# Portfolio Project Proposal

Mehmet Ege OGUZMAN
StudentId: 20257961
Data Mining & Machine Learning 1, MSc in Data Analytics
National College of Ireland Dublin, IRELAND
Email: x20257961@student.ncirl.ie. URL: www.ncirl.ie

## I. Motivation

The field of sports analytic has become quite common with the popularization of the concept of data science. When we look at investments and circulating capital in the field of sports,teams have also become businesses. A lot of commercial,player or team-based data is produced about the teams and this data plays an important role in determining the strategies of the teams. Analyzing and processing this data has become a major contributor to teams winning matches, becoming champions in their own leagues, and even determining their financial strategies.

In this report, college men's basketball teams will be analyzed for the NCAA (National Collegiate Athletic Association) league, which is very popular in the United States. By analyzing the match statistics and general conditions of the teams in the season and tournaments, and using the right machine learning algorithms, the champion of the March Madness tournament, the biggest tournament of the league, will be tried to be predicted.

The motivations behind writing this report are to test and compare different machine learning models, to gain experience in understanding and processing multiple complex data sets, and in sports analytic and since sports analytic, a sub-field of data science, is the field I want to work in, I chose this project to shape my career and expand my portfolio.

## II. Research Question

The purpose of the report is to analyze different types of data sets such as season and tournament matches played in the past, team statistics and to make a probability prediction about which teams will win in the tournament in the future. These winning probability predictions will allow us to predict the final bracket of the tournament like in Figure 1. In addition, since winning and losing between teams is a binary classification problem, different types of classification algorithms will be trained with historical data and the results will be examined.

## III. Datasets

Data sets have been published for the March Machine Learning Mania 2021 - NCAAM competition published on Kaggle [1]. This Data set consists of many csv files. From these structured data sets, the data of the matches of the teams in the past season and the matches of the past March Madness tournaments will be used.

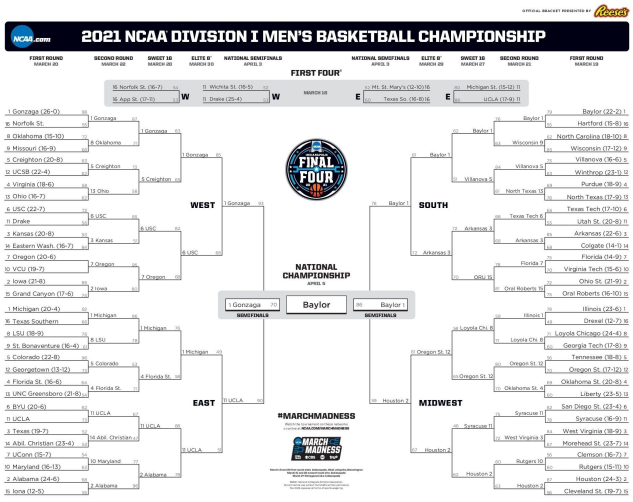


Fig. 1. 2021 NCAA Men's basketball March Madness initial bracket

TABLE I
Dataset descriptions

| Dataset Name | Number of Sample | Number of Column |
|---|---|---|
| Historical Season | 92832 | 34 |
| Historical Tournament | 1115 | 34 |
| Historical Tournament Seed | 2284 | 3 |
| ELO Ranks 2021 | 356 | 4 |

It will also use data where team IDs are kept. Because team names will be sent to machine learning algorithms in integer format, not in string format.

In addition to this data, there will be ranking data determined for the teams each season. Kaggle gives us the first ranking parameter, "seed" values for each team and each season. These values are given for the 4 sides of America (East, West, South, Midwest).

Another team ranking parameter is ELO. The ELO system will be used ready-made and the existing ranking values will be used [2]. A new ELO calculation will not be made for this project.

## IV. Literature Review

When the literature is searched for this project, we see that Logistic Regression algorithm is generally used as classification algorithms. As the input of the model, ratios such as

Turnover, Free throw percentages, shooting percentages are calculated and predictions are made according to the average performance of the teams [3] [4]. In addition, it is mentioned in the literature that the Random Forest and Decision Tree machine learning algorithms also gives a successful result [5].

In this project, SVM, KNN and XGBoost algorithms will also be included to be compared with these two algorithms.

## REFERENCES


[1] March Machine Learning Mania 2021 - NCAAM Dataset [Online]. Available: https://www.kaggle.com/c/ncaam-march-mania-2021/

[2] NCAA Men's Basketball 2021 ELO and Ranks [Online]. Available: https://www.warrennolan.com/basketball/2021/elochess

[3] Bryce Brown. Predictive Analytics for College Basketball: Using Logistic Regression for Determining the Outcome of a Game. Peter. T Paul College of Business and Economics University of New Hampshire. May 2019.

[4] Zifan Shi, Sruthi Moorthy, Albrecht Zimmermann. Predicting NCAAB match outcomes using MLtechniques – some results and lessons learned. KU Leuven, Belgium. 14 Oct 2013.

[5] Cody Kocher  Tim Hoblin. Predictive Modelfor the NCAA Men's Basketball Tournament. Ball State University Muncie, IN. April 2017.