testing and analysis.

As I get a deeper understanding and knowledge of the modules there may be also opportunities to experiment with other software and statistical techniques.

Definitions, Acronyms, and Abbreviations

Actor - An actor defines a coherent set of roles that users of the system can play when interacting with it. An actor can be played by either an individual or an external system
ATP - Association of Tennis Professionals
WTA - Women's Tennis Association
The tennis Grand Slam - Is composed of four tournaments: the Australian Open, the French Open, Wimbledon and the US Open
Location = Venue of tournament
Tournament = Name of tournament
Data = Date of match
Series = Name of ATP tennis series (Grand Slam, Masters, International or International Gold)
Tier = Tier (tournament ranking) of WTA tennis series.
Court = Type of court (outdoors or indoors)
Surface = Type of surface (clay, hard, carpet or grass)
Round = Round of match
Best of = Maximum number of sets playable in match
Winner = Match winner
Loser = Match loser
WRank = ATP Entry ranking of the match winner as of the start of the tournament
LRank = ATP Entry ranking of the match loser as of the start of the tournament
WPts = ATP Entry points of the match winner as of the start of the tournament
LPts = ATP Entry points of the match loser as of the start of the tournament[25]

User Requirements Definition

First of all it's important to define the users of this application.

- Coaches
- Players
- Sponsors

[25] Tennis Betting, Results & Livescores Portal (2014). The Tennis-Data Betting, Results and Livescores Portal. Available at: http://www.tennis-data.co.uk/alldata.php [Accessed 22th February 2014].

- Bookmakers
- Sport Journalists

The coaches and players want to use the match statistics to assess performance. Therefore, it is important to interpret match statistics correctly as misinterpretation could affect future performance.

The statistical analysis will provide the players and coaches with another tool to use in training and in developing match strategies.

The bookmakers require the analysis so they can work out their margins for every event. There are variations in margins that bookmakers hold, as this is what determines the value of their odds and a higher ranking player would yield fewer odds than a lower ranking player.

When Sponsors are planning their budgets they need maximum exposure in order to increase brand loyalty, create awareness and increase sales. They want to know who they need to invest in. Is it safe to back the high ranking players? Do the higher ranking players typically win the match? Which tournaments do they prefer playing in? (Surface type, court type).

Sports journalist need to capture headlines. Using historic data they can get statistics on how a player performs and competes with their opponent, certain playing dynamics attract larger crowds and media attention.

The stakeholder's require the system and statistical analysis to be both reliable and robust.

Requirements Specification

There is no requirement for hardware or operating system from the stakeholder's point of view. The output will be presented via a presentation tool and/or provided in a hard copy format.

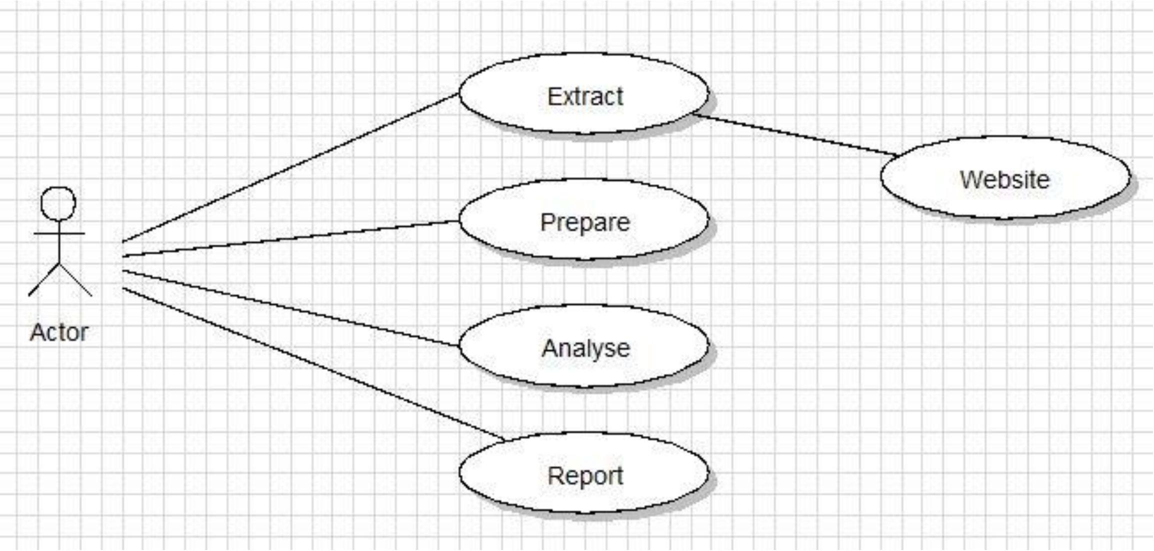
All analysis carried out will be documented so that any user can understand and replicate if needs be. The coding, programming and testing will be commented and each step of the process and procedure will be filed, signed off and documented.

Functional requirements

The functional requirements describe how the system should behave.

Use Case Diagram

The Use Case Diagram provides an overview of all functional requirements.



Requirement 1

The first functional environment of my project is Extract.

Description & Priority

Choosing the appropriate data set is crucial for the success of this project. This requirement is a high priority. This requirement is about sourcing and downloading a data set. It is the main requirement because without a data set the user cannot use my application.

Use Case

Use Case 1 – Extracting the data.

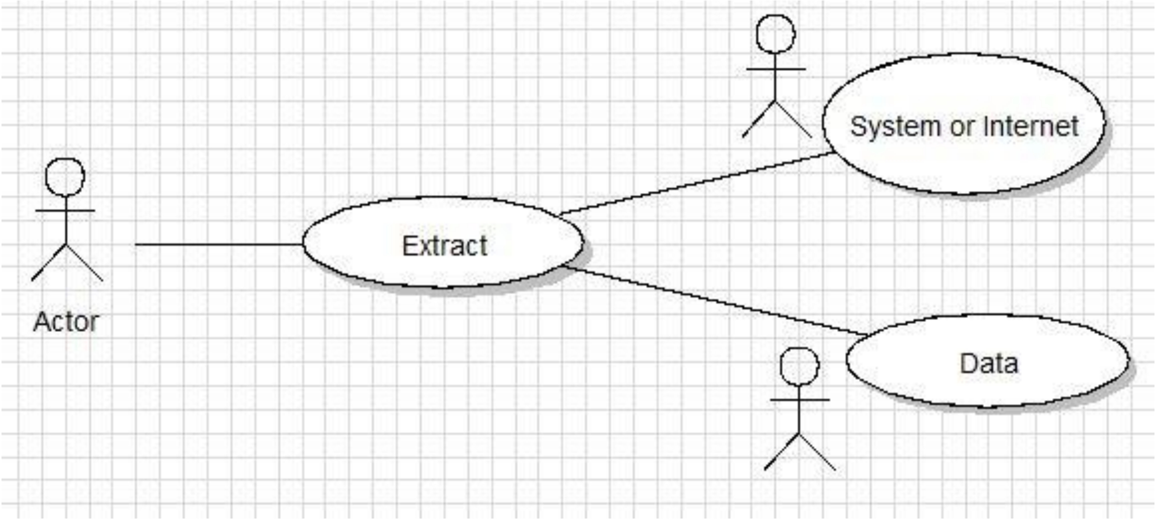
Scope

The scope of this use case is to search the World Wide Web for appropriate data sets on Tennis which contain the appropriate variables and number of observations.

Description

This use case describes the process to retrieving the data set.

Use Case Diagram



Flow Description

Precondition

Before this use case can be initiated, the Actor must have internet access, broadband, wireless and is able to browse Online Website.

Activation

This use case starts when an Actor logs on to the internet and searches for data through several web channels.

Main flow

1. The system identifies and presents a list of all categories.
2. The Actor refines it search and conducts a search for the category ‘tennis data set’.
3. The system displays a list of data sets and what file format the data set is stored as.
4. The Actor downloads the dataset.

Alternate flow

A1:

1. The system requests user identification to access website.
2. The Actor provides details and carries out a search for ‘tennis data set’.
3. The use case continues at position 3 of the main flow.

Exceptional flow

E1:

4. The system requires specific/additional software to be downloaded in order to read the files (open source software).
5. The Actor downloads additional software.
6. The use case continues at position 4 of the main flow.

Termination

The system goes into a wait state.

Requirement 2

The second functional environment of my project is Preparation.

Description & Priority

For the purpose of data processing the data sets will need to be cleaned up, reshaped, transformed and aggregated. This is a priority in order for the output to be a success.

Use Case

Use Case 2 – Preparing the data.

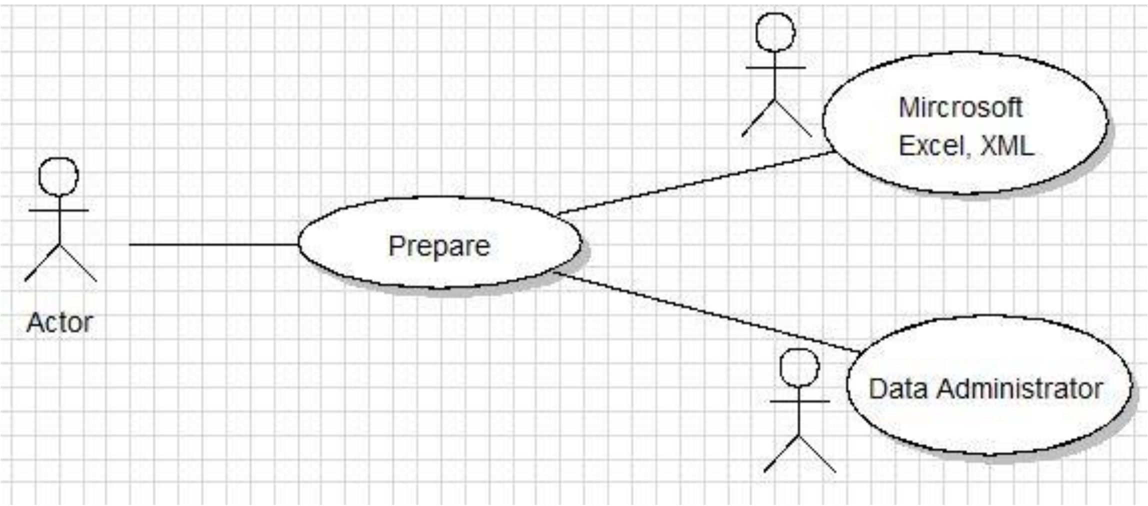
Scope

The scope of this use case is to import the data into an appropriate environment such as Microsoft Excel and store it on the hard drive.

Description

This use case describes the process to preparing the data set for data processing.

Use Case Diagram



Flow Description

Precondition

The system must have data sets available and downloaded.

Activation

This use case starts when an Actor having the appropriate software installed on their personal computer i.e. Microsoft Excel.

Main flow

1. The system recognises the file format and displays the data set in Microsoft Excel.
2. The Actor examines the fields and variables.
3. The system is capable of storing and saving the data.
4. The data administrator cleans, reshapes, transforms and aggregates the data.

Alternate flow

A1:

5. The system does not have the appropriate software to open file.
6. The Actor must get a licensed copy of Microsoft Excel.
7. The use case continues at position 2 of the main flow.

Exceptional flow

E1:

8. The system is unable to download software because of application error, 'there is not enough memory available'.
9. The Actor needs to update the system with additional RAM.
10. The use case continues at position 1 of the main flow.

Termination

The system is ready for data processing.

Post condition

The system goes into a wait state.

Requirement 3

The third functional environment of my project is Analyse.

Description & Priority

A number of hypothesis tests will be carried out. This a priority in order to gather statistical analysis for the stakeholders.

Use Case

Use Case 3 – Analysing the data.

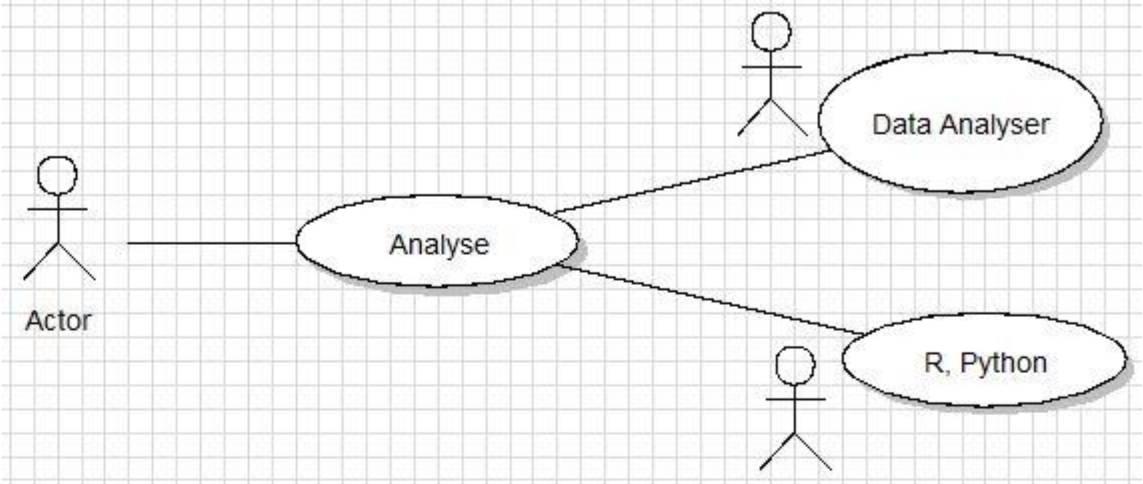
Scope

The scope of this use case is to identify the relevant variables, independent and dependent in the dataset that will provide the correct analysis that is requested by the stakeholder.

Description

This use case describes the process to analysing the data.

Use Case Diagram



Flow Description

Precondition

The system must be capable of analysing the cleaned, reshaped and transformed data.

Activation

This use case starts with an Actor having the appropriate cleansed data.

Main flow

1. The system uses the appropriate software, R and/or Python to carry out the analysis.
2. The Actor writes the code.
3. The system is capable of carrying out statistical analysis methods such as, t test, regression, correlation.
4. The data analyser identifies the independent and dependent variables and outputs the results.

Alternate flow

A1:

6. The system does not have the R and /or Python installed.
7. The Actor installs the software.
8. The use case continues at position 2 of the main flow.

Exceptional flow

E1:

9. The system is unable to download software because of application error, 'there is not enough memory available'.
10. The Actor needs to update the system with additional RAM.
11. The use case continues at position 1 of the main flow.

Termination

The system is ready to interpret the date, produce graphs, plot results.

Post condition

The system goes into a wait state.

Requirement 4

The fourth functional environment of my project is Report.

Description & Priority

In order to properly communicate statistical analysis it is important that an appropriate visualisation tool is used. Communication of the model used and being able to report the results of the analysis is a priority for stakeholders.

Use Case

Use Case 4 – Reporting the statistical analysis.

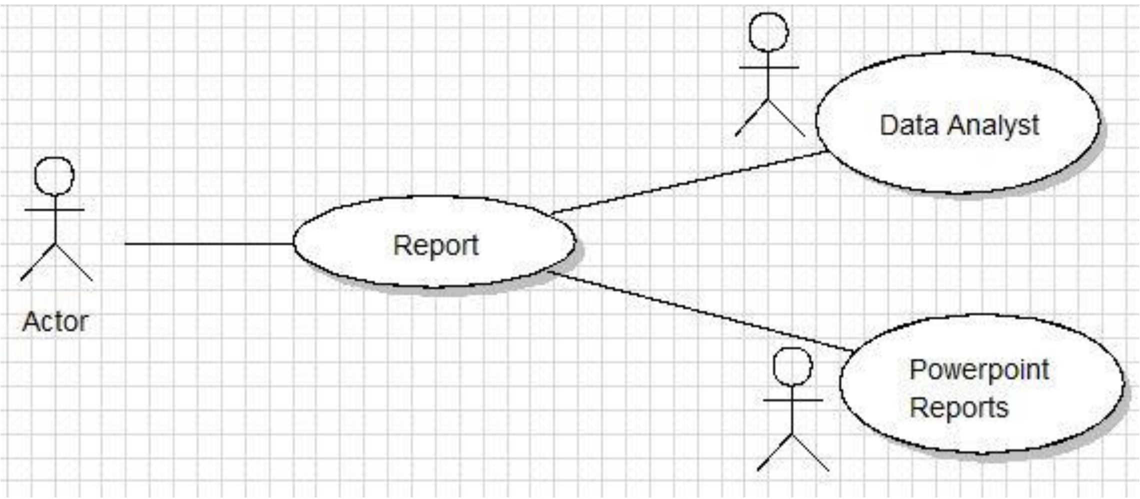
Scope

The scope of this use case is to provide and present the output of the statistical analysis of the relevant variables, independent and dependent in the dataset to the stakeholders.

Description

This use case describes the process to presenting the statistical results.

Use Case Diagram



Flow Description

Precondition

There must be statistical results from the analysed data.

Activation

This use case starts with an Actor having the appropriate results from the analysed data.

Main flow

1. The system has Microsoft PowerPoint installed.
2. The Actor can create a PowerPoint presentation and is capable of understanding the analysed data.
3. The system is capable of saving the documents that where created.
4. The data analyst present the results to a technical and non-technical audience.

Alternate flow

A1:

5. The system does not have PowerPoint installed.
6. The Actor installs the software.
7. The use case continues at position 2 of the main flow.

Exceptional flow

E1:

8. The system is unable to download software because of application error, ‘there is not enough memory available’.
9. The Actor needs to update the system with additional RAM.
10. The use case continues at position 1 of the main flow.

Termination

The system is ready to present the results.

Post condition

The system goes into a wait state.

Non-Functional Requirements

Performance/Response time requirement

In order to provide accurate statistical analysis as per the stakeholder requests there is no tolerance for failure and as a result the system should not freeze or crash during use or cause any incomplete analysis. The application must be reliable and the rate of failure occurrence should mirror any other systems on the market.

Availability requirement

There is not a requirement to have any functions available as the results will be presented in PowerPoints and hard copy. During the process of this project the system will need to be available for feedback on progress.

Recover requirement

If the event there is a network or hardware failure the stored data and related documentation, programmable code will be saved as back up on a portable hard drive.

Robustness requirement

As per users requirements

Security requirement

During the project the computer will have approved log in credential and users in order to access the relevant folder. From the clients perspective there is no need for access to this folder.

Reliability requirement

As per users requirements

Maintainability requirement

The only maintenance would be if there were other tools required outside the scope of the requirements specification to analyse the data but for present there is no maintenance required.

Portability requirement

Not applicable to this project

Extendibility requirement

Not applicable to this project

Reusability requirement

Not applicable to this project

Resource utilization requirement

Not applicable to this project

Interface requirements

There is no requirement for an interface.

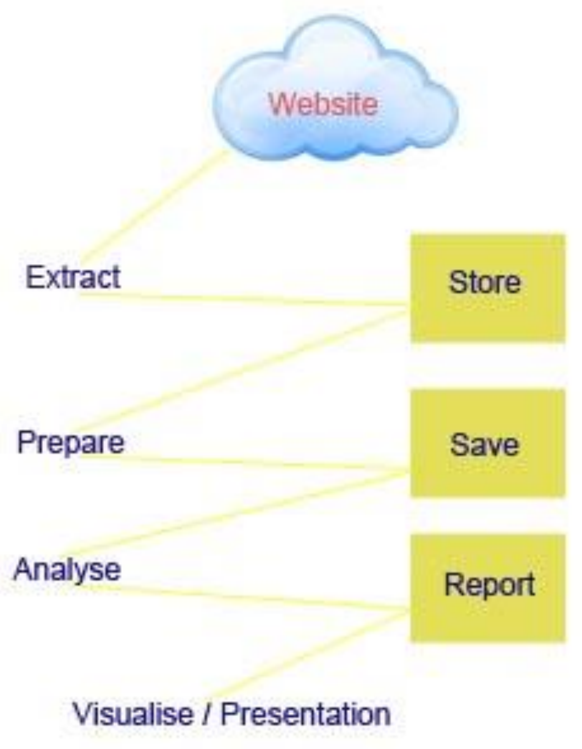
GUI

The results of the analysis will be too light to invest in extreme costs such as a Graphical User Interface.

Application Programming Interfaces (API)

The results of the analysis will be too light to invest in extreme costs such as a Application Programming Interface.

System Architecture



The system architecture clearly outlines each step of the analysing process as illustrated in the User Case Diagram, from downloading the dataset to presenting the analysis to the stakeholders. This is a clear visual process, after extracting the data you store it, once the preparation is complete you save the new version of the data and after analysing the data it is presented to the stakeholders.

System Evolution

The data consists of 38 variables for the past three years of all the tournaments from the men's ATP and women's WTA tours. There are 365 tournaments in total and 14,220 matches for 2011, 2012 and 2013. In order to fulfil the requirements there is only a certain number of variables that will be used so it is feasible to look into other variables and carry out other analysis like association, classification and clustering techniques.

Also another option is to look at the data prior 2011 and carry out historically trends and statistics. This system will be documented which will allow future tournaments to be added and analysed.

11.4. Management Progress Reports

11.4.1. Management Progress Report 1

An Investigation into the Statistical Properties of Ranking in the Sport of Tennis

Management Progress Report 1

Pauline Kildunne

Computing Department

Higher Diploma in Science in Data Analytics

Submitted to:

Dr. Ioana Ghergulescu

Higher Diploma in Science in Data Analytics

Highlight Report History

Document Location

This document is the first of three Management Progress Reports. It is stored in the Project Folder on the H drive of the student section on the National College of Ireland's server. H:\Project

Revision History

Date of this revision: 15/03/2014
Date of Next revision: 29/03/2014

Revision date	Previous revision date	Summary of Changes	Changes marked
15/03/2014		First issue	

Approvals

This document requires the following approvals.
Signed approval forms are filed in the Management section of the Project Folder.

Name	Signature	Title	Date of Issue	Version
Dr. Ioana Ghergulescu		Lecturer	15/03/2014	1

Distribution

This document has been distributed to:

Name	Title	Date of Issue	Version
Dr. Ioana Ghergulescu	Lecturer	15/03/2014	1
Stakeholders (as per requirements specification)	Various	15/03/2014	1

Highlight Report

Purpose of Document

The purpose f this document is to update the stakeholders on the progress of the Project, what has been completed, what the next steps are, potential problems or risks and update on project plan.

Date of Report

15/03/2014

Period Covered

01/02/2014 - 15/03/2014

Budget Status

At this stage we can confirm the project is on budget.

Schedule Status:

The project is on schedule. The stakeholders have informed the Sponsor that they do not require a Preliminary Presentation; this will allow extra time to carry out further analytical testing.

All academic deliverables have been completed on time and we are awaiting approval from the Project Proposal submission and Requirements Specification submission.

The Project Plan has been updated to reflect the tasks that were carried out prior to Management Progress Report. This is illustrated in Figure 1. These include the following;

1. Review Project Feasibility

As the project is carefully planned the time scale is tight but achievable, however because we are still awaiting approval for Project Proposal and Requirements Specification, there is potential cause for 'Scope Creep'. By this I mean new requirements may be requested. This can impact the schedule of the project so it is important that the focus is not removed from the original scope of the project.

2. Merge Data for ATP players

The data will be saved as a single file format; after extracting, cleaning and formatting the files/data sets for 2011, 2012 and 2013 for the Association of Tennis Professionals for Men, it was then saved to a single file format.

3. Further Research Data Analytical Techniques

Additional time has been spent on reviewing the R packages that are relevant for the project; this was necessary and useful to gain more knowledge of the many libraries available in R.

4. Test Sample Data & Run Analysis

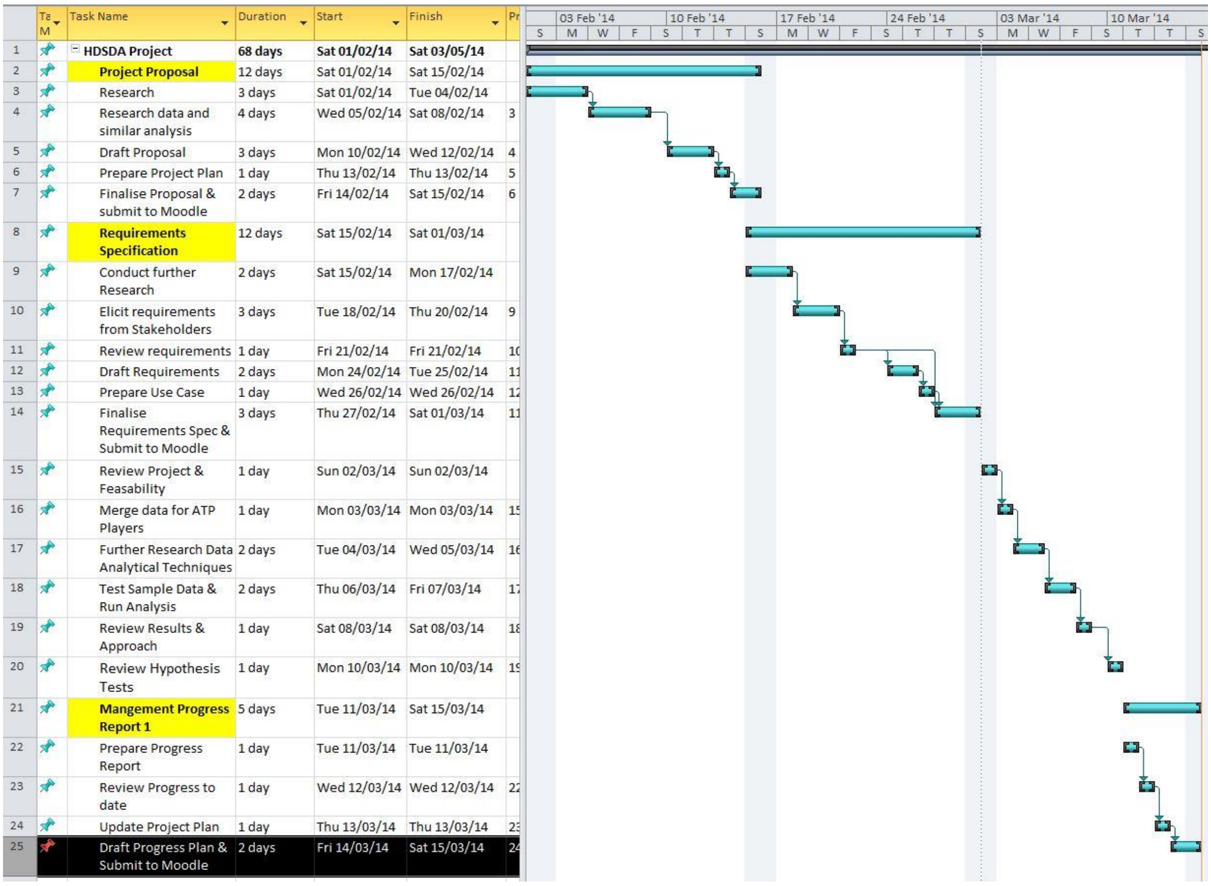
An independent t-test was carried out on a sample of the Men tennis players for 2011, using the variables, WRank (ATP Entry ranking of the match winner as of the start of the tournament) and LRank (ATP Entry ranking of the match loser as of the start of the tournament).

For further analysis we still need to define an independent and dependent variable.

5. Review Hypothesis Tests

In order to remain focus on the goals and objectives of the project we still need to carry out further statistical hypothesis tests to identify the following questions;

- Is it possible for a player to become rank number 1 by entering into more less ranking tournaments?
- Do the higher ranking players typically win the match?
- Does court type and surface type influence their performance and in turn influence their ranking?
- Explore the possibility of analysing the previous three years of data and predict the winning players for each Grand Slam in 2014.



	Ta M	Task Name	Duration	Start	Finish	Pr
1		HDSDA Project	68 days	Sat 01/02/14	Sat 03/05/14	
2		Project Proposal	12 days	Sat 01/02/14	Sat 15/02/14	
3		Research	3 days	Sat 01/02/14	Tue 04/02/14	
4		Research data and similar analysis	4 days	Wed 05/02/14	Sat 08/02/14	3
5		Draft Proposal	3 days	Mon 10/02/14	Wed 12/02/14	4
6		Prepare Project Plan	1 day	Thu 13/02/14	Thu 13/02/14	5
7		Finalise Proposal & submit to Moodle	2 days	Fri 14/02/14	Sat 15/02/14	6
8		Requirements Specification	12 days	Sat 15/02/14	Sat 01/03/14	
9		Conduct further Research	2 days	Sat 15/02/14	Mon 17/02/14	
10		Elicit requirements from Stakeholders	3 days	Tue 18/02/14	Thu 20/02/14	9
11		Review requirements	1 day	Fri 21/02/14	Fri 21/02/14	10
12		Draft Requirements	2 days	Mon 24/02/14	Tue 25/02/14	11
13		Prepare Use Case	1 day	Wed 26/02/14	Wed 26/02/14	12
14		Finalise Requirements Spec & Submit to Moodle	3 days	Thu 27/02/14	Sat 01/03/14	11
15		Review Project & Feasability	1 day	Sun 02/03/14	Sun 02/03/14	
16		Merge data for ATP Players	1 day	Mon 03/03/14	Mon 03/03/14	15
17		Further Research Data Analytical Techniques	2 days	Tue 04/03/14	Wed 05/03/14	16
18		Test Sample Data & Run Analysis	2 days	Thu 06/03/14	Fri 07/03/14	17
19		Review Results & Approach	1 day	Sat 08/03/14	Sat 08/03/14	18
20		Review Hypothesis Tests	1 day	Mon 10/03/14	Mon 10/03/14	19
21		Mangement Progress Report 1	5 days	Tue 11/03/14	Sat 15/03/14	
22		Prepare Progress Report	1 day	Tue 11/03/14	Tue 11/03/14	
23		Review Progress to date	1 day	Wed 12/03/14	Wed 12/03/14	22
24		Update Project Plan	1 day	Thu 13/03/14	Thu 13/03/14	23
25		Draft Progress Plan & Submit to Moodle	2 days	Fri 14/03/14	Sat 15/03/14	24

Figure 1

Completed during Period

Project Proposal	The Project Proposal has been submitted and awaiting feedback.
Project Requirement Specification	The Project Requirements Specification has been submitted and awaiting feedback
Project Research	Although we are satisfied with the data sets, there will be continual research carried out during this project.

Risk Assessment

Actual

Lack of progress in analysis: It is critical for project delivery that further analysis and tests are carried out.

Ways to address this risk: Factor more test between this report and next progress report. Save the data from the ATP - Association of Tennis Professionals for Men, 2011, 2012, 2013 and the WTA - Women's Tennis Association for 2011, 2012 and 2013 to a single file format.

Potential Risk

Data Analysis: That the data analysis does not clearly answer the project objectives and/or it could possibly take longer than

	forecasted to deliver.
Project Proposal & Requirements Specification	Potential risk is that these reports are not approved, additional requirements or change of scope is requested to the project.
Future Plans	Complete a full data analysis cycle from identifying the variables, writing the code, running the code through an appropriate programme, testing the code, and finally analysing results produced and documenting the results.
Scheduled Date	29/03/2014 – due date for next progress report

Project Issue Status

The stakeholders have requested that we do not include the Preliminary Presentation; this will allow extra time to be spent on the statistical analysis which should resolve our current risk. There will be further developments as a result.

I have logged these issues in the RAID report.

11.4.2. Management Progress Report 3

An Investigation into the Statistical Properties of Ranking in the Sport of Tennis

Management Progress Report 3

Pauline Kildunne

Computing Department

Higher Diploma in Science in Data Analytics

Submitted to:

Dr. Ioana Ghergulescu

Higher Diploma in Science in Data Analytics

Highlight Report History

Document Location

This document is the third Management Progress Reports. It is stored in the Project Folder on the H drive of the student section on the National College of Ireland's server.
H:\Project

Revision History

Date of this revision: 04/05/2014
Date of Next revision: 14/05/2014

Revision date	Previous revision date	Summary of Changes	Changes marked
04/05/2014	29/03/2014	Third issue	

Approvals

This document requires the following approvals.
Signed approval forms are filed in the Management section of the Project Folder.

Name	Signature	Title	Date of Issue	Version
Dr. Ioana Ghergulescu		Lecturer	04/05/2014	3

Distribution

This document has been distributed to:

Name	Title	Date of Issue	Version
Dr. Ioana Ghergulescu	Lecturer	04/05/2014	3
Stakeholders (as per requirements specification)	Various	04/05/2014	3

Highlight Report

Purpose of Document

The purpose f this document is to update the stakeholders on the progress of the Project, what has been completed, what the next steps are, potential problems or risks and update on project plan.

Date of Report

04/05/2014

Period Covered

15/03/2014 - 04/05/2014

Budget Status

At this stage we can confirm the project is on budget.

Schedule Status:

The project deadline has been extended to the 29/05/2014. This is to accommodate other projects which timelines coincide too closely. The stakeholders have agreed with the Sponsor the final date of submission and require that two copies are printed for review; this will allow extra time to carry out further analytical testing.

All academic deliverables have been completed and we have received approval for the Project Proposal submission and Requirements Specification submission.

There has been no update made to the last Project Plan (Figure 1) this will need to be addressed week 15. The following tasks are ongoing. These include the following;

6. Review Project Feasibility

The Gant chart needs to be updated to reflect the timelines for project and presentation. All tasks which will attribute to the success and delivery of the project needs to be documented. There is still a lot of work to be done but the project is carefully planned the main focus is not removed from the original scope of the project.

7. Merge Data for ATP players

The data has been saved as a single file format for data sets for 2011, 2012 and 2013 for the Association of Tennis Professionals for Men. There were a number of N/A values for which I replace these with a value of ‘0’. We will need to merge the data for the Women's Tennis Association (WTA) for datasets 2011, 2012, 2013; after extracting, cleaning and formatting the files.

8. Further Research Data Analytical Techniques

Additional time has been spent on Weka, which is open source software and is used for a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing and data mining tasks such as classification, regression, clustering, association rules, and visualisation. It is also well-suited for developing new machine learning schemes.

9. Test Sample Data & Run Analysis

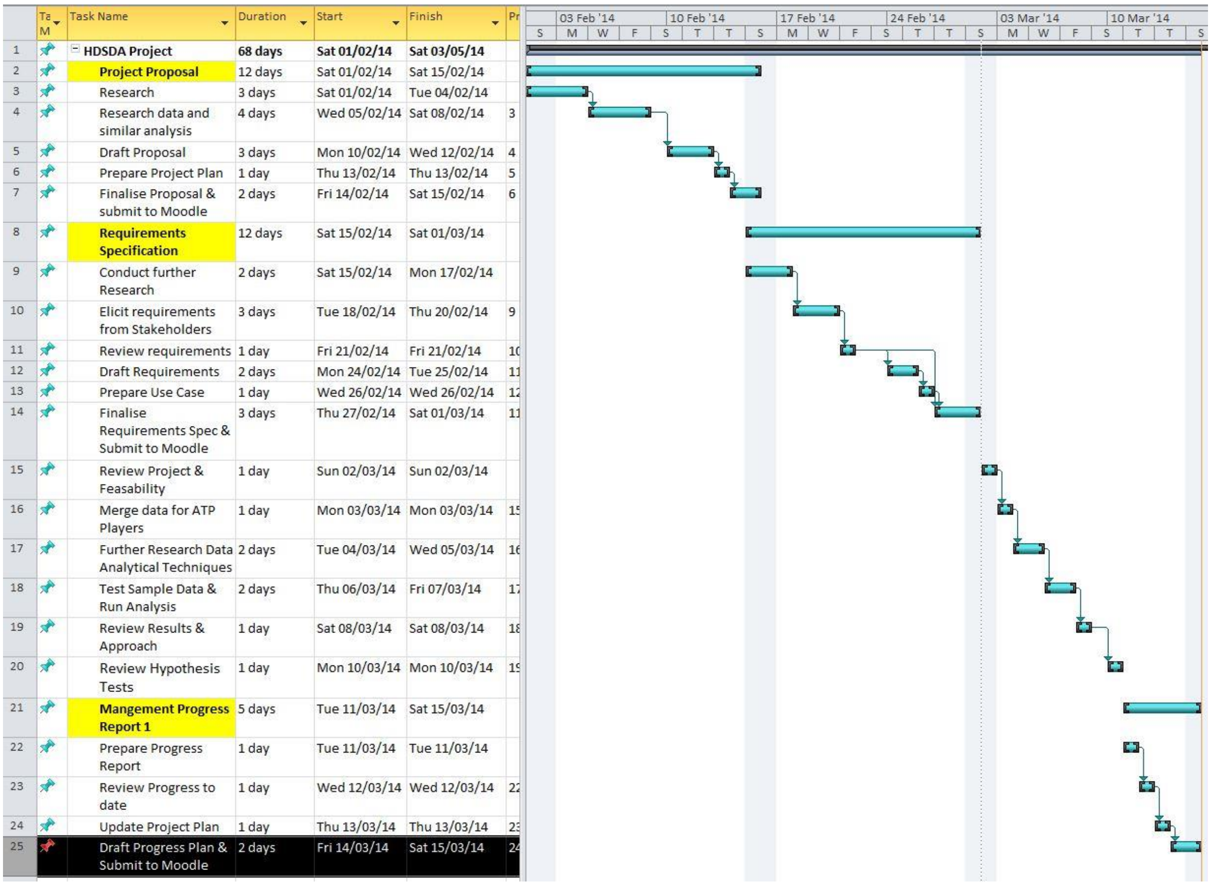
A number of classification techniques in data mining (Naïve Bayes and Bayesian Networks) was carried out on the dataset Men tennis players, using the variables, Surface, Court, Tournament, Series, Winner, Loser, WRank (ATP Entry ranking of the match winner as of the start of the tournament) and LRank (ATP Entry ranking of the match loser as of the start of the tournament), to explore their relationship in the context of solving practical classification problems.

Further analysis needs to be carried out and a decision on which variables to used..

10. Review Hypothesis Tests

In order to remain focus on the goals and objectives of the project we still need to carry out further statistical hypothesis tests to identify the following questions;

- Is it possible for a player to become rank number 1 by entering into more less ranking tournaments?
- Do the higher ranking players typically win the match?
- Does court type and surface type influence their performance and in turn influence their ranking?
- Explore the possibility of analysing the previous three years of data and predict the winning players for each Grand Slam in 2014.



	Ta M	Task Name	Duration	Start	Finish	Pr
1		**HDSDA Project**	**68 days**	**Sat 01/02/14**	**Sat 03/05/14**	
2		**Project Proposal**	12 days	Sat 01/02/14	Sat 15/02/14	
3		Research	3 days	Sat 01/02/14	Tue 04/02/14	
4		Research data and similar analysis	4 days	Wed 05/02/14	Sat 08/02/14	3
5		Draft Proposal	3 days	Mon 10/02/14	Wed 12/02/14	4
6		Prepare Project Plan	1 day	Thu 13/02/14	Thu 13/02/14	5
7		Finalise Proposal & submit to Moodle	2 days	Fri 14/02/14	Sat 15/02/14	6
8		**Requirements Specification**	12 days	Sat 15/02/14	Sat 01/03/14	
9		Conduct further Research	2 days	Sat 15/02/14	Mon 17/02/14	
10		Elicit requirements from Stakeholders	3 days	Tue 18/02/14	Thu 20/02/14	9
11		Review requirements	1 day	Fri 21/02/14	Fri 21/02/14	10
12		Draft Requirements	2 days	Mon 24/02/14	Tue 25/02/14	11
13		Prepare Use Case	1 day	Wed 26/02/14	Wed 26/02/14	12
14		Finalise Requirements Spec & Submit to Moodle	3 days	Thu 27/02/14	Sat 01/03/14	11
15		Review Project & Feasability	1 day	Sun 02/03/14	Sun 02/03/14	
16		Merge data for ATP Players	1 day	Mon 03/03/14	Mon 03/03/14	15
17		Further Research Data Analytical Techniques	2 days	Tue 04/03/14	Wed 05/03/14	16
18		Test Sample Data & Run Analysis	2 days	Thu 06/03/14	Fri 07/03/14	17
19		Review Results & Approach	1 day	Sat 08/03/14	Sat 08/03/14	18
20		Review Hypothesis Tests	1 day	Mon 10/03/14	Mon 10/03/14	19
21		**Mangement Progress Report 1**	5 days	Tue 11/03/14	Sat 15/03/14	
22		Prepare Progress Report	1 day	Tue 11/03/14	Tue 11/03/14	
23		Review Progress to date	1 day	Wed 12/03/14	Wed 12/03/14	22
24		Update Project Plan	1 day	Thu 13/03/14	Thu 13/03/14	23
25		Draft Progress Plan & Submit to Moodle	2 days	Fri 14/03/14	Sat 15/03/14	24

Figure 1

Completed during Period

Project Proposal	The Project Proposal has been submitted and approved.
Project Requirement Specification	The Project Requirements Specification has been submitted and approved
Project Research	Although we are satisfied with the data sets, there will be continual research carried out during this project.

Risk Assessment

Actual

Lack of progress in analysis:	It is critical for project delivery that further analysis and tests are carried out.
Ways to address this risk:	Looking at data mining tasks in Weka and making sure that the results are understood and relevant. Save the data from the WTA - Women's Tennis Association for 2011, 2012 and 2013 to a single file format.

Potential Risk

Data Analysis:	That the data analysis does not clearly answer the project objectives and/or it could possibly take longer than forecasted to deliver.

Future Plans	Complete a full data analysis cycle from identifying the variables, writing the code, running the code through an appropriate programme, testing the code, and finally analysing results produced and documenting the results.
Scheduled Date	Complete updated Gant chart by 05/05/2014

Project Issue Status

Although the stakeholders have postponed the deadline date they need to agree a date for the project presentation.

I have logged these issues in the RAID report.