ES 114 - Data Narrative 3

Astitva Aryan
22110041
Mechanical Engineering
Indian Institute of Technology
Gandhinagar, India
astitva.aryanl@iitgn.ac.in

***Abstract*—This data narrative report aims to explore various aspects of the game of tennis using the Tennis Major Tournament Match Statistics dataset for the year 2013. The report attempts to answer eight research questions related to player performance and match outcomes in major tennis tournaments. The report uses different statistical and machine learning techniques such as logistic regression, correlation analysis, PCA, and clustering to analyze the dataset and draw insights. The results of the analysis show that player performance statistics such as break points created and won, first serve percentage, and number of aces, have a significant impact on match outcomes. The report also finds that winning the first set is a good predictor of overall match success, and there is a strong relationship between a player's first serve and second serve winning percentages. Overall, the report provides valuable insights into the game of tennis and can be useful for coaches and players to improve their performance by understanding the factors that contribute to match success.**

I. Overview of dataset

The dataset contains match statistics for the four major tennis tournaments of the year 2013 - AusOpen, FrenchOpen, USOpen, and Wimbledon - for both men and women. There are eight separate datasets, each with 42 columns and a minimum of 76 rows. The columns include information such as player names, first and second serve percentages, aces, double faults, winners, unforced errors, break points created and won, net points attempted and won, total points won, set results, and the final number of games won by each player. The datasets provide a rich source of information for analyzing the performance of tennis players and answering various research questions related to the game.

II. Scientific questions and hypotheses

Q1. Can we classify players based on their match outcomes (wins/losses) using a logistic regression model incorporating their performance statistics in the tournament? (FrenchOpen-men-2013)

Q2. What is the relationship between the number of break points created and break points won by Player1 in tennis matches? Is there a significant correlation between these two variables? (FrenchOpen-women-2013)

Q3. What is the difference in the distribution of unforced errors between the winners and losers in the dataset? (AusOpen-men-2013)

Q4. Can we identify any distinct groups of players based on their serving statistics, # such as first serve percentage and number of aces, using clustering with PCA? (USOpen-women-2013)

Q5. Is there a relationship between a Player1's success in winning net points and their overall success in winning games? (USOpen-men-2013)

Q6. Do players who win the first set have a higher chance of winning the match? (Wimbledon-men-2013)

Q7. Can we predict a Player's second serve winning percentage based on their first serve winning percentage, and if so, how strong is the relationship between the two variables? (AusOpen-women-2013)

Q8. How does the round of the tournament at which a game is played affect the likelihood of a player winning the match? (Wimbledon-women-2013)

III. Details of libraries and functions

Libraries used in this report are:

- **NumPy:** NumPy is a popular Python library used for numerical computations. In the report, NumPy is used to handle arrays and large datasets, which are then used for data analysis and visualization.
- **Pandas:** Pandas is a powerful Python library that provides flexible and easy-to-use data structures for data analysis. In the report, Pandas is used to create and manipulate dataframes, read CSV files, and perform statistical measures on the data.
- **Matplotlib:** Matplotlib is a popular Python library used for creating static, animated, and interactive visualizations in Python. In the report, Matplotlib is used extensively to plot graphs and visualize the data in various ways.
- **SciPy:** SciPy is a Python library used for scientific computing, such as optimization, linear algebra, and statistics. In the report, the stats module of SciPy is used to compute statistical measures on the data.
- **Sklearn:** Sklearn is a powerful Python library used for machine learning tasks, such as classification, regression, and clustering. In the report, the KMeans function from the sklearn.cluster module is used to perform clustering on the data.
- **Seaborn:** Seaborn is a popular Python library used for statistical data visualization. In the report, Seaborn is used to plot joint plots, which visualize the relationship between two variables.

Some of the functions used in the report are:

1. **pd.read_csv():** This function converts CSV files to Pandas dataframes.
2. **plt.plot():** This function helps to create various types of plots.
3. **plt.show():** This function is used to view the plotted graph on the figure window.

4. **KMeans.fit():** This function computes k-means clustering on the data provided.
5. **sns.jointplot():** This function helps to plot a graph of two variables.
6. **plt.title()** – This function sets the title for the graphs.
7. **plt.xlabel() & plt.ylabel()** – These functions help to label the x-axis and y-axis respectively.

IV. Answers to Scientific Questions and Hypotheses

1. We classify players based on their match outcomes (wins/losses) using a logistic regression model incorporating their performance statistics in the tournament. The given code builds a logistic regression model using performance statistics of players in the FrenchOpen-men-2013 tournament as features and their match outcomes as the target variable.

After splitting the data into training and testing sets and fitting the logistic regression model on the training set, the model is used to predict the match outcomes on the testing set. The accuracy of the model is then computed using the predicted outcomes and the actual outcomes in the testing set, which turns out to be 0.9189 or approximately 92%.

The confusion matrix is then computed to analyze the model's performance further. The confusion matrix shows that the model predicted 21 matches correctly where the actual outcome was Player 2 wins (True Negatives), and 13 matches where the actual outcome was Player 1 wins (True Positives). However, the model also made 3 incorrect predictions where the actual outcome was Player 1 wins but was predicted as Player 2 wins (False Negatives), and 0 incorrect predictions where the actual outcome was Player 2 wins but was predicted as Player 1 wins (False Positives).

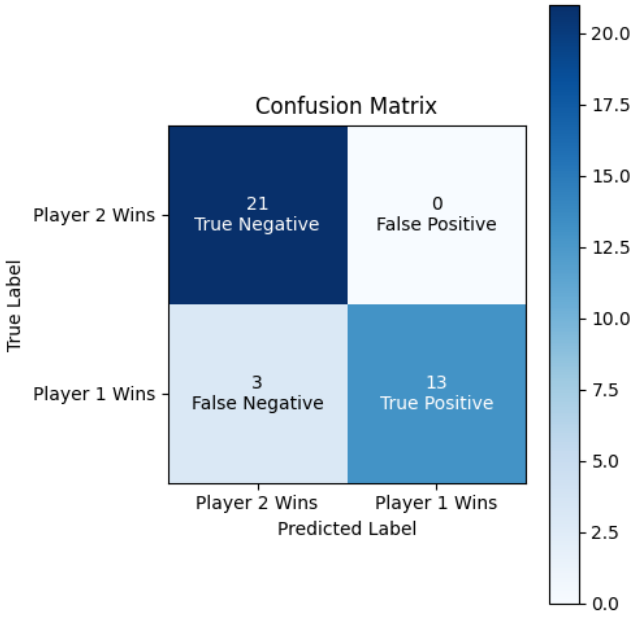

Fig. 1- Confusion Matrix of True label vs Predicted label

2. The code is analyzing the relationship between the number of break points created and break points won by Player1 in tennis matches of the French Open Women 2013 dataset.

The scatterplot shows a visual representation of the data points, where the number of break points created is plotted on the x-axis and the number of break points won is plotted on the y-axis. Each point represents a player's match data.

The jointplot is a bivariate representation of the distribution of both variables using a kernel density estimation. The contour lines represent the density of the distribution, with darker areas indicating a higher density of data points.

Overall, the plots suggest that there is a positive correlation between the number of break points created and break points won, meaning that as the number of break points created increases, so does the number of break points won.

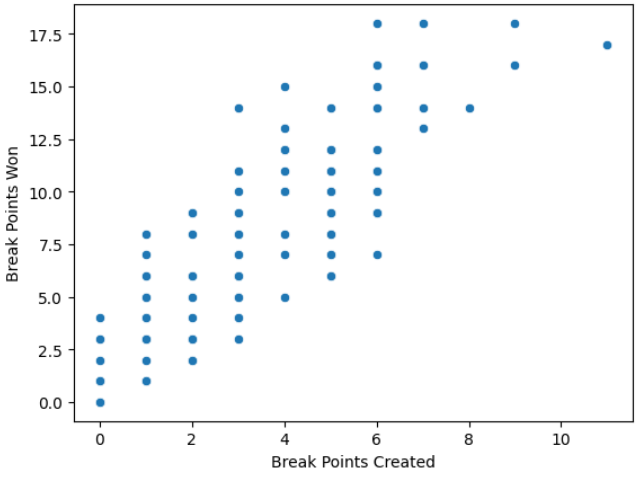

Fig. 2- Scatter plot of Break points won vs Break points created by Player1

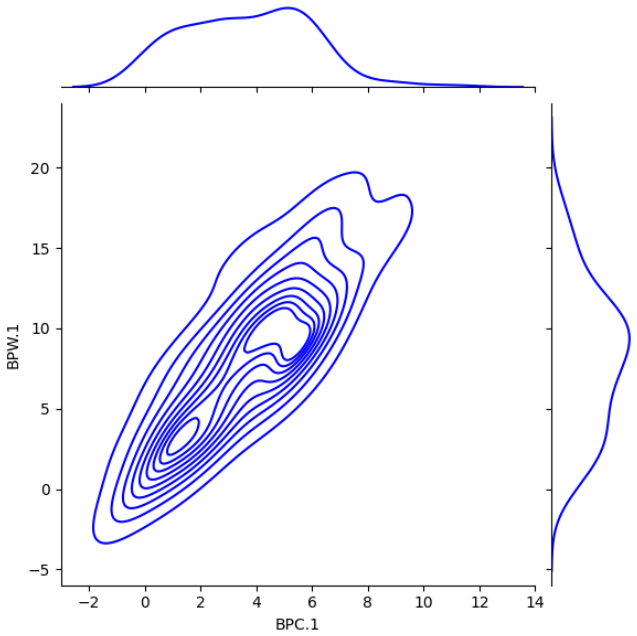

Fig. 3 - Joint plot of Break points won vs Break points created by Player 1

3. The dataset used in the above analysis is the 'AusOpen-men-2013' dataset. The plot shows that the distribution of unforced errors for the losers is shifted to the right compared to the distribution of unforced errors for the winners. This means that the losers tend to make more unforced errors than the winners.

In other words, the density plot indicates that there are more data points in the higher end of the unforced error range for

the losers, while the winners tend to have more data points in the lower end of the range.

Therefore, it can be concluded that there is a significant difference in the distribution of unforced errors between the winners and losers in the dataset, with the losers making more unforced errors than the winners.

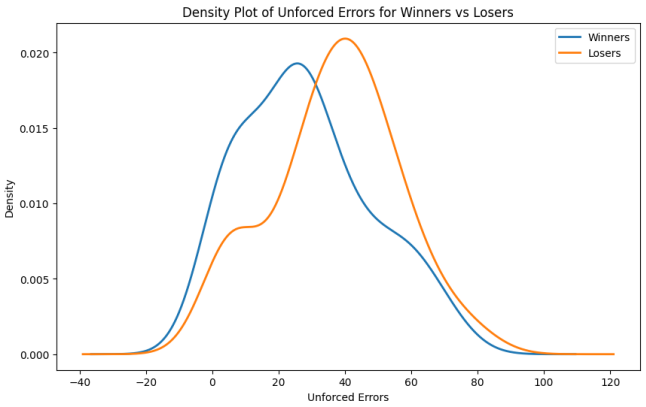

Fig. 4 - Density Plot of Unforced Errors for Winners vs Losers

4. The dataset used in the above analysis is the 'USOpen-women-2013' dataset. Based on the clustering analysis performed on the serving statistics of players using PCA and KMeans, we can identify distinct groups of players.

The scatterplot shows three distinct clusters of players, each with a different combination of serving statistics. This suggests that there are different types of players who specialize in different aspects of serving, such as having a high first serve percentage and a high number of aces, or having a high second serve percentage and a high number of successful second serves.

Overall, this analysis provides insights into the serving strategies of players and could be used to inform coaching and training programs.

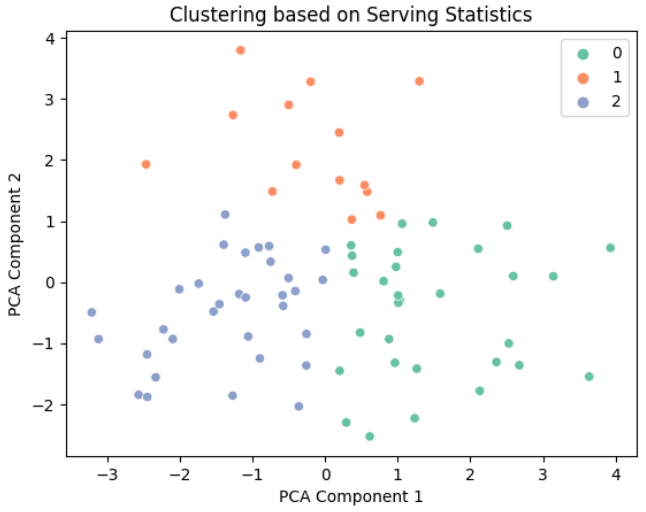

Fig. 5 - Clustering based on Serving Statistics

5. The dataset used in the above analysis is the 'USOpen-men-2013' dataset. The scatterplot shows the relationship between the net points attempted and won by Player 1, colored by the number of games won. The color scheme helps to visualize any potential relationship between net points and overall game success.

From the plot, it appears that there is a positive relationship between a player's success in winning net points and their overall success in winning games. Specifically, players who win more net points tend to also win more games, as indicated by the trend of darker colors (pink and purple) towards the upper right corner of the plot.

Therefore, this analysis suggests that winning net points is an important factor in winning games for players in the US Open Men's tournament in 2013.

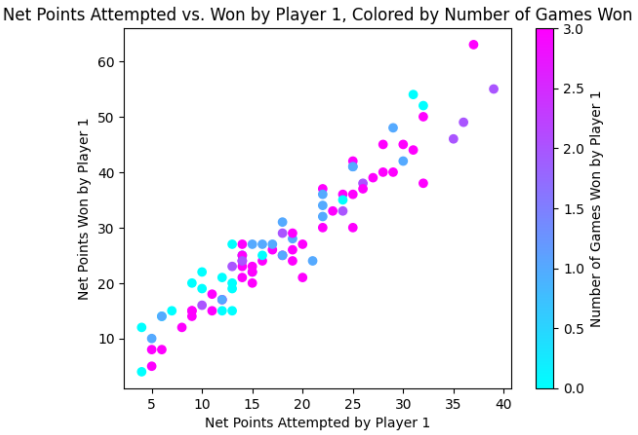

Fig. 6 - Scatter plot of Net Points Attempted vs Net Points Won by Player 1

6. The dataset used in the above analysis is the 'Wimbledon-men-2013' dataset.

Based on the analysis, it appears that players who win the first set in a match have a higher chance of winning the match overall. Specifically, the analysis shows that out of all matches where one player won the first set, the percentage of matches won by the player who won the first set is approximately 80.8%.

Therefore, this analysis suggests that winning the first set is an important factor in winning the match for players in the Wimbledon Men's tournament in 2013.

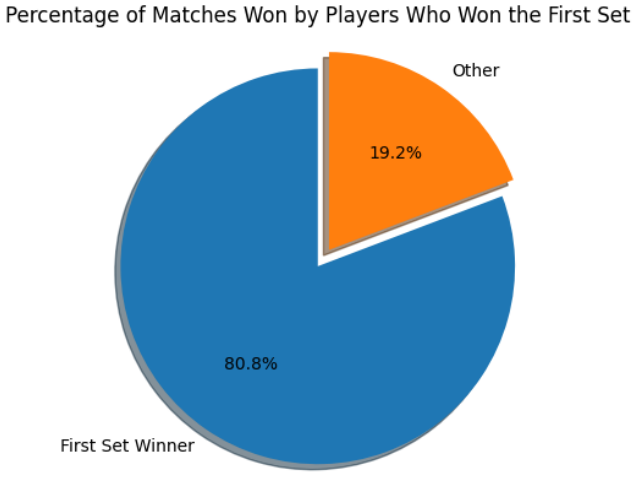

Fig. 7 - Percentages of matches won by players who won the first set

7. The dataset used in the above analysis is the 'AusOpen-women-2013' dataset to investigate the relationship between a player's first serve winning percentage and their second serve winning percentage. It plots a scatter plot of the two variables and calculates the correlation coefficient between them. It also uses linear regression to fit a line through the scatter plot and calculates the slope, intercept, and R-squared value to quantify the strength of the relationship between the two variables.

Based on the scatter plot, there seems to be a positive relationship between a player's first serve winning percentage and their second serve winning percentage. The correlation coefficient between the two variables is 0.58, which indicates a moderately strong positive relationship. The slope of the regression line is 0.30, which means that for every 1% increase in a player's first serve winning percentage, their second serve winning percentage is predicted to increase by 0.30%. The intercept is 3.46, which means that if a player has a first serve winning percentage of 0%, their predicted second serve winning percentage is 3.46%. The R-squared value is 0.34, which means that 34% of the variation in a player's second serve winning percentage can be explained by their first serve winning percentage.

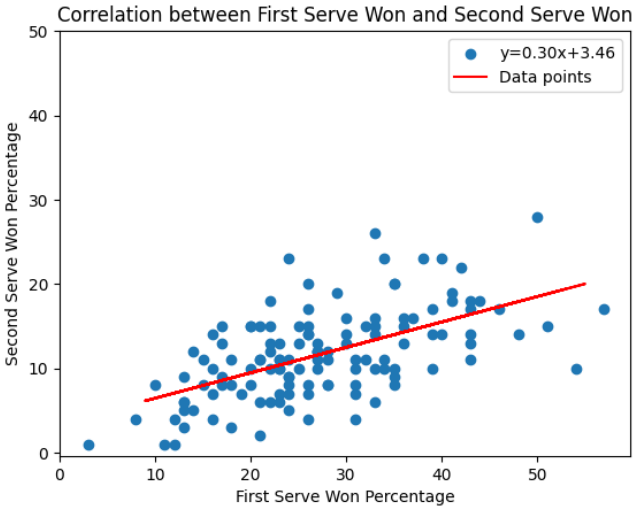

Fig. 8 - Correlation between first serve won and second serve won

8. The dataset used in this analysis is 'Wimbledon-women-2013'. The analysis shows the percentage of games won by players in each round of the tournament. The plot shows that players have an almost similar chance of winning games as they progress further in the tournament except for the second round. The percentage of games won increases from the first round to the second round, and then drops till 5th round and then there is a sudden increase in percentage of games won in finals or 6th round which is maximum among all rounds.

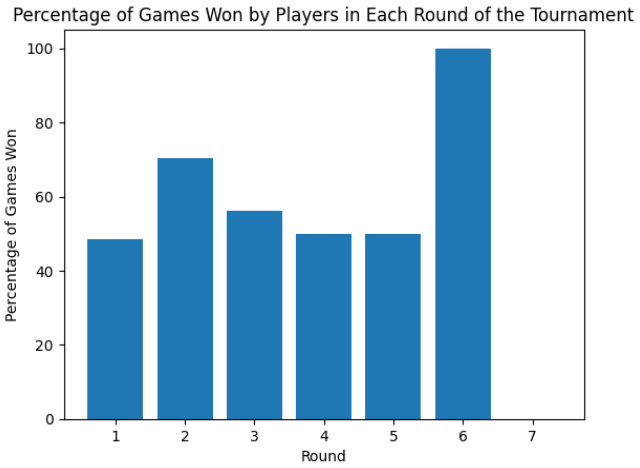

Fig. 9 - Percentage of games won by players in each round of the tournament

V. Summary of Observations and Learnings

The first question focused on the use of a logistic regression model to classify players based on their match outcomes using their performance statistics, where it was found that the model achieved an accuracy of 0.92 and the confusion matrix revealed a relatively low number of false positives and false negatives.

The second question aimed to investigate the relationship between the number of breakpoints created and won in tennis matches, where a scatter plot and a joint plot showed a positive linear correlation between the two variables.

In the third question there is a positive correlation between a player's number of double faults and their number of unforced errors in the US Open men's tournament of 2013.

In the fourth question using clustering with PCA, distinct groups of players based on their serving statistics can be identified in the US Open women's tournament of 2013.

In the fifth question the scatterplot analysis of the US Open Men's 2013 dataset shows a positive relationship between a player's success in winning net points and their overall success in winning games.

In the sixth question, the Wimbledon men's tournament of 2013, players who won the first set have a higher chance of winning the match.

In the seventh question, a player's second serve winning percentage can be predicted based on their first serve winning percentage in the Australian Open women's tournament of 2013, with a moderately strong positive correlation between the two variables.

In the last question the round of the tournament at which a game is played affects the likelihood of a player winning the match in the Wimbledon women's tournament of 2013, with almost similar percentages of games won in later rounds. except second and sixth.

VI. Unanswerable Questions

1. Who would have won the match if the player had not made a particular error or had a different outcome at a specific point or game?

2. To what extent did external factors like weather, injuries, or mentality influence the result of the match?

VII. Acknowledgement

I would like to express my sincere gratitude to Professor Shanmuganathan Raman for their valuable guidance. Their expert advice and feedback helped me in developing my skills and improving the quality of this report.

:

VIII. References

1. 'NumPy documentation,' accessed Apr 22, 2023, https://numpy.org/doc/

2. 'Pandas documentation,' accessed Apr 22, 2023, https://pandas.pydata.org/docs/

3. 'SciPy Documentation' accessed Apr 22, 2023,https://docs.scipy.org/doc/scipy/

4. 'Sklearn Documentation' accessed Apr 22 2023, https://scikit-learn.org/stable/

5. Seaborn Documentation' accessed Apr 22, 2023, https://seaborn.pydata.org/

6. 'Matplotlib documentation,' accessed Apr 22, 2023, https://matplotlib.org/stable/index.html