ISyE 7406 Data Mining and Statistical Learning Final Project

**Group Members:**

* Arjun Goyal
* Richard Hardis
* Fred Sackfield
* Anthony Temeliescu

**Project Description:**

Every year, millions of Americans turn on their televisions and tune into the NCAA men’s basketball tournament, commonly known as March Madness. This tournament pits 68 of the best men’s teams from every conference across Division I against each other in a single elimination style bracket. This tournament is the culmination of a full season’s worth of training and games and the athletic, and financial, stakes could not be higher for every program. Not only does March madness generate a tremendous media buzz, it also is the topic statistical analysis on every level from amateurs betting in March Madness brackets to individuals with Ph.D.’s attempting to predict the outcome of the games. Our group is proposing a novel approach to take on this same task. Using data gathered from various sources on the internet, our project goal is to predict the outcome of each game in the NCAA tournament using prior years as training data with the hope of using the 2020 tournament results as a real-world test of our methodology.

**How and Where the Data Were Obtained:**

The main portion of our data is taken from the Kaggle dataset called College Basketball Dataset by Andrew Sundberg ([https://www.kaggle.com/andrewsundberg/college-basketball-dataset#cbb.csv](about:blank#cbb.csv)). It contains data on 24 different variables measuring relevant team performance metrics such as adjusted tempo, tournament seed, and defensive rebound percentage. In addition to this dataset, the website Kenpom.com provided additional statistical data at the team level. Beyond these two large datasets, we were interested in exploring the commonly held belief that certain coaches give their team a strong advantage in the tournament, regardless of how well the team has performed during the regular season. To account for this potential effect, we have gathered data on each team’s coach’s number of years coaching his team and the number of prior NCAA tournament wins that coach has. Another potential feature that we can analyze is related to the effect of traveling on team performance.

**Scientific Research Questions Addressed:**

Our goal is to predict the number of games won in a tournament by every team in the bracket field. In doing so, we hope to both build an accurate algorithm and to uncover some variables that are strong drivers of tournament success.

**Proposed Statistical Methods and Models:**

Because our goal is to predict the number of games that each team will win, we will use this as our response variable for modeling purposes. The response variable is continuous, so we should choose our models accordingly. Since we don’t have much prior knowledge about which specific technique is most suitable for the problem, our methodology will include testing different types of models, and ultimately choosing the best one for prediction based on training/validation accuracy, as well as interpretability. Most of the techniques we utilize will be some variation of Regression Analysis, such as Kernel Regression, Principal Component Regression, Regression Trees, or Clustering-based methods.