



## **ASHESI UNIVERSITY COLLEGE**

**PREDICTING INJURY IN FOOTBALL USING PITCH QUALITY,  
PLAYER'S FUNCTION, PLAYER'S AGE AND MATCH INTENSITY:  
A CASE STUDY OF THE 2017 AFRICAN CUP OF NATIONS**

### **APPLIED PROJECT**

B.Sc. Management Information Systems

**Ayeley Commodore-Mensah**

**2017**

**ASHESI UNIVERSITY COLLEGE**

**Predicting injury in football using pitch quality, player's function,  
player's age and match intensity: A case study of the 2017 African Cup  
of Nations**

**APPLIED PROJECT**

Applied Project submitted to the Department of Computer Science, Ashesi  
University College in partial fulfilment of the requirements for the award of  
Bachelor of Science degree in Management Information Systems

**Ayeley Commodore-Mensah**

**April 2017**

## DECLARATION

I hereby declare that this Applied Project is the result of my own original work and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature:

.....

Candidate's Name:

.....

Date:

.....

I hereby declare that preparation and presentation of this Applied Project were supervised in accordance with the guidelines on supervision of such laid down by Ashesi University College.

Supervisor's Signature:

.....

Supervisor's Name:

.....

Date:

.....

## **Acknowledgement**

To my supervisor Dr Charles Jackson, who encouraged, supported and guided me from the beginning to the final stage of this project. I am also thankful to Mr Jon Boafo and Mr Nathan Quao for their assistance in getting this project completed. I am indebted to my family and colleagues for the support I received from them.

Above all, to God, the source of knowledge and wisdom, for seeing me through this project.

## **Abstract**

Football is the most popular sport in the world, and many individuals have taken advantage of it to earn a living and improve upon their standards of living. Injuries are also unfortunate incidents that occur in daily life and in sports, which affect an individual's ability to make good use of his sporting talent to earn a living for himself and his family.

In this project, modifiable risk factors that affect a player's likelihood of getting an injury are identified, and their individual contributions to injury of a player is assessed. A predictive model for determining important risk factors for determining injuries in football is generated using the identified risk factors: pitch quality, match intensity, player function and player age.

## Table of Contents

DECLARATION .....	i
Acknowledgement .....	ii
Abstract .....	iii
List of Figures .....	vi
List of Abbreviations .....	vii
Chapter 1 .....	1
1.1 Introduction and Background.....	1
1.2 Motivation .....	1
1.3 Problem Description.....	2
1.4 Benefits.....	3
1.5 Objectives.....	3
1.6 Outline of Project .....	4
Chapter 2: Related Works.....	5
2.1 Background .....	5
2.2 Relevance of Sports Injury Prediction .....	5
2.3 Factors that Result in Injuries.....	6
2.4 Related Works .....	7
2.4.1 Oslo Sports Trauma Research Center.....	7
2.4.2 NSW Waratahs .....	8
2.4.3 SAP SE .....	8
2.4.4 Team from the University of Birmingham and Southampton Football Club.....	8
2.4.5 Kitman Labs .....	9
2.4.6 Sports Injury Predictor.....	9
2.4.7 Researchers from Dow Jones and Wall Street Journal.....	9
2.5 Findings and Proposed Solution.....	10
2.6 Proposed Solution .....	10
Chapter 3: Requirements.....	11
3.1 Functional Requirements.....	11
3.1.1 Business Understanding .....	11
3.1.2 Data Inventory and Understanding.....	12
3.1.3 Data Preparation .....	14
3.1.4 Data modelling .....	14
3.1.5 Testing, Evaluation, Interpretation, Understanding.....	15

3.1.6 Deployment .....	15
3.2 Non-Functional Requirements .....	15
3.2.1 Product Requirements.....	15
3.2.2 Security Requirements.....	16
Chapter 4: High Level Architecture.....	17
4.1 Data Mining Algorithm.....	17
Chapter 5: Implementation .....	18
5.1 Data sources .....	18
5.1.1 AFCON Data .....	18
5.1.2 Match intensity data.....	22
5.2 Tools.....	25
5.3 Language .....	25
5.4 Implementation.....	25
5.4.1 Dataset 1 .....	26
5.4.2 Dataset 2 .....	27
5.4.3 Logistic regression.....	30
Chapter 6: Testing.....	37
6.1 Testing results .....	37
Chapter 7: Conclusions and Recommendations .....	39
References.....	41
Appendix.....	43

## List of Figures

Figure 2.1 The Meeuwisse model for classifying injury risk factors .....	7
Figure 5.1 Injury distribution based on periods in the match .....	20
Figure 5.2 Number of matches played on the different pitches .....	22
Figure 5.3 Match intensity levels for all matches .....	24
Figure 5.4 Linear regression results on Dataset 1 (very serious injuries).....	27
Figure 5.5 Linear regression results on Dataset 2 (all injuries) .....	28
Figure 5.6 Plot of player function and injuries sustained .....	29
Figure 5.7 Plot of pitch quality and injuries sustained.....	30
Figure 5.8 Head view of dataset .....	31
Figure 5.9 Plot of all variables considered in the analysis.....	32
Figure 5.10 Plot of player function on different pitch qualities.....	33
Figure 5.11 Correlation output.....	34
Figure 5.12 Logistic regression output .....	35
Figure 6.1 Result of testing in R .....	38
Figure 7.1 Proposed interface for club administrators to work with .....	39



## **List of Abbreviations**

The Fédération Internationale de Football Association – FIFA

Confederation of African Football CAF

National Football League-NFL

Major League Soccer - MLS

Major League Baseball – MLB

African Cup of Nations- AFCON

Confederación Sudamericana de Fútbol –CONMEBOL

Union of European Football Associations-UEFA

# **Chapter 1**

## **1.1 Introduction and Background**

Football is regarded as the most popular sport in the world. It is an enjoyable form of exercise, and helps develop agility, balance, coordination and sense of team work Stopsportsinjuries.com (n.d.). It is a contact sport and as is common with all contact sports, there is the likelihood of an injury occurring in the life of a player. Players, football teams, player agents, physiotherapists and the country are some important stakeholders in the sport who are affected by injuries sustained by a footballer.

Stephen Appiah, Michael Essien, Junior Agogo, Kwadwo Asamoah and Kevin Prince Boateng are all players who have represented Ghana in football at various times in the past. That's not all they have in common. These players had their careers destroyed by injuries they sustained while playing football (Pulse.com.gh, 2016). These players were fortunate to have played to play in well-established leagues outside Ghana, which meant they got the right treatment for their injuries. They recovered but not to the level they were before they got injured. For a player who plied his career on the local scene, the story is different. A serious injury means his career is over. Former Accra Hearts of Oak and Kumasi Asante Kotoko player, Charles Taylor, is one whose career was ended by injuries (Goal.com, 2012). There are other players who could have also become stars like Charles but injuries cut their careers short before it even began.

## **1.2 Motivation**

In Ghana, there are no specialized sports hospitals to take care of the sports injuries. Typically, when a player sustains an injury while on duty for his team he is let go with no hope for treatment. When a player sustains an injury, the impact is not felt by him

alone. First, he can no longer play and his only means of supporting himself or making money has been lost. His family is affected, as the person they look up to for financial support is no longer available to assist them. The most affected stakeholder is the team he plays for. The team that doled out huge sums of money to buy him in the first place must contend with being without their player for some weeks, months or in some cases years. One player who comes to mind in this case is Andre Ayew, a Ghanaian player who sustained an injury on his first outing for English Premier League club, West Ham United. This was after they bought him for a club record fee. His injury meant that he would be out for close to four months, during which time West Ham's value for their money would be lost (Forbes.com, 2015).

This project seeks to develop a solution that uses the risk factors for sustaining an injury to create a predictive model that will reduce the likelihood of a player sustaining an injury.

### **1.3 Problem Description**

The most common injuries sustained by the sportsman are the lower extremity injuries such as sprains and strains, cartilage tears and anterior cruciate ligament sprains in the knee, overuse lower extremity injury being soreness in the calf (shin splints), pain in the knee or the back of the ankle (Achilles tendinitis), upper extremity injuries, and head, neck and face injuries (Stopsportsinjuries.com, n.d.). Injury has been known to be a major cause of derailing the careers of major sports men and women all over the world. These effects may be physical or psychological. Some emotions that are associated with injuries include sadness, anger, and frustration. In 2015, it was estimated that the average cost of player injuries in the top four professional leagues in Europe was \$12.4 million per team (Goal.com, 2016).

Technology has been used in advanced countries to mitigate the risk factors leading to injuries. However, these practices use sophisticated technology which are expensive and not readily available in Ghana to be used by the team and players that ply their trade here.

#### **1.4 Benefits**

Being able to predict an injury means it can be prevented to an extent. When the risk factors that are most likely to result in a player getting injured are identified, focus can be placed on eliminating these unfavourable risk factors, which ultimately results in less injuries occurring. It will be useful to coaches to identify when their players are more susceptible to injury, and inform their decision to rest them or play them. This saves the team money that would have been lost in treating the player. The player in turn earns money that would have been lost if he were out due to injury and makes sure his family is well taken care of. With respect to the sports entertainment industry, this will be extremely useful in fantasy football team selections for sports fans.

#### **1.5 Objectives**

This project seeks to generate a model using multiple logistic regression analysis that will assist in the prediction of injuries and the possible prevention of injuries in football. Significant objectives that will be achieved in this process include:

- Identifying modifiable risk factors that affect a player's likelihood of getting injured.
- Identifying how these different factors contribute to a player getting injured.
- Design a model to predict the probability of an injury occurring.
- Outlining steps to be taken by stakeholders involved to reduce the occurrences of injuries.

## **1.6 Outline of Project**

This paper contains six chapters and will be outlined as follows:

Chapter 1 introduces readers to the project and the problem it is trying to solve, highlighting the motivation behind undertaking this project in the first place. Chapter 2 reviews related works in sports injury prediction, highlighting scholarly work emphasizing the importance of injury prediction, as well as identifies major risk factors to be considered in injury analysis and prediction. Furthermore, it analyses existing models to find out what new feature can be added or what can be done differently. Chapter 3 describes the functional and non-functional requirements using the the Cross Industry Standard Process for Data Mining (CRISP-DM) 1.0 data mining process model. Chapter 4 focuses on the architecture and factors that will be used in predicting a model. Chapter 5 deals with the implementation of the project, describing the tools and technology and platforms that will be used and explore the reasons why they were chosen. Chapter 6 will cover the testing of the model and results from testing will be discussed. Finally, Chapter 7 will make conclusions and recommendations. Limitations encountered and suggestions for further work are made.

## **Chapter 2: Related Work**

This chapter tackles related work in the field of sports injury prediction. It discusses the relevance of predicting injuries, and then highlights instances where injury prediction was effectively in sports.

### **2.1 Background**

Predicting and possibly avoiding injuries is touted as the next big thing in sports data. For this chapter, a background study into the importance of sports injury prediction was covered in section 2.2. It further analysed the different risk factors that could be assessed in injury prediction. Under section 2.4 a study into the different risk factors was discussed, as well as existing solutions in injury prediction were analysed. Finally, section 2.5 summarises the major solutions currently in place and section 2.6 gives recommendations on how the existing measures can be adapted and improved upon in this project.

### **2.2 Relevance of Sports Injury Prediction**

Gabett (n.d.) in his article, “Injury prevention and performance enhancement in team sports: Train smarter and harder”, discusses how injuries can be prevented and performances enhanced in team sports, basically through training smarter. The paper weighs the argument of the correlation between training loads and injuries from three angles, the first being that suggesting that the harder these athletes train the more injuries they will sustain, and the second that the if training loads exceeded a planned ‘threshold’, athletes were ‘managed away’ from potential injury and finally that insufficient training may lead to increased injury risk.

In their paper, Colston and Wilkerson looked at physiological factor that could lead to a player developing an injury. It used a 3-factor prediction model that looked at injury

risk factors that could be used to identify injuries. The research that accompanied the paper was designed in the form of a cohort study. The purpose of this study was to investigate the relationship between physical workload and injury risk in elite youth football players. The researchers used the workload data and injury incidence of 32 players, monitored throughout two seasons. This approach to injury prediction relied on multiple regression to compare cumulative loads between injured and non-injured players for specific GPS and accelerometer-derived variables. It was discovered that higher accumulated and acute workloads were associated with a greater injury risk. However, progressive increases in chronic workload may develop the players' physical tolerance to higher acute loads and resilience to getting injured.

### **2.3 Factors that Result in Injuries**

Murphy, Connolly and Beynnon in October 2002 undertook a study which investigated risk factors among athletes and military recruits aged between 14 and 39 years for lower extremity injuries. The results of this study were published a paper titled Risk factors for lower extremity injury: a review of the literature, were divided into extrinsic and intrinsic risk factors. The extrinsic risk factors considered were level of competition, skill level, shoe type, ankle bracing and playing surface. Intrinsic factors studied included age, sex, phase of the menstrual cycle (for women), previous injury and inadequate rehabilitation, aerobic fitness, body size, muscle strength, imbalance and reaction time. Their research concluded that there was an increased incidence of injuries on artificial turfs than on grass or gravel. Additionally, there was an increased likelihood of injuries occurring in less skilled players as compared to highly skilled players who can easily maneuver away from an imminent tackle. With respect to intrinsic factors, the study revealed that there was an increased incidence of injuries for players older than 25 years.

## 2.4 Related Work

### 2.4.1 Oslo Sports Trauma Research Center

Roald Bahr and Ingar Holme of the University of Sport and Physical Education at the Oslo Sports Trauma Research Centre Education conducted research into the various factors that could lead to injuries in sports. This was published in a paper titled Risk factors for sports injuries — a methodological approach. Using a multivariate statistical approach, they investigated potential risk factors for injuries. These risk factors were classified using the Meeuwisse model which divides risk factors into intrinsic and extrinsic, and measured the impact these factors had on injuries. These researches used the linear logistic regression model.

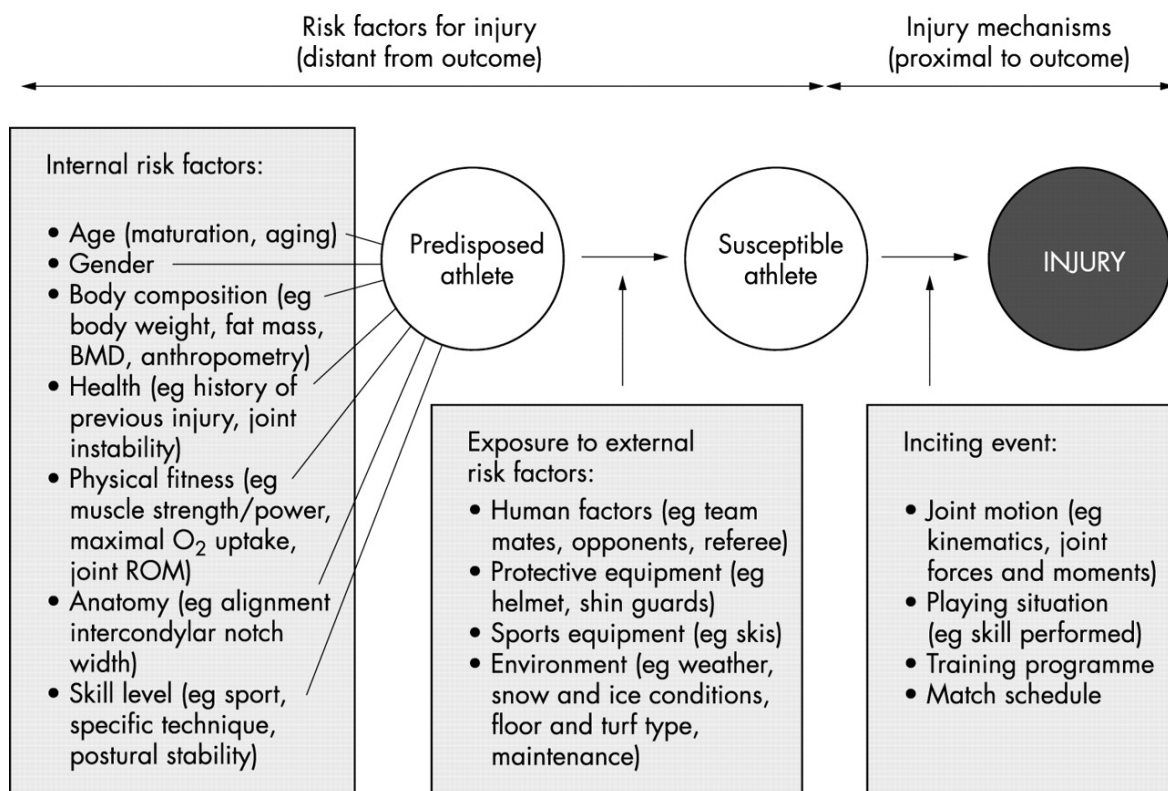


Figure 2.1 The Meeuwisse model for classifying injury risk factors



#### 2.4.2 NSW Waratahs

In Australia, the NSW Waratahs Rugby Team's use of IBM's Predictive Analytics helped reduce player injury. This in turn optimized team performance. The analysis model used predicted the likelihood of a player being injured, informing the coaching team to monitor each player's training program and minimize their chance of getting injured (IBM, 2013).

#### 2.4.3 SAP SE

German company SAP SE, uses sensors and cloud computing through its Injury Risk Monitor to predict and prevent football injuries. It uses huge catalogues of data and its HANA cloud platform to make injuries less likely – and even preventable. Players wear sensors which gathers data while they play, and along with statistics collected from a player's entire career and held on SAP's HANA Cloud Platform. The Injury Risk Monitor then gives a percentage to indicate how likely each player is to injure themselves in their next match. The system considers how fit each player is, based on their diet and exercise regime, along with the date of their last injury, and how long they usually take to recover from a variety of injuries (Ibtimes.co.uk, n.d.).

#### 2.4.4 Team from the University of Birmingham and Southampton Football Club

A team from the University of Birmingham and Southampton Football Club used GPS technology to analyze the performance of youth team players, to study the link between training activity and rates of injury. These GPS trackers monitored their speed, distance travelled and total forces experienced by their bodies on the pitch during games and training. This data was cross referenced against any recorded injuries which caused players to miss training activity – and classified as mild, moderate and severe (Sciencedaily.com, 2016).

#### **2.4.5 Kitman Labs**

Kitman Labs, analyses data about players and can predict when a player might get injured with an unprecedented degree of accuracy. Partnering with the Leinster and Irish rugby teams, they gathered data from athletes via sensors like GPS vests and heart rate monitors ([www.thejournal.ie](http://www.thejournal.ie), 2014). It pulls together data recorded on the player's sleep pattern, work done on the pitch, heart rate variability and other metrics. Kitman is looking to move away from rugby into the United States market where they will look at opportunities in MLB, MLS and every other track and field sport.

#### **2.4.6 Sports Injury Predictor**

Sports injury predictor is an algorithm that determines the probability of a player being injured. It uses an injury database, considering every injury that has taken place, type of injury and kind of treatment required ([Sportsinjurypredictor.com](http://Sportsinjurypredictor.com), 2017). It uses an injury correlation matrix to determine the statistical probability of an injury occurring based on previous injury. It also considers biometrics data like age, height and weight, play by play data, position and how many times player is likely to touch the ball. This is used by the National Football League (NFL) in the United States of America. This algorithm is however pending a patent.

#### **2.4.7 Researchers from Dow Jones and Wall Street Journal**

Researchers from Dow Jones and Wall Street Journal applied advanced machine learning to predict the probability of an injury for a player in the NBA. This was revealed at the MIT SLOAN Sports Analytics Conference. The model they created was based on play-by-play game data, player workload and measurements, and team schedules covering a period of two years. Their approach enabled team management and decision-makers to identify the best time for a team to rest their star players and reduce the risk of long-term injuries, while optimizing team strategies (Talukder & Vincent, 2016).

## **2.5 Findings and Proposed Solution**

From the study of related works, it is seen that most existing models used sophisticated technology to monitor a player's likelihood of injury. These were in the form of wearables like heart rate monitors and GPS vests. Advanced machine learning and multivariate logistic regression were used in some cases by researchers to predict and prevent injuries, using the data collected from monitoring the player's vitals. In some instances, the data was collected in game, which would immediately prompt the medical staff if any change was necessary based on risk factors encountered. It is evident that the use of these technology in predicting injury significantly led to a reduction in the incidence of injuries in various sports ranging from football, basketball and rugby.

## **2.6 Proposed Solution**

The model will take into consideration a mix of intrinsic and extrinsic factors, based on the Meeuwisse model. Considering the limited technology available in Ghana to carry out this prediction, multiple logistic regression is a statistical tool that can be used to determine the probability of an injury occurring.

## **Chapter 3: Requirements**

In this chapter, section 3.1 covers functional requirements using the Cross industry Standard Process for Data Mining. Section 3.2 covers non-functional requirements.

### **3.1 Functional Requirements**

For this project, the Cross Industry Standard Process for Data Mining (CRISP-DM) 1.0 data mining process model will be used (Chapman et al, 2000). Major stages in that model are outlined below in this requirements plan. These stages include business understanding, data understanding, data preparation, modelling, evaluation, and deployment.

#### **3.1.1 Business Understanding**

This section seeks to provide an overview of the project context. It covers the problem that exists and how data mining can be used to provide a solution. Resources that are used in the project are identified as well as constraints. Finally, the criteria by which how success will be measured is outlined.

##### **3.1.1.1 Background**

In sports, an injury is basically an event that causes absence from one or more games or practice sessions. For this analysis on the 2017 African Cup of Nations, an injury is any event which led to a stoppage in play, and required the presence of the medical personnel of the team on the pitch to treat the player. This may have resulted in a player being absent from subsequent games.

Throughout the course of the 2017 African Cup of Nations, there were injuries in almost every game, with most players having their tournament cut short as a result. The main objective of this project is to assess the likelihood of a player getting injured at the

2017 African Cup of Nations tournament and create a model for predicting the likelihood of an injury, using data collected on players who were at the tournament.

#### **3.1.1.2 Resources and Constraints**

Data will be collected from primary sources online. Data sourced from online is open source and will be adequately referenced. A sports expert who was present in Gabon during the African Cup of Nations tournament will be consulted. Data mining tools like R and Excel will be used. A constraint on the data used will be the small sample size involved, as the Cup of Nations tournament lasted for only three weeks and involved only 368 players, out of which less than 250 actively participated in the tournament.

#### **3.1.1.3 Success Criteria**

Success will be measured in the short term if it is identified that there exists a relationship between injury and risk factors identified. Medium term success will be determined by improvement in the states of the significant risk factors that have been identified to have strong relationship with injuries. In the long term, success will be measured by a reduction in the injuries in footballers.

### **3.1.2 Data Inventory and Understanding**

The next step in this project would be to collect the data that is available for this project. A description of database acquired is given. Steps are then taken to verify data quality.

#### **3.1.2.1 Description of data**

Proprietary data in this case is obtained from the 2017 African Cup of Nations and interviews with experts who were in Gabon for the tournament. External data sources on match intensity would be generated from FIFA.com.

There are different factors that can cause an injury and these are classified as intrinsic or extrinsic. Intrinsic are those factors that are peculiar to the player while extrinsic are environmental factors the player may encounter. These environmental factors include weather, pitch surface and football boot type. The factors likely to cause injuries may further be classified as modifiable and non-modifiable. Examples of non-modifiable factors may be gender and age, while pitch surface quality are modifiable factors. These factors may be further classified as continuous, for example age or categorical. Pitch surface quality is regarded as categorical.

#### **3.1.2.2 Understanding data**

In identifying the risk factors to use in the analysis, different variables were collected. The next step gives an indication of which variables were selected to be used as risk factors. Minutes played was identified as a factor/variable that could be used but this is an after the fact variable, as every player would want to play 90 minutes, given the absence of injury. Fouls suffered is also an indication of how a player is susceptible to injury but the player does not determine how often he is fouled in a match situation. However, the number of fouls a player suffers can be high or low based on the function he plays on the field.

After the data understanding process, player's age, pitch surface quality, player function and match intensity were selected as risk factors to be analysed in the data mining process.

#### **3.1.2.3 Data quality**

Data collected online is verified across different sources. The data is passed through the sports expert to be the final check on the quality of data obtained from online.

### **3.1.3 Data Preparation**

The next step in this would be analysing the data with the objective of determining whether the available data can be used to come up with a model for predicting injury. If the data is found not to be suitable, then the data will be modified or manipulated to suit the objectives. If this is still not possible, objectives may be scaled back to reflect what is possible with the data given.

#### **3.1.3.1 Data Selection**

Data collected online included the tackles a player made, duels a player was involved in, minutes played, total passes completed. However, these risk factors were not used as there was no way these could be predicted before a game.

#### **3.1.3.2 Data Cleaning and Construction**

The huge amount of data received will be prepared and put in a format which is much easier to work with so that it can be successfully modelled. This would involve a straightforward examination of the main components of the data expected to be relevant to modelling. Steps to be taken include scanning the data for extreme and/or invalid values, locating oddities in the data as well as ensuring the data provides appropriate values. Data collected from different data sources are merged into one dataset to be used in this analysis. Data is then stored in the .csv format that can be used in the R software. To run the logistic regression, the dependent variable in the model, injury, is converted into binary format, with 0 for no injury and 1 for injury, irrespective of the number of injuries sustained in the game.

#### **3.1.4 Data modelling**

To test this project to see how viable it is, a test run will be done on a handful of participants. At the very least, analytics would be conducted on a radically slimmed down

version of the data set to accelerate modelling run times. Afterwards, this would be expanded to the entire dataset. This stage will involve automatically running and summarizing a series of experiments. Using a logistic regression model, data will be analysed to find a relation between the key factors identified and the likelihood of injury.

In creating the model, a statistical approach is taken using the multiple logistic regression.

### **3.1.5 Testing, Evaluation, Interpretation, Understanding**

This stage will take advantage of the model created to decide on whether injuries can be prevented or not. The model will be interpreted in the sense of what it can do to help prevent injuries and preserve the playing lives of players. As compared to the previous stage that will make use of logistic regression, this stage will just involve analysing the results of the previous stage.

### **3.1.6 Deployment**

This is done once the project is tested and provides the model required. Data is deployed in the test environment. A varied selection of possible variables that would have existed should another match have been played at the tournament will be run through the model.

## **3.2 Non-Functional Requirements**

### **3.2.1 Product Requirements**

#### **3.2.1.1 Performance**

The model should be able to perform the basic function of predicting the possibility of an injury when it is fed with data. It should possess the ability to handle large datasets.



### **3.2.2 Security Requirements**

The system should keep player data secure and confidential.

## **Chapter 4: High Level Architecture**

This chapter discusses the algorithm that will be used in analysing the data.

### **4.1 Data Mining Algorithm**

The data mining algorithm used in this prediction model generation is multiple logistic regression. This is used because the model will try to determine binary outcomes.

The logistic regression approach, was used because it gives the user a reliable way of assessing the model and assessing predictors. It assists the user to predict an outcome variable that is categorical from predictor variables that are continuous and/or categorical (Ziegler-Hill, n.d.). The presence of age (continuous variable) and pitch quality (categorical variable) made it essential to use logistic regression.

Additionally, logistic regression was used because it could answer the following questions which were highlighted in the project objectives.

- Can modifiable risk factors that affect a player's likelihood of getting injured be identified?
- How do these different factors contribute to a player getting injured?
- Do some of the risk factors interact with each other?

## **Chapter 5: Implementation**

This implementation section will cover the datasets needed to complete this project and sources where it will be obtained from, as well as tools and languages used. For this analysis, the external factors considered were pitch quality, match intensity and player position. Age was considered as an intrinsic factor. It also indicates how the data was entered into the Excel workbook.

### **5.1 Data sources**

Data was collected from BBC.com, which gave a play by play commentary on Match proceedings. This data was corroborated with data from Guardian.com, which gave a play by play commentary as well. A third data source used was Sofascore.com which provided match and player statistics, as seen in appendix A. FIFA.com also provided data on match intensity.

#### **5.1.1 AFCON Data**

The first step of the implementation involved collecting data on player injuries at the 2017 African Cup of Nations. This competition provided the best avenue for gathering information as there was a widespread criticism on the high number of injuries that occurred, with multiple attributed to the bad quality pitches. Characteristics of this tournament are given in the following sentences. 16 nations qualified for this tournament hosted in Gabon, with each country presenting a 23-man squad of players. There were four stadia used with four teams stationed permanently at each of these venues throughout the group stages. Each team played 3 matches during the group stages in a round robin format, with the top two teams from each group qualified for the quarter finals. From this stage onwards, all games were played in a knockout format till the finals. In all, there were 32 games played at the tournament. A total of 52 goals were scored across all 32 games

indicating an average of 1.6 goals per game. The low number of goals scored on the average was an indication of the defensive minded approach most teams took towards the games. 84 yellow cards were given out in total, an indication of how aggressive and physical the matches were.

The nature of data gathered from the AFCON is found below:

- Player injury data
- Player's function on the pitch (forward, midfielder, defender, goalkeeper)
- Player's age (continuous data- integer value)
- Pitch quality data

#### **5.1.1.1 Player injury data**

All instances of a player being substituted because of an injury were noted by 1. Searching for this data was made easy using Sofascore.com, which indicated the players who were substituted because of an injury. This resulted in 24 unique observations of players who had to be substituted. This data was corroborated with the play by play accounts on both BBC.com and Guardian.com. Using the web browser's inbuilt search function, the keyword injury from bbc.com was searched for, and all observations noted.

A different record of injuries was also obtained which noted the different times a medic team had to come onto the football pitch to treat an injury. For anytime play had to be stopped for the medics to treat a player or for the player to be stretchered off the field, it was with 1, and all other instances with a 0. For this record that was taken, there was a total of 104 injuries. For the analysis performed on this dataset, the second record of injuries requiring medics was used. This was because it was observed that because some teams had exhausted their substitution options, some players had to play through their injury. This would then make using instances where players were substituted inaccurate.

From Figure 5.1 below it is observed that most injuries happened in the second half, with a total of 92, as compared to 12 in the first half.

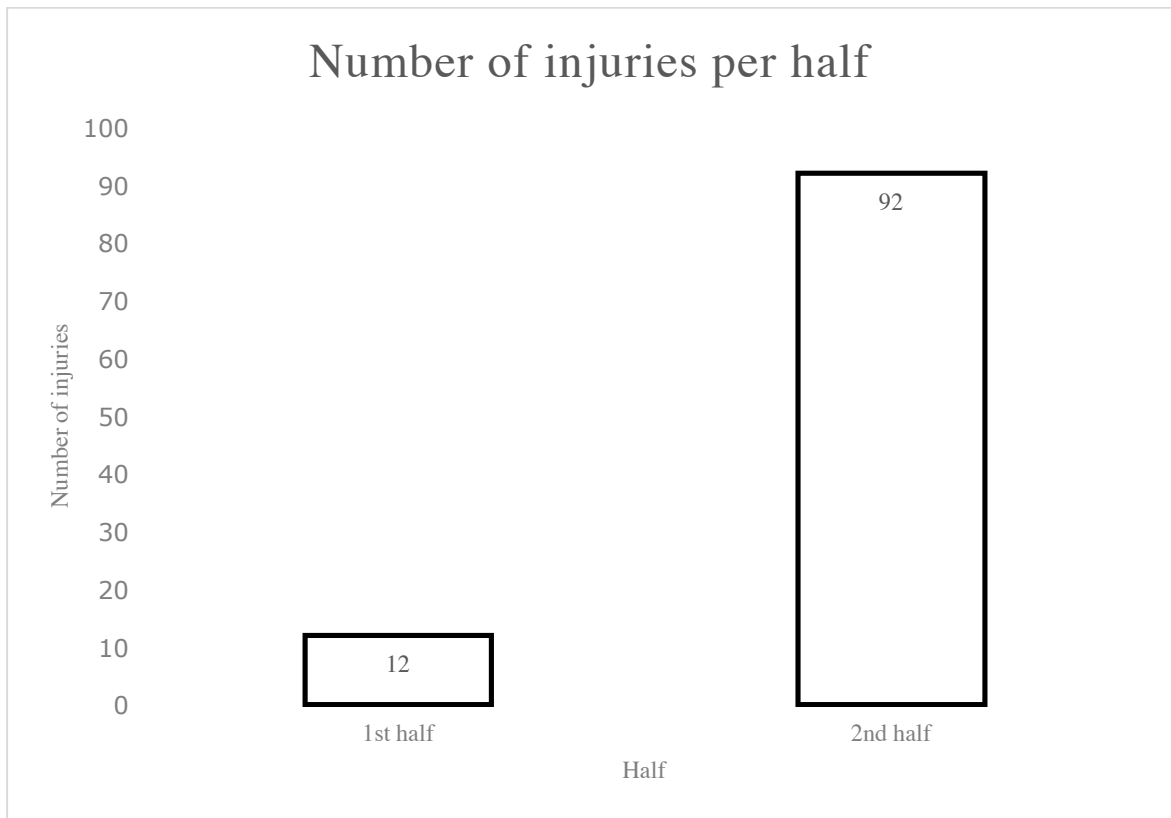


Figure 5.1 Injury distribution based on periods in the match

#### 5.1.1.2 Player's function on the pitch

For any football match, a player can assume the position of goalkeeper, defender, midfielder or forward. In all, there were 48 goalkeepers at the tournament, 119 defenders, 114 midfielders, and 87 forwards. This data was obtained from the official Confederation of African Football website, from the official list of players submitted. In entering the data into Excel, goalkeepers were denoted with 1, defenders with 2, midfielders with 3 and forwards with 4.

#### **5.1.1.3 Player's age**

The player's age represented continuous data. This data was gathered from transfermarkt.com which is an online database of professional football leagues, clubs and players. Each player's name was entered into the search panel on the transfermarkt.com website which gave the players date of birth and age. This assisted in deriving the player's age as at the time the tournament started. The oldest player was 44 years old while the youngest player was 18 years old. The mean age was 26 years. 108 out of the 368 players present at the tournament were aged 24, making it the modal age.

#### **5.1.1.4 Pitch quality**

Pitch surface quality was selected as an external factor to be analysed after coaches and doctors attributed the high injury rate to the pitches. Frank Boahene, a pitch expert and managing director of Green Grass Technology (GGT), a pitch development solutions firm, revealed that the pitch could be blamed for the knee joint, ankle and hamstring injuries that players suffered at the tournament (Boahene, 2017).

There were four different stadia used at the 2017 AFCON, with their pitches possessing varying degrees of quality. The stadium in Franceville was viewed as having the best pitch and was given a rank of 1. Group B played its group matches here. This was followed by the stadium in Oyem, ranked 2<sup>nd</sup>. This was a new one built specifically for the tournament. Group C matches were played here. Notwithstanding its good quality, drainage on the pitch was poor and it was known to get soggy whenever there was very heavy rain. The stadium in Libreville hosted Group A. This pitch was regarded as sandy and was given a rank of 3. The last match venue was the stadium in Port Gentil, which was one of the two new stadia built for the tournament. Unlike Oyem, this was regarded as the worst pitch in the tournament. It was very sandy, which made balls bounce awkwardly. This stadium hosted group D and was ranked 4th. This ranking was done based on insight

gained from an interview with Citi FM sports editor Nathan Quao (2017). Figure 5.2 below shows the number of matches played on different pitches at the tournament. Most matches were played on pitch quality 2 which was the best at the tournament.

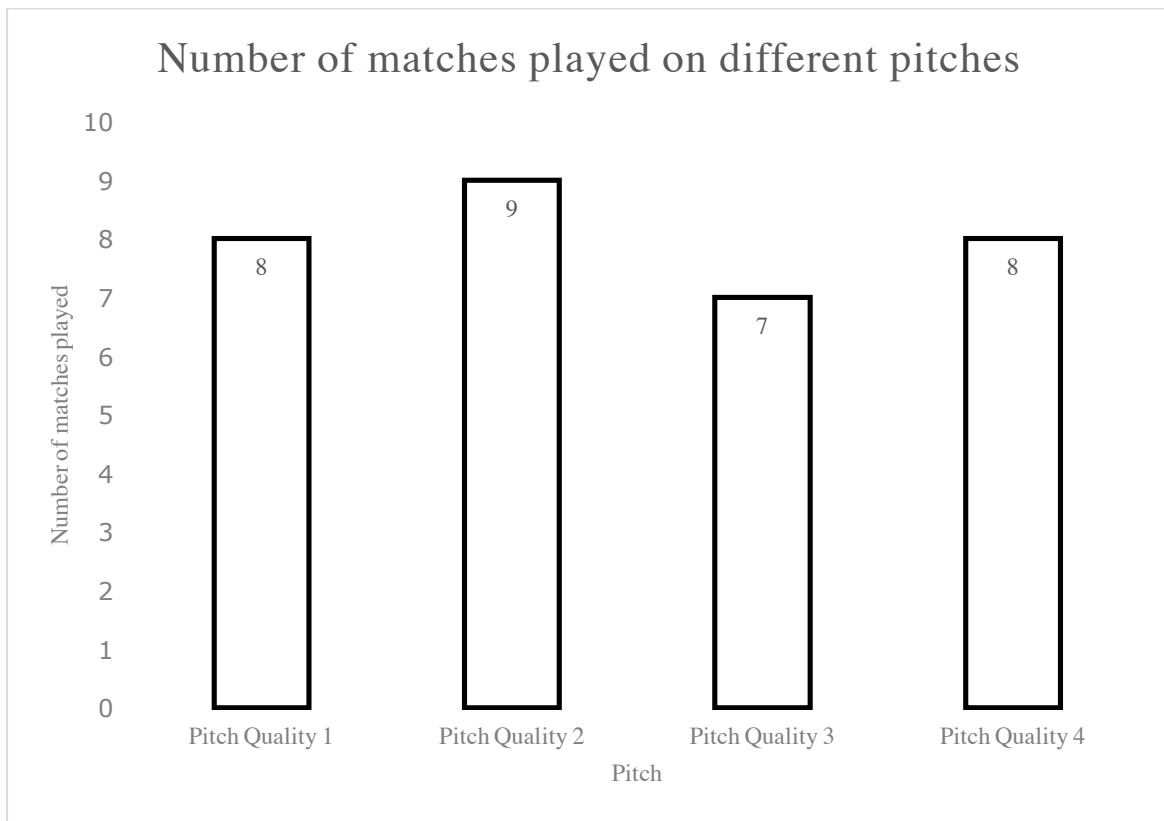


Figure 5.2 Number of matches played on the different pitches

### 5.1.2 Match intensity data

Data for determining match intensity was sourced from the FIFA Coca Cola world ranking found on the FIFA website. This is because any team that performs in football gains points which enable it to rise in the rankings. Countries that are ranked close to each other have a similar performance record, while countries that are far apart on the rankings have varying performance levels. Matches involving countries close on the rankings indicate that they have the same level of skill and so it would be an intense match involving two teams that are well equipped to win. This would make a match between these two teams an intense one, as compared to one with teams far away from each other

on the ranking. Comparison of the different countries is seen in Appendix B and Appendix C.

The ranking is determined by the Fédération Internationale de Football Association (FIFA). The world football governing body releases a ranking monthly on all male senior national teams. To ensure that the team's standing reflects current performance, a team's total number of points over a four-year period is determined by adding the average number of points gained from matches during the past 12 months and the average number of points gained from matches older than 12 months which depreciates yearly.

#### **5.1.2.1 FIFA's calculation of rankings**

The rankings are calculated as follows based on the FIFA formula (2017):

$$\text{Points} = \mathbf{M} \times \mathbf{I} \times \mathbf{T} \times \mathbf{C}$$

where:

‘**M**’ is the points for match result.

‘**I**’ is the importance of the match.

‘**T**’ is the strength of opposing team

and ‘**C**’ is the Strength of confederation

Figure 5.3 below shows the different levels of match intensity that was witnessed at the tournament. There were 14 matches that were extremely intense, with 5 moderately intense.



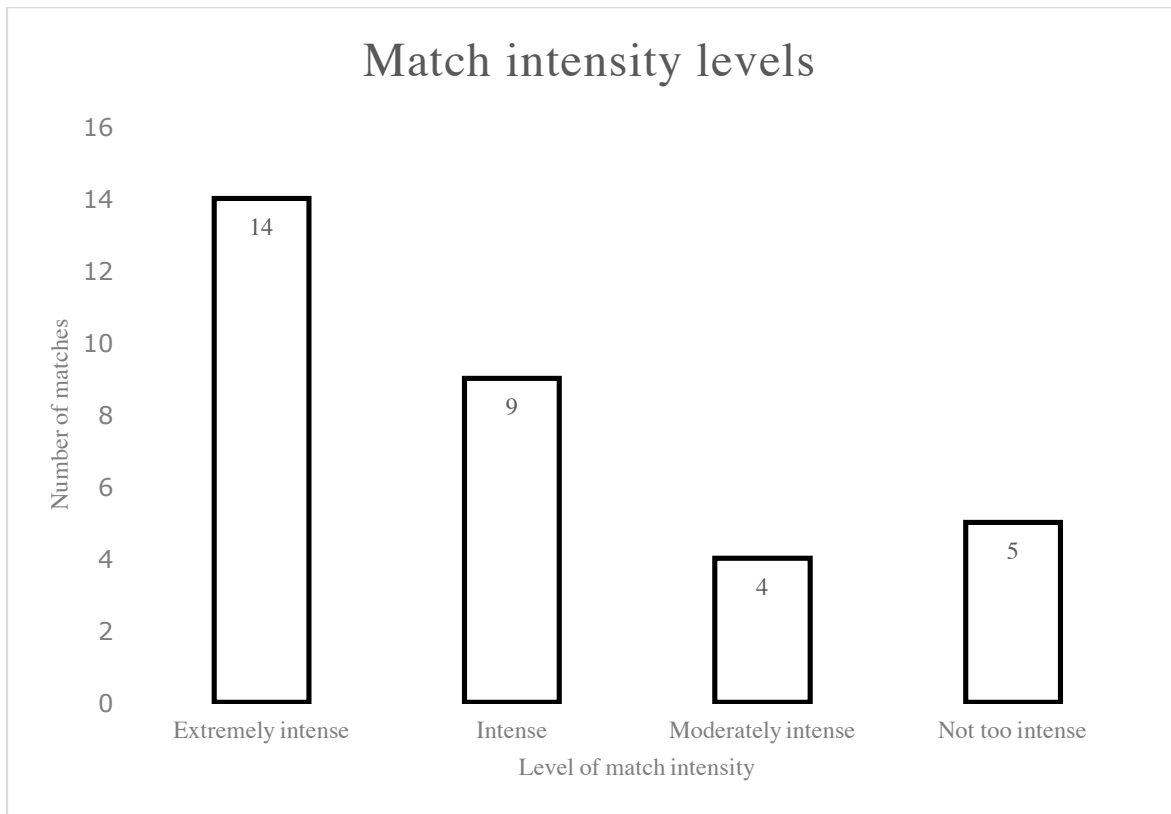


Figure 5.3 Match intensity levels for all matches

#### 5.1.2.2 Calculating match intensity

The rankings for January 2017 was scraped from the FIFA website using, DataMiner, a web based add-on. This was then placed in an Excel workbook after which it was manipulated into a format that would be easy to work with.

First, the list of over 210 member countries of FIFA was pruned down to the 16 countries that qualified for the AFCON tournament, with their respective ranking on the log. Next, the difference in ranking for all the participating nations were determined. For instance, if Senegal was ranked 1<sup>st</sup> and Ghana was ranked 9<sup>th</sup>, then the difference in their ranking would be 9 places. This was done for all countries, as is seen in Appendix A. It was observed that the highest difference in ranking was 70 and the lowest difference in ranking was 1. This difference in ranking was used to determine match intensity as follows:

- All 32 matches that took place were noted down with their corresponding differences in ranking.
  - With the varying differences in ranking, all matches with differences greater less than 17.25 were given a rank of 1. Denoting that the match was very intense. Matches with a difference score ranking between 17.25 and 34.50 were given a rank of 2. For matches with difference score between 34.50 and 51.75, a score of 3 was given. 4 was reserved for matches with difference score of 51.75 and above.
- Appendix B.

## **5.2 Tools**

For data collection DataMiner, a web based add on for the Google Chrome browser was used. Excel was used in putting the data together into a format that would be easy to manipulate in R. R statistical software was used for data analysis.

## **5.3 Language**

The language used in the generation of this model is R. A multiple logistic regression was run on the dataset which included 5 variables: Injury, Pitch quality, Match intensity, Player function and the Player's age.

## **5.4 Implementation**

Both linear and logistic regression was run on the dataset to observe the factors that were very significant in the model. The linear regression model was run on two datasets. The first dataset had players whose injuries were so serious to the extent that they had to be substituted. The second dataset involved injuries that had medics coming onto the pitch, including those that led to substitutions. Afterwards, the logistic regression was run on the second dataset which was a more accurate representation of the injury statistics. Different stages with the implementation and results at each phase is explained below. The

implementation will seek to confirm whether risk factors that affect a player's likelihood of getting injured can be identified, and the level at which rate these factors contribute to a player getting injured. Appendix D shows a cross section of the data that was collected and used in the analysis.

#### **5.4.1 Dataset 1**

This recorded injury of players who had to be substituted. In all there were 24 unique observations of this. Other columns which were included were player's age, Player function, Match intensity and pitch quality. A regression was run on this data, and the result seen as follows:

The only hint of a slightly significant value was with match intensity with a significance level of 0.1. All other variables were insignificant. This prompted the use of Phase B data. This was because it was identified that some players had to play through their injury due to selection constraints on the match. Thus, analysing this data without them would have rendered the results inaccurate.

```

> fit

Call:
lm(formula = Injury ~ PQ + Age + Match.intensity + Player.position,
    data = data2)

Coefficients:
    (Intercept)              PQ              Age  Match.intensity
    0.007996         0.001513         0.003227         0.008624
Player.position
   -0.003820

> summary(fit)

Call:
lm(formula = Injury ~ PQ + Age + Match.intensity + Player.position,
    data = data2)

Residuals:
    Min       1Q   Median       3Q      Max
-0.05612 -0.03023 -0.02175 -0.01464  0.98766

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.007996   0.026371   0.303   0.7618
PQ             0.001513   0.004547   0.333   0.7393
Age            0.003227   0.007671   0.421   0.6740
Match.intensity 0.008624   0.004772   1.807   0.0711 .
Player.position -0.003820   0.005637  -0.678   0.4982
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.153 on 871 degrees of freedom
Multiple R-squared:  0.004634, Adjusted R-squared:  6.332e-05
F-statistic: 1.014 on 4 and 871 DF,  p-value: 0.3992

```

Figure 5.4 Linear regression results on Dataset 1 (very serious injuries)

### 5.4.2 Dataset 2

This accounted for all cases a medic team had to come onto the football pitch to treat an injury. There were 104 times this happened. From the regression output, it is observed that Pitch quality and Player position are significant in determining injuries. From a table plot generated of player position and injuries sustained, it is observed that goalkeepers stand a higher chance of getting injured based on historical data from the 2017 AFCON.

```

> fit

Call:
lm(formula = Injury ~ PQ + Age + Match.intensity + Player.position,
    data = data2)

Coefficients:
    (Intercept)              PQ              Age  Match.intensity
    1.465e-01      1.809e-02     -8.814e-05      1.595e-02
Player.position
   -3.682e-02

> summary(fit)

Call:
lm(formula = Injury ~ PQ + Age + Match.intensity + Player.position,
    data = data2)

Residuals:
    Min       1Q   Median       3Q      Max
-0.2260 -0.1408 -0.1045 -0.0678  0.9689

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.465e-01  9.777e-02   1.498  0.13440
PQ             1.809e-02  9.871e-03   1.833  0.06712 .
Age           -8.814e-05  3.055e-03  -0.029  0.97699
Match.intensity 1.595e-02  1.010e-02   1.579  0.11466
Player.position -3.682e-02  1.190e-02  -3.093  0.00204 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3216 on 871 degrees of freedom
Multiple R-squared:  0.01716,    Adjusted R-squared:  0.01264
F-statistic: 3.801 on 4 and 871 DF,  p-value: 0.004516

```

Figure 5.5 Linear regression results on Dataset 2 (all injuries)

Given the significant factors identified, this prompted a bar plot to be drawn to show the relationship between player function and injury, and find which function was more likely to get the player injured.

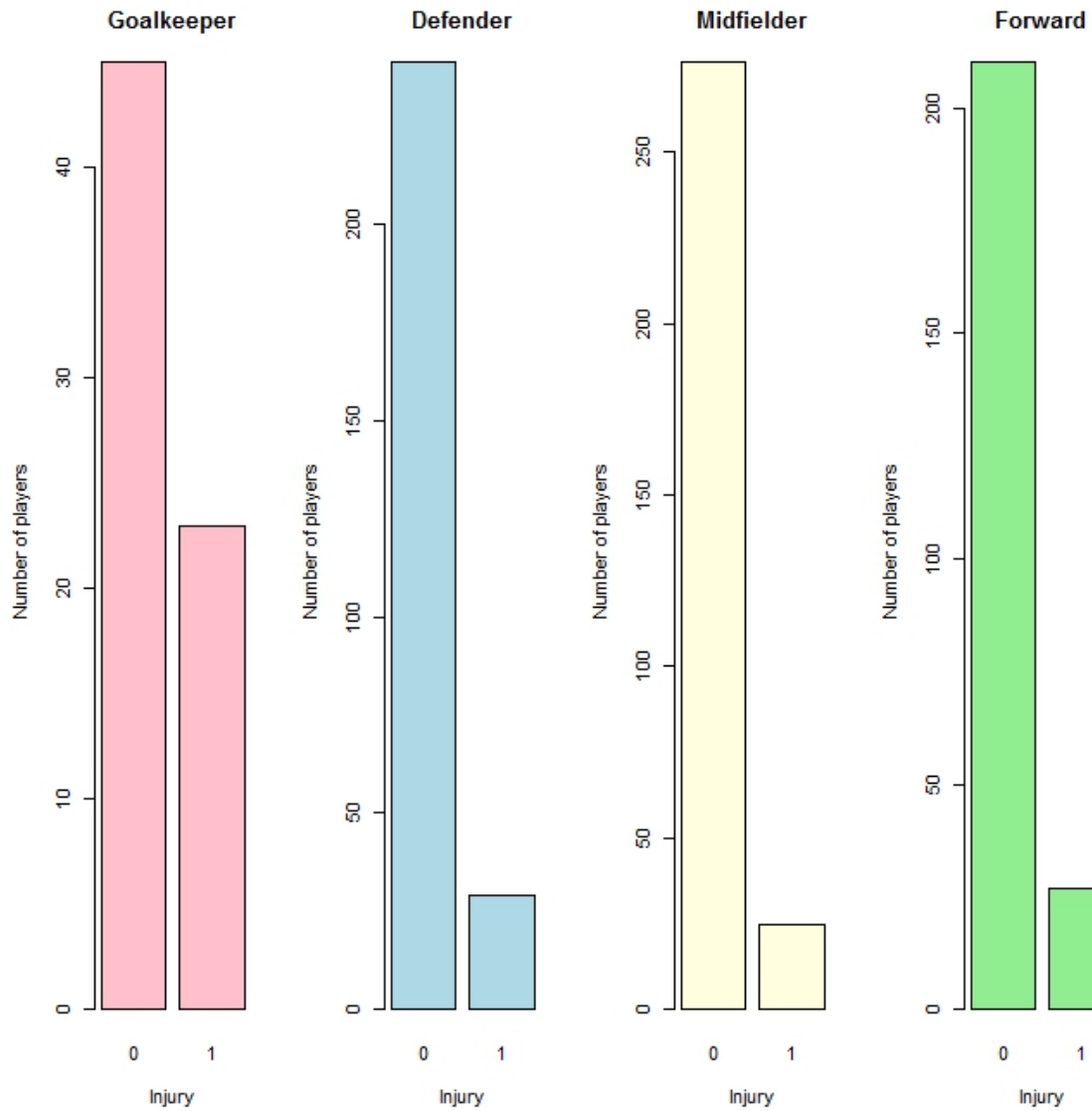


Figure 5.6 Plot of player function and injuries sustained

Additionally, a bar plot to be drawn to show the relationship between pitch quality and injury, and give a graphical view of which pitch quality was more likely to get the player injured.

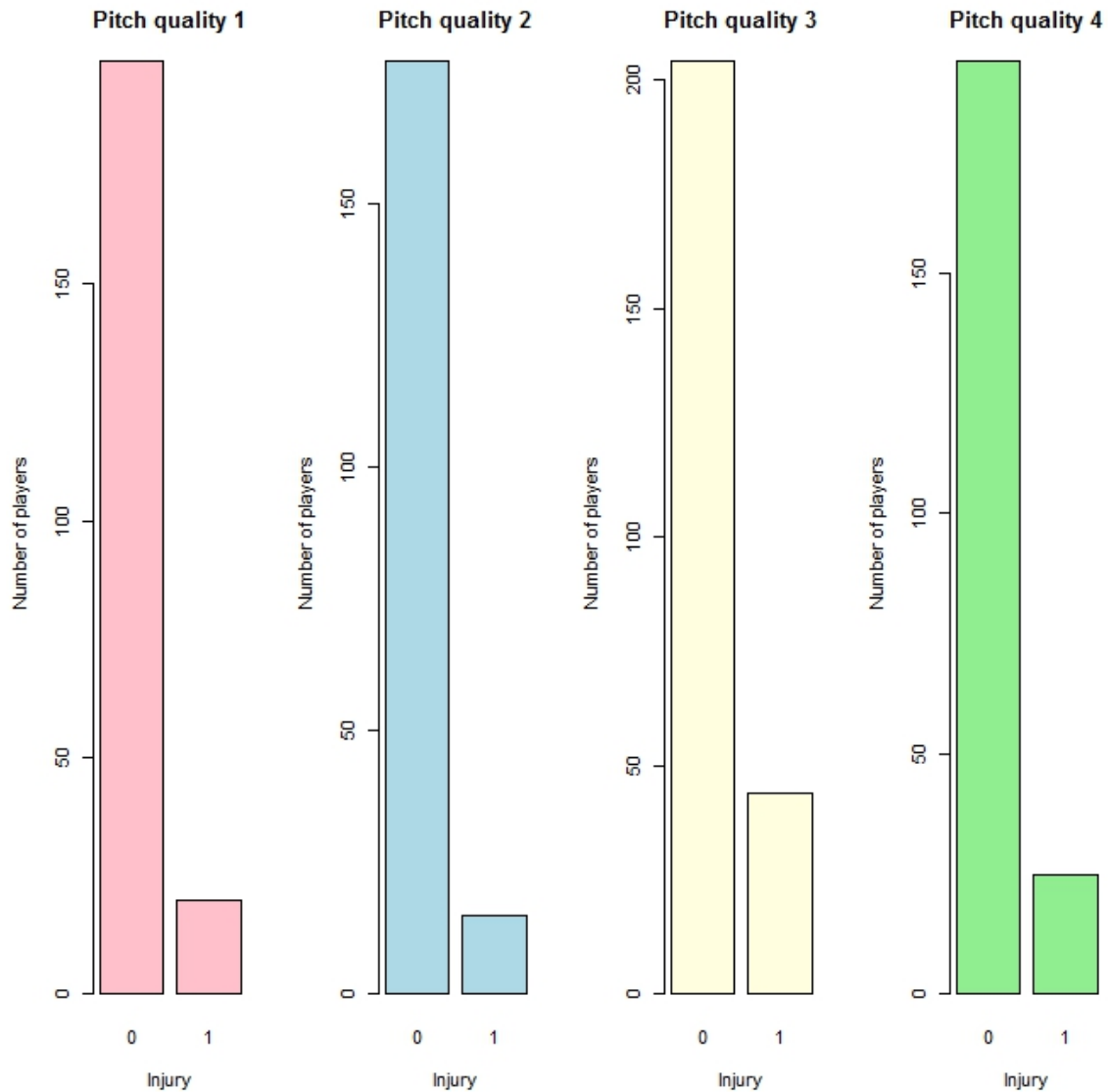


Figure 5.7 Plot of pitch quality and injuries sustained

### 5.4.3 Logistic regression

The Data being analysed consist of injury data for 876 instances a player stepped onto the pitch, and a binary variable coded as a 1 if a player gets an injury and a 0 if not. For each player, it is also indicated if he is a goalkeeper (1), defender (2), midfielder (3) or forward (4). The pitch surface quality they played on where very good is coded as 1, good is coded as 2, okay is coded as 3 and poor is coded as 4. A very intense match is coded 1

and a match with very low intensity is coded 4. Shown below is the data for the first 6 players.

```
> head(data)
  Age PQ Injury Match.intensity Player.position
1  34  3     0                3              1
2  24  3     0                3              2
3  23  3     0                3              2
4  28  3     0                3              2
5  23  3     0                3              2
6  28  3     0                3              2
```

Figure 5.8 Head view of dataset

A plot of the data was then generated to show the relationship between the sets of data.



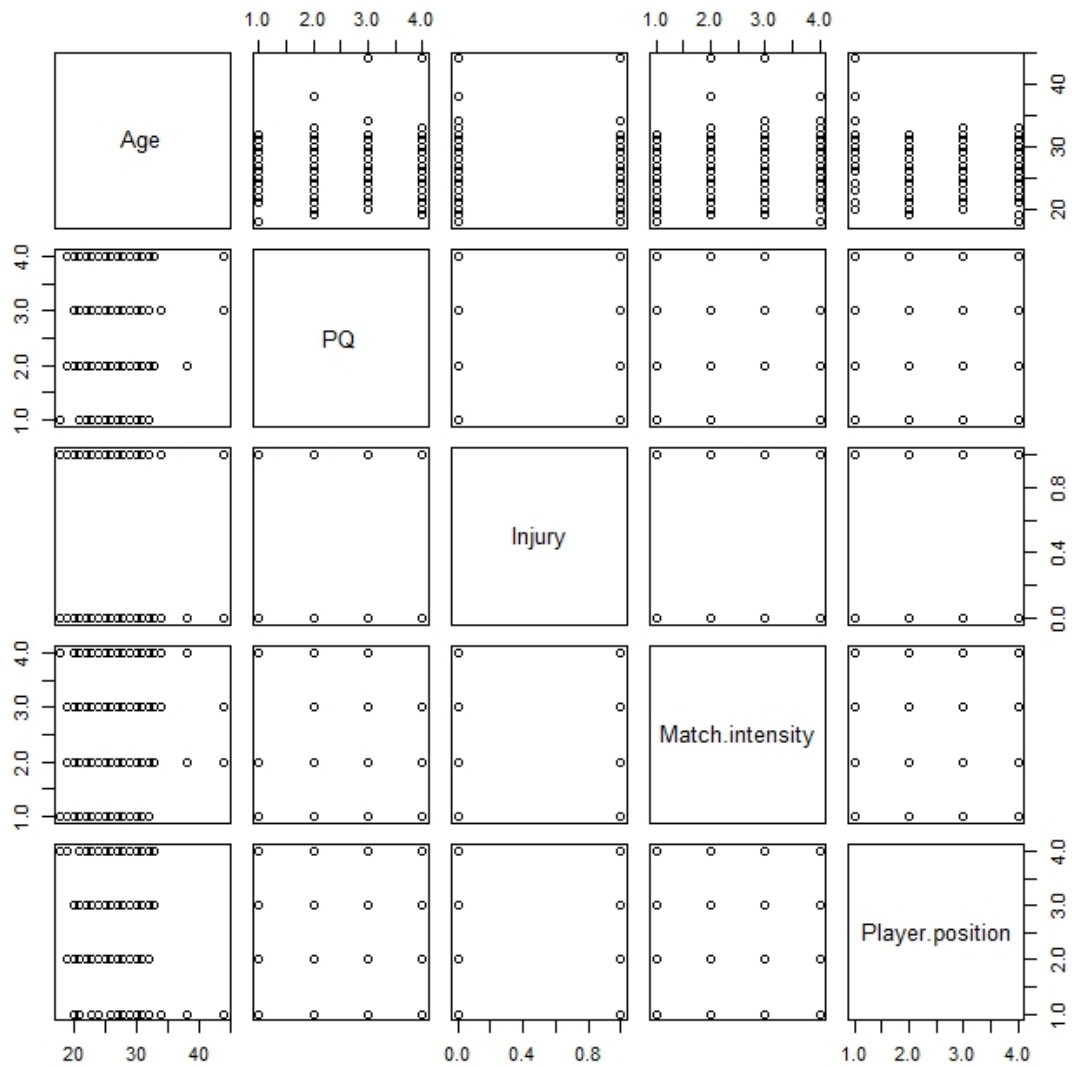


Figure 5.9 Plot of all variables considered in the analysis

A bar plot was then generated to display categorical data in a bid to assess if there are any interesting trends to consider.

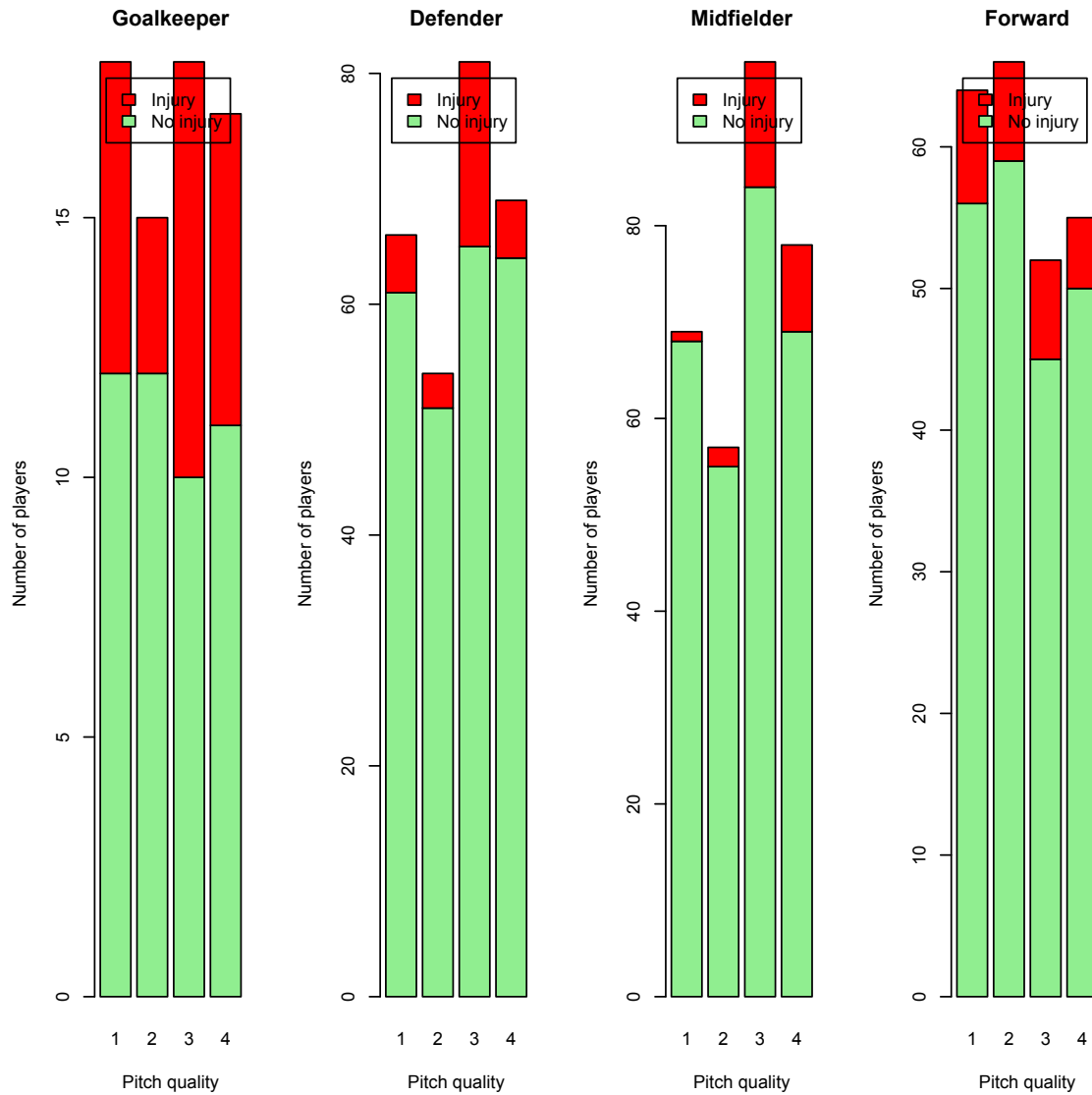


Figure 5.10 Plot of player function on different pitch qualities

The findings from the linear regression is enforced, that goalkeepers stand a higher chance of getting injured. Also, pitches with quality 3 and 4 which were deemed to be poor had more injuries occurring on it. For goalkeepers, they would prefer to play on pitch quality 2, as the number of injuries that occurred on this pitch were relatively less than the rest. Additionally, a closer look at the injury plot for Midfielders show they sustained much injuries on pitch quality 3 and 4. Also, defenders suffered more injuries on pitch quality 3. However, for forwards, the numbers are relatively equal.

Correlation.

A correlation matrix was run to display the correlations between each of the variables with each other. It is observed that there is positive correlation between Match intensity and Player Position and a negative correlation between Match intensity and Pitch quality.

	PQ	Match.intensity	Age	Player.position
PQ	1.00000000	-0.153664720	-0.03812075	-0.036054876
Match.intensity	-0.15366472	1.000000000	0.05690210	0.006766718
Age	-0.03812075	0.056902104	1.00000000	-0.141648213
Player.position	-0.03605488	0.006766718	-0.14164821	1.000000000

Figure 5.11 Correlation output

A logistic regression model was then run on the dataset.

```

> fit

Call:  glm(formula = Injury ~ PQ + Match.intensity + Age + Player.position,
          family = binomial(logit), data = data)

Coefficients:
      (Intercept)              PQ  Match.intensity              Age
      -1.77223          0.18377          0.15981         -0.00326
Player.position
      -0.35568

Degrees of Freedom: 875 Total (i.e. Null);  871 Residual
Null Deviance:      638.4
Residual Deviance: 623.2      AIC: 633.2
> summary(fit)

Call:
glm(formula = Injury ~ PQ + Match.intensity + Age + Player.position,
    family = binomial(logit), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7753  -0.5386  -0.4571  -0.3837   2.4433

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.77223    0.91466  -1.938  0.05267 .
PQ              0.18377    0.09794   1.876  0.06060 .
Match.intensity 0.15981    0.09841   1.624  0.10440
Age            -0.00326    0.02795  -0.117  0.90715
Player.position -0.35568    0.11614  -3.063  0.00219 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 638.38  on 875  degrees of freedom
Residual deviance: 623.17  on 871  degrees of freedom
AIC: 633.17

Number of Fisher Scoring iterations: 5

```

Figure 5.12 Logistic regression output

The above gives us regression coefficients estimates with standard errors. The player position variable is significantly different from zero with a 0.01 significance level. None of the other coefficients are significantly different from zero with only Pitch Quality having significance of 0.1. This reveals that out of age, player function, pitch quality and match intensity, player function and pitch quality are the most likely to make a player at the African Cup of Nations get an injury, with the function a player performs on the pitch

being the likeliest factor to cause an injury. This logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable (Ziegler-Hill, n.d.). For example, from figure 5.4.4, for every one unit change in age, the log odds of sustaining an injury decreases by 0.00326. The logistic regression also confirms the results from the bar plots and linear regression, that pitch quality and player function play a part in the likelihood of a player getting injured.

### **Implications of Results**

Having identified pitch quality and player function as serious risk factors, these can be worked on and modified to reduce the possibilities of an injury occurring. Out of these two, player function however has a higher impact on a player getting injured. Pitch quality is the most modifiable of all the factors, and regarding this, stadium administrators can place more emphasis on upgrading pitches to grade A pitches. Players should also take precautionary measure when playing in conditions that pre-exposes them to injury risk factors.

## Chapter 6: Testing

For this testing, the model generated will be fitted with dummy variables to determine if it is accurate. This will be compared to results from the predict function in R. For instance, from the observed historical data, it is seen that goalkeepers stand a greater risk of getting injured as compared to player in other positions. Thus, if a goalkeeper's details are to be keyed into the model, the likelihood of an injury occurring should be high.

### 6.1 Testing results

The inbuilt R prediction model was used to test the testing variables generated. Unsurprisingly, the testing data that generated the highest percentage of an injury belonged to a goalkeeper playing on poor playing surface. Using 876 testing variables generated, the likelihood of an injury occurring ranged between 0.04962 and 0.26840. For this analysis, a cut-off point was decided on for determining the occurrence of injuries. If the likelihood of getting an injury was above 18%, then there was sure to be an instance of a player getting injured, and any percentage that fell below 12% indicated that the likelihood of an injury was low.

When this result from the test was compared with the result from the logistic regression and bar plots, goalkeepers had injury percentages above the cut-off point. This was seen in high injury percentages for pitches with poor surface quality.

Figure 6.1 shows the results of the test carried out in R, with 382 unique probabilities generated.

```

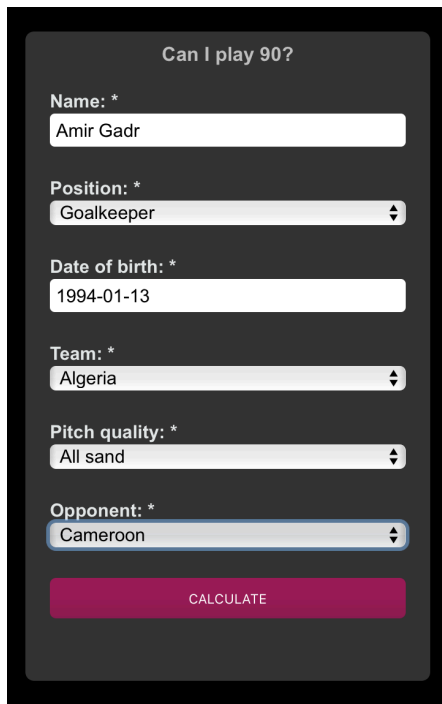
[1] 0.23005010 0.17783260 0.17830981 0.17593376 0.13235920 0.13086863 0.13198522
[8] 0.09656949 0.09600207 0.09515653 0.23411764 0.17640697 0.17405091 0.13123993
[15] 0.13161212 0.09487616 0.09628541 0.18463523 0.13502579 0.13426599 0.13540705
[22] 0.09801894 0.09744392 0.07227297 0.07183696 0.07161986 0.07140337 0.18512658
[29] 0.13578922 0.13655628 0.10005538 0.09773105 0.06990479 0.07097221 0.13440230
[36] 0.09697450 0.09783446 0.09726040 0.07019354 0.07040664 0.06955782 0.05165881
[43] 0.05118171 0.04992974 0.05070878 0.19738351 0.14904590 0.15070761 0.14863284
[50] 0.14658170 0.10742239 0.11026860 0.10899555 0.07824015 0.07847561 0.07942397
[57] 0.07918591 0.13139657 0.09812261 0.07083463 0.07126503 0.07062034 0.05023999
[64] 0.05055205 0.05039579 0.20207366 0.14780955 0.14739933 0.10867931 0.10994912
[71] 0.07777116 0.07894849 0.15596132 0.11496415 0.11730686 0.11364382 0.11397265
[78] 0.08292849 0.08194201 0.08317680 0.08243391 0.05940439 0.05995330 0.05922247
[85] 0.06125259 0.22927075 0.17576787 0.17529601 0.17435528 0.17482514 0.12889690
[92] 0.12599637 0.13036839 0.12853125 0.09394203 0.09311276 0.09256357 0.09201729
[99] 0.09283780 0.11629782 0.11463282 0.11596315 0.11562931 0.08268087 0.08392576
[106] 0.06013732 0.05904107 0.06069256 0.05976980 0.22356019 0.17388641 0.17202090
[113] 0.12816648 0.12963087 0.12563776 0.12743959 0.09174524 0.09366487 0.23726053
[120] 0.18136404 0.18088045 0.18184863 0.13286592 0.13475543 0.13249072 0.13437572
[127] 0.09955430 0.09638506 0.09553650 0.17991629 0.17943572 0.13590000 0.13361901
[134] 0.13324201 0.09695450 0.09781430 0.18233423 0.18039787 0.17847759 0.17895615
[141] 0.13399692 0.09724035 0.09610146 0.09839126 0.23903521 0.22852386 0.17752346
[148] 0.17800003 0.13174301 0.13513604 0.13211641 0.09810240 0.25956657 0.20241060
[155] 0.20293749 0.19979149 0.20031327 0.15181288 0.15014114 0.15055764 0.11143855
[162] 0.11079447 0.10983447 0.18170827 0.13275723 0.13617229 0.09946973 0.09888714
[169] 0.09917805 0.07075753 0.09976217 0.26843532 0.20188474 0.20346540 0.14931096
[176] 0.14848458 0.15223319 0.14889730 0.20050110 0.14822072 0.10773542 0.08086615
[183] 0.08014210 0.13214261 0.09899163 0.09754705 0.09612131 0.06934710 0.05008464
[190] 0.13176914 0.17817159 0.13225087 0.13489215 0.13076133 0.13113237 0.09592021
[197] 0.09479516 0.09535620 0.09820616 0.06898801 0.06961892 0.07004248 0.07111184
[204] 0.15553260 0.11663333 0.08493391 0.06032187 0.05958683 0.17298042 0.13262549
[211] 0.09648719 0.09763016 0.09451575 0.09677181 0.06836240 0.06919772 0.06983040
[218] 0.26738085 0.20630764 0.20737747 0.20684204 0.15280259 0.15238099 0.15449858
[225] 0.15407315 0.15196033 0.11482451 0.11122921 0.20956268 0.15839758 0.15753022
[232] 0.15709799 0.11617741 0.11584304 0.11820127 0.08383569 0.08459005 0.08535055
[239] 0.08308746 0.15970589 0.15796342 0.15623642 0.15666672 0.11651262 0.11718555
[246] 0.08408646 0.08509636 0.25791010 0.20524191 0.20206924 0.20418027 0.20259547
[253] 0.15112189 0.15492498 0.15578065 0.15154064 0.11219986 0.11317791 0.17546155
[260] 0.23764363 0.17688118 0.17735639 0.12939234 0.09571948 0.09685433 0.09859698
[267] 0.10034935 0.13478207 0.06976913 0.05086596 0.04962131 0.19686748 0.14945990
[274] 0.10963046 0.10931260 0.07800533 0.17769467 0.13375476 0.13300102 0.13337744
[281] 0.09705717 0.06940802 0.06857035 0.20278489 0.15337729 0.15127232 0.15295439
[288] 0.15253244 0.11199201 0.11006119 0.08168199 0.08143775 0.08046745 0.08022653
[295] 0.20331250 0.15211145 0.15043700 0.11134514 0.11264216 0.10910677 0.11070154
[302] 0.07926941 0.08095128 0.07998627 0.21064490 0.15883272 0.15926882 0.11550951
[309] 0.11517681 0.11684866 0.08611726 0.08333619 0.13551757 0.15283217 0.11254787
[316] 0.11189819 0.11287394 0.08210132 0.08112381 0.08259409 0.08063904 0.08136719
[323] 0.05879681 0.05952272 0.05897751 0.15410292 0.11222262 0.08161123 0.05807921
[330] 0.05772350 0.17929682 0.13464544 0.07032996 0.07118749 0.23844264 0.18379707
[337] 0.13137050 0.09666941 0.09752695 0.15510484 0.08367543 0.08417678 0.08342577
[344] 0.06050695 0.18330845 0.13554433 0.06998104 0.21045463 0.15395019 0.15522854
[351] 0.15565658 0.15651554 0.11308323 0.11539261 0.11506020 0.11275663 0.08103648
[358] 0.08300029 0.13251697 0.09870120 0.09640496 0.09668936 0.07104953 0.19774646
[365] 0.15352600 0.15480147 0.15225914 0.11210590 0.11472863 0.11406798 0.11572585
[372] 0.11243085 0.20991337 0.11439789 0.11341065 0.08275247 0.08374781 0.21446490
[379] 0.11786185 0.11752328 0.08283940 0.21282162

```

Figure 6.1 Result of testing in R

## Chapter 7: Conclusions and Recommendations

It is proven that pitch quality and player position has a role to play in a player sustaining an injury. This model generated meets the expected requirement of being able to predict an injury. However, the dataset and variables used till date were not too predictive because of its limited size. For future development, the formula will be fine-tuned to provide a more accurate model. A much larger dataset would also be needed to get a more perfect model. Getting a user interface for coaches and club administrators to work with during game play will also be considered. This will influence their substitution decisions.



Can I play 90?

Name: \*  
Amir Gadr

Position: \*  
Goalkeeper

Date of birth: \*  
1994-01-13

Team: \*  
Algeria

Pitch quality: \*  
All sand

Opponent: \*  
Cameroon

CALCULATE

Figure 7.1 Proposed interface for club administrators to work with

The study has shown that pitch surface quality, player function, player age and match intensity are important variables that can be used to predict injury, especially with pitch surface quality and player function being particularly significant. There is however



room for more variables, such as individual player characteristics, to be included, which will be tackled in future implementations of the project.

## References

- Bahr, R., & Holme, I. (2003). Risk factors for sports injuries — a methodological approach. *British Journal of Sports Medicine*, 37, 384-392.
- Boahene, F. (2017). Pitch expert blasts Gabon surfaces. Retrieved from myjoyonline.com: <http://www.myjoyonline.com/sports/2017/January-20th/pitch-expert-blasts-gabon-surfaces.php>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *crisp-dm.pdf*. Retrieved from <https://www.the-modeling-agency.com/crisp-dm.pdf>
- FIFA. (2017). The FIFA/Coca Cola World Ranking - Men's Ranking Procedure. Retrieved from fifa.com: <http://www.fifa.com/fifa-world-ranking/procedure/men.html>
- Forbes.com. (2015). The crippling cost of Sports Injuries. Retrieved from Forbes.com: <http://www.forbes.com/sites/sap/2015/08/11/the-crippling-cost-of-sports-injuries/#3ea9f9241d0d>
- Gabett, T. (n.d.). Injury prevention and performance enhancement in team sports: Train smarter and harder. *Aspetar sports medicine journal*, 218-223.
- Goal.com. (2012). Charles Taylor set to quit football as injuries hamper his career. Retrieved from Goal.com: <http://www.goal.com/en-gh/news/4389/ghanaian-football/2012/08/27/3333650/charles-taylor-set-to-quit-football-as-injuries-hamper-his>
- Goal.com. (2016). Slaven Bilic issues Andre Ayew injury update. Retrieved from Goal.com: <http://www.goal.com/en-gh/news/4392/ghanaians-abroad/2016/08/26/26906442/slaven-bilic-issues-andre-ayew-injury-update>

- IBM. (2013). IBM predictive analytics reduces player injury and optimises team performance for NSW Waratahs rugby team. Retrieved from [www-03.ibm.com: http://www-03.ibm.com/press/au/en/pressrelease/42613.wss](http://www-03.ibm.com/press/au/en/pressrelease/42613.wss)
- Ibtimes.co.uk. (n.d.). injury risk monitor sap uses sensors cloud computing predict prevent football injuries. Retrieved from ibtimes.co.uk: <http://www.ibtimes.co.uk/injury-risk-monitor-sap-uses-sensors-cloud-computing-predict-prevent-football-injuries-1514453>
- Pulse.com.gh. (2016). Black stars list of Ghanaian players destroyed by injury. Retrieved from Pulse.com.gh:<http://pulse.com.gh/sports/features/black-stars-list-of-ghanaian-players-destroyed-by-injury-id4211632.html>
- Sciencedaily.com. (2016). Study uses GPS technology to predict football injuries. Retrieved from sciencedaily.com: <https://www.sciencedaily.com/releases/2016/08/160802222248.htm>
- Talukder, H., & Vincent, T. (2016). Preventing in game injuries for NBA players. MIT Sloan Sports Analytics Conference. Boston.
- The Journal (2014). Predict injuries make a business big data. Retrieved from [www.thejournal.ie: http://www.thejournal.ie/predict-injuries-make-a-business-big-data-1541702-Jun2014/](http://www.thejournal.ie/predict-injuries-make-a-business-big-data-1541702-Jun2014/)
- Zeigler-hill. (n.d.). psy\_512\_logistic\_regression.pdf. Retrieved from [http://www.zeigler-hill.com/uploads/7/7/3/2/7732402/psy\\_512\\_logistic\\_regression.pdf](http://www.zeigler-hill.com/uploads/7/7/3/2/7732402/psy_512_logistic_regression.pdf)

# Appendix

## Appendix A

SofaScore

EN

Odds

Search...

Favorites

Profile

Football

Tennis

Basketball

Hockey

Volleyball

Handball

Formula 1

Cricket

Rugby

American Football

Baseball

More

Football

Africa

Africa Cup of Nations

Ghana Uganda Live Score, video stream and H2H results

All

Live (27)

Ghana

1 - 0

(1 - 0)

Uganda

FOLLOW

FOLLOW

Started at 16:00

Ended

FT 1 - 0

Additional time 3'

85' Afriyie Acquah Out: Jordan Ayew

72' Emmanuel Badu Out: Asamoah Gyan

Out: Isaac Isinde Geoffrey Sserunkuma 70'

Out: Kizito Muhammad Shaban 57'

Out: Michael Azira Moses Oloya 46'

HT 1 - 0

Additional time 2'

39' Frank Acheampong Out: Baba Rahman

32' 1 - 0 André Ayew

Foul Isaac Isinde 31'

bet365 Live Streaming

00:00 / 90:00

bet365

TO WATCH LIVE STREAM AT BET365:

1. Sign in or Register (it's free) to watch Live Stream\*

2. Open live stream player and select the game. bet365 offers over 70,000 live streaming events per year. Unfortunately, live streaming is not available for this match on bet365.

3. Open live stream player to see streaming schedule. Live Streaming is available to customers with credit on their account.

PLAYERS

Lineups

Player statistics 

New!

Summary

Attack

Defence

Passing

Duels

Goalkeeper

#		Shots (on target)	Tackles	Passes (acc.)	Duels (won)	Minutes played	Position	Rating
	Christian Atsu	3 (2)	2	41 (73%)	9 (7)	90'	M	7.6
	Brimah Razak	0 (0)	0	28 (53%)	0 (0)	90'	G	7.5
	Thomas Partey	2 (0)	1	63 (76%)	13 (10)	90'	M	7.4

43

## Appendix B

<div> <div> <div></div> <div></div> <div></div> </div> <div> <div></div> <div></div> <div></div> </div> <div> <div></div> <div></div> <div></div> </div> <div> <div></div> <div></div> <div></div> </div> <div> <div></div> <div></div> <div></div> </div> <div> <div></div> <div></div> <div></div> </div> </div>					
Home Insert Page Layout Formulas Data					
J24					
	A	B	C	D	E
1			Difference in rankings		Rankings on
2	Egypt	<b>Cameroon</b>	27	70	2
3	<b>Burkina Faso</b>	Ghana	1	1	1
4	<b>Cameroon</b>	Ghana	8	34.5	1
5	Burkina Faso	Egypt	18	17.25	2
6	<b>Egypt</b>	Morocco	22	51.75	2
7	D.R. Congo	<b>Ghana</b>	5		1
8	Senegal	Cameroon	29		2
9	<b>Burkina Faso</b>	Tunisia	17		1
10	<b>Egypt</b>	Ghana	19		2
11	Uganda	Mali	9		1
12	<b>Morocco</b>	Ivory Coast	23		2
13	Togo	<b>D.R. Congo</b>	41		3
14	Senegal	Algeria	6		1
15	Zimbabwe	<b>Tunisia</b>	67		4
16	Cameroon	Gabon	46		3
17	Guinea Bissau	<b>Burkina Faso</b>	15		1
18	<b>Egypt</b>	Uganda	38		3
19	<b>Ghana</b>	Mali	10		1
20	<b>Morocco</b>	Togo	33		2
21	Ivory Coast	D.R. Congo	15		1
22	<b>Senegal</b>	Zimbabwe	70		4
23	Algeria	<b>Tunisia</b>	3		1
24	<b>Cameroon</b>	Guinea Bissau	6		1
25	Gabon	Burkina Faso	55		4
26	Mali	Egypt	29		2
27	<b>Ghana</b>	Uganda	19		2
28	<b>D.R. Congo</b>	Morocco	8		1
29	Ivory Coast	Togo	56		4
30	Tunisia	<b>Senegal</b>	3		1
31	Algeria	Zimbabwe	64		4
32	Burkina Faso	Cameroon	9		1
33	Gabon	Guinea Bissau	40		3

Appendix C

Final Project-q																			Search Sheet	Share
Home Insert Page Layout Formulas Data Review View																			A1	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1				33	34	35	36	39	49	53	54	57	62	64	68	73	90	103	108	
2				Senegal	Cote d'Ivoire	Egypt	Tunisia	Algeria	Congo DR	Burkina Faso	Ghana	Morocco	Cameroon	Mali	Guinea-Bissau	Uganda	Togo	Zimbabwe	Gabon	
3	33	Senegal		0	1	2	3	6	16	20	21	24	29	31	35	40	57	70	75	
4	34	Cote d'Ivoire			0	1	2	5	15	19	20	23	28	30	34	39	56	69	74	
5	35	Egypt				0	1	4	14	18	19	22	27	29	33	38	55	68	73	
6	36	Tunisia					0	3	13	17	18	21	26	28	32	37	54	67	72	
7	39	Algeria						0	10	14	15	18	23	25	29	34	51	64	69	
8	49	Congo DR							0	4	5	8	13	15	19	24	41	54	59	
9	53	Burkina Faso								0	1	4	9	11	15	20	37	50	55	
10	54	Ghana									0	3	8	10	14	19	36	49	54	
11	57	Morocco										0	5	7	11	16	33	46	51	
12	62	Cameroon											0	2	6	11	28	41	46	
13	64	Mali												0	4	9	26	39	44	
14	68	Guinea-Bissau													0	5	22	35	40	
15	73	Uganda														0	17	30	35	
16	90	Togo															0	13	18	
17	103	Zimbabwe																0	5	
18	108	Gabon																	0	
19																				
20																				
21																				

Appendix D: Excel view of data

Final Project-q (Compatibility Mode) - Excel																			Ayely Commadore-Mensah									
File Home Insert Page Layout Formulas Data Review View Tell me what you want to do																			Share									
Clipboard Font Paragraph Styles Cells																			Insert Delete Format Clear Sort & Find & Filter Select									
Q10																												
A																			B C D E F G H I J K L M N O P Q R									
Position Age MD1 PQ Injury MD2 PQ Injury MD3 PQ Injury MD4 PQ Injury MD5 PQ Injury																												
26	Guinea Bissau	Goalkeeper	27	90	1	2	90	1	1	90	1	0																
27	Jonas Asvedo Mendes	Goalkeeper	31	0	1	0	0	1	0	0	1	0																
28	Papa Masse Fall M'baye	Goalkeeper	22	0	1	0	0	1	0	0	1	0																
29	Rui Suleimane Camara Dabo	Goalkeeper	27	90	1	0	0	1	0	0	1	0																
30	Agostinho Soares	Defender	27	90	1	0	0	1	0	0	1	0																
31	Emmanuel Henri Gomis Mendy	Defender	26	0	1	0	0	1	0	0	1	0																
32	Juary Marinho Soares	Defender	24	90	1	0	90	1	0	90	1	1																
33	Mendes Umpeca Eridson	Defender	26	0	1	0	0	1	0	0	1	0																
34	Muhammad Youssef Cande	Defender	26	0	1	0	90	1	1	90	1	0																
35	Rudnilson Brito E Silva	Defender	32	90	1	0	90	1	0	90	1	0																
36	Tomas Soares Dabo	Defender	23	90	1	0	90	1	0	90	1	0																
37	Aldair Adulal Djalo Balde	Midfielder	25	16	1	0	0	1	0	15	1	0																
38	Bocundji Ca	Midfielder	30	0	1	0	0	1	0	0	1	0																
39	Brito E Silva Toni	Midfielder	23	90	1	0	66	1	0	90	1	0																
40	Francisco Santos Da Silva Junior	Midfielder	25	74	1	1	90	1	0	75	1	0																
41	Idrissa Camara	Midfielder	24	0	1	0	39	1	0	0	1	0																
42	Josse Luis Mendes Lopes	Midfielder	24	90	1	1	90	1	0	90	1	0																
43	Lassana Camara	Midfielder	25	0	1	0	0	1	0	27	1	0																
44	Nanislao Justino Mendes Soares	Midfielder	25	90	1	0	90	1	1	63	1	0																
45	Piqueti Djassi Brito E Silva	Midfielder	23	29	1	0	51	1	0	65	1	0																
46	Abel Issa Camara	Forward	27	66	1	0	9	1	0	25	1	0																
47	Frederic Mendy	Forward	28	24	1	0	81	1	0	90	1	0																
48	Joao Mario Nunes Fernandes	Forward	23	61	1	0	0	1	0	0	1	0																
49	Leocisio Julio Sami	Forward	28	0	1	0	24	1	0	0	1	0																
50	Cameroon																											
51	Bokwe Motase George	Goalkeeper	27	0	1	0	0	1	0	0	1	0		120	2	0	0	2										
52	Goda Stephanex Jules	Goalkeeper	27	0	1	0	0	1	0	0	1	0		0	2	0	0	2										
53	Joseph Fabricie Ondoa Ebogo	Goalkeeper	21	90	1	0	90	1	0	90	1	1		0	2	1	90	2										
54	Adolphe Teikeu Kamgang	Defender	26	90	1	0	90	1	0	90	1	0		120	2	0	90	2										
55	Benjamin Bile Moukandjo	Defender	28	90	1	1	90	1	0	90	1	0		120	2	0	90	2										
56	Djetei Mohamed	Defender	22	0	1	0	0	1	0	0	1	0		0	2	0	0	2										
57	Ernest Olivier Bienvenu Mabouka Massoussi	Defender	28	90	1	0	0	1	0	0	1	0		0	2	0	0	2										
58	Joseph Jonathan Nguem II	Defender	25	0	1	0	0	1	0	0	1	0		0	2	0	0	2										
59	Ngadeou Ngadjui Michael	Defender	26	90	1	0	90	1	0	90	1	0		120	2	0	90	2										
60	Nicolas Ndoubena Julio Nkoulou	Defender	26	0	1	0	27	1	0	90	1	0		0	2	0	0	2										
Ready																												

## Appendix E

```
#setting working directory

setwd("/Users/ayeleycommodore-mensah/Desktop/Spring
Senior/COMMODORE-MENSAH, Ayeley - 61402017/data")

getwd()

list.files()


#read in dataset to be used.

data=read.csv("data_log_reg.csv", header=TRUE)

head(data)

plot(data)


#draw boxplots to show relationship between pitch quality and
player function and injury

par(mfrow=c(1,4))

barplot(table(subset(data,data$Player.position==1)$Injury,
subset(data,data$Player.position==1)$PQ),
col=c("red","lightgreen"),

legend.text=c("No injury","Injury"),main="Goalkeeper",ylab="Number
of players",

xlab="Pitch quality")

barplot(table(subset(data,data$Player.position==2)$Injury,
```

```

subset(data,data$Player.position==2)$PQ),
col=c("red","lightgreen"),

legend.text=c("No    injury","Injury"),main="Defender",ylab="Number
of players",

xlab="Pitch quality")

barplot(table(subset(data,data$Player.position==3)$Injury,

subset(data,data$Player.position==3)$PQ),

col=c("red","lightgreen"),

legend.text=c("No injury","Injury"),main="Midfielder",ylab="Number
of players",

xlab="Pitch quality")

barplot(table(subset(data,data$Player.position==4)$Injury,

subset(data,data$Player.position==4)$PQ),

col=c("red","lightgreen"),

legend.text=c("No injury","Injury"),main="Forward",ylab="Number of
players",

xlab="Pitch quality")

cor(data[,c("PQ","Match.intensity","Age","Player.position")])

#logistic regression

fit <- glm(formula=Injury ~ PQ + Match.intensity + Age +
Player.position,

data=data, family=binomial(logit))

```



```

fit

summary(fit)


#prediction dataset

newdata=read.csv("data_predict.csv", header=TRUE)

newdata2=predict(fit, newdata, type="response")

write.csv(newdata2,file = "prediction_output.csv",row.names=FALSE,
na="")

tail(newdata2)

summary(newdata2)


#draw boxplots to show relationship between player function and
injury

par(mfrow=c(1,4))

barplot(table(subset(data,data$Player.position==1)$Injury),
col="pink", main="Goalkeeper",ylab="Number of players",
xlab="Injury")

barplot(table(subset(data,data$Player.position==2)$Injury),
col="lightblue",main="Defender",ylab="Number of players",
xlab="Injury")

barplot(table(subset(data,data$Player.position==3)$Injury),
col="lightyellow",main="Midfielder",ylab="Number of players",
xlab="Injury")

```

```

barplot(table(subset(data,data$Player.position==4)$Injury),
col="lightgreen",main="Forward",ylab="Number of players",
xlab="Injury")

#draw boxplots to show relationship between pitch quality and
injury

par(mfrow=c(1,4))

barplot(table(subset(data,data$PQ==1)$Injury),          col="pink",
main="Pitch quality 1",ylab="Number of players",
xlab="Injury")

barplot(table(subset(data,data$PQ==2)$Injury),
col="lightblue",main="Pitch quality 2",ylab="Number of players",
xlab="Injury")

barplot(table(subset(data,data$PQ==3)$Injury),
col="lightyellow",main="Pitch quality 3",ylab="Number of players",
xlab="Injury")

barplot(table(subset(data,data$PQ==4)$Injury),
col="lightgreen",main="Pitch quality 4",ylab="Number of players",
xlab="Injury")

#linear regression

data2=read.csv("data_lin_reg.csv", header=TRUE)

```

```
fit <- lm(Injury ~ PQ + Age + Match.intensity +Player.position,  
data=data2)  
  
fit  
  
summary(fit)
```