*Github repository: https://github.com/Edvard132/TimoVillemEdvard*
Team members: Timo Kaasik, Edvard Notberg, Villem Paabo

# Team D20 - An analysis and predictions of NBA game outcomes

## Business understanding

### Business goals

#### - Background

Our group doesn't have any prior experience with this kind of databases outside of this course.

#### - Business goal

Our primary goal is to develop a model which could predict the odds of an NBA team winning a match, and to find out whether a team has a considerable edge when playing on its home field. Also, we plan to provide a prediction of different game aspects, such as expected most-valuable-players, rebounds, assists, blocks, three-pointers, free-throws and steals.

#### - Business success criteria

All the predicted results will be compared to actual scores afterwards. The project will be judged a success if the outcome of an event can be predicted correctly at least 50% of the time.

# Current situation

## - Inventory of resources

Our resources are sufficient, the data of NBA matches are publicly downloadable. As a software we will be using Excel tables to store the data and Jupyter notebook to process it.

## - Requirements, assumptions and constraints

The project is scheduled to be completed before the poster session. Our poster must be ready and will be presented on Thursday, December 15, 2022 at 14:00-17:00.

## - Risks and contingencies

If including statistics from several years to an algorithm that predicts a winner just of a particular match-up, predictions will get less relevant as different factors such as team members, coaches etc. probably have changed over seasons.

## - Terminology

Data-mining (DM) - discovering patterns from data
Machine-learning (ML) - algorithms that learn from data
Data-science (DS) - science about how to operate with data

## - Costs and benefits

As we are doing this project for school, we do not have any budget and financial costs. Our only cost will be our time. Accordingly, we will not receive any money for our work, and our only benefit will be the experience and the grade at the end of the course.

# Data-mining goals

## - Goals for the project

1. Try building a model for predicting NBA game outcome details.

2. Visualize a chance of a team winning and game in-depth statistics indicators (such as three-pointers) at the home-court
3. Finally, create a report concluding all our work.

### - Data-mining success criteria

Data should be large enough to predict an accurate outcome of the match-up for each team.
Dataset should have no or as few as possible missing values.

## Data understanding

### - Gathering data

The data can be easily found and downloaded as a preformatted csv file from many websites. They provide different measurements from which it is up to us to decide which of them to use. Some of them go too in depth whilst others remain too shallow. We are most interested in some basic, but key measurements, like team pace and team effective field goal percentage, but we also need some data to support these, as these in themselves are not sufficient enough to make any credible predictions.

### - Describing data

| **Attribute** | **Descritpion** |
|---|---|
| game_ID | The ID for the game, is a number based on the date and home team |
| game_date | The date when the game was played |
| team_score | The score of the home team |
| player | The name of the player |
| playerid | The ID of the player |
| ast | The number of assist made by the player |
| stl | The number of steals made by the player |
| blk | The number of blocks made the player |
| pts | The number of points scored by the player |

| | |
|---|---|
| fg_pct | Basket scored percentage on any shot or tap other than a free throw |
| fg3_pct | 3-point field goal percentage |
| ft_pct | Free throws percentage |
| plus_minus | A player's impact on the game, represented by the difference between their team's total scoring versus their opponent's when the player is in the game. |
| mp | The number of minutes played by the player |

For our data, we decided to go with the data from https://www.advancedsportsanalytics.com/nba-raw-data as it has a lot of data, starting from the team stats that we were looking and also including all of the players stats that were playing in any given game in our time range. It has 89500 rows and 55 columns.

As the player data is very thorough, we will probably opt to cut some of it, as some of it too specific and might offer miniscule benefit to our goals.

## - Exploring data

As mentioned, there is a lot of data and some of it is very thorough. It will be interesting to see which useful patterns can be found from all of this. There does not seem to be any real quality problems with the data that we are interested in other than that the dates are not sorted and some fields do have NA values which will end up being cut.
Some of the most important fields are still the team pace and team effective field goal percentage, but in addition to those, a lot of help comes from team turnover percentage and team offensive rating. The problem for the team offensive rating is, that as we have a full season's stats, that rating is based on the entire season, but for us, to try and predict something, it would have to not include the games that have not been played, that is something for us to figure out. Then there is the importance of player stats, whilst they are the cornerstones of the project, they still impose the underlying problem in the topic that we are dealing with, and that is the fact that humans are not machines and these are to change in unpredictable ways due to external factors. All in all, everything is like we pictured it to be and we are looking to dive into the work ahead.

- Verifying data quality

As the data has been carefully selected to fit our needs, there aren't currently any poor data points we can think/see of. Of course this could change in the future when exploring the data more thoroughly. Everything right now is in correct format and readable by both humans and computers.

# Project Plan

1. Collecting the data
    - Estimated time: 5 hours
    - Methods, tools: web-search
    - Details: objective is to gather only the data that we need, as there are lots of different NBA datasets available on the internet.

2. Data preparation
    - Estimated time: 5 hours
    - Methods, tools: Pandas dataframe on Jupyter notebook
    - Details: preparing data before operating with it. It is possible that we will drop values that we won’t need .

3. Discovering patterns from data
    - Estimated time: 15 hours
    - Methods, tools: pandas, seaborn
    - Details: discover the most frequent and relevant patterns, based on statistics and group them

4. Visualization of data
    - Estimated: hours 8 hours
    - Methods, tools: matplotlib, pandas, seaborn
    - Details: Visualize relations of the attributes

5. Clustering data
    - Estimated: hours 8 hours
    - Methods, tools:
    - Details: Applying clustering algorithms

6. Apply machine learning
    - Estimated: hours 20 hours
    - Methods, tools: Jupyter notebook sklearn.
    - Details: Trying to build predictive models for classification, regression.

7. Poster
    - Estimated: hours 8 hours
    - Methods, tools: Microsoft Word
    - Details: Create a poster to present at the poster session