DSA-210 PROJECT REPORT

ALI YILMAZ
32281

PREDICTABILITY OF TENNIS METRICS

SUPERVISED BY

SELIM BALCISOY

Introduction

This project aims to determine whether the tennis metrics of match winners can be predicted. To explore this question, the report analyzes the predictability of five tennis metrics (e.g., Serve Point Won Percentage) over the years, considering factors such as surface type and tournament level. ATP Tour match data from 2000 to 2024 has been sourced from Kaggle, and five new metrics have been created by manipulating existing ones.

The project begins with a data analysis and enrichment phase, where the five metrics are defined and their behaviors over the years are explored using visualization techniques. Next, the predictability of these metrics is evaluated through hypothesis testing using ANOVA. Finally, various machine learning techniques are applied to assess prediction accuracy, and the best-performing model is used to generate sample predictions, testing the findings from the hypothesis analysis.

Metrices of the report

-Serve point won percentage

-Ace per game

-Break point save percentage

-Return point won percentage

-Break point conversion rate

Hypotheses

Null Hypothesis:

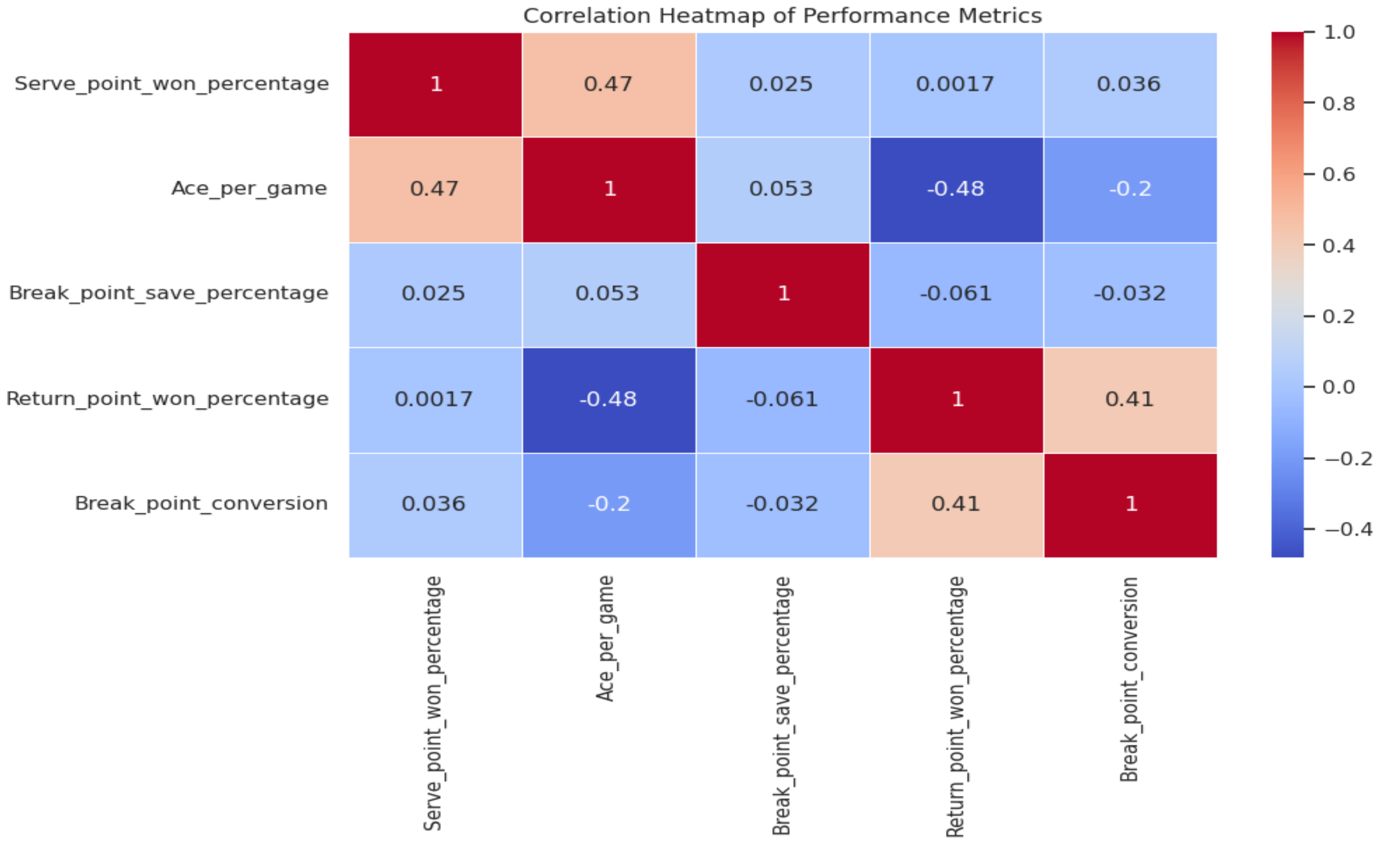
The performance metrics of ATP Tour tournament winners remain consistent over time, allowing for the prediction of future winners' match statistics based on historical trends.

Alternative Hypothesis:

The performance metrics of ATP Tour tournament winners fluctuate unpredictably over time, making it impossible to reliably estimate future winners' match statistics.

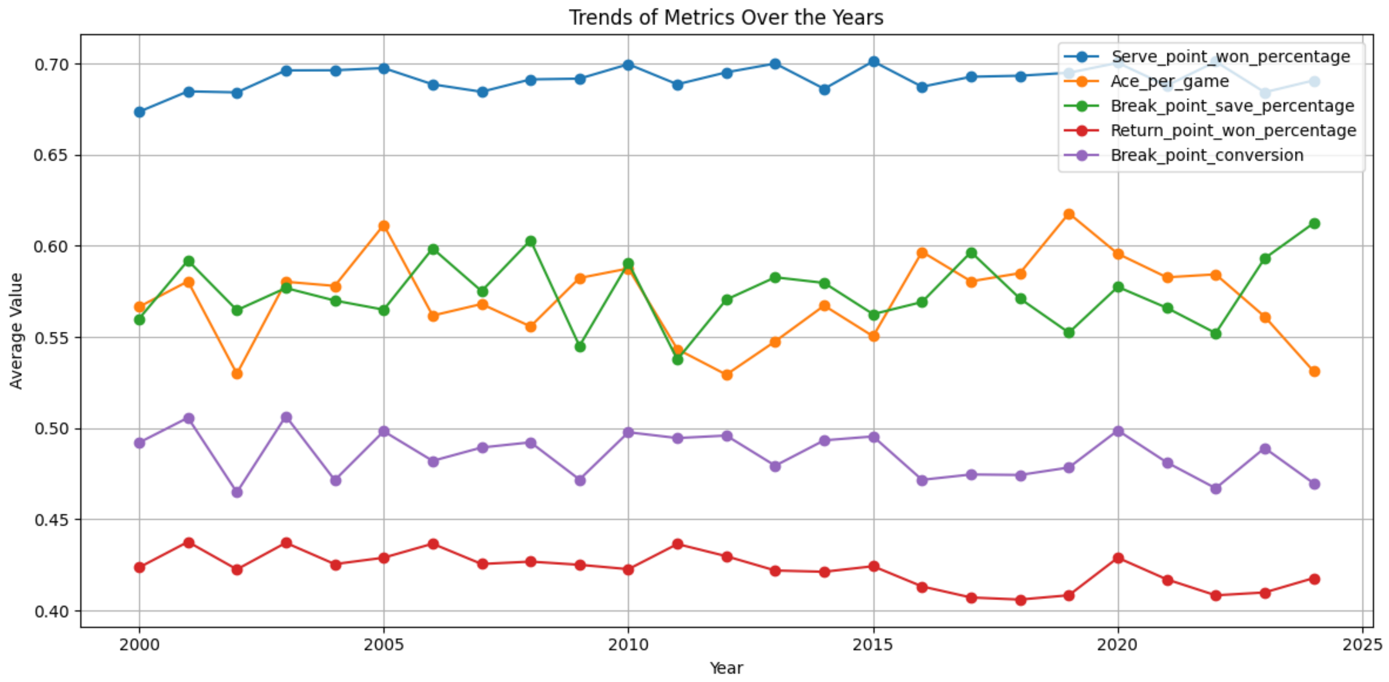
Data Analysis and Enrichment

To identify the most important metrics in tennis matches and to better visualize them through graphs, I decided to create metrics that generally fall within the 0 to 1 range. Four of the metrics are directly expressed as percentages, so they naturally lie within this interval. Additionally, the "Aces per Game" metric is almost always between 0 and 1 when averaged across a tournament. To represent the overall performance in a match, I combined the Serve Point Won and Return Point Won percentages. To capture performance in critical moments, I used three other metrics, such as Break Point Conversion Rate. After defining these metrics, I examined the relationships between them to gain deeper insights into their interactions.



From this heatmap I have understood that there is no strong correlation between metrices they are showing different parts of the game which is in coherent with my purpose while creating them.

Then I have looked their changes through the years:



Some of the metrices like Serve point won and Return point won percentages look like they stay consistent through years, but other 3 are fluctuating at some moments, so that has shown me they can result in alternative hypothesis.

Hypothesis Testing

For hypothesis testing, I conducted ANOVA in three different ways. First, I applied it directly to the five metrics across all years. Second, I performed the test separately for different surfaces, as surface type is a major factor in tennis that significantly influences both metrics and match outcomes. Third, I conducted ANOVA based on tournament levels, since major tournaments are often dominated by a few players—such as Nadal, Federer, and Djokovic—who have consistently won throughout the 21st century. In contrast, smaller tournaments tend to have more variability in winners. To prevent bias and ensure fair comparison, I accounted for this distinction as well.

In total, I obtained 45 different ANOVA results: 5 from the general test across all years, 20 from surface-based tests (across 4 surfaces: Clay, Hard, Grass, and Carpet), and 20 from tournament-level tests (across 4 categories: Grand Slam, ATP Finals, Masters, and ATP 250/500). These tests allowed me to rigorously evaluate my hypotheses regarding the predictability of the tennis metrics. Critical value of p is used as 0.05 for these tests.

Result of Hypothesis Testing

In 45 results 42 of them Fail to reject Null Hypothesis and only 3 of them Reject the Null Hypothesis. Rejecting ones are:

-On Clay courts Return Point Won percentage has shown significant difference with p value of 0.0018

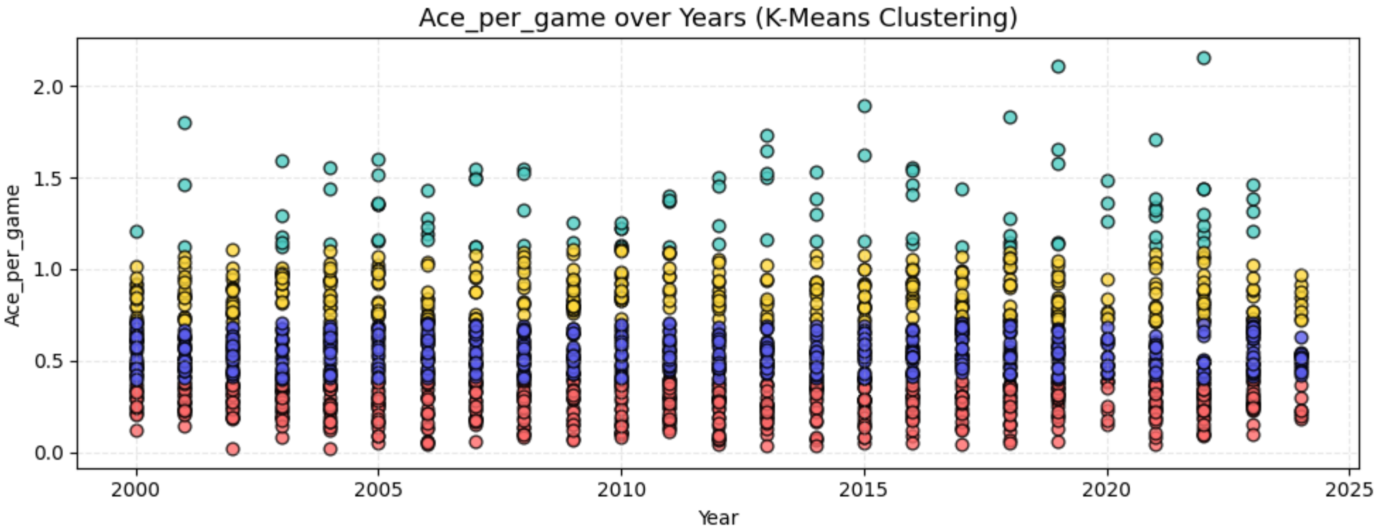
-On ATP 250/500 tournaments Return Point Won percentage has shown significant difference with p value of 0.0121

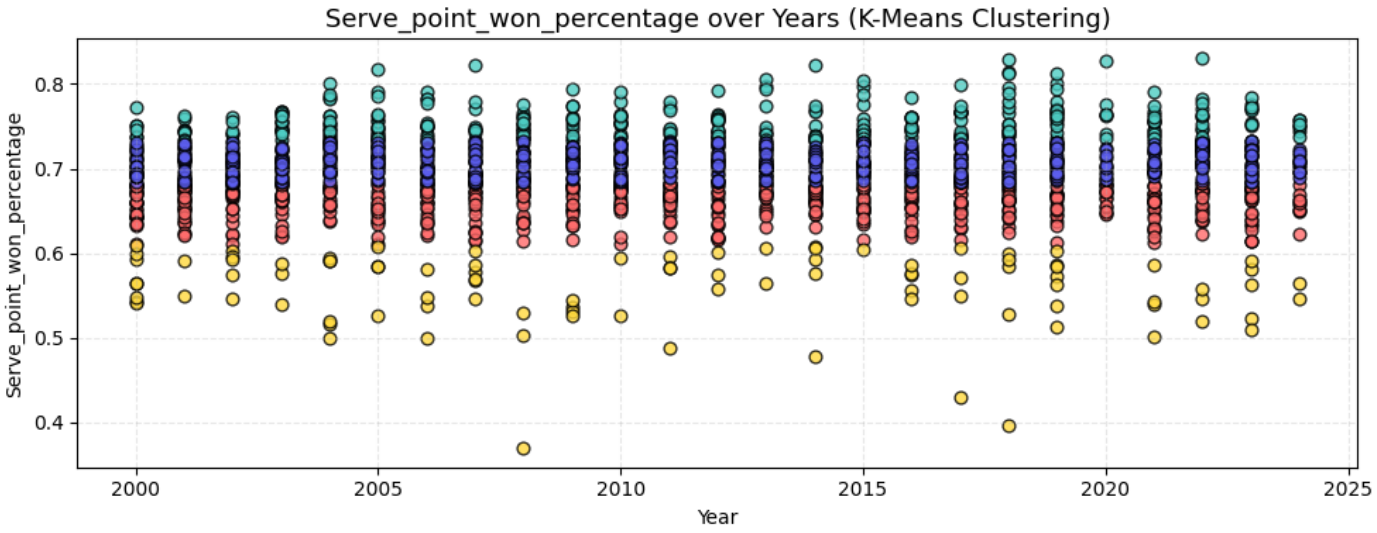
-On general test Return Point Won percentage has shown significant difference with p value of 0.0008

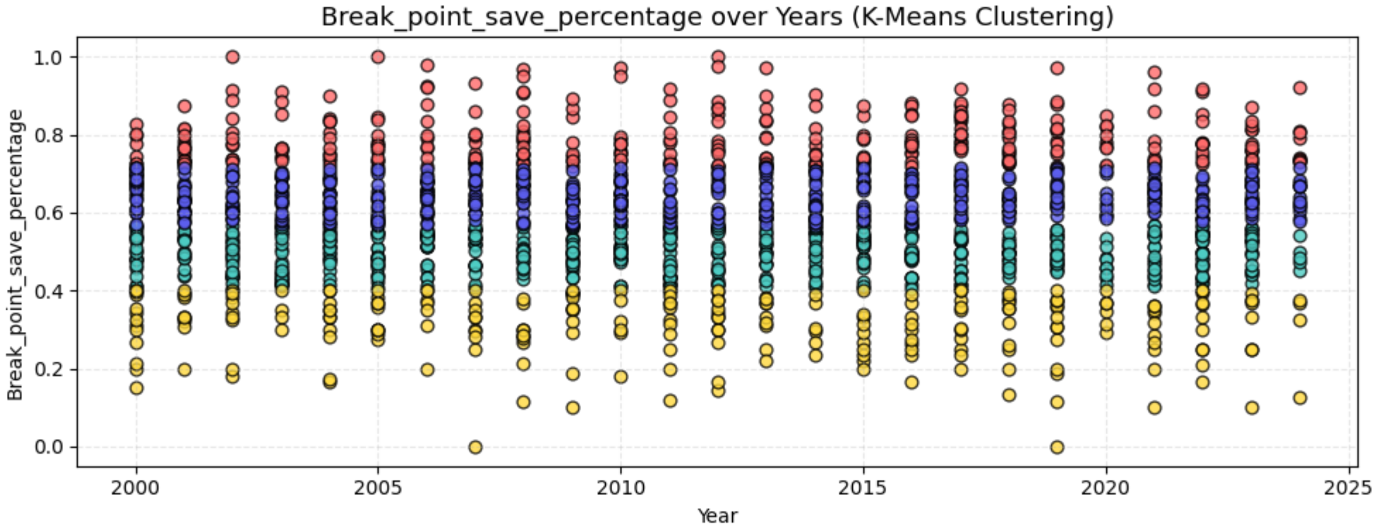
Return Point Won percentage is only metric, which is fluctuating, but even that metric did not fluctuate every time, so predicting metrices of the tournament winners in Tennis is possible and I have used some kind of Machine Learning methods to test it.

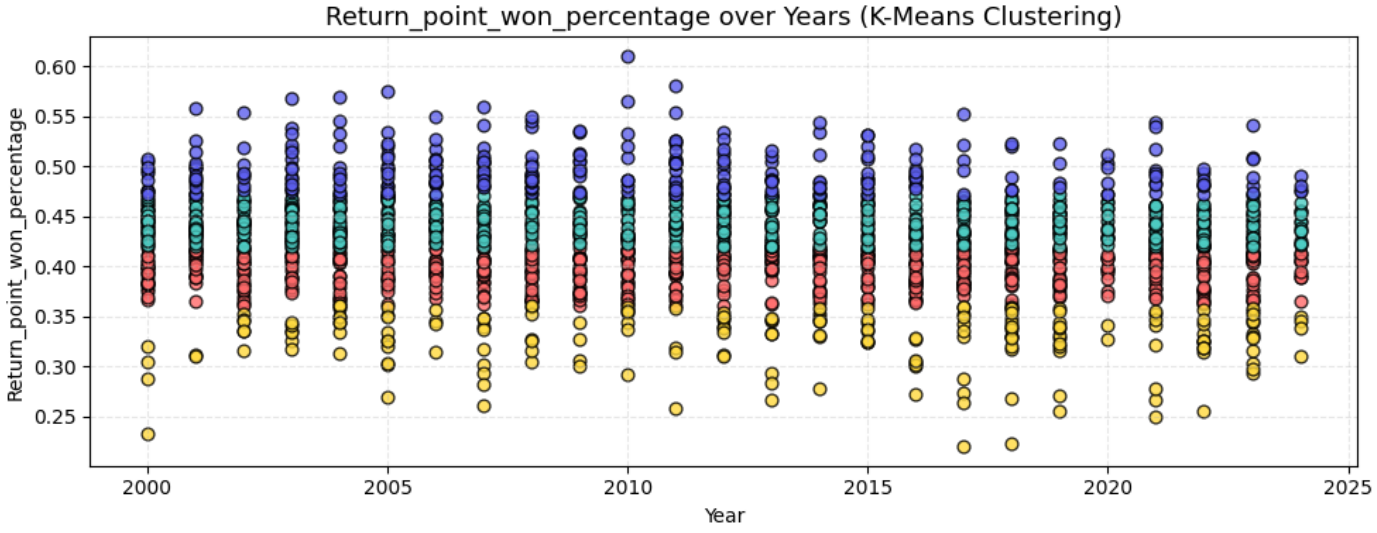
Unsupervised Method (K means Clustering)

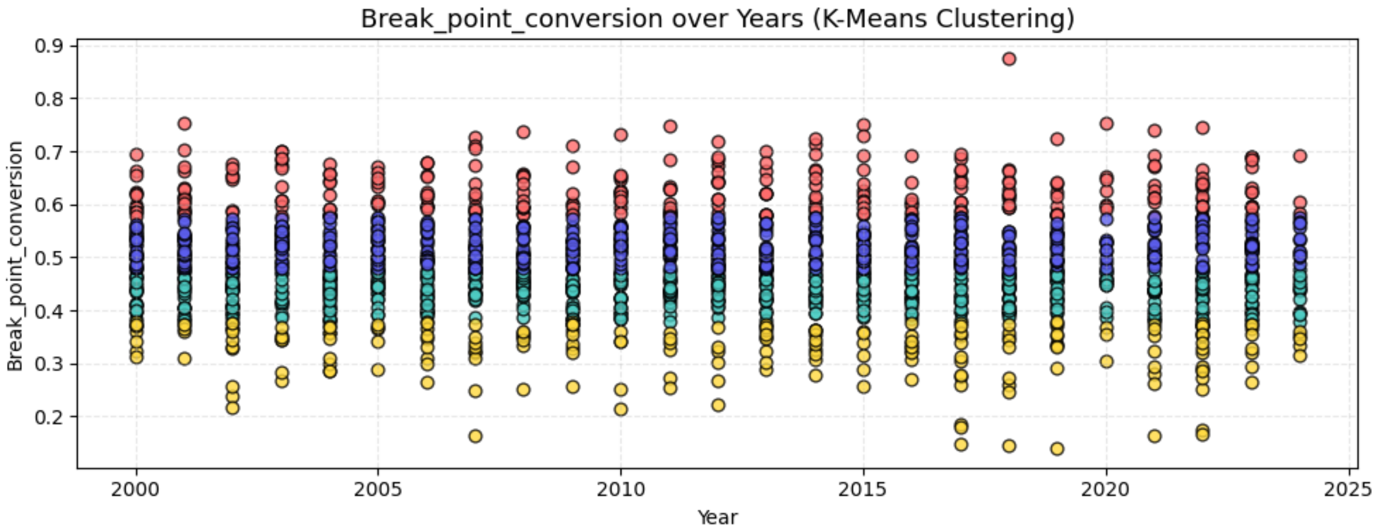
I first decided to use an unsupervised learning method to analyze which machine learning techniques might be most compatible with my data. For this purpose, I chose K-Means Clustering due to its simplicity and ease of implementation. I created four different clusters, corresponding to the four surface types and four tournament levels, to explore how the data naturally groups and whether these categories influence the patterns in the metrics.




Serve_point_won_percentage over Years (K-Means Clustering)
Serve_point_won_percentage
0.4
0.5
0.6
0.7
0.8
2000
2005
2010
2015
2020
2025
Year

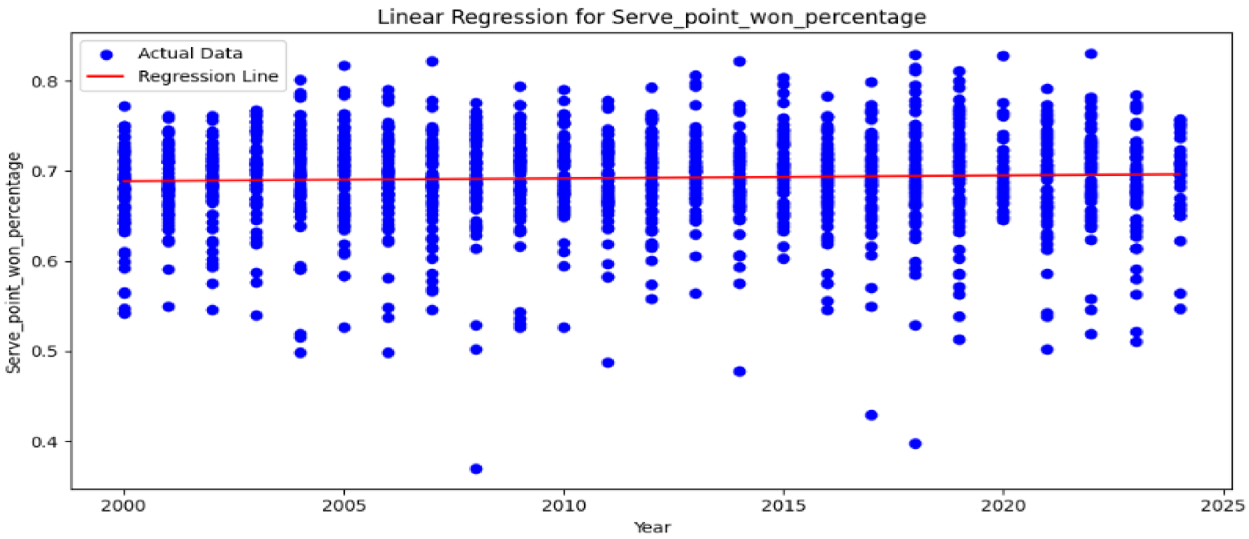

Break_point_save_percentage over Years (K-Means Clustering)
Break_point_save_percentage
0.0
0.2
0.4
0.6
0.8
1.0
2000
2005
2010
2015
2020
2025
Year

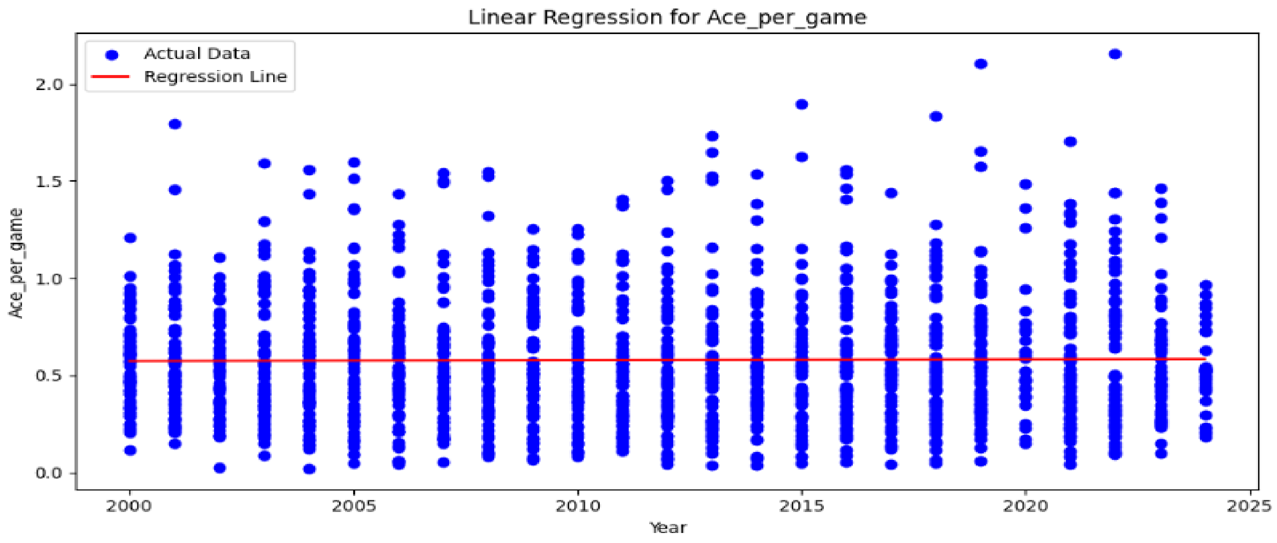

Return_point_won_percentage over Years (K-Means Clustering)
Return_point_won_percentage
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
2000
2005
2010
2015
2020
2025
Year

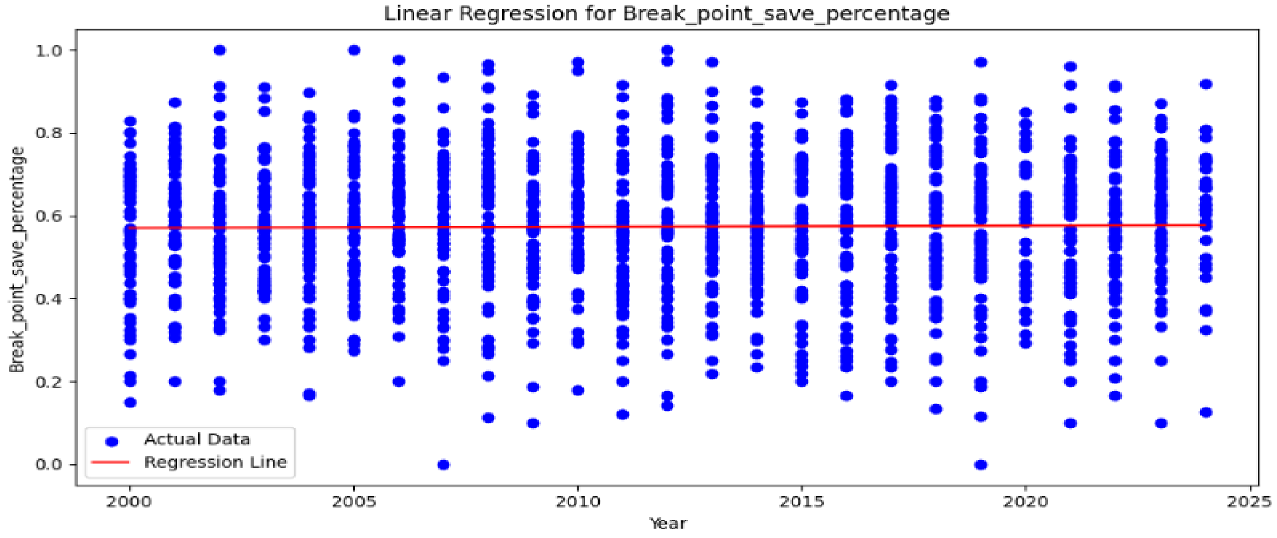


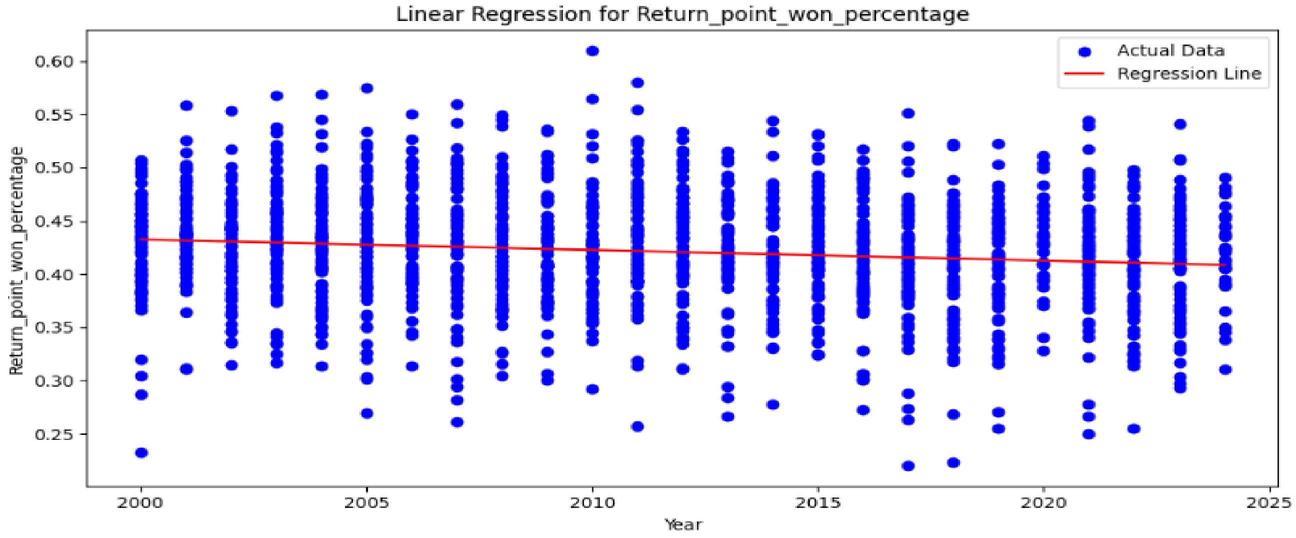
Based on the four different clusters, I observed a general sense of stability and consistency within the data. This led me to conclude that regression models could be suitable for predicting next year's metrics. To test this, I implemented three different regression models. I began with a simple Linear Regression model, then moved to a more complex approach using a Decision Tree with XGBoost. Finally, I applied a K-Nearest Neighbors (KNN) regression model. I compared the error rates of these models to determine which one performed best for the prediction task.

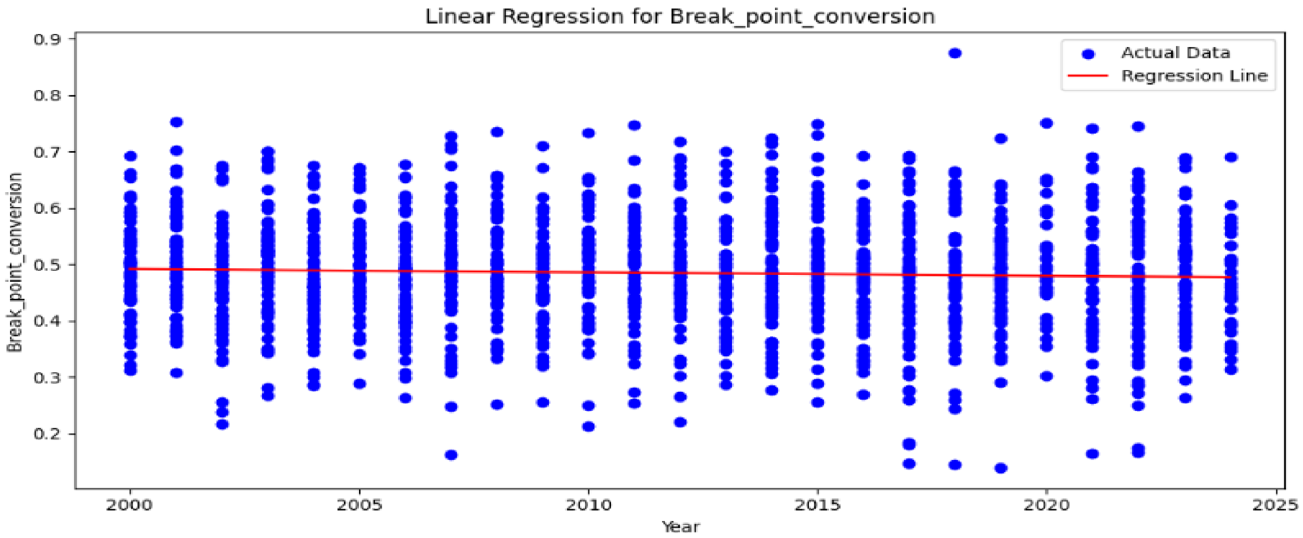
Linear Regression




Linear Regression for Ace_per_game
Actual Data
Regression Line
Ace_per_game
0.0
0.5
1.0
1.5
2.0
2000
2005
2010
2015
2020
2025
Year


Linear Regression for Break_point_save_percentage
Actual Data
Regression Line
Break_point_save_percentage
0.0
0.2
0.4
0.6
0.8
1.0
2000
2005
2010
2015
2020
2025
Year

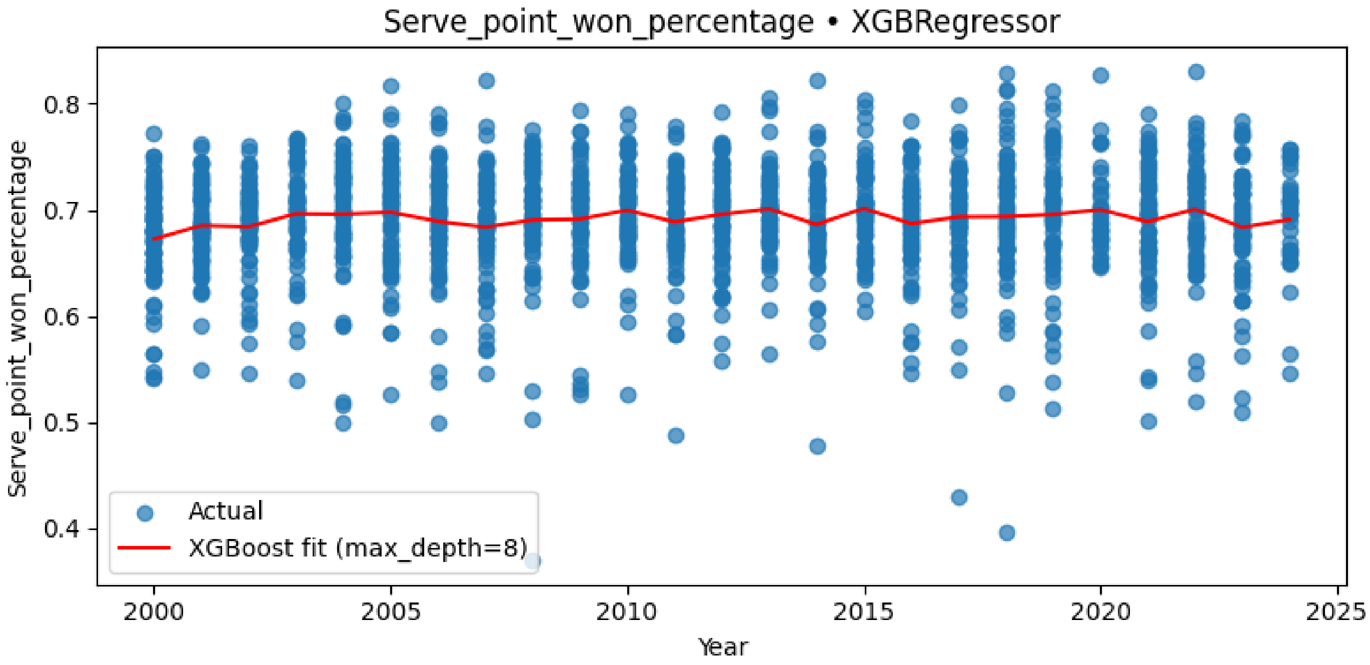

Linear Regression for Return_point_won_percentage
Actual Data
Regression Line
Return_point_won_percentage
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
2000
2005
2010
2015
2020
2025
Year

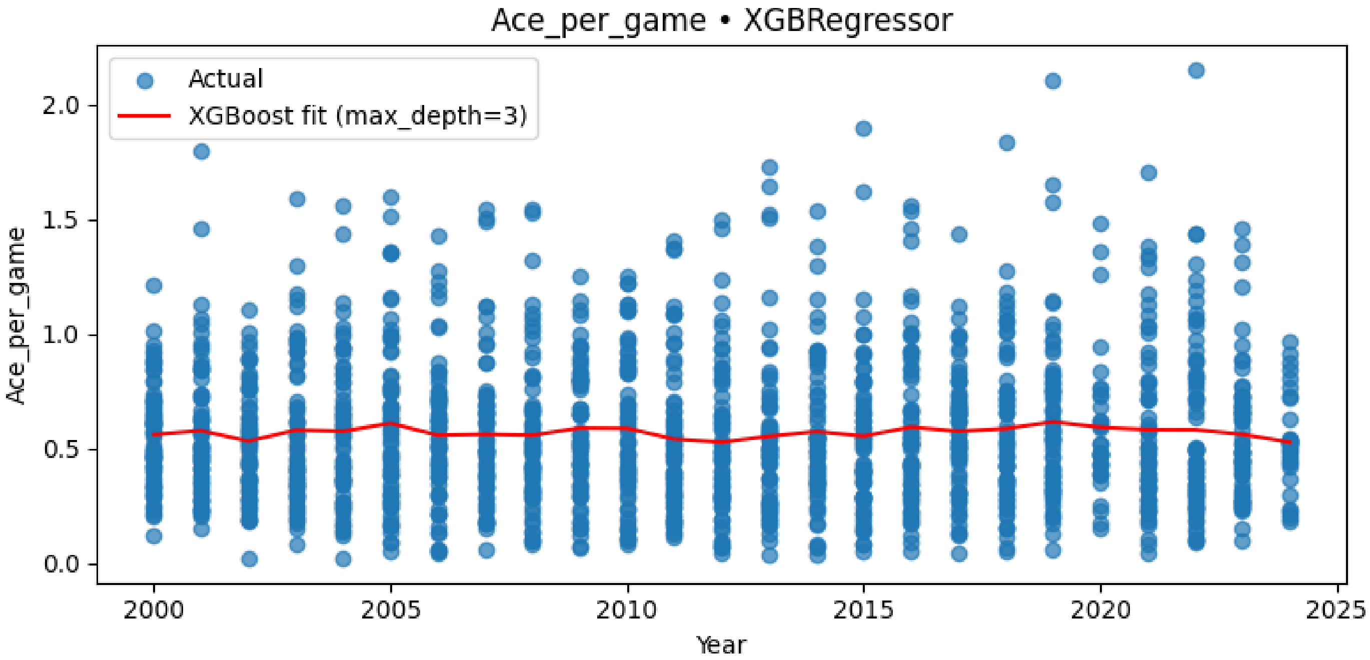


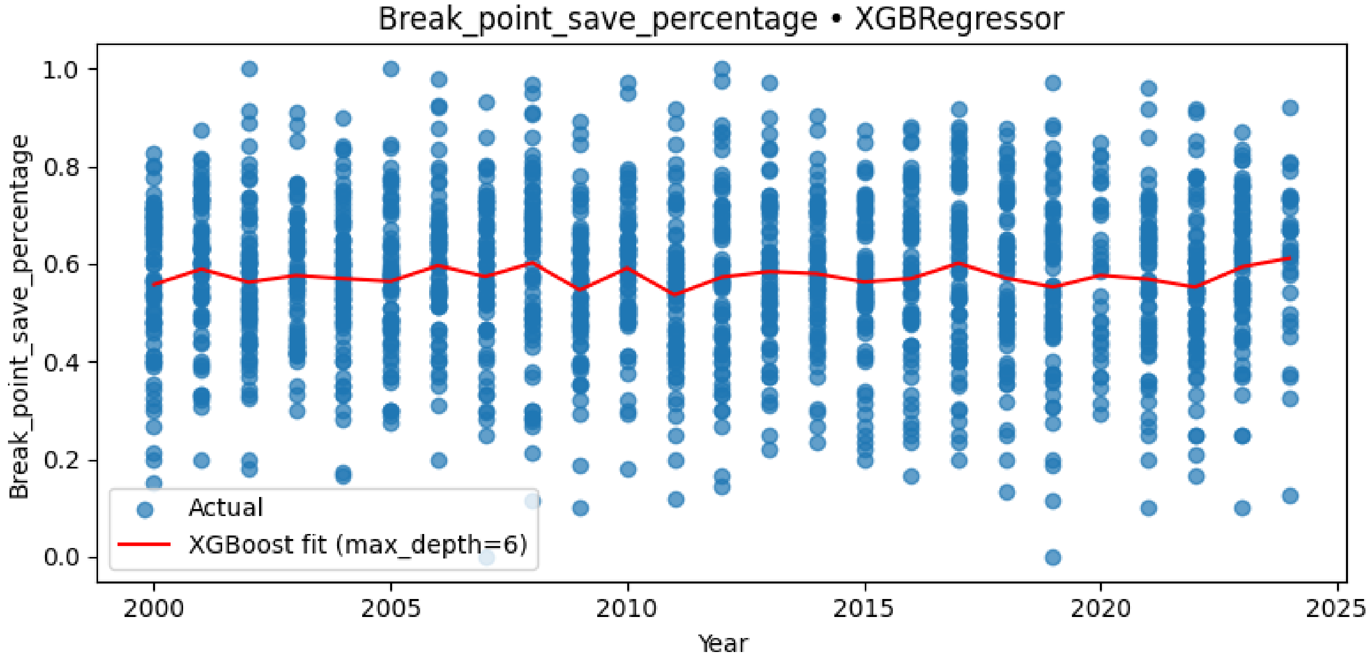
Serve_point_won_percentage - MSE: 0.0029, RMSE: 0.0537
Ace_per_game - MSE: 0.1079, RMSE: 0.3285
Break_point_save_percentage - MSE: 0.0261, RMSE: 0.1616
Return_point_won_percentage - MSE: 0.0032, RMSE: 0.0567
Break_point_conversion - MSE: 0.0091, RMSE: 0.0956

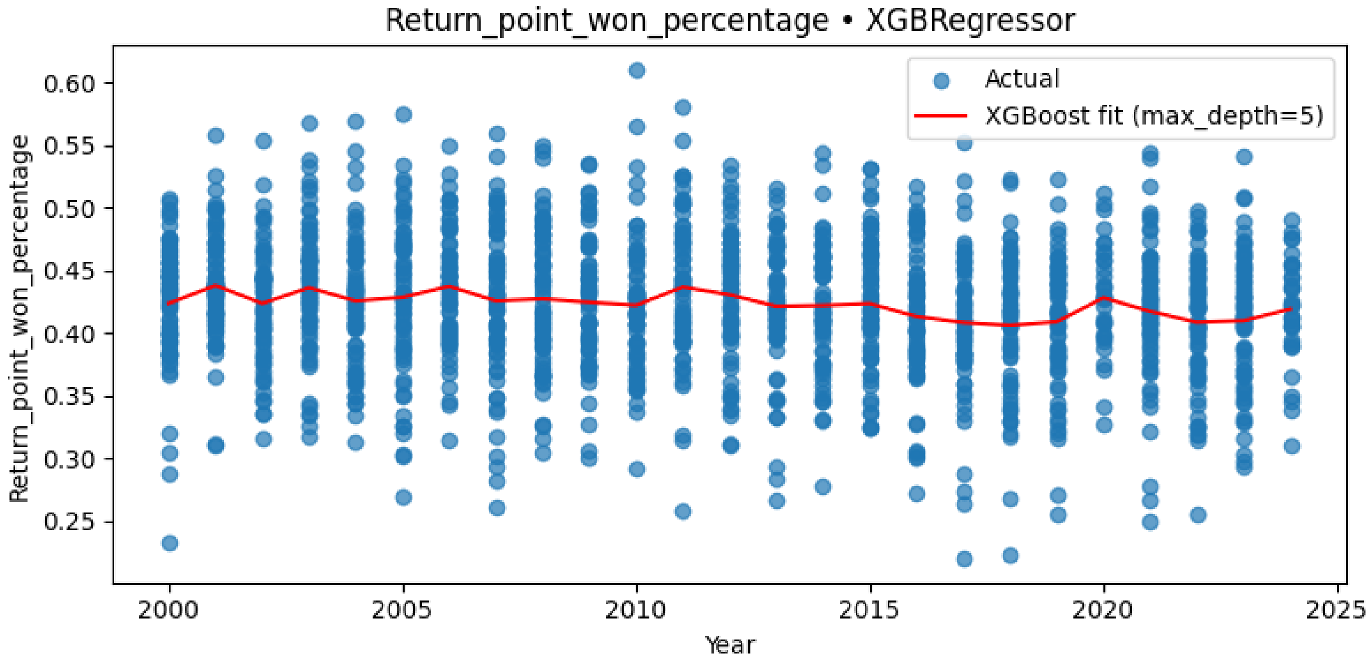
These are my error rates in Linear Regression Model I think it works very well with low error rate because to predict mean of the point usually it stays same and Linear Regression can draw a line without a little slope and it makes him good candidate to use. However, complex models could surpass it and predict some noises better than Linear Regression, so I tried other models.

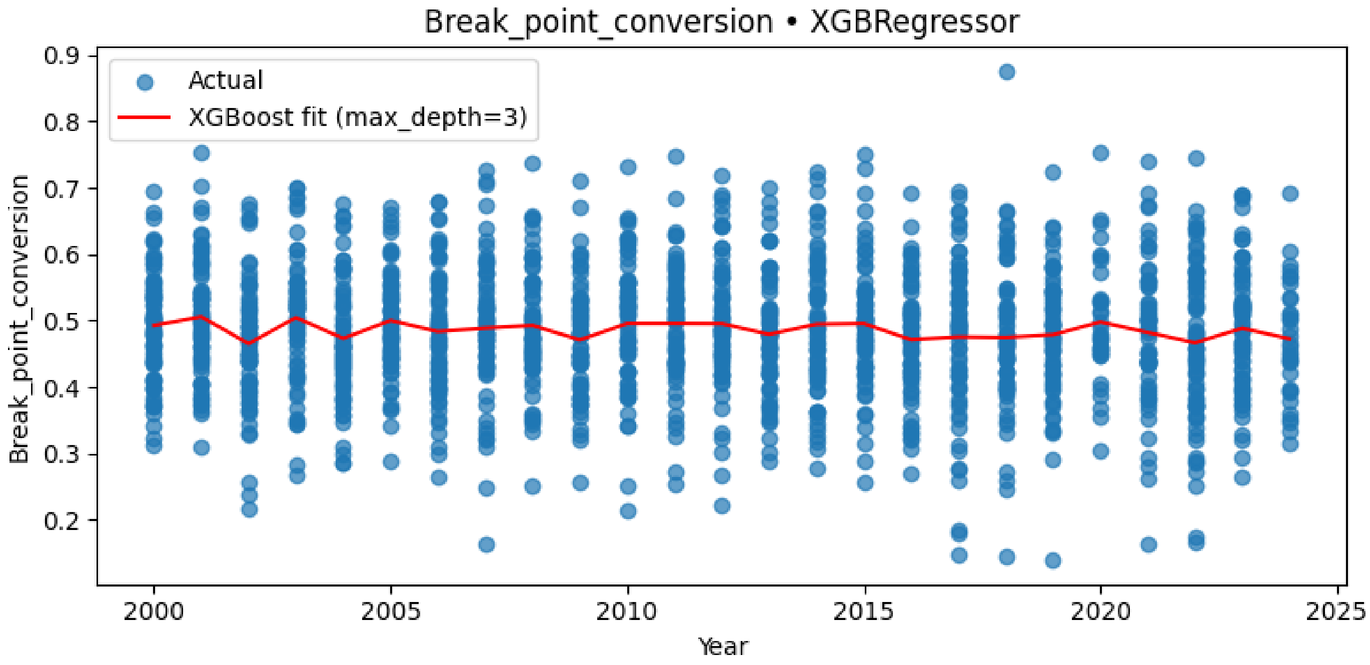
Decision Tree




Ace_per_game • XGBRegressor
Actual
XGBoost fit (max_depth=3)
Ace_per_game
0.0
0.5
1.0
1.5
2.0
2000
2005
2010
2015
2020
2025
Year


Break_point_save_percentage • XGBRegressor
Actual
XGBoost fit (max_depth=6)
Break_point_save_percentage
0.0
0.2
0.4
0.6
0.8
1.0
2000
2005
2010
2015
2020
2025
Year



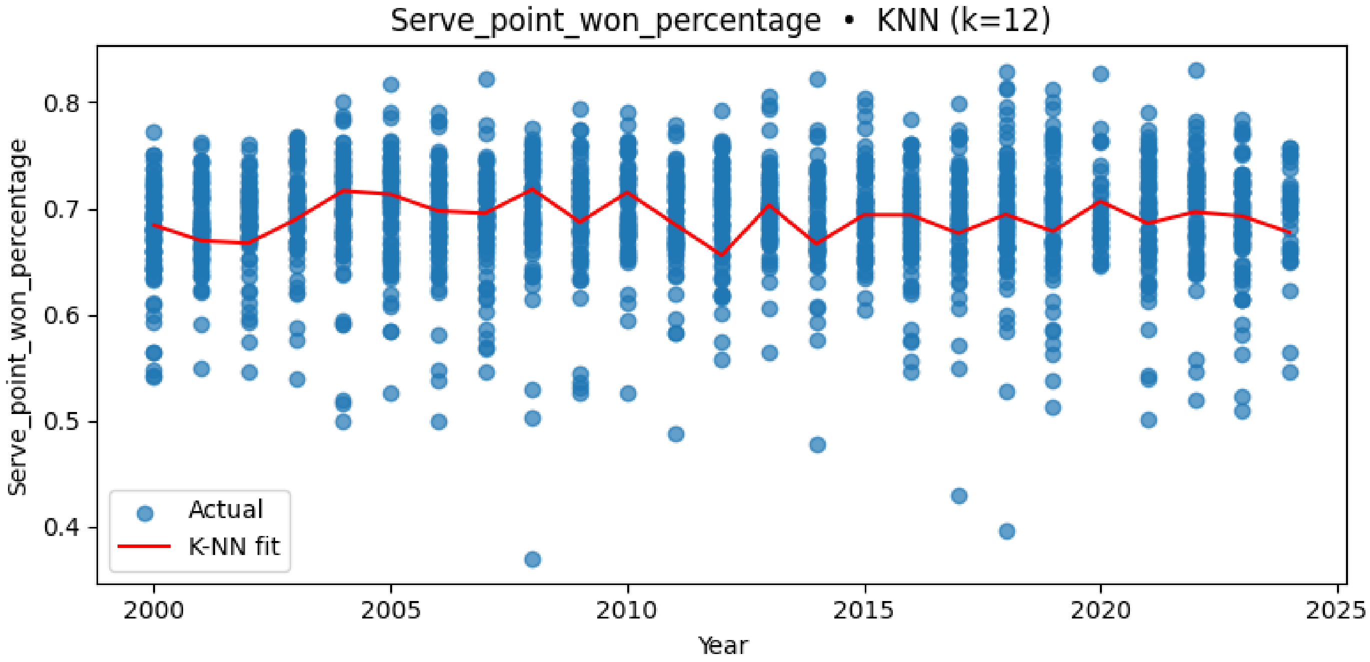


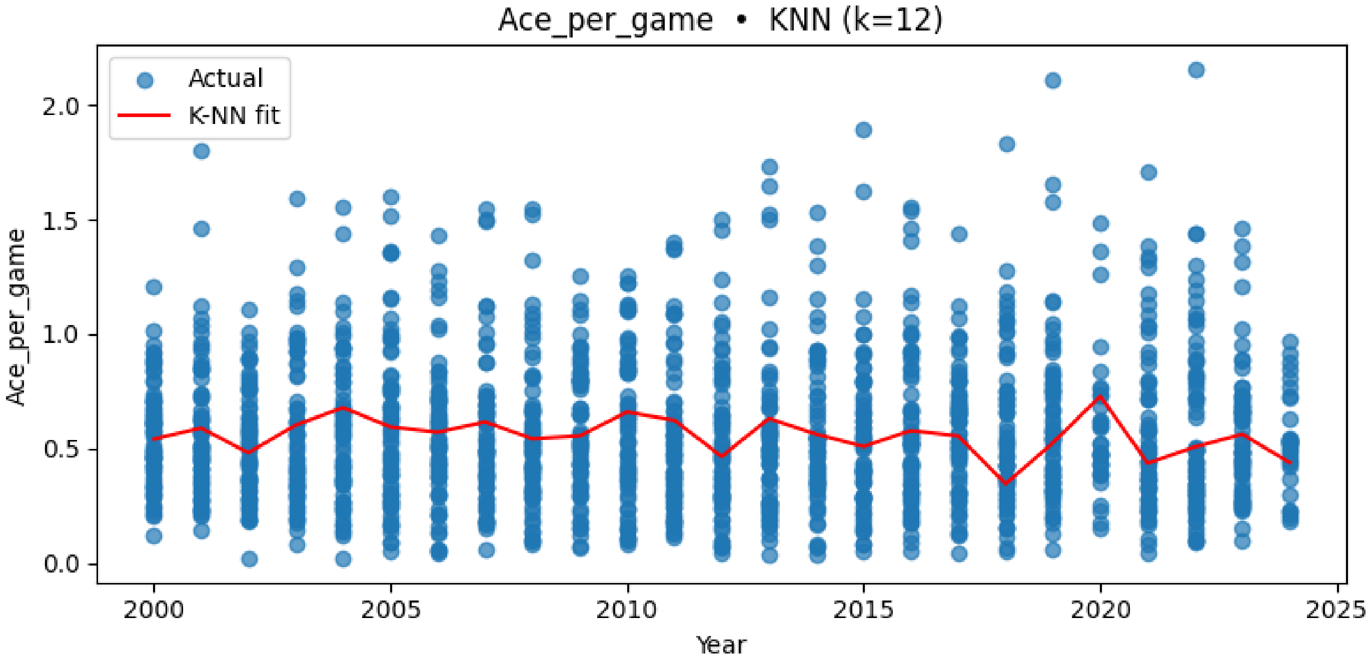
For Decision Tree I have found best max depth and use it to get better result and my error rate results are:

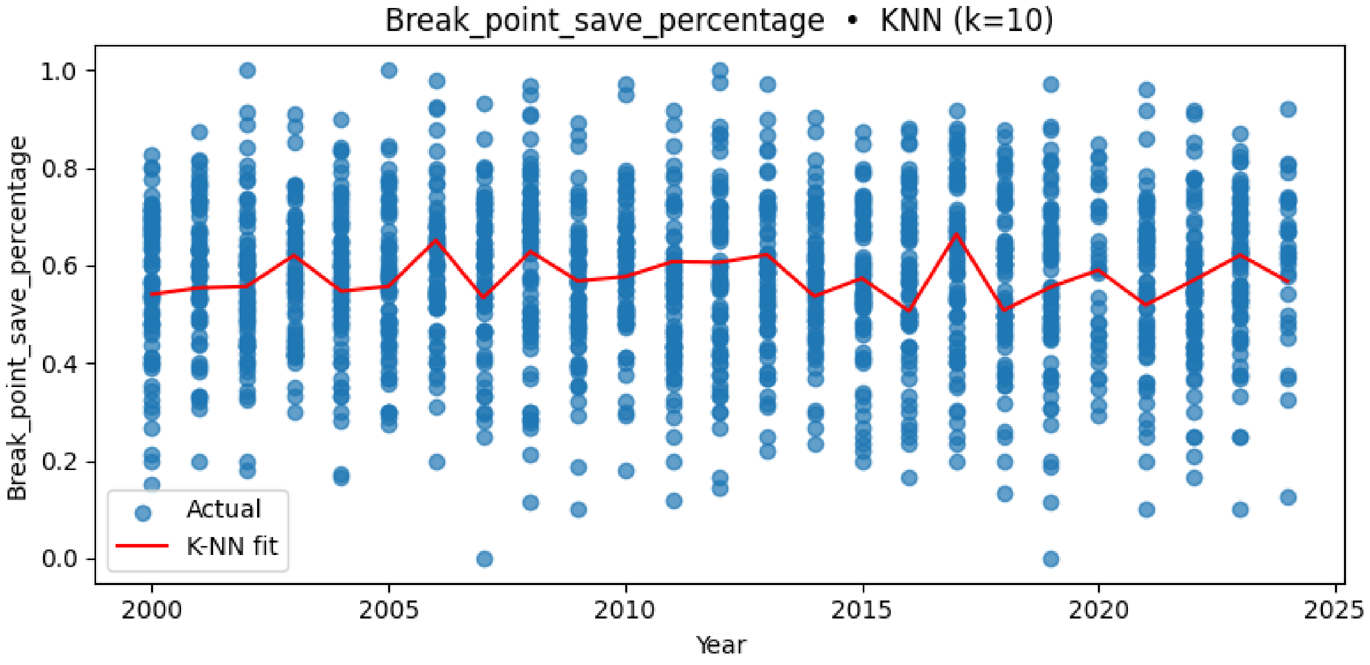
	MSE	RMSE
Serve_point_won_percentage	0.018	0.018
Ace_per_game	0.174	0.174
Break_point_save_percentage	0.068	0.068
Return_point_won_percentage	-0.002	-0.002
Break_point_conversion	0.032	0.032

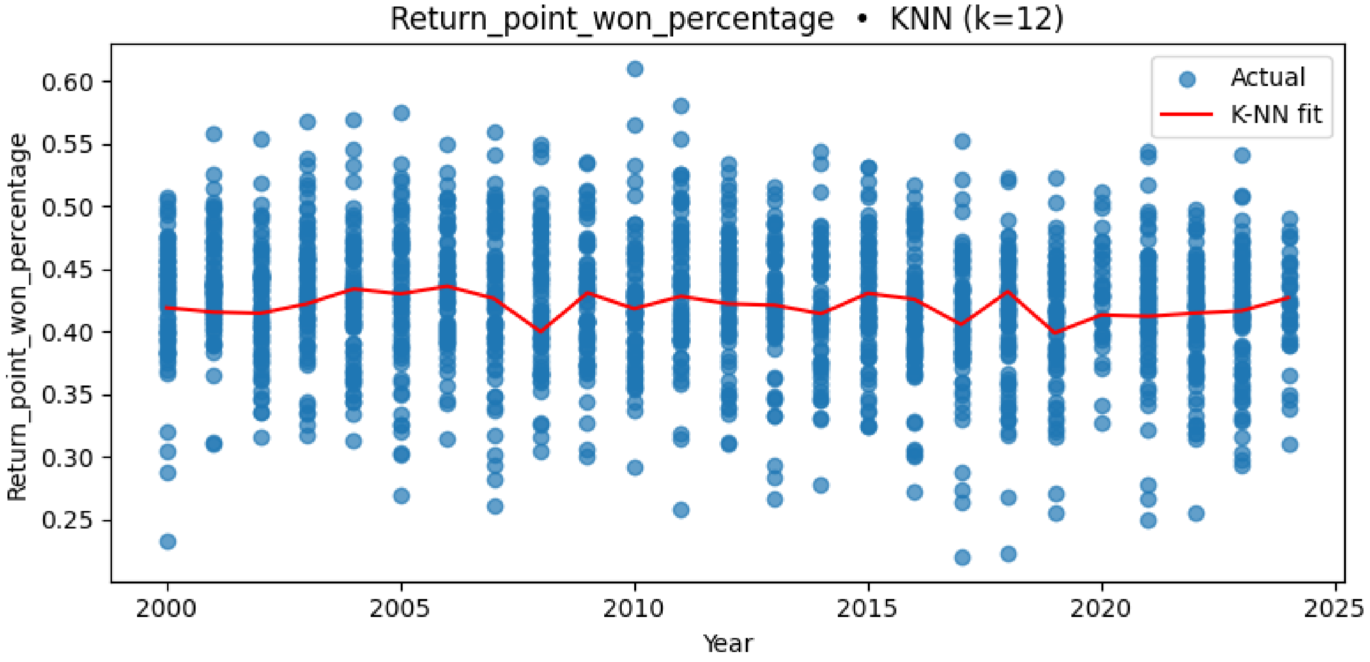
These results look worse than Linear Regression Decision Tree could not surpass simpler model, so I tried KNN regression which is more complex than Linear Regression, but less complex than Decision Tree to be able to gain best method I can use.

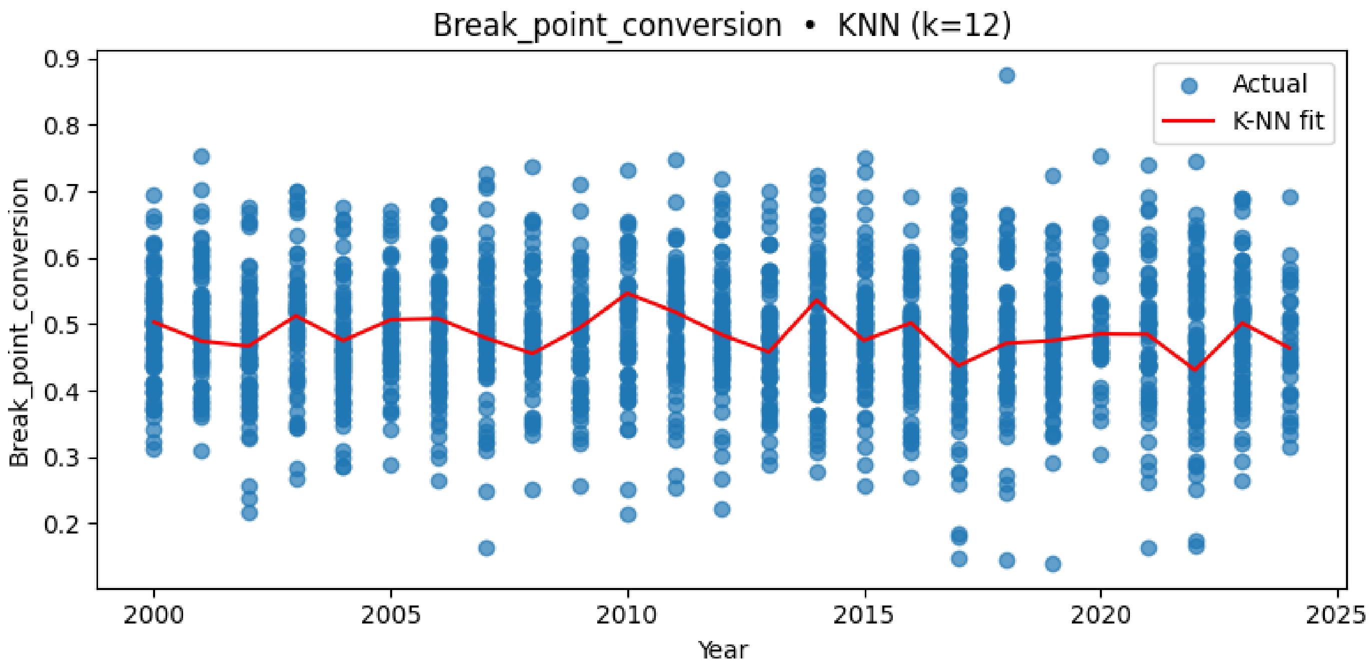
KNN Regression






Break_point_save_percentage • KNN (k=10)
Break_point_save_percentage
1.0
0.8
0.6
0.4
0.2
0.0
2000
2005
2010
2015
2020
2025
Year
Actual
K-NN fit


Return_point_won_percentage • KNN (k=12)
Return_point_won_percentage
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
2000
2005
2010
2015
2020
2025
Year
Actual
K-NN fit



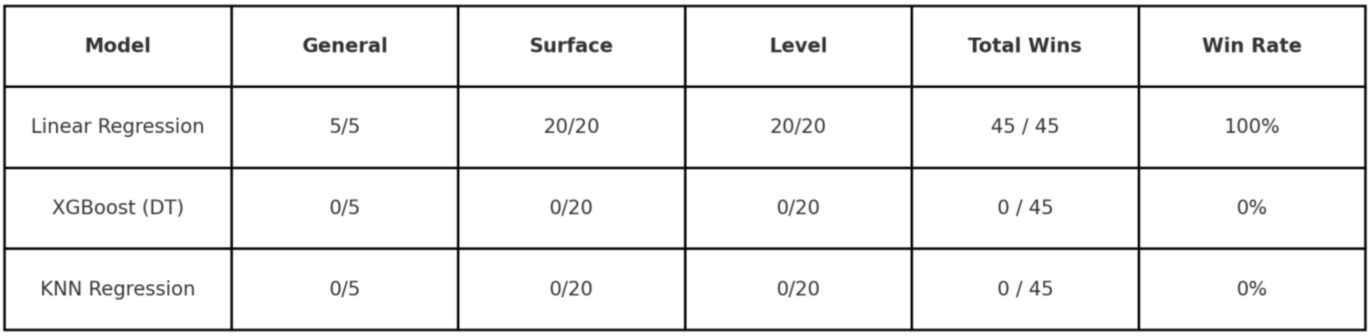
For KNN regression firstly, I have found best k value between 1 to 12 and use it and my error rates:

	MSE	RMSE
Serve_point_won_percentage	0.003	0.003
Ace_per_game	0.170	0.170
Break_point_save_percentage	0.072	0.072
Return_point_won_percentage	0.011	0.011
Break_point_conversion	0.035	0.035

These error rates are generally closer to Linear Regression, but still, it could not get better results than Linear Regression.

Machine Learning Results

I have done comparison also for each surface and each tournament levels one by one and all the results of Machine Learning error rates are:



Model	General	Surface	Level	Total Wins	Win Rate
Linear Regression	5/5	20/20	20/20	45 / 45	100%
XGBoost (DT)	0/5	0/20	0/20	0 / 45	0%
KNN Regression	0/5	0/20	0/20	0 / 45	0%

For all comparisons Linear Regression gives least error, so I decided to use this simple but effective model on my predicting next year purpose.

Prediction Results

I have done prediction firstly for 2023 and 2024 to see how accurate they are then I have done prediction of metrices for 2025.

```
--- Results for year 2023 ---
Real Data Mean:
- Serve_point_won_percentage: 0.684
- Ace_per_game: 0.561
- Break_point_save_percentage: 0.593
- Return_point_won_percentage: 0.410
- Break_point_conversion: 0.489
Predicted Data:
- Serve_point_won_percentage: 0.695
- Ace_per_game: 0.576
- Break_point_save_percentage: 0.574
- Return_point_won_percentage: 0.411
- Break_point_conversion: 0.479

--- Results for year 2024 ---
Real Data Mean:
- Serve_point_won_percentage: 0.691
- Ace_per_game: 0.531
- Break_point_save_percentage: 0.612
- Return_point_won_percentage: 0.418
- Break_point_conversion: 0.470
Predicted Data:
- Serve_point_won_percentage: 0.695
- Ace_per_game: 0.577
- Break_point_save_percentage: 0.574
- Return_point_won_percentage: 0.410
- Break_point_conversion: 0.478

--- Results for year 2025 ---
Predicted Data:
- Serve_point_won_percentage: 0.696
- Ace_per_game: 0.577
- Break_point_save_percentage: 0.574
- Return_point_won_percentage: 0.409
- Break_point_conversion: 0.478
```

For 2023 and 2024 predictions they are very close the real ones and Linear Regression is the best model to predict that I tried, so I can trust in average tournament in 2025 a player should be able to do these metrices on their stats to win tournament.

Can it be better?

In this project, I used only ATP Tour data and focused exclusively on matches from the 21st century. If structured data for WTA (women's tennis) and matches from the 1990s becomes available, the project could be significantly enhanced. Additionally, incorporating more metrics would likely improve the accuracy and reliability of the analysis.

Conclusion

Tennis is a suitable sport for forecasting next year's metrics due to the stability of its statistical averages and the compatibility of its metrics with machine learning methods. While some individual metrics may produce unstable results, these cases are relatively rare. Overall, the general trends indicate that it is possible to estimate the average metrics required to win tournaments with a reasonable degree of accuracy.