How Can Machine Learning Determine Whether a Women's Tennis Player Will Make it to Top 100?

Saina Deshpande[1] and Vanessa Klotzman[#]

[1]Vikhe Patil Memorial School, Pune, India
[#]Advisor

ABSTRACT

There are a lot of speculations within and outside of the tennis community about whether factors like height, age, and nationality play a role in the success of a tennis player. For this study, 'success' is defined as making it to the Top 100 ranked list. There have been studies in the past associating height of a tennis player with success, but this has primarily been done for men's tennis players. In this study, we establish the relation between height and success of women tennis players. We also consider two additional factors: age and nationality. We mathematically conclude using Pearson's correlation coefficient whether there is any statistical correlation between these three factors and success. Once we establish the relationship, we develop an Artificial Intelligence model to predict future successful players based on historical tennis data. Since earlier studies have already considered height as one of the success factors, our machine learning model uses Naïve Bayes' to determine the probability of success using all three factors to predict success with an accuracy of 0.67 for dataset used. The individual Pearson correlation coefficients for height and age with success, demonstrating the applicability of factors in identifying a player's potential for success are 0.23 and 0.19 respectively. Further research can be conducted by using more factors or larger dataset and could foster greater understanding of female success in tennis.

Introduction

A lot of tennis players wonder if they will make it to the Top 100 rank. Nowadays, machine learning can come up with an answer to questions like this when given enough data. Machine learning can constantly learn recent trends based on historical and current data to determine a player's odds of success.

Machine learning is being used in sports more and more by the day. In tennis, there is an effort to improve the tactical game by applying machine learning algorithms towards analyzing different parts of a match. But machine learning is mostly used to predict the outcome of a match ((Sipko, M. (2015). Machine Learning for the Prediction of Professional Tennis Matches.)).

This is used widely in the betting industry. However, not many have used machine learning to analyze the success factors of tennis players. One reason could be that it is difficult to predict success. Just the collection of data for so many factors is a daunting task.

There are very few studies done to predict success for female players. Over the last few years, we have seen an increasing number of female athletes competing in lawn tennis. The Women's Tennis Association (WTA) is the official women's tennis organization and posts a ranking list that is constantly updated and archived. Currently, there are 1576 women on the list from over 85 nations. How do they or anyone know who will make it to the Top 100?

Since lawn tennis is a physical sport, previous research has generalized the success of a player to physical attributes like height, muscle mass, and body build. However, this method does not hold true most of the time, as can

be seen in the current rankings. For example, the top-ranked women's player, Ashleigh Barty, is 5 feet 5 inches tall, which is well below the average height for the top 100 players (5 feet 9 inches) ((Wood, R. (2016). Height of Wimbledon Players Over Time. *Topend Sports Website*)). This poses a problem in understanding how certain attributes affect the probabilistic success of female lawn tennis players.

This paper analyzes the relation between the ranking of a professional women's lawn tennis player and some of her attributes, specifically age, height, and nationality, and shows the dependence of the ranking on each of these factors using machine learning. Hence, we are attempting to predict success with a combination of three factors. To achieve the stated goals of our research, the rest of the paper is organized as follows. First, in section 2, we establish our data sources, the dataset used, and the correlation between factors and success. Then in section 3, we outline the usage of Bayes' based machine learning and Pearson's correlation coefficient for our machine learning algorithm. We then present the results of the algorithm in section 4 followed by a brief section on the limitations in section 5. We conclude the paper in section 6 with a discussion on related and further research.

Data

Since lawn tennis is a widely recognized sport, there are multiple databases and reference websites available for application in this study. Data is available on each professional player according to different categories. However, the data we needed was not easily available in a singular, centralized database. Therefore, the data was collected, cleaned, and processed manually. The primary site of reference was the official Women's Tennis Association (WTA) singles ranking website. This website has rank, country, and age information for professional women tennis players. The WTA website classifies this information by month and year from 2000 up to the current month.

For this study, we defined 'success' as reaching the Top 100 ranked list in any given year. The following data was collected accordingly:

Top 100 ranking

We got the year-end rankings list as of 31st December for each of the last 10 years - 2012-2021 - for the Top 100 rank players. We obtained this list from the WTA website. This list was needed to train the machine learning model for successful players. A numerical value of '1' was assigned for the Top 100 players. In the case of players who appeared in the Top 100 ranking for multiple years, we only used the data on each player in their first year in the Top 100 in this time period.

Players ranked 101-300

This list was needed to train the machine learning models for unsuccessful (for the purpose of this research paper) players. We pulled this list from the WTA website as of 31st December for the last 10 years- 2012 to 2021- for Top 100 rank players. A numerical value of '0' was assigned for these players. Since there was a high possibility that a player could be part of the unsuccessful list in multiple years, we used the data from only one instance of each player in the first year their 101- 300 ranking appeared. If a player was included in the top 100 list, all instances of the player were deleted from the 'unsuccessful' list.

Factor 1 – Age of player

Calculating the age of the player as of that ranking year was crucial and this data is not readily available. WTA gives only the current age of the player, not their age during the ranking year. So, we had to search for the birth year of each player manually from the WTA player profile and then subtract it from the ranking year to get the age of the

player. ((Women's Tennis Association (n.d), Active WTA Players))

Factor 2 — Height of the player

The WTA ranking list does not provide height of the player, so we had to manually obtain the height of each player from the WTA player profile. ((Women's Tennis Association (n.d), Active WTA Players))

Factor 3 — Nationality

We got this information from the WTA site, but we had to give a distinct numerical value to each country so that the machine learning algorithm can consider it in Bayes' theorem. Players whose information was not available were excluded from this list. Despite these challenges, 1633 records spread over the past 10 years were collected.

Methodology

For our research question, 'How can machine learning determine the probability of professional women lawn tennis players making it to the Top 100 rank based on attributes like height, age, and nationality?', we used Pearson's Correlation Coefficient to determine if factors selected influenced success, and then used Bayes' Theorem to calculate the probability of success as can be computed using these factors. The following section outlines the steps of this methodology in greater depth:

Data assembly

A data file containing the information of all players was created. All the factors including Nationality and Success were converted to Numerical values.

Using Pearson's correlation coefficient

The next step was to find Pearson's correlation coefficient to prove there is a relation between the factors and success. The correlations are allotted values between 1 and -1, where 1 is an absolute positive correlation, -1 is an absolute negative correlation and 0 signifies no correlation between the data. This means a positive value shows that the data is in direct proportion and a negative value shows that it is in inverse proportion. ((Glen, S. (n.d). Correlation Coefficient: Simple Definition, Formula, Easy Steps. *Statistics How To*)). The larger the absolute value of the coefficient, the stronger the relationship between the data is. The Pearson's Correlation Coefficient is calculated using Equation 1:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$

Equation 1: Pearson's Correlation Coefficient

Pearson's correlation coefficient for height and weight was calculated. Nationality could not be included in the calculation since the values used were the three-digit standard country codes which have no numerical logic.

Applying Naive Bayes

The next step is to use Naïve Bayes as our machine learning algorithm. Naïve Bayes uses Bayes' Theorem to predict whether the player will be in the top 100 or not based on the values of the three features. Naïve Bayes draws influence from the Bayes theorem which describes the probability of a given event based on prior knowledge of the conditions

that are related to that event. ((Joyce, J. (2008). *Bayes' Theorem*)) . The Bayes Theorem can be used to generate the following classification model:

$$P(y|X) = \frac{P(X|y)P(X)}{P(y)}$$

Equation 2: Bayes' Theorem

Equation 2 can be further extended as follows:

$$P(y|x1, x2, x3 \dots xN) = \frac{P(x1|y).P(x2|y).P(x3|y) \dots P(xN|y).P(y)}{P(x1).P(x2).P(x3) \dots P(xN)}$$

Equation 3: Further classification of Bayes' Theorem

In Equations 2 and 3, X = x1, x2, x3, …xN are a list of independent predictors and y is the class label. Naive Bayes' theorem assumes that all predictors (features) have an equally important impact on the outcome's probability. Since there is no verified dependence of a player's success on factors like height, nationality, and age, it was assumed that the result is equally dependent on all three factors. We will specifically use the Gaussian Naïve Bayes since the feature values are continuous and are expected to follow a Gaussian distribution ((Sharma, P. (2021). *Implementation of Gaussian Naïve Bayes in Python Sklearn*)). The algorithm was coded in Python. The following python modules were used in the program:

Pandas

((Mckinney, W. (2010). *Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference*, 51-56))This is used for data science/data analysis and machine learning tasks

Sklearn.model_selection

((Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python. Journal of machine learning research*, 12(October), 2825-2830)). To use functions like train_test_split to split the data into train, test sets for training.

Sklearn.naive_bayes

((Scikit-Learn. (2007). 1.9. Naive Bayes.)); ((Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python. Journal of machine learning research*, 12(October), 2825-2830)). GaussianNB implements the Gaussian Naive Bayes algorithm for classification.

Scipy.stats

((Virtanen, P., Gommers, R. and Oliphant, T. E. (2020). *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods*, 17, 261-272)). Pearsonr helps to calculate the Pearson correlation coefficient and the r-value for testing correlation.

Matplotlib

((Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment. Computing in science & amp; engineering*, 9(3), 90-95)). Matplotlib is a cross-platform, data visualization and graphical plotting library for Python.

Seaborn

((Waskom, M., Botvinnik, O., O'kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., Ruiter, J. D., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G.,

Quintero, E., Bachant, P., Martin, M., … Qalieh, A. (2017). *Seaborn: Statistical Data Visualization. Journal of Open Source Software*, 6(60))). It is used to make statistical graphics in Python and builds on Matplotlib.

This algorithm can predict whether a player will be in the top 100 or not based on values inputted for the three features. The factors taken into consideration are height, age, and nationality. The train-test split technique was used for evaluating the performance of the machine learning algorithm.

Results

Based on the machine learning algorithm, our results indicated that we were able to train the model based on collected data and test the model using the test data. The below subsections outline the individual correlation of height and age with success and also explain the output and accuracy of the Naïve Bayes model in detail:

Pearsonr Outputs

When we executed the first step of our algorithm, we mathematically prove that there is a correlation between height and success, as well as between age and success. To glean a tentative understanding of how much each factor individually influenced the outcome, Pearson's correlation constant was calculated separately for height and age factors. The heatmap (Figure 1), which is an output of the algorithm, shows the correlation between each factor and success.

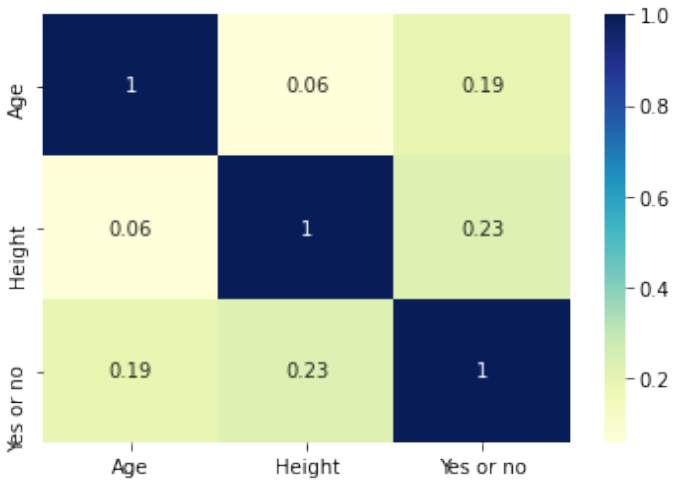

Fig 1. Heatmap showing correlation between various factors and success

For this research, let us ignore the first two rows of the heatmap. The third row of the heatmap clearly shows that height and age have positive correlations of 0.23 and 0.19 respectively with reference to success (Yes/No). This means success weakly correlates with height and age. Since the correlation coefficient for height (0.23) is greater than that of age (0.19), the preliminary dependence of success is greater on height than on age.

Naive Bayes Outputs

Next, the algorithm is able to predict the success of the player using three factors. Half of the dataset was used to train the model and the other half was used to verify if the prediction based on the training set was correct. The algorithm

gave an output of '1' or '0' based on whether the player could be in the Top 100 or not.

Accuracy

The accuracy of this model was calculated by comparing the algorithm outcome to the actual value of success. It had an accuracy of 0.67 when the training set used was 50% of the total data set, i.e., 816 data points.

Limitations

We identified some limitations which are highlighted below:

Size of dataset

The accuracy of the algorithm can be improved by ensuring a larger dataset is available. We considered unsuccessful player rankings from 200-300. We can increase the dataset from 200-1000 ranking. However, since a dataset with all factors was not readily available, the data had to be manually compiled, processed, cleaned, and analyzed. This motivated the choice of dataset for our paper.

Number of Factors

This research considered three factors based on which success is dependent. ((Bosscher, V. D., Knop, P. D. and Heyndels, P. (2004). Comparing tennis success among countries. *JCMS Journal of Common Market studies*, 25(1), 49-68)) ; ((Li, P., Weissensteiner, J. R., Pion, J. and Bosscher, V. D. (2020). Predicting elite success: Evidence comparing the career pathways of top 10 to 300 professional tennis players. *International Journal of Sports Science & Coaching*, 15(5-6))); ((Ramamonjisoa, 2020)) Some other factors could also be considered. We decided to consider simple factors which can be mathematically proven

Dependence on time

Over the past few years, the average height of the top 100 professional women's tennis players has increased slightly. However, the data set does not take this difference into account and predicts the outcome based on the value of the height.

Discussion

The previous research studies compared the height of male successful players manually to analyze the advantage height would give to certain tennis players. ((Ramamonjisoa, S. (2020). *How height matters in professional tennis?*)) Some previous research has also calculated the mean age of players to prove that age of top-ranked players has increased. ((Gallo-Salazar et al., 2015)) The basis of this research is to create a predictive machine learning mathematical model to analyze the dependency of multiple factors in determining success.

The results show that it is indeed possible for an algorithm to predict whether a player could be in the top 100 or not with the test data accuracy of 0.67. This is significant since we can estimate the player's potential based on her age, height, and nationality.

As shown by the Pearson correlation coefficient calculation, the height of the player has slightly more of an impact on the ranking than the age. Both have a positive correlation, which means that the chance of being in the top

100 rankings increases with height and age up to 25 years. ((Gallo-Salazar, C., Salinero, J. J., Sanz, D., Areces, F. and Coso, J. D. (2015). Professional tennis is getting older: Age for the top 100 ranked tennis players. *International Journal of Performance Analysis in Sport*, 15(3)))

This research can be taken further by adding more features to the dataset to make the algorithm more accurate. It can also be built on a larger data set spanning a greater timeline. Further research on this topic could foster a greater understanding of female success in tennis, leading to greater gender equity in tennis.

Acknowledgments

Thank you for the guidance of Vanessa Klotzman mentor from the University of California, Irvine in the development of this research paper.

References

Bosscher, V. D., Knop, P. D. and Heyndels, P. (2004). Comparing tennis success among countries. *JCMS Journal of Common Market studies*, 25(1), 49-68, Retrieved from https://www.researchgate.net/ publication/239844205_Comparing_Tennis_Success_Among_Countries

Burns, E. (2021). Retrieved from https://www.techtarget.com/ searchenterpriseai/definition/machine-learning-ML

Gallo-Salazar, C., Salinero, J. J., Sanz, D., Areces, F. and Coso, J. D. (2015). Professional tennis is getting older: Age for the top 100 ranked tennis players. *International Journal of Performance Analysis in Sport*, 15(3), Retrieved from https://doi.org/10.1080/24748668.2015.11868837 doi: doi.org/10.1080/24748668.2015.11868837

Glen, S. (n.d). Correlation Coefficient: Simple Definition, Formula, Easy Steps. *Statistics How To*, Retrieved from https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & amp; engineering*, 9(3), 90-95, Retrieved from https://ui.adsabs.harvard.edu/link_gateway/2007CSE.....9...90H/ doi:10.1109/MCSE.2007.55 DOI: doi:10.1109/MCSE.2007.55

Joyce, J. (2008). *Bayes' Theorem.,*. Retrieved from https://philpapers.org/ rec/JOYBT

Li, P., Weissensteiner, J. R., Pion, J. and Bosscher, V. D. (2020). Predicting elite success: Evidence comparing the career pathways of top 10 to 300 professional tennis players. *International Journal of Sports Science & Coaching*, 15(5-6). DOI: doi:10.1177/1747954120935828

Mckinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, 51-56, Retrieved from 10.25080/Majora-92bf1922-00a

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(October), 2825-2830, Retrieved from https://www.jmlr.org/papers/volume12/pedregosa11a/ pedregosa11a.pdf

Ramamonjisoa, S. (2020). *How height matters in professional tennis?*. Retrieved from https://www.siskoramamonjisoa.com/post/how-height-matters-in-professional-tennis

Scikit-Learn. (2007). *1.9. Naive Bayes.,*. Retrieved from https://scikit-learn.org/stable/modules/naive_bayes.html

Sharma, P. (2021). *Implementation of Gaussian Naïve Bayes in Python Sklearn.,*. Retrieved from https://www.analyticsvidhya.com/blog/2021/ 11/implementation-of-gaussian-naive-bayes-in-python-sklearn/

Sipko, M. (2015). Machine Learning for the Prediction of Professional Tennis Matches. , Retrieved from http://www.doc.ic.ac.uk/teaching/ distinguished-projects/2015/m.sipko.pdf

Virtanen, P., Gommers, R. and Oliphant, T. E. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat Methods*, 17, 261-272. DOI: https://doi.org/10.1038/s41592-019-0686-2

Waskom, M., Botvinnik, O., O'kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., Ruiter, J. D., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., … Qalieh, A. (2017). Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, 6(60), Retrieved from https://doi.org/10.5281/ zenodo.883859 doi: doi:10.21105/joss.03021

Women's Tennis Association (n.d), available: https://www.wtatennis.com/ [Accessed 21 Oct 2021]

Women's Tennis Association (n.d), Active WTA Players, available: https://www.wtatennis.com/players

Wood, R. (2016) *Height of Wimbledon Players Over Time*, Topend Sports Website, available: https://www.topendsports.com/sport/tennis/anthropometry-wimbledon.htm