Predicting Tennis Match Outcomes using Machine Learning Techniques

Yuwei Fu, Zhuocheng Han, Tianyi Niu, Alexander Zheng

1. Introduction

In recent years, sports analytics have gained significant attention, and the use of data-driven techniques has become increasingly prevalent in various sports, including tennis. This project aims to apply machine learning methods to predict tennis match outcomes, specifically tournament finals and individual player performance. Our approach involves analyzing a comprehensive dataset of historical tennis match results, conducting exploratory data analysis (EDA), and building predictive models to achieve two main objectives: predicting tournament finals outcomes based on tournament performance and predicting match outcomes based on player historical performance.

Our project represents a stepping stone towards more advanced and comprehensive models for predicting tennis match outcomes. In the future, we plan to explore additional factors, such as player fatigue and weather conditions, to further improve the predictive accuracy of our models. Additionally, we aim to investigate the application of our models in other individual sports and expand our analysis to a broader range of competitive scenarios.

In this report, we will delve into the details of our dataset, the data analysis we conducted, and the models we developed to achieve our objectives. Through this comprehensive analysis, we hope to demonstrate the potential of machine learning techniques to predict tennis match outcomes based on tournament performance and player historical performance. By refining and expanding upon these models, we can contribute to the growing field of sports analytics and drive advancements in sports betting and player performance analysis.

2. Dataset and EDA

2.1 Dataset Introduction

Our primary dataset is Jack Sackmann's comprehensive tennis WTA and ATP match dataset, which covers match data from 1920 to the present day. This dataset is free and publicly available on GitHub, making it an ideal resource for our analysis. The dataset is organized into yearly files with match data, providing an easy-to-navigate and understand structure. For our study, we decided to use the ATP and WTA match dataset from 2009 to 2019. After data cleaning, we were left with a total of 16,590 matches to analyze.

2.2 Data Cleaning and Processing

Before diving into the data analysis, we first needed to clean and preprocess the dataset. This process involved handling missing values, inconsistencies, and errors in the data. Additionally, we removed any duplicate entries and irrelevant columns, such as columns containing only redundant information or those with an excessive number of missing values. We first removed all matches that contained a player who had played fewer than 100 matches, since they have a lot of NAs in their observations. We also decided to remove all matches that ended in retirement, which means a player forfeited due to injury. This step ensured that our dataset was accurate, consistent, and ready for analysis. Since we are predicting the winner and loser of each match, we need to mask the winner of each match. So

we changed all occurrences of the strings “winner”, “loser”, “w”, and “l”, among the feature names to "player_1" and "player_2," with player_1 referring to the player in each row whose name comes first alphabetically.

During the data preprocessing phase, we also addressed the issue of high collinearity among certain columns in our dataset. High collinearity can lead to unstable and unreliable predictive models. To mitigate this issue, we strategically merged some variables, reducing complexity and improving the interpretability of our model while maintaining the integrity of the data.

There are some new variables that we want to create to indicate a player’s historical performance. To address the challenge of creating cumulative and chronological features in the presented data, we organized all matches in chronological order and developed several new features that summarized a player's past performance leading up to their current match. These features included surface win percentage, level win percentage, head-to-head record, and overall win percentage. The surface win percentage feature was particularly relevant because tennis has four different types of court surfaces, each favoring different playing styles. Considering a player's level win percentage was also important, as some players tend to perform better in higher-profile or larger tournaments, such as grand slams, due to their ability to handle pressure. Below are the variables we decided to use for our models.

```
[1] "tourney_id"        "tourney_name"      "surface"           "draw_size"         "tourney_level"     "tourney_date"
 [7] "match_num"         "winner_id"         "winner_seed"       "winner_entry"      "winner_name"       "winner_hand"
[13] "winner_ht"         "winner_ioc"        "winner_age"        "loser_id"          "loser_seed"        "loser_entry"
[19] "loser_name"        "loser_hand"        "loser_ht"          "loser_ioc"         "loser_age"         "score"
[25] "best_of"           "round"             "minutes"           "w_ace"             "w_df"              "w_svpt"
[31] "w_1stIn"           "w_1stWon"          "w_2ndWon"          "w_SvGms"           "w_bpSaved"         "w_bpFaced"
[37] "l_ace"             "l_df"              "l_svpt"            "l_1stIn"           "l_1stWon"          "l_2ndWon"
[43] "l_SvGms"           "l_bpSaved"         "l_bpFaced"         "winner_rank"       "winner_rank_points" "loser_rank"
[49] "loser_rank_points"
```

The original variable lists

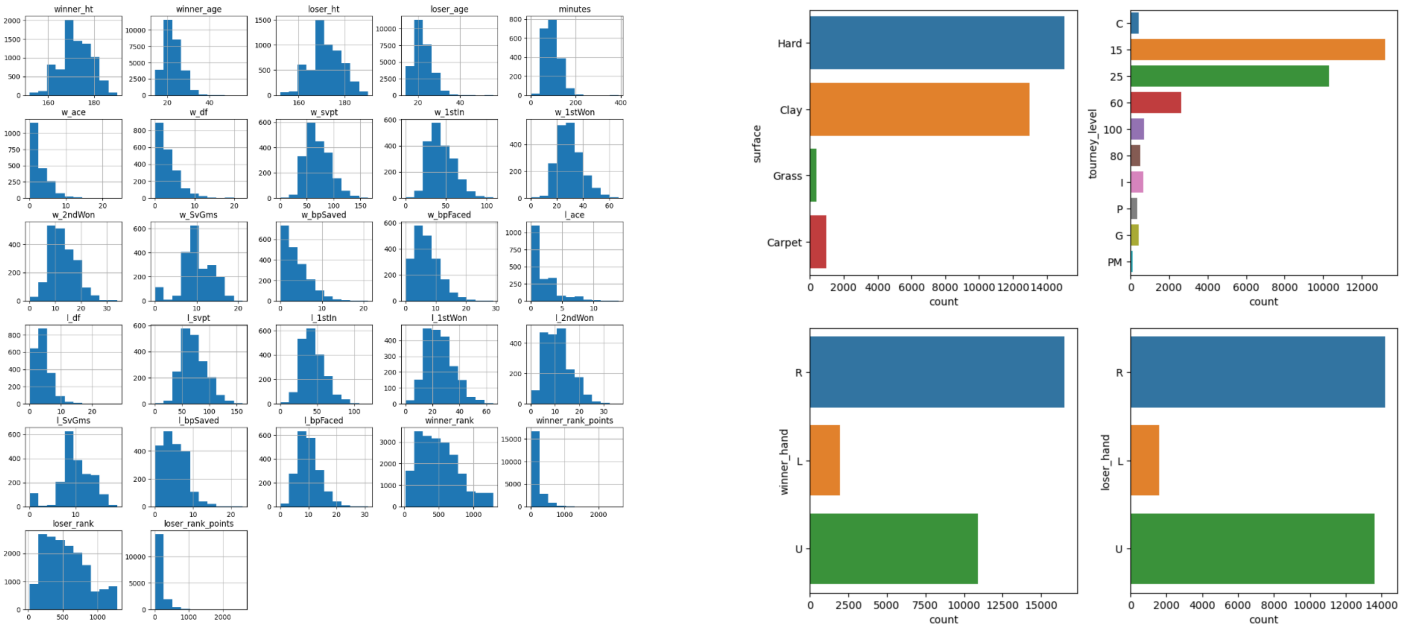
```
 [1] "surface"               "player_1_age"          "player_2_age"          "player_1_ht"           "player_2_ht"
 [6] "player_1_rank"         "player_2_rank"         "player_1_ace"          "player_2_ace"          "player_1_df"
[11] "player_2_df"           "player_1_1stWon"       "player_2_1stWon"       "player_1_1stIn"        "player_2_1stIn"
[16] "player_1_bpFaced"      "player_2_bpFaced"      "player_1_2ndWon"       "player_2_2ndWon"       "player_1_svpt"
[21] "player_2_svpt"         "player_1_bpSaved"      "player_2_bpSaved"      "player_1_h2h"          "player_2_h2h"
[26] "player_1_win_pct"      "player_2_win_pct"      "player_1_surface_win_pct" "player_2_surface_win_pct" "player_1_level_win_pct"
[31] "player_2_level_win_pct" "winner_binary"        "tourney_level"
```

After data cleaning and feature extraction

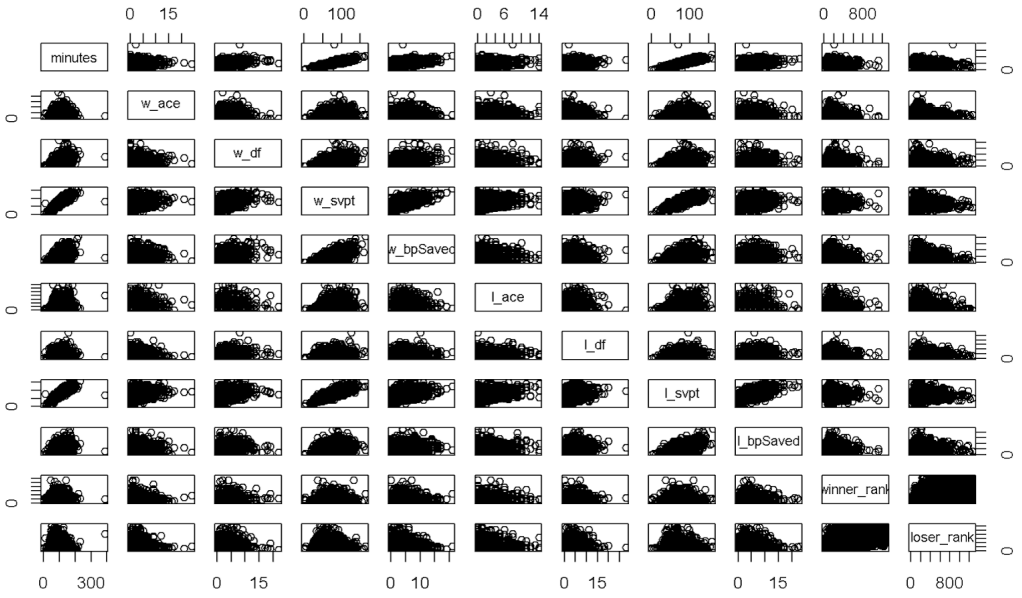
2.3 Data Visualization and Exploratory Data Analysis

With the cleaned and preprocessed dataset, we proceeded with data visualization and exploratory data analysis (EDA) to gain insights into the data and identify potential patterns, trends, and relationships among the variables.

We visualize the distribution of player rankings, ages, heights, and other relevant variables [2]. This allowed us to understand the overall composition of the dataset and identify any outliers or unusual data points that may require further investigation. We also examined the relationships between various variables, such as the correlation between player ranking and match outcome, and the impact of age and height on player performance [3]



Distribution of players' information



Relationships between various variables

These visualizations, along with other analyses conducted during the EDA process, helped us identify patterns, trends, and relationships among the variables. This information guided our model selection and feature engineering processes. Our preliminary test results demonstrate the effectiveness of these preprocessing techniques and EDA. After making these modifications, we observed significant improvements in the test metrics, indicating that our approach has enhanced the overall predictive capability of our machine learning models. In the following sections, we will discuss the models we used to achieve our objectives and provide a detailed analysis of their performance.

3 Models

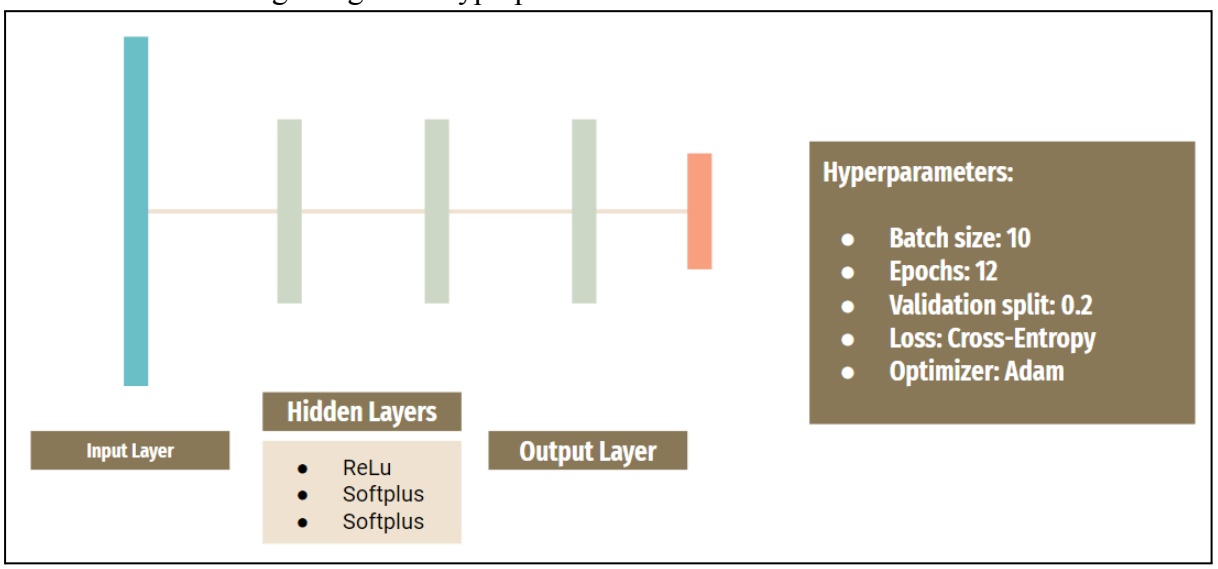
3.1 Model Selection and Evaluation

We cross-examined multiple models, including: Logistic Regression, Naive-Bayes, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Tree, Random Forest, Bagging, and Feed-Forward Network.

To evaluate the performance of these models, we split our dataset into training and testing sets using 80-20 split. We used accuracy, precision, recall, and F1-score as our evaluation metrics to assess the performance of each model.

3.2 Neural Network

We used the following design and hyperparameters for our neural network model:



Our model is composed of an input layer, three hidden layers, and an output layer. Each hidden layer is [60x1] while the output layer is [2x1]. As our project involves different objectives, which will be further described below, the dimension of the input layer is not consistent.

4 Predicting Tournament Finals Outcome Based on Tournament Performance

4.1 Methodology

Our first objective was to predict the outcome of a tournament's final match based on the performance of the two finalists throughout the elimination stages. Using data from both players only available from the current tournament, we aimed to answer the following questions:

1. Can we reasonably predict a match's outcome given match statistics (excluding match score and player points)?
2. Are the statistics from each finalist's previous elimination matches (typically around 4) enough to indicate performance?

4.2 Experimental Results and Analysis for Question 1:

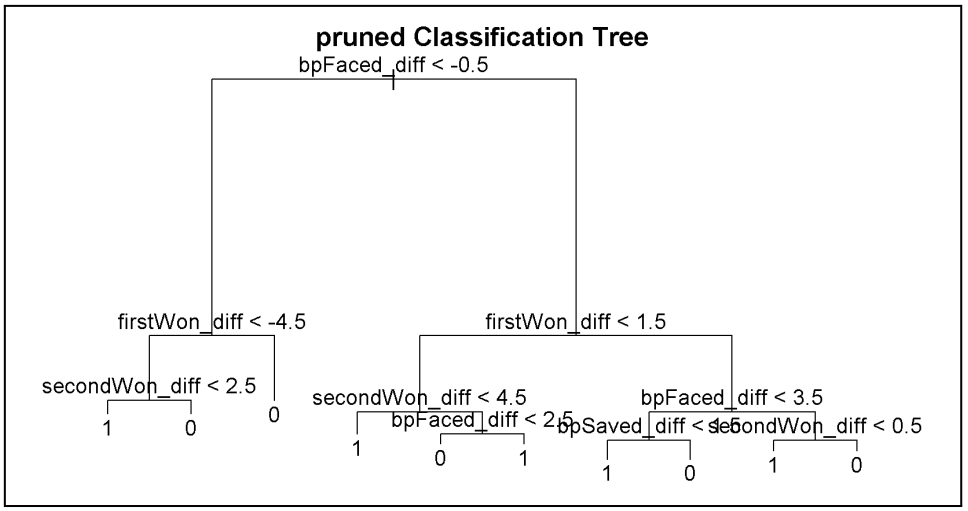
To address question 1, we employed the following methodology. For each match in our dataset, we obtained the statistics for both players (player 1 and player 2) and calculated the difference between them. This difference served as the match encoding, which was then fed into our models to obtain a prediction. Note that to faithfully use data only available at the current stage of the tournament, we discarded each player's historical data calculated earlier, such as "surface win percentage" and "head-to-head win percentage."

We evaluated the performance of our models using the precision, recall, and F1-score metrics. The results are as follows:

	LR	**LDA**	**QDA**	**NB**	**FFN**
Accuracy	0.97	0.97	0.94	0.8	0.95
Precision	0.97	0.97	0.94/0.95	0.82/0.79	0.93/0.98
Recall	0.97	0.97	0.95/0.94	0.76/0.84	0.98/0.92
F1	0.97	0.97	0.94	0.79/0.81	0.95

(Left statistic describes label/class 0, right statistic describes label/class 1 in a binary classification. Only one statistic displayed if both are equal.)

Classification Decision Tree: (Accuracy: 0.8692)



From looking at the nodes in the pruned classification tree, bpFaced diff and secondWon diff are some of the most used features. A breakpoint faced means an opportunity for a player serving to lose their service game. As serving is considered as an advantage in tennis, it is expected that the serving player will win the set most of the time. If a player is facing a breakpoint, they are at a disadvantage. So if player 1 has a bpFaced less than player 2 by more than 0.5, unless their number of first serves won is less than player 2 by more than 4.5, they would win the game according to the tree. The important match statistics variable we need to note is bpFaced, and firstWon, since they are at the top of the tree, since they have the greatest impact on the final classification.

Our experimental results indicate that Logistic Regression and Linear Discriminant Analysis (LDA) achieved the highest accuracy, precision, recall, and F1-score among all models. While our neural network model (Feed-Forward Network) achieved slightly lower performance compared to Logistic Regression and LDA, it still demonstrated strong predictive capabilities, with an accuracy of 0.95. These experiments suggest that the statistics recorded for each match provides enough information to determine the winner.

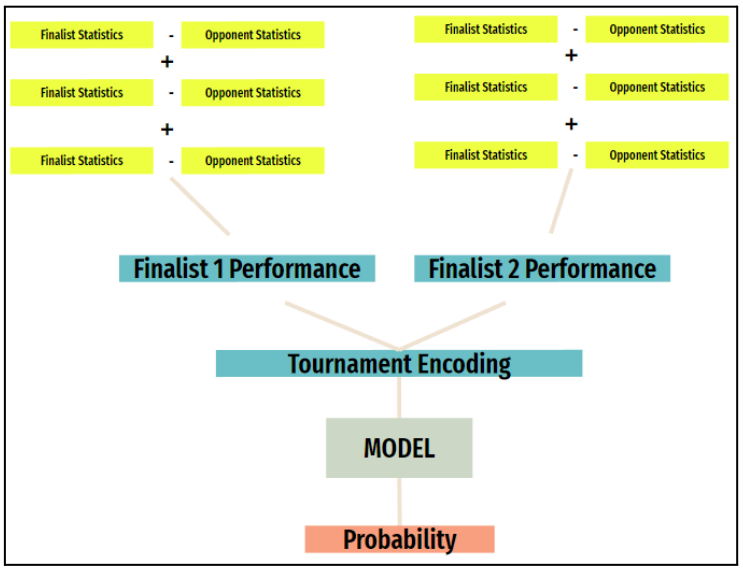
4.3 Experimental Results and Analysis for Question 2

Since the previous results indicate the model can accurately determine the winner of a match by match statistics, we turn our attention to predicting the outcome of a tournament's finals match. We hope to determine the winner of a tournament by looking at each finalist's previous elimination stage matches. We again deploy the difference-encoding strategy utilized earlier, albeit with a few changes.

For each tournament in our dataset:

1. Determine the two finalists.
2. For each elimination match involving finalist 1, obtain the finalist (winner) and opponent (loser) statistics and save the difference as the match encoding.
3. Pool together all match encodings into finalist 1's tournament performance encoding.
4. Repeat steps 2-3 for finalist 2.
5. Pool (subtract/concatenate) both performance encodings.
6. Feed the pooled performance encodings into our models to obtain a prediction.

This pipeline can be described by the following graph.



After processing each tournament from the original dataset into a single vector according to the pipeline described above, we performed experiments with the same models. The results are as follows:

	LR	LDA	QDA	NB	FFN
Accuracy	0.548	0.519	0.452	0.509	0.54
Precision	0.56/0.54	0.53/0.51	0.45	0.52/0.5	0.54
Recall	0.57/0.54	0.51/0.53	0.38/0.53	0.53/0.49	0.64/0.43
F1	0.56/0.53	0.53	0.41/0.49	0.52/0.5	0.59/0.48

Our experimental results indicate that the Logistic Regression and the Feed-Forward network achieved the best performance among the models. However, the results are not satisfactory. Given a baseline majority-vote approach of 0.51, these results are only slightly better than guessing. These results indicate that the statistics of the 4 elimination matches from both finalists are not sufficient to predict the outcome of a final's match without further information.

5 Predicting Match Outcome Based on Player Historical Performance

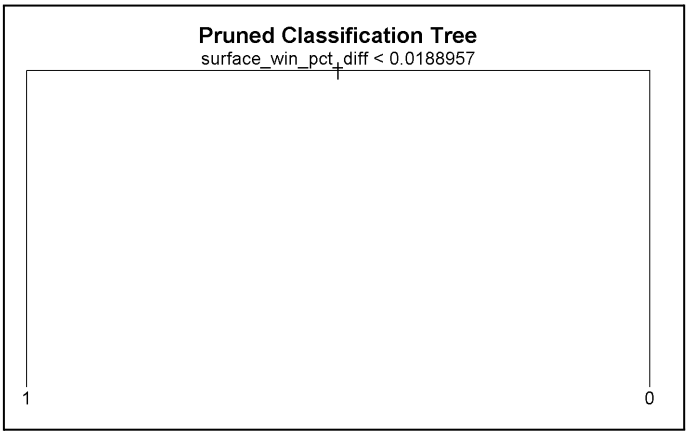
5.1 Methodology

Our second objective was to predict the outcome of a match based on the historical performance of the players. We realized that even though having match data such as first serve won difference, break points faced difference, ace, and other match data variables can improve the accuracy of predicting the winner of the match, the match data variables are not available until the end of the match, when the winner would already be known. We want to be able to predict the winner of a match before the match starts, which means we need to focus on variables that are known before the match, such as age, height, surface, tourney level, height, and ranking. We also considered new variables that we created, such as win percentage, surface win percentage, tourney level win percentage, and head-to-head records.

5.2 Experimental Results and Analysis

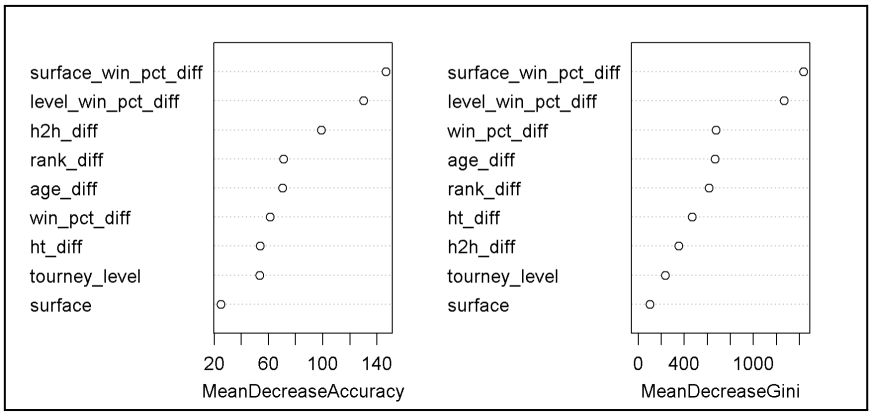
We noticed that 66.13% of the matches were won by the higher-ranked player, which can serve as a good baseline. We evaluated the performance of our models using the precision, recall, and F1-score metrics. The results are as follows:

Classification Tree: (Accuracy 0.7014)



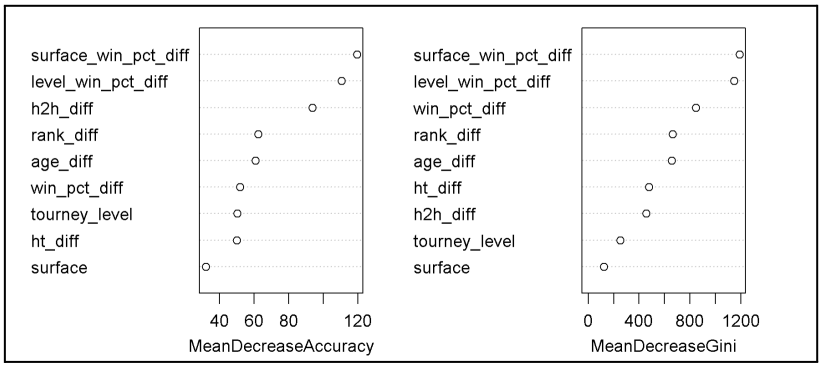
The pruned decision tree is very simple with the only criteria for final classification is surface win percentage. If player1 has a surface win percentage 0.0188957 higher than player 2, player 1 would win the game, if not, player 2 wins the game. It seems that the only useful variable is the surface win percentage if you only use variables available before the match.

Bagging: (Accuracy 0.7597)



The bagging model includes 500 trees, and at each split in the tree, a random subset of 9 predictor variables was considered. The most important variables are surface win percentage difference and level win percentage difference.

Random Forest: (Accuracy 0.7639)



The random forest model includes 500 trees, and at each split in the tree, a random subset of 3 predictor variables was considered. The most important variables are surface win percentage difference and level win percentage difference, which is the same as the bagging model. The accuracy slightly improves by not considering all predictors for each split of the tree.

Logistic Regression: (Accuracy 0.7424)

Our experimental results show that Bagging and Random Forest achieved the highest accuracy among all models, with accuracies of 0.7597 and 0.7639, respectively, which is higher than the baseline 66.13%. These results indicate that it is possible to predict match outcomes based on a player's historical performance to some extent. However, the overall accuracy is still not very high, which suggests that there is room for improvement in our prediction models.

Our analysis demonstrates that predicting match outcomes based on match performance (Objective 1 Question 1) yields better results compared to predictions based on

player historical performance (Objective 2), which performs better than tournament performance(Objective 1 Question 2). This suggests that the player's match performance is a more reliable indicator of a player's chances of winning a match than their overall historical performance.

The models for both objectives still have room for improvement. Future work could involve incorporating more features, such as player fatigue, mental strength, and other contextual factors that might influence match outcomes. Exploring more advanced machine learning algorithms or fine-tuning the existing models might lead to better prediction accuracy.

6 Future Directions

In light of our research findings, there are several potential directions for future work in this area. One promising avenue is to acquire more extensive and higher-quality datasets. Access to larger and more diverse data can help improve the performance of our models and enable more nuanced analyses. Incorporating additional variables related to player performances, such as average serve speed, RPM, physical conditioning, mental strength, and on-court strategies, may enhance the prediction accuracy of our models. Some of the new variables we can create for a player's historical performance include the player's breakpoint faced percentage and first serve won percentage, since we saw from above analysis that these two are the more important match statistics variables.

Another direction for future research is to explore other machine learning techniques that have not been applied in this study. By testing a wider range of models and algorithms, we may discover new approaches that yield better results for predicting tennis match outcomes. Additionally, it would be interesting to investigate the potential benefits of combining both tournament performance and player historical performance information in our models. By integrating these two aspects, we might be able to create more comprehensive and accurate predictions and account for a player's recent form in the tournament. Our models might thus become more accurate and robust in predicting tennis match results.

7 Conclusion

In this study, we have presented a comprehensive analysis of tennis match outcome prediction using statistical models based on tournament performance and player historical performance. Our project has demonstrated the potential of these approaches in providing accurate predictions of match outcomes, with some models achieving high levels of accuracy.

Our first objective focused on predicting tournament final outcomes based on tournament performance, ignoring player historical/career statistics. We explored several machine learning models, including logistic regression, LDA, FFN, Naive-Bayes, QDA, classification trees, bagging, and random forests. This objective involves two questions. Firstly, we want to ensure that given any matches' statistics, a model can accurately predict the winner of the match. Secondly, given the elimination matches for each finalist (around 4 each, total 8 matches), and after combining these 4 matches into a "finalist performance vector," would a model be able to predict the final's match winner using these two vectors. For the first question, most of these models demonstrated impressive accuracy, with logistic regression and LDA achieving an accuracy of 0.97. However, for our second question, the

accuracy was not satisfactory. The best model, logistic regression, achieves an accuracy of only 0.54, slightly better than the baseline majority-vote approach of 0.51.

Our second objective was to predict match outcomes based on the player's historical performance. We employed a similar set of models, with random forests achieving the highest accuracy of 0.7639. We were able to achieve a high level of accuracy with a relatively primitive data set. This result highlights the potential of using player historical performance data for predicting match outcomes and the potential improvement with a more sophisticated and accurate data set.

We also investigated the potential future directions for this research. These directions include acquiring more extensive and higher quality datasets, introducing new variables for player performance, exploring other machine learning techniques, and combining both tournament performance and player historical performance information in our models to predict the winner of the tournament final. Pursuing these future directions could further enhance the prediction accuracy and robustness of our models.

Our project has provided insights into the prediction of tennis match outcomes using statistical models based on tournament performance and player historical performance. While there is still room for improvement, our findings demonstrate the potential of these approaches to contribute to the broader field of sports analytics. We believe that by continuing to explore these ideas and pushing the boundaries of our models, we can make significant progress in understanding and predicting tennis match outcomes, benefiting players, coaches, and analysts alike.

8 References

1. Sackmann, J. (n.d.). Tennis WTA: Professional women's tennis match results, stats, and more. GitHub. Retrieved from https://github.com/JeffSackmann/tennis_wta

2. Sackmann, J. (n.d.). Tennis ATP: Professional men's tennis match results, stats, and more. GitHub. Retrieved from https://github.com/JeffSackmann/tennis_atp

3. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer-Verlag New York.

4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer Series in Statistics.