Su Kara



**Decoding the Djokovic-Nadal Rivalry: A Statistical Journey into Predicting Tennis Match Outcomes**

**Introduction**

In the realm of professional tennis, the conclusion of each match unveils a series of statistics, offering a glimpse into the players' performance. These statistics, which include winners, unforced errors, and serve percentages, have long intrigued me. They raise a question: can these figures alone, without directly watching the match, predict the winner? This report is an exploration of this hypothesis, delving into the predictive power of match statistics in tennis, particularly focusing on the renowned rivalry between Novak Djokovic and Rafael Nadal.

**Background and Motivation**

My interest in this analysis was sparked by a common feature in tennis broadcasts: the display of a chart comparing players' post-match statistics. This led me to ponder the depth of the story these numbers could tell. *Are they mere reflections of the game, or do they hold the key to revealing the winner?* To explore this, I experimented with the Tennis Major Tournament Match Statistics from the UCI ML repository, concentrating on the 2013 men's Grand Slam tournaments: Australian Open, French Open, Wimbledon, and US Open. This comprehensive dataset listed a wealth of variables for each match, including player names, match result, first and second serve percentages, aces, winners, unforced errors, break points created and won, net points attempted and won, total points won, set results, and final number of games won, among others.[[1]](#footnote-0) It provided the perfect ground to investigate if match winners could be predicted based solely on statistical performance, examining not just the outcome but the nuanced interplay of skills and strategies as quantified in these variables.

**General Match Statistics Analysis**

My initial goal in this study was to explore the relationship between various player performance metrics and match outcomes in tennis. Specifically, I aimed to determine whether a higher value in key metrics such as First Serve Percentage (FSP), Second Serve Percentage (SSP), First Serve Won (FSW), and Second Serve Won (SSW), as well as a metric I developed called 'Consistency' (Winners minus Unforced Errors), generally corresponded with winning a match. To make the analysis more manageable, I transformed these continuous variables into discrete ones. For example, if Player 1 had a higher FSP than Player 2, Player 1's FSP was assigned a value of 1 and Player 2's FSP was given a value of 0, and vice versa. To check the correspondence between these binary variables and match outcomes, I calculated the frequency with which the player with the higher value of a given metric won the match. This was done by finding the number of times the player with the higher examined variable also won, and then dividing this by the total number of matches. This approach was crucial in identifying which metrics were most indicative of winning a match. To code and execute this analysis, I utilized programming libraries such as 'os', 'pandas', and 'numpy', which facilitated the manipulation and analysis of the dataset.

I found the following percentages indicating the frequency with which the player with the higher statistic won: 0.57 for FSP, 0.39 for SSP, 0.69 for FSW, 0.58 for SSW, and notably, 0.86 for Consistency. These percentages were instrumental in revealing the relative importance of each metric. Notably, the 'Consistency' metric, which I had calculated, emerged as the most significant indicator of match victory, with the highest percentage of 0.86.

**Djokovic-Nadal Rivalry**

After establishing that 'Consistency' was the most pivotal factor in predicting match outcomes, I was intrigued to test this finding in a specific context. I decided to analyze a long, historical rivalry between two incredible players: Djokovic and Nadal. Interestingly, according to statistics from tennisabstract.com[[2]](#footnote-1), their head-to-head record stood almost even, with Djokovic winning 30 and Nadal winning 29 of their 59 total matches. This made the battle an ideal case study for my analysis, given its unpredictability.

To delve deeper, I embarked on compiling a tailored dataset of their matches. I cross-referenced the match history from Wikipedia with the records on Tennisabstract, ensuring accuracy and completeness. For each encounter, I created a detailed spreadsheet shown in **Figure 1** capturing not only the match names but also the 'Consistency' and 'First Serve Won' (FSW) stats for each player.

Note: The inclusion of FSW was based on its notable association with winning outcomes in the previous section. However, I decided to focus mainly on Consistency in the next part of my study because it was the most crucial factor. It's worth noting that the analytical method I describe next, while applied to Consistency, could be equally adapted to examine FSW.

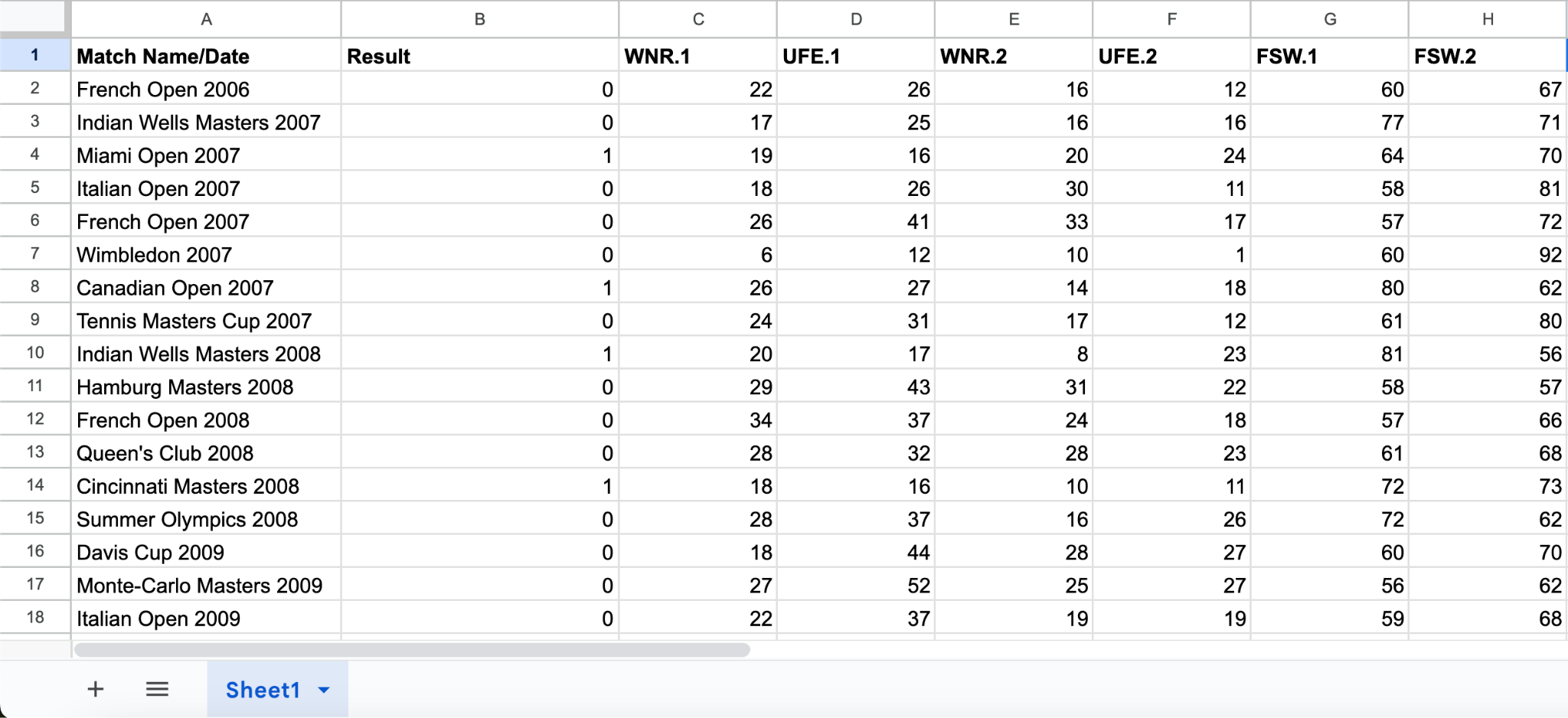


Figure 1: The above spreadsheet shows the data I compiled on Djokovic-Nadal match statistics

The next step was converting these metrics into a more analyzable form. Drawing from my previous approach, I transformed 'Consistency' from a continuous to a discrete variable. In this binary format, the player with the higher consistency in a match was assigned a value of 1, while the other was assigned 0. Similarly, the match result was coded as 1 for a Djokovic win and 0 for a Nadal win. This binary representation aimed to provide a clearer picture of the relationship between a player’s consistency and their success in these specific matches.

**The Magic of Bayes'**

I used Bayes’ Theorem to explore the relationship between consistency and success for each player separately, as shown in the two equations above. Let's break down the first equation for Djokovic:

: This was the main probability I wanted to find. To calculate it, I needed the other probabilities in the equation.

: This measures how often Djokovic is more consistent in the matches he wins. To find this, I looked at all the matches Djokovic won and checked how many times he was more consistent than Nadal in those matches.

: This is the overall chance of Djokovic winning a match, regardless of any consistency statistics. It's calculated by taking the total number of matches played and seeing what fraction of them Djokovic won (which I mentioned in ‘Djokovic-Nadal Rivalry’ was 30/59).

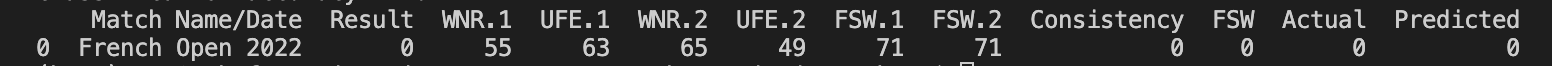
: This is how often Djokovic is more consistent than Nadal in all their matches, which is found by looking at all their encounters and counting the times Djokovic had better consistency.

By plugging these values into Bayes’ Theorem, we are given the likelihood of Djokovic winning a match under the condition that he is more consistent – . The same method applies to Nadal in the second equation.

**Final Results and Analysis**

The outcomes of the Bayesian analysis were quite revealing. For Djokovic, was . In contrast, for Nadal, was slightly lower at . These results align with the ‘General Match Statistics Analysis’ conducted earlier, where I found that the probability of a player winning when they were more consistent was approximately 0.86, a number in between the range of the values found for Djokovic and Nadal. This underscores the critical role of ‘Consistency’ in determining match outcomes. It seems that maintaining a balance between aggressive shots and errors, as captured by the 'Consistency' metric, is a key determinant of success on the tennis court.

(**Note**: I had this project idea a month ago, before learning about Naive Bayes and how it can be used to find and calculate classification accuracy. After implementing Naive Bayes on PSet 6, I wanted to try it on this project. I split the train dataset to contain the first 58 matches between Djokovic and Nadal, and put their 59th match in the test set to see if the model would accurately predict the winner of the next match. It correctly classified the outcome of the match as a win by Nadal. 🙂While I could extend this project to do more research using Naive Bayes – testing more features, and with a larger train and test set – and using Logistic Regression with the initial continuous values of variables, I am hoping to continue this journey in a future project…)



**Code**

(I did use ChatGPT to help provide me with debugging advice when I was stuck for a long time.)

Here is a link to my code: <https://drive.google.com/drive/folders/1gOacQSwj3_QLYED6Xuo4Ct-EDPJ8GCmU?usp=sharing>

**Discussion and Conclusion**

This analysis does more than just shed light on the dynamics of one of tennis’s greatest rivalries; it also highlights the potential application of these methods in other contexts. Similar statistical approaches could be used in other sports or competitive scenarios to identify key performance indicators and understand their impact on success.

In the high-stakes world of professional tennis, particularly in the matches between Djokovic and Nadal, consistency is not just a virtue but a strategic weapon. It's the thread that weaves through their encounters, often determining the victor in their epic battles. Just as their rivalry has captivated tennis fans worldwide, the story their data tells is equally compelling, providing a deeper, numbers-driven narrative of their on-court strategies and triumphs.

1. <https://archive.ics.uci.edu/dataset/300/tennis+major+tournament+match+statistics> [↑](#footnote-ref-0)
2. <https://www.tennisabstract.com/cgi-bin/player-classic.cgi?p=NovakDjokovic&f=ACareerqq&q=RafaelNadal&q=RafaelNadal> [↑](#footnote-ref-1)