March Madness Predictions

By: Ryan Oscherwitz and Chris Lee

The division I March Madness tournament is one of the yearly highlights in the yearly sports world. With the widespread integration of data analytics in various sports leagues, a prevalent endeavor is utilizing statistical models to forecast the annual winner of March Madness. With this popular goal in mind, we focused our project into two main areas. The first was to attempt to group teams together based on statistics (box score and advanced) using k-means clustering and see if we can use those groups to predict which teams went further in the bracket based on which group they were placed in. The second area was to try to predict the winners of the tournament using Support Vector Machines (SVM).

We began our analysis by finding a dataset (from Kaggle) that includes several statistics for all teams that played at least one game in the tournament from 2008 to 2023. Each observation was a team that season. It is important to note that the data was from both the regular season as well as all of the tournament games. We chose the response variable to be the farthest round the team made it to in the tournament. Some of the predictor variables we selected included tempo, offensive and defensive efficiency, and a statistic that describes the probability of the team beating an average division 1 team.

From there, we began our dive into k-means clustering, which is an unsupervised algorithm that groups similar observations into clusters. While the similarity between two observations can be calculated in several different ways, we choose the distance between any two pair of observations to be given by the Euclidean distance (also known as the straight line distance). For k-means clustering, we must select the number of clusters that we want. This can be done in several ways. Sometimes there is a clear number of groups you would expect to see based on the data/response variable. If there is no clear number of clusters based on the data, oftentimes the number of clusters is set to the square root of half the number of observations in the data. An additional method to determine the number of clusters is what is called the elbow method, which is what we chose to use. From the elbow plot we found, we decided to have four clusters.

Once we ran the k-means clustering code in R, we created a confusion matrix that showed the number of teams in each cluster based on how far that team made it in the tournament. The results were pretty good as it was clear which clusters contained the best and worst teams based on both their season statistics and how far they made it in the bracket.

We then began to model our data using SVMs. Specifically, we created a model using SVM that predicted the round a team would be eliminated from. Outputting a confusion matrix using this model, we obtained an accuracy rate of 51.2%; however, using predictions from the test data, the accuracy rate was 46.4%. Our model was greatly under-predicting the round.

In an attempt to remedy some of the problems with many predicted outputs, we used SVMs to predicate whether or not a team would make it to the round of 16 – a binary output. This approach yielded an accuracy rate of 81.9%, and 83.6% accuracy when using the testing data. This is certainly an improvement; however, the model still overpredicts the number of teams that don’t make it to the round of 16. Had we had more time, it would have been interesting to try and force the model to have a certain number of teams that make the sweet sixteen and those that don’t.

Overall, through k-means clustering, we were able to categorize teams into certain clusters based on various statistics over the season. Additionally, SVM modeling allowed us to somewhat accurately predict if a team would make it to the round of 16 or not.