# 

# 

March Madness Final Report

Ridge Primus, Simon Holmes, Lindsay Koplitz, Raf Machado

## **Executive Summary**

Each year the tournament starts with four play in games. Four of these teams are fighting for two different 11 seed spots, and the four other teams are fighting for two different 16 seed spots. After these four games are played and the winners are named, it leads us into the start of the 64 team tournament. Subsequently, millions of people across the world try to predict the outcome of this tournament with “March Madness’ brackets. In all the years of this tournament's existence, only a small number of brackets have been predicted perfectly. By taking these 64 teams, we are able to choose a team that we think will make it far in the tournament based on specific statistics that we decide to use to determine our winner. The motivation to fill out brackets is not just based on the feeling of competition, but also the exorbitant cash prizes that are offered as incentives. Warren Buffett offers a $1 billion grand prize for anyone that is able to fill out a perfect bracket. The formula for determining the winner can vary from person to person. Some people just look at how the teams are seeded, the average points scored per game, how a team has done throughout the rest of the season, or they just stick with their household favorites. These are all possible qualities that can be used in the magical formula to derive an ultimate winner. But, how do we determine which variables are the best to use? Which formula is the most successful for predicting a winner? Our group is setting out to determine what components are the best and the most accurate at predicting the outcome of the March Madness NCAA tournament.

Our goal is to predict the winners, the losers, and eventually the tournament champion. Hopefully to an accuracy level of 70%.

Our preliminary conclusion was that the University of North Carolina (UNC) and Kansas were going to be formidable contenders in the tournament. We did this simply by looking at their rosters and records. We have also noticed a few issues. It is difficult picking out which data is valuable, and also what variables are useful. There is a vast amount of data archived over decades of the tournament. Other issues included picking which learning tools to use, and how to decipher the data from using these tools.

Table of Contents Page

[Executive Summary](#h.13gj5ymhat21)

[Corrected Contents from report 2:](#h.duq38tyeg8ij)

[Data Understanding](#h.con60bk43kxn)

[Data Preparation](#h.mntn0k55zpfh)

[Descriptive Analytics](#h.w1d84evwilza)

[Modeling and Descriptive Analytics](#h.tdvqdysc4dry)

[Expanded Modeling and Predictive Analytics](#h.x2mduopycfds)

[Evaluation](#h.6hdkuhhao59q)

[Prescriptive Analytics](#h.wg5gy7k5bo5b)

[Comparison With Documented Results](#h.uahyy96qv3pa)

[Deployment](#h.ktrfw3ybs0gw)

## **Corrected Contents from report 2:**

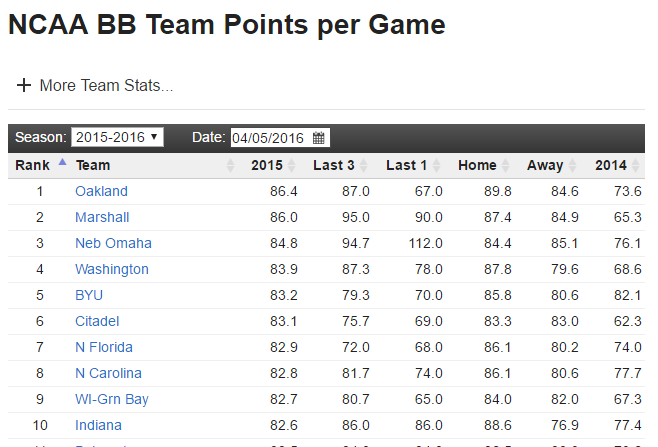
### Data Understanding

When determining what kind of data to use, we took the obvious approach and decided to use quantitative research only. Quantitative research is information that is measured and written down with numbers, so besides team names, sports statistics are always looked at in terms of numerical data and percentages. We also knew that using these numeric variables would allow us to convert our spreadsheet into a CSV to show us different models and predict analytics later down the road.

We pulled almost all of our statistics from espn.com and teamrankings.com but we also looked through data on Kaggle.com. Because the data from Kaggle went all the way back through 1985, we saw it more as a sample dataset for the creation of our own. We wanted to base our first datasheet on the previous two seasons only (2014-15 and 2015-16) so we pulled our statistics from the updated sports websites. Shown below is a sample of the data that we used from:

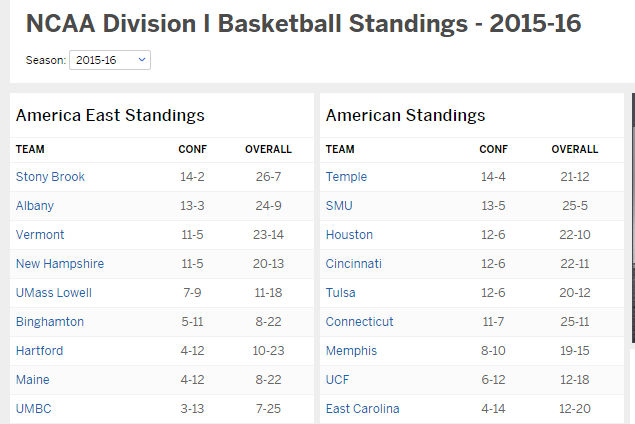
<https://www.teamrankings.com/ncaa-basketball/stat/points-per-game>

As you can see, the team names are listed on the left side under “Team”. The statistics we used from this website were the team’s average points per game for their 2015 and 2014 seasons.



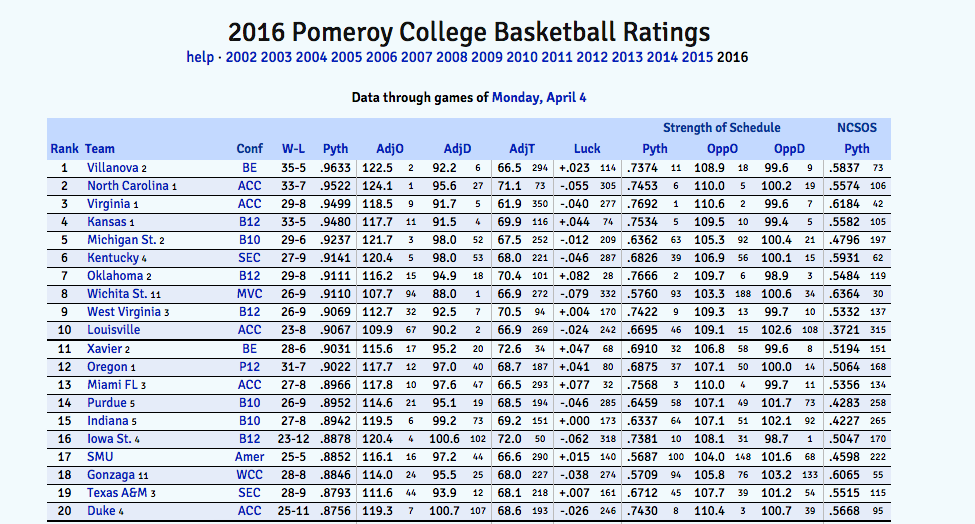
Another website that we collected statistics from was: <http://espn.go.com/mens-college-basketball/standings>

Shown in the picture below, you can see that we were able to use each team's overall wins and losses. The team names are listed to the left under “Team” and each team was found under their specific conference name (ex. America East Standings, American Standings, Big Ten ect.) To find each team’s wins and losses, we looked under the section called “Overall”, with wins as the first number and losses as the number following the dash. Having these two stats in hand, we were also able to calculate another variable. A team's winning percentage is computed by dividing their number of wins by their total number of games. For example, the team Stony Brook shown below has 26 wins and 7 losses. This means that they played a total of 33 games and their winning percentage is 26/33= .788%



All of the other variables and their statistics that are to be explained in depth in our the data preparation section below, were also found on these two websites.

After our groups status report we decided that it would be to our benefit to add some more variables to our data set. We came across a website called kenpom.com that was created by a man named Ken Pomeroy. This website stuck out among all of the rest because of the many kinds of advanced basketball analytics it included. We knew this website was reliable because Ken Pomeroy had included one interesting statistic called the luck-ratio. Created by Dean Oliver, this stat was nothing like we had ever seen before. With a little bit of our own research we found out that Dean Oliver was extremely well-known in the analytics world. As the first full time statistical analyst in the NBA and working with the analytics team at ESPN, Oliver contributed to creating one of our previously used statistics, the Basketball Power Index. With all of this information, we thought this would be a good contribution to include in our dataset. Shown below is a sample of the 5 new variables that we pulled from: <http://kenpom.com/>

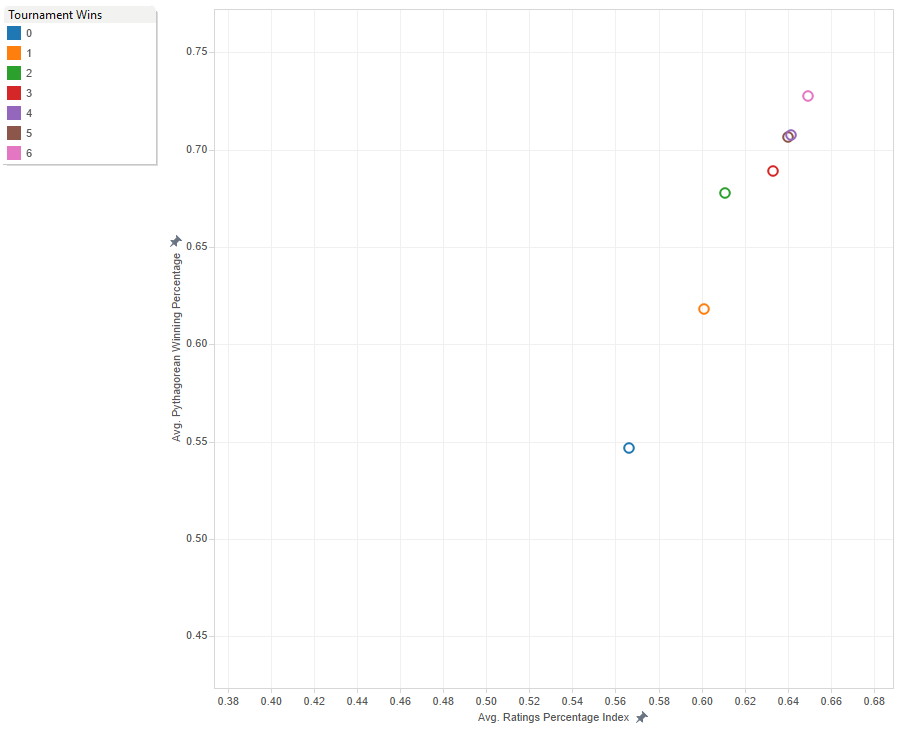


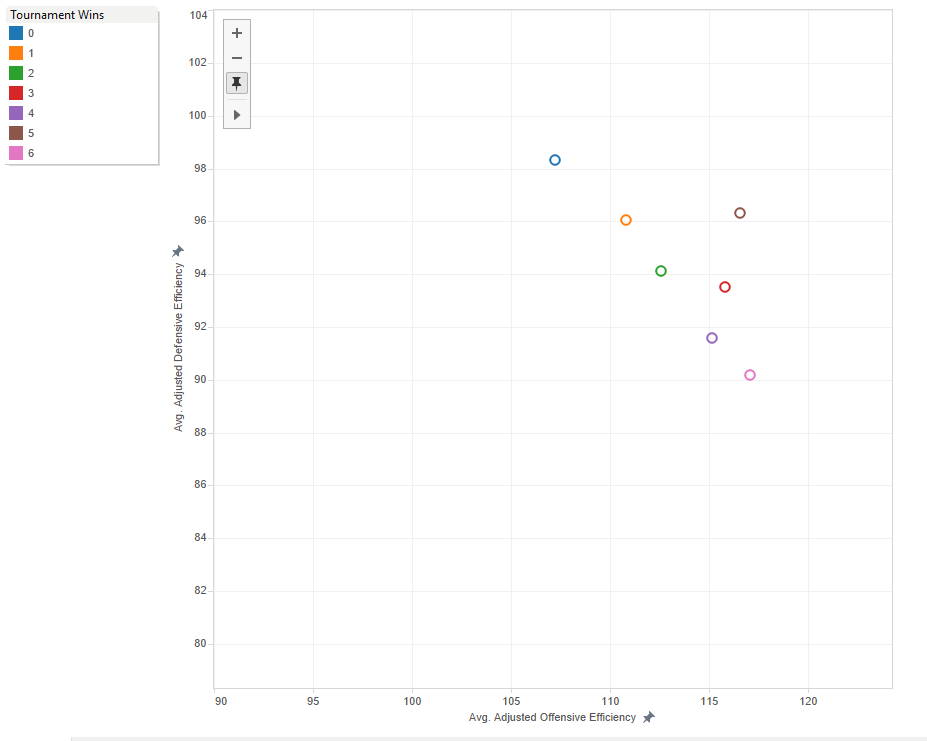
### Data Preparation

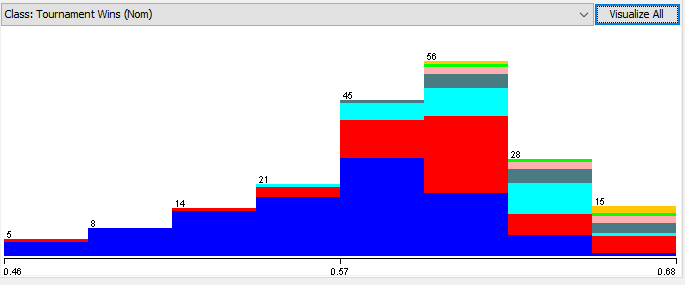
To begin the data preparation process, we started off by downloading many different datasets from a previous March Madness model on Kaggle.com. We studied these sample data sets several times to get a broad idea on how we should create our own, personal datasets. The first step in creating our own datasets was to choose the variables we thought were appropriate for a testing. We started by choosing each variable, one by one, based on how important we thought they would be in determining a winner of the tournament. When deciding exactly which variable to use, we started off using team’s wins, losses, and winning percentages. These were the first variables added to our data set and they were the wins, losses, and winning percentage from each tournament team’s regular season record. We chose these three right away because we felt that they were the most important aspect in deciding how to rank a team. Next we decided to choose RPI (ratings percentage index) which is a statistic comprised of three factors; a team’s winning percentage, their opponents winning percentage, and the winning percentage of all the other opponents that their opponent plays. RPI is the main statistic that the tournament selection committee uses to select where teams should be seeded within the tournament so we thought it would be an important variable to include. Each team's winning percentage accounts for 25 percent of their RPI in which each win and loss is weighted. Home wins and road losses are both weighted at 0.6, and home losses and road wins are weighted at 1.4. Neutral site wins and losses are also both weighted at 1.0. The second factor of RPI is opponent’s winning percentage discounting the main team in question. This accounts for 50 percent of the RPI. Lastly, the third factor of RPI is the winning percentage of all the other opponents that their opponent plays. This counts for the last 25 percent of the RPI. The next variable we decided to use is strength of schedule. A team’s strength of schedule is defined as the combination of the winning percentage of a team’s regular season opponents from their entire schedule. We used this because it tells us whether a team with a large amount wins is winning because they are actually good (playing harder teams) or if they are winning because they have only played the weak/bad teams. This could also have an effect on team’s rankings for the tournament and predict the chances of an upset. The next variables we chose were points for and points against. We used these in our data set simply because the other models we analyzed all included these as variables. The last variable we included in our data set, only from the previous season, was the amount of tournament wins. We thought this was a relevant statistic because it can show if a high seed from the previous year got upset, the tournament wins would show that. Once we chose all of the variables that we wanted to use, we manually inserted each variable into an excel spreadsheet. We split the spreadsheet into multiple datasheets depending on year of each season. We then looked up all of the stats for these variables and found them on the previously listed websites. After finding all of the stats from the 2014-15 and 2015-16 seasons, we inserted them under their correct variables and across from their correct teams in the spreadsheet. This process of inserting all of the statistics into the excel spreadsheet was difficult and extremely time consuming. Once our two separate datasets were completed, we then needed to convert them into applicable CSV files to be run in Weka and Tableau. To do this, we first needed to clean the datasets. With the amount of quantitative statistics we manually entered, some human errors were made with decimal placements and missing data. By deleting all of the unneeded information and entering the variable’s averages for any statistics that we could not find, we were able to correct our datasets to the best of our ability. To actually convert the datasheets into CSV files, we had to copy sets into a text editor application and then continue with more cleaning. This cleaning was a little bit different. We had to delete all of the apostrophes, periods, and percentage signs in the text editor version, like for example, in the team name St. Joe’s the apostrophe needed to either be deleted or switched to /n in order for the conversion to work. Once the final cleaning process was completed, we saved the text editor file as a .CSV and were able to upload them in our machine and visual learning tools for further analysis.

As previously stated, after our status report we knew that we needed some more uncommon variables to help with the accuracy level of our model. Narrowing between Kenpom.com, Sagarin.com, and masseyratings.com, we decided to use the advanced metrics from Kenpom.com as they’re stats was similar to the two other sites but they made their stats easy to see all of the tournament teams as a whole, which made it easy for us to add their stats to our dataset. From Kenpom.com we added their five main advanced metrics to our dataset. The first metric we added was kenpom’s “Pythagorean” rating, which is a fancy way to describe a team’s expected winning percentage. It is calculated from this algorithm (AdjO^11.5)/(AdjO^11.5+AdjD^11.5). The next two variables from kenpom that we chose were adjusted offensive and defensive efficiency. These are the estimate of the points scored and points allowed against the statistical average division I offense and defense respectively. The next metric we added from Kenpom.com is adjusted tempo. Adjusted tempo is an estimate of the amount of possessions a team would have per 40 minutes of basketball. The next variable we added from kenpom.com was their luck ratio. We were mostly intrigued by this metric out of the five new ones because it is calculated by the deviation between kenpom’s pythagorean expected winning percentage and a team’s actual winning percentage and it shows how likely a team is to win all or none of the close games they play, deciding whether or not a team is lucky. These metrics as a set were chosen not necessarily because they are they are advanced, but because they were metrics that we noticed from our research, that weren’t used in other models.

### Descriptive Analytics

Once our dataset was prepared we imported the data into Tableau to help us visualize the statistics’ relationships. Since tournament wins is our class variable we converted it into a dimension rather than leave it as a measure. The more and more graphs and scatterplots we created the more we began to see a common trend amongst the data. Each one of the variables we had included in the dataset was related to each other in some way, shape, or form. Most of them being positive with the exception of losses, which makes sense since the more you lose the less likely you are to succeed in the tournament, and adjusted defensive efficiency, which is calculated as points allowed per one hundred possessions (less is good!).

In the graph above we see average ratings percentage index by pythagorean winning percentage. If a team has a high RPI they will also have a high Pythagorean winning percentage. All of our scatterplots showed similar data. First round knockouts (represented by the blue circle) and the second round knockouts (represented by the orange circle) are easily identifiable as stand outs from the rest of the teams across all dimensions. But as you approach the three, four, five, and six win teams they begin to cluster up extremely close and it make it hard to differentiate the teams from one another. 

In the above graph we see adjusted offensive efficiency versus adjusted defensive efficiency. This was one of the two negative correlations we had which were easily explained that having a lower stat was actually a good thing. The better your offensive efficiency the lower(also better) defensive efficiency the team has. We can see that while champions and runners-up (six and five wins, respectively) were pretty similar in offensive efficiency but the teams with five wins seem to struggle with their defense and most likely lost the championship game because of it. 

Finally we used Weka’s histogram feature to try to give us a different visualization to see the distribution of RPI amongst the different winners. Looking at distribution on a histogram can make it hard to determine a winner. Thankfully our scatter plots made in Tableau helped us tremendously in seeing the relationships between our variables. But since our variables all have a correlation that is roughly the same, how do we choose which variables to use in determining our final bracket? That’s where machine learning kicks in.

## Modeling and Descriptive Analytics

We imported our dataset into Weka and began to get to work. We decided to use a J48 decision tree classifier to help us find out which variables held the most weight in determining how far a team can go in the NCAA March Madness tournament. Our J48’s first few splits were on RPI, pythagorean winning percentage, and adjusted offensive efficiency, respectively. Since these seemed to give us the most information on teams that went farther in the tournament, we determined that those three variables were the ones to decide the outcomes of each matchup. But relying on a single model to determine our variables seemed a bit crazy. So, we decided to take it a step further to ensure we weren’t overfitting the data and to ensure we had the right variables.

## Expanded Modeling and Predictive Analytics

We ran a NaiveBayes because we also had a lot of experience using this classifier in class. NaiveBayes classifiers are easy to implement and the models aren’t as complex. This model showed that Pythagorean Winning Percentage and the Ratings Performance Index were very informative in regards to the tournament wins class. We decided to add one more classifier just be safe. During our first bit of research we’ve noticed a lot of other groups, that also analyzed the NCAA March Madness tournament, had used something called a RandomForest classifier. RandomForest is an unpruned classification or regression trees and it generally shows a performance improvement over single tree classifiers like C45 or J48. By creating these other models using NaiveBayes and RandomForest we can see whether the three variables we got from the J48 still stand out as the best determinants of a tournament winner. Our conclusion remained the same after running these other two classifiers; RPI, pythagorean winning percentage and adjusted offensive efficiency were the three variables to look at to determine our winner.

## Evaluation

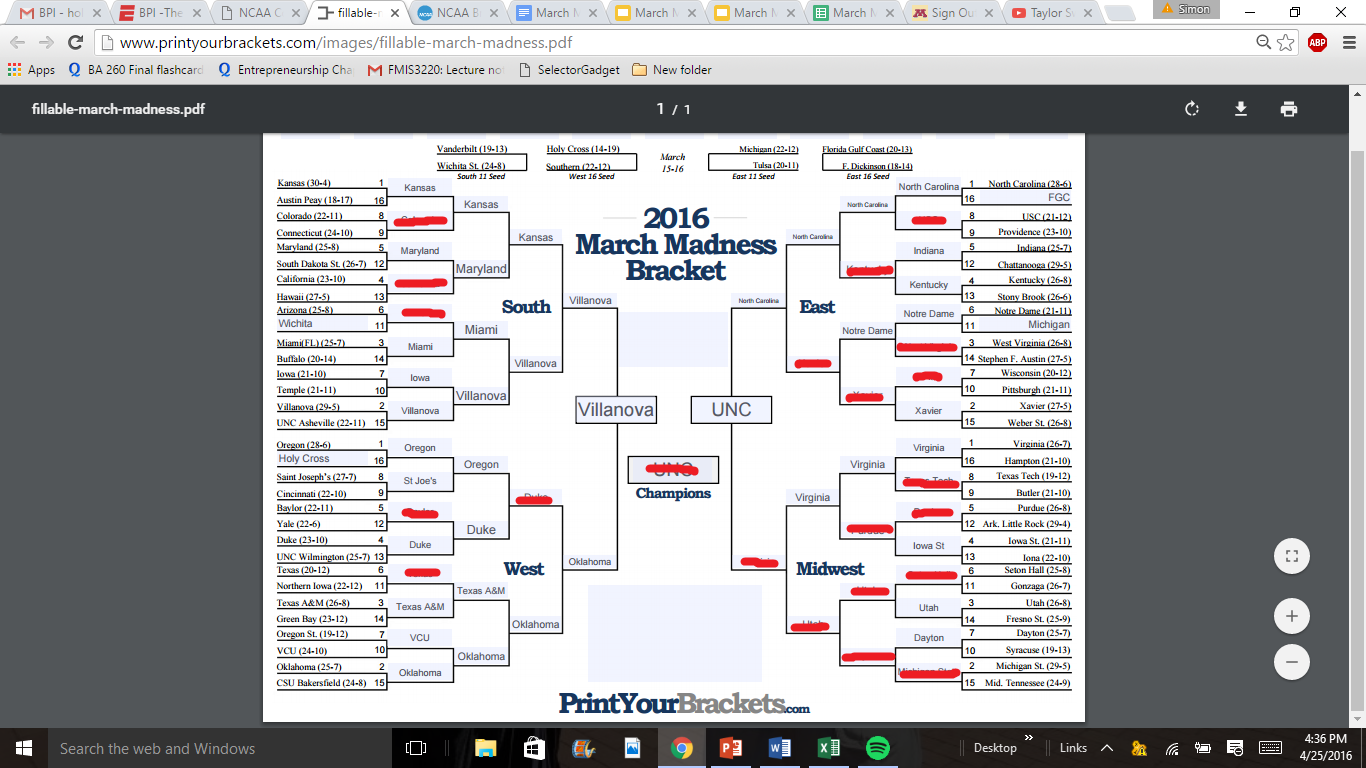
Evaluation of the model is based upon meeting our set goal of 70% accuracy. Because we met our goal dead on, we believe the model is magnificent. Admittedly we did want to pick the correct national title winner, but that was not our main goal. Other ways our model could be evaluated are by accuracy per round. A higher accuracy in the first round could lead to a much higher accuracy overall because there are more games in round 1 by far. If we wanted to do well in an ESPN challenge bracket, we would want our model to be more accurate in later rounds because the later rounds are worth more points. An important part of our model is how we prevented overfitting. We did this by holding out 30% of the data in the J48, Naive Bayes, and Random Forest tests. We also prevented overfitting by using pre pruning, which stopped the tree early before it perfectly classified the training set. We also post pruned the tree as another method of avoiding overfitting. Another way our model could be evaluated is by the Area Under the Curve in the J48, Naive Bayes, and Random forest classifiers. Typically AUC needs to be .9 or above to be considered good, but we’ve come to realize that in the insane randomness of college basketball that .6 or above is pretty decent. The AUC on our J48 tree was .67, which is pretty good. This AUC is a good indicator of the strength of our model, and the variables associated with it. The AUC of our Naive Bayes was .6, and the AUC of the Random Forest was also about .6. The combination of these three made us pretty confident in the evaluation of our model being solid. An interesting inference in these AUC’s however is the discrimination between the AUC of the J48 and the Naive Bayes near the middle of the graph. This could simply be because they are different classifiers. To improve analysis of our model we would want to run many more tests with different variables and classifiers. With more data we could for example run an Apriori or even check out clusters to further our model.

## Prescriptive Analytics

Alternatives for our model included 3 different scenarios. The original model was based upon the variables RPI, Pythagorean Winning Percentage, and Adjusted Offensive Efficiency. These three variables were chosen after running our data sheet through Weka in our three different classifiers, and then using our own comparative discretion to single out variables. With these three variables we then ranked them in order of importance, and gave them weights. RPI was substantially the most important variable, but then Pythagorean Winning Percentage and AOE were about the same. We then cross analyzed our spreadsheets with the March Madness Tournament Bracket and were able to pick winners round by round. A winning team was typically a team that had the higher of two of the three variables, or sometimes all three variables. There were certain cases where a team had only RPI higher, but the other two variables were so close to the opposing teams that we were able to pick the team with the higher RPI. The end result of our first scenario was a 62% accuracy. We correctly picked 39 of the 63 games, but most astounding of all was our 100% accuracy of rounds 2 and 3 in the South and West. This led to a 75% overall accuracy in rounds 2 and 3. The disappointing section of this scenario was that we only picked 2 of the 4 teams in the Final Four, and had zero of the teams correctly picked for the national title game. (the crossed out teams are our incorrect picks)



Looking at this data we decided to incorporate a new variable called BPI into our bracket for rounds 4 to the end. BPI is a team rating system that accounts for the final score, pace of play, site, strength of opponent and absence of key players in every Division I men's game. Starting in round 4 we then hand picked winners by whoever had the higher BPI. The results kept our accuracy percentage at around 62%, but increased our final four accuracy to 75%. It also gave us the correct national title game, but we still were unable to pick the correct winner.



To continue our pursuit the most accurate bracket, we decided to incorporate BPI into all rounds. To do this we took out AOE, in made it the most important variable, along with RPI and Pythagorean in lesser importance. We then hand picked winners with a secret combination of these variables to get an accuracy of 70%. We were very pleased with this result.



There is still definitely room for improvement. If we were to include a closer to home variable, we believe that more of the first round upsets could be predicted. Our next step would be to search through even more data and eventually craft a spreadsheet with these new variables to be run through Weka to see if they could be of any importance. If said data ended up being valuable it would be tested in our hand picking method to conclude if accuracy could be increased in the tournament bracket.

## Comparison With Documented Results

Because we got most of our data from Kaggle, we were able to compare our results with many other people. In comparison to the winner of the event, we didn’t use geographic location or seed. We also ran as many tests as we could in the first round like they did. The winner uploaded their data into SQL and then used R, but we did not use either. The winner had Villanova beating North Carolina, but because this winner was on the private portion of the competition, we could not see any of their script. With that being said, so we had no further insight from them.

In a separate documented result done by Paul Bessire (http://www.analytics-magazine.org/blogs/39-blogs/1249-method-to-march-madness), different key attributes were used. His key attributes were offensive rebounds, how often a team gets to the foul line, and the depth of their roster. He also stated that trying to make upset picks is the biggest mistake people make. His final pick was Indiana, which was a 2 seed.

In a study done by BracketOdds (http://bracketodds.cs.illinois.edu/), a group from the University of Illinois, they created algorithms similar to those of logistic regression and linear distribution in Weka. Using their algorithms, a number 1 seed is chosen to win almost 80% of the time. This is because number 1 seeds have won the tournament almost 60% of the time since 1985 (19 of 32). They used data from the past 32 tournaments, while in comparison, we only used data from the last few. The most important thing talked about in their report was “probabilistic statistics”. This is identifying a distribution that models the probability of certain seed combinations playing each round, accomplished by determining the frequency that each seed reaches a specific round, then fitting the data in a truncated geometric distribution. In comparison to what this group has done, something we have similar is that we also picked a number 1 seed to win. We picked different key attributes than this group because we had different algorithms.

## Deployment

After compiling all of our results together we came to a few conclusions in what we could do with this information. Our first idea is to create our own personal website. This website would link to a copy of the dataset we created as well as the models we used. The user will be able to run their own classifiers with different data and build a bracket of their own, or see the one that we built to compare with. We could also choose to charge a subscription fee to use our website or allow everyone to access it for free. One problem we could possibly encounter if we chose to create our own website, is the cost of hosting the site. We would either have to host the site ourselves, which can be spendy due to equipment costs, or outsource the hosting to someone else. We can reduce the cost of hosting by either charging a subscription fee, or using GoogleAd services. But if online consumers choose not to subscribe, then costs will outweigh the benefits. Another idea we came up with was giving our tools to sports channels such as ESPN or CBS Sports. They could take our findings and expand on them, even improve them with the help of their department of specialized analysts. Similar to hosting it ourselves, everyone can have access to our findings. If all of ESPN and CBS Sports’ readers use the same findings we did, the number of identical NCAA brackets would sky rocket. This leads to our third option, we can keep it for ourselves. By withholding our findings we can enter our brackets into the many NCAA pools to hopefully win some money. With an accuracy level of 70% we could easily win our money back and potentially win some more!

## 