Master thesis plan

**Department of Data Science and Knowledge Engineering  
Board of Examiners**

# Personal Details

Name: *(Student’s family name and first names)*    
Osewa Redencio Jozefzoon

Student number: *(Student’s ID number as provided by university)*   
6075910

Program: *(Choose the master program that applies to this thesis)* Master Artificial Intelligence  
 Master Data Science for Decision Making

Start date of thesis research  
28 February 2018

Expected graduation date: *(Expected date when the thesis will be finished)*  
18 July 2018

# Supervision details

*There should be two thesis examiners that approve of this thesis plan. At least one of the examiners acts as a thesis supervisor. The examiners must be DKE staff members with a PhD degree.*

Name proposed thesis examiner 1: Signature:

Jerry Spanakis (G.) ………………………………………………

Name proposed thesis examiner 2: Signature:

Evgueni Smirnov ………………………………………………

*The thesis supervisors are DKE staff members that give process based and content-based supervision. At least one of the thesis supervisors acts as an examiner. External supervisors are listed in the next section.*

Name DKE thesis supervisor 1: examiner 1  
Name DKE thesis supervisor 2: optional  
Name DKE thesis supervisor 3: optional

Frequency of contact between DKE thesis supervisor(s) and student:  
Weekly

# External collaboration

This thesis involves an external collaboration *(check if an external party if actively involved)*

In case this thesis is not part of an external collaboration proceed to the section “assessment”.

Name external institute: *(Name the institute of external collaboration. In the case of multiple institutes, name the most important one and add an appendix with an overview of external parties)*

This is:

Location external institute:

Name external supervisor 1:        
Name external supervisor 2:        
Name external supervisor 3:

Frequency of contact between the external supervisor(s)/advisor(s) and student:

Frequency of contact between the external supervisor(s)/advisor(s) and DKE thesis advisor(s):

The student will be hosted by/have a secondment at the external institute *(Check this if the student will work as an internal at the external institute or be visiting the external institute for a period larger than two weeks)*

In case the student is not hosted by the external institute proceed to the section “assessment”.

The student will be hosted by the external institute for:  
 The entire period  
 A period of       weeks

During this period, the work times of the student are as follows:

And the student will spend       hours per week on thesis related work.

There is a confidentiality agreement with the following clauses: *(check when applicable)*  
  The thesis shall remain under embargo for a period of       months  
  The code developed will remain property of the external institute  
  Other *(add the confidentiality agreement as an attachment)*Note that all thesis must be send to the DKE education office for storing, where they can be labelled “confidential”, but DKE committees and assessment committees (visitatie commissies) will always have (confidential) access for auditing purposes.

The student will receive from the external institute:  
 No financial compensation  
 An internship fee amounting to       per month gross  
 Salary amounting to       per month gross  
 Travel costs reimbursement  
 Other financial compensation of in       total

On behalf of the external institute,       consents with this thesis plan

Signature: ………………………………….

# Assessment

The assessment will be based on:   
*(Specify the terms of assessment that at least should contain the thesis and the presentation)*  
 The contents and form of the thesis, which must at least be sufficient for passing the thesis  
 A (software) product that accompanies the thesis  
 The presentation  
 (optional)  
 (optional)

# Thesis details

Thesis title: *(Working title of the thesis)*  
Modelling Career Path For Professional Basketball Players.

Description of the assignment: *(A single page, describing the assignment and indications of the techniques that will be explored)*

**Moneyball is a book (later turned into a film) that describes a team’s analytical, evidence-based, sabermetric approach to assembling a competitive baseball team. Nevertheless, not many such scientific approaches have been applied to basketball (or other sports), despite all statistics being available (for example, see accumulated statistics for a basketball player in Euroleague and the National Basketball Association (NBA)). There are many data available like number of points, free throw percentage, turnovers, etc, organized per season.**

**This thesis requires the following steps:**

**(1) Given a sport, (basketball) collect & assemble the data**

**(2) Label accordingly the players and detect career paths.**

**(3) Apply state-of-the-art (or new!) machine learning algorithms (SVM, decision trees, etc) to predict career paths.**

**Predicting the career path of basketball players**

**This thesis proposes the application of machine learning (ML) algorithms (traditional like decision trees and support vector machines and modern like neural network architectures, namely LSTM) in order to solve the problem of accurately predicting the career path of basketball players.**

**In comparison to basketball it is notable that much research has been conducted to predict the career path of Major League Baseball (MLB) players. Therefore, those studies could be used as an outline for basketball. Baseball has made much progress with the application of advanced statistics to the available data. These studies focus on the usage of labeled data, with labels such as salaries and physical attributes of a player. Linear regression models were applied on this data in order to predict the career path, i.e success, of a player. This led to models that were able to predict the trajectory of a MLB player to a reasonable extent. These types of analysis are novel to the domain of basketball. Recently, an increased availability of basketball data has caused a growth in interest for prediction models. Consequentially, the National Basketball Association (NBA) decided to publish much of their available data.**

**Recent research has attempted to achieve similar results for the career path of basketball players (NBA) by focusing on the usage of labeled data, such as number of seasons played. However, poor results have been achieved with applying nearest neighbor classifications to this data[1]. Other research has attempted to use salaries as a label, which led to numerous difficulties. Not only did earlier research show that the salaries [2] are not a good indicator of success [3], the salaries appeared to be less uniform and available than expected by the researchers. Due to the absence of salaries for many players in the dataset, used by Shah et. al [3], the models achieved low accuracy in predicting the trajectories. In their future work Shah et. al proposed to solve the scarcity of the salary data before applying the models. Research conducted by Moxley et al. (2014) [4] defines career success as a consequence of past performance.**

**This thesis defines a career path in terms of the team a player belongs to, the salary this player earns, the development of the physical attributes and the retirement age, as seen throughout the years. Not only will this thesis focus on the use of salaries as an indicator of success, but other features are considered as well. These features entail, for example, residence, draft year [5][6][7], seasons played[3] and, as was the case with the MLB, physical attributes that might influence performance and the career path[5]. Neural networks have been used on basketball data prior in order to predict the outcomes of NBA games [8]. However, most research on basketball is conducted using statistical methodologies and less is done with ML algorithms [9]. This research aims to demonstrate the strength of applying neural network architectures to basketball data by predicting and classifying the career path of any given NBA or Euroleague player.**

**In order to evaluate the outcome, the results of the sequence models are compared to previous work where only statistical methods were used. This comparison is done on precision, recall and accuracy of the predictions. The data consists of labeled basketball data, scraped and collected from several sources such as Euroleague.com, Nbaminer.com and NBA.com from 1985 onwards. The availability of the data of several players throughout the year allows for sequential patterns which can be exploited to improve the accuracy of the classifiers. By separating the data into different timeframes the train set, test set and validation set have as little overlap as possible. Given the labeled data, the aim of this thesis is to classify players according to the team(s) they play for, their salary, retirement age and physical attributes on an annual basis. In doing so an analysis is made to determine which factors are the cause of transfers.**

**Finally, to assess the accuracy of the models, a comparison of the chosen algorithms is made to aid in the process of career path prediction for basketball players. Euroleague players and NBA players usually have minimal overlap in their career paths. Therefore, the models can be compared to evaluate the difference between predicted career paths of NBA and Euroleague players and the transferability of the models within these different domains.**

Planning: *(Provide a complete planning for the thesis period with a reasonable time path for completion, including time for writing the thesis. The preferred format is a Gantt chart that you can attach separately)***See Redencio - i6075910 - Master thesis - Gantt project planner.pdf**

****

Problem statement: *(single sentence describing the overall challenge faced by the student)*  
How well could the career path be predicted for an individual professional basketball player (ipbp)?

Research questions: *(provide two to five research questions of a single sentence that by answering them, jointly provide a path for addressing the problem statement)*

1. Provided the available data (from NBA and Euroleague), what is the accuracy of the prediction for a career path of an ipbp?
2. Using the results of the prediction of past career paths, is it possible to predict future career paths?
3. Which features from a given set are the most suitable to predict a career path?
4. Which algorithm fit the predictions the best, given the selected feature set?
5. (optional)?

# Signature

The student would like to make the following additional statements to the board of examiners regarding this thesis plan:  
(optional)

Ensure that you: *(Ensure that you check all)*

are familiar with the Education and Examination Regulations (EER) that are published on student portal, as is expected of each student

are familiar with the Rules and Regulations that are published on student portal, as is expected of each student

are familiar with the concept of plagiarism and will refrain from committing it

are aware of academic standards, in particular relating to proper citing

have filled out the complete form

have included a planning and/or attached a Gantt chart with a planning for your thesis

have collected the signatures of both examiners and if applicable, the external institute

have obtained 40 ECTS in year 1 of your master programme already and have attached a transcript of your grades

Date: 24 January 2018

Student’s signature ………………………………………………..