# Proposal: March Madness Predictor

**DATA 450 Capstone**

Joseph Bellani

2/5/23

## 1 Introduction

For my project I want to explore the world of predicting sports, specifically the NCAA Men's basketball Tournament. To do so my goal is to create an algorithm that can predict the chances that a team has to win against any given team. Using all sorts of metrics and data about tournaments past. I want to see if I can build a predicted model that will perform better than average human being, and would like to put it to the test in a month and a half, given all goes well.

The NCAA Tournament is a single elimination basketball tournament that includes 68 Division 1 college teams. One of the most popular events of the year for sports fans, many of whom fill out their predictions of what will happen. The single elimination format creates a lot of variance within the tournament, most notably the fact that there has yet to be a bracket that has correctly predicted each game, with a vast majority failing to remain perfect through the first day of the tournament. That is what inspired me to conduct this project, to see if a statistic based model can help us move towards creating better brackets.

## 2 Dataset

TeamRankings: TeamRankings - Accessed 2/11/2023

kenpom: kenpom ratings - Accessed 2/5/2023

From: TeamRanking * Team: The name of the team * wins: The team's wins in the season * losses: The team's losses in the season * ppg: The team's points scored per game * papg: The team's points allowed per game From KenPom: * AdjEM: The team's Adjusted Efficiency Margin * AdjO: The team's Adjusted Offensive Efficiency * AdjD: The team's Adjusted Defensive Efficiency * AdjT: The team's Adjusted Tempo * Conf: The team's Conference * SOS: The team's Strength of Schedule in the season

(Note: Each variable will have a 'vs' variant that will be used when a team is considered the opponent, the data and meaning will not change.) (Note: Every variable will be for the team's regular season peformance for that year.) ]

## 3 Data Acquisition and Processing

The way I will be collecting the data will be through WebScraping from both TeamRankings and KenPom. After scraping is complete I will need to go through the data and see if anything must be cleaned before proceeding. It is likely that some variable names will need to be adjusted before use.

The way TeamRankings formats their tables is by putting one table for each statistic. This allows for easy websraping, but tedious data processing as I will need to combine mutiple dataframes into one. Overall that process is easy as after dropping and renaming columns all it takes is a simple merge command.

The more difficult half is going to be designing the final dataframe. The idea currently is to have 2 rows for each game, one for the perspective of the winner, the other for the loser. Each set of 2 rows will contain the same data points but flipped so the model can attribute the statistics on each side for each game. The last column will be an alternation set of 1s and 0s that represent Wins vs Losses and will be the column that we will try to predict.

This dataframe frame will be fairly large as it should contain 1260 rows with data for 630 unique matchups. The current design is quite plausible in functionality as taking the frame of a game and a team allows the model to analyze both the statistics and the result of the tournament games.

## 4 Research Questions and Methodology

1. Can a model based on data predict the March Madness tournament better than the average person? For this I will create a logistic regression model that will predict singular games which then will be combined to form a whole bracket. I will compare the results of the generated bracket to that of the ESPN "The People's Bracket" which is a culmination of every bracket submitted on ESPN, whichever bracket scores more points higher will be deemed the better bracket.

Once the bracket is released to the public before the start of the tournament you can begin the process of building a bracket. All of the matchups will then be predetermined thus you may input the statistics from the teams playing each other. Once a matchup is put in you can receive the win probabilty from the algorithm and you advance the team that the model gave the higher chance to win. Rinse and repeat until the bracket is completed.

2. Does the model prefer offensive teams or defensive teams? To analyze this I can look at the equation that the model uses and see which side of the game the model holds more weight on. Since each of the stats that compare offensive and defensive efficiency are on equal scales it will be comparable. For instance, the model will be using offensive and defensive efficiency as 2 separate variables, once the model is completed I can look at which of the two efficiencies hold a higher weight.

3. What is the difference in conferences that make the tournament versus conferences that make the Final Four? Every conference gets at least one team into the field of 68 teams, but most do not make it very far. I think it would be interesting to look at which conferences dominate the attendance at the pretigious Final Four.

4. Another exploratory question can come from looking at teams grouped by seeds. One interesting graphic will be to show how often each seed has won over the last 10 years. I would expect the 1 seeds to have the highest win rate and for the win rates to gradually decrease as seed increases, but any anomalies would be an interesting find.

5. How has average points per game and average points per game of tournament teams changed over the last 10 years? To accomplish this I can make a dual bar chart comparing both statistics for each year, averaging it out for every team that made the field.

6. Which games project to be the most one-sided? For this question I can make a list that will show off which games the model gave one team a much higher chance than the other. This question will obviously have to be answered after the model is created as opposed to the previous few asked.

7. What is the distribution of offensive and defensive officiencies amongst teams who made the 2023 Tournament? To accomplish this I can make a scatterplot with one statistic on the x-axis and the other on the y-axis and it will show where each team fell in both statistics. Time permitting, I can also color them to indicate where each team lost in the tournament.

8. Which states send the most teams to the NCAA Tournament? Similar to the conference question I think this question could be very interesting. I could use geocoding to separate the teams into their respective states and use a color gradient to determine how many teams are from each state.

9. What is the relationship between Strength of Schedule and Seed? It is well known that you have to play good teams to earn a high seed (as seen by 31-3 Murray State getting a 7 seed last year). I would like to see how much a team's schedule affects where a team will be placed in the tournament.

10. Do some conferences earn higher seeds on average then others, if so what is the differnce? I want to look at the historic difference between seeding within conferences. I can build a horizontal bar chart or simply make a list to answer this question.

11. Is there a relation between Adjusted Tempo and Points per game? In theory there should be as the faster a team makes the game the more points should be scored. To check this assumption I can scatterplot with the two variables and find the true conclusion.

12. How much does strength of schedule affect wins and losses? I could make a new variable for win percentage so that I could again make a scatterplot to find the answer to this question. I would expect a downward trend such that as Strength of Schedule increases win percentage decreases.

## 5 Work plan

[Fill in the list below with a plan for what you will do each week. You should have around 7 hours worth of work each week. Writing work counts. Several tasks have already been filled in for you.]

**Week 4 (2/6 - 2/12):** [Just an example:

- Webscraping Data (4 hours)
- Rework Questions (2 hours)
- Build Blueprint for Dataframe (2 hours)

**Week 5 (2/13 - 2/19):**

- Data Preprocessing (2 hours)
- Dataframe Construction (8 hours)
- Continue Webscraping (4 hours)

**Week 6 (2/20 - 2/26):**

- Continue Preprocessing (If necessary)
- Create Model for Question 1 (4 hours)
- Create visualizations for other Questions (4 hours)

**Week 7 (2/27 - 3/5):**

- Presentation prep and practice (4 hours)
- Apply Model to a previous Tournament (4 hours)
- Continue making visualizations (2 hours)

**Week 8 (3/6 - 3/12):** *Presentations in class on Thurs 3/9.*

- Presentation peer review (1.5 hours)
- Insert Data and Predict for 2023 Tournament (6 hours) (Done between 12th and 16th)
- Complete any visualizations left (2 hours)

**Week 9 (3/20 - 3/26):**

- Poster prep (4 hours)
- Gather Results and Assess for 2023 Bracket (2 hours)

**Week 10 (3/27 - 4/2):** *Poster Draft 1 due Monday 3/27. Peer feedback due Thursday 3/30.*

- Peer feedback (2.5 hours)
- Poster revisions (2 hours)

**Week 11 (4/3 - 4/9):** *Poster Draft 2 due Monday 4/3. Final Poster due Saturday 4/8.*

- Poster revisions (1 hour).

**Week 12 (4/10 - 4/16):**

**Week 13 (4/17 - 4/23):** [All project work should be done by the end of this week. The remaining time will be used for writing up and presenting your results.]

- Draft blog post (5 hours).

**Week 14 (4/24 - 4/30):** *Blog post draft 1 due Monday 4/24. Peer feedback due Thursday 4/27. Blog post draft 2 due Sunday 4/30.*

- Peer feedback (2.5 hours)
- Blog post revisions (2 hours)
- [Do not schedule any other tasks for this week.]

**Week 15 (5/1 - 5/7):** *Final blog post due Tuesday 5/2.*

- Final presentation prep and practice.
- [Do not schedule any other tasks for this week.]

**Final Exam Week (5/8):** *Final Presentations during final exam slot, Monday May 9th 3:20-6:40pm.* [Do not schedule any other tasks for this week.]

## 6 References

I'm not sure if we are meant to repost the same references from earlier, the description said it would autogenerate a list, but it did not so I put the websites I will be scraping from down here as well.

TeamRankings: TeamRankings - Accessed 2/11/2023

kenpom: kenpom ratings - Accessed 2/5/2023