Exploring ATP Tennis Match Statistics

Shreyas Shetty

202350975

MSc Data Analytics

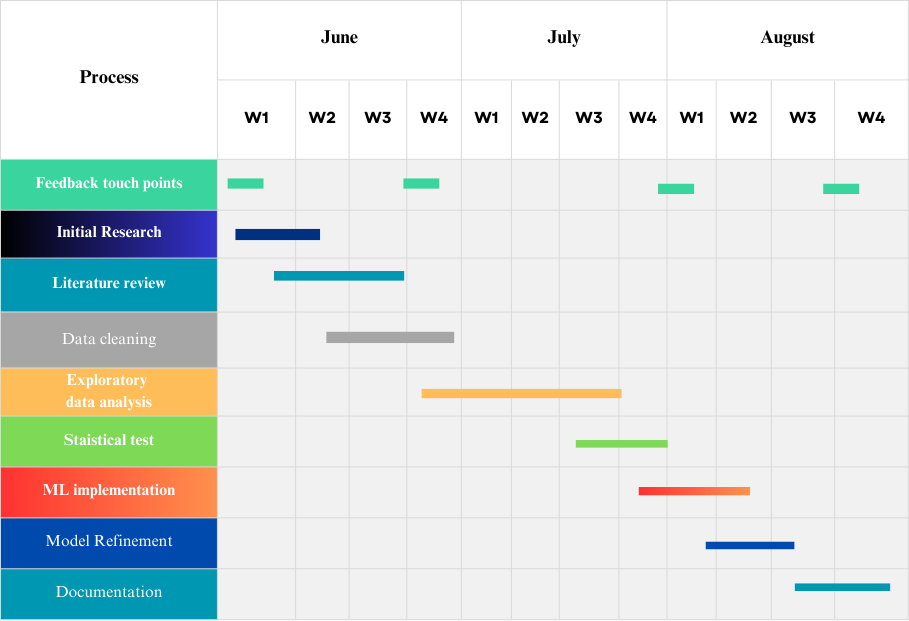
Academic Supervisor: Ruaraidh McPike

**Project Setting**

This project was provided by University of Strathclyde as an internal project. will provide a thorough analysis of ATP World Tour match data, combining traditional statistical methods with modern machine learning techniques to offer valuable insights into match outcomes. This project will contribute to the growing body of knowledge in sports analytics, demonstrating the power of data-driven approaches in professional tennis. By understanding the factors that drive success on the court, this research aims to enhance the strategic decisions of players and coaches, while also engaging fans and stakeholders in the evolving narrative of the sport.

Ever since the modern days of sports have been integrated with the application of Data Analytics, every aspect of the game from players performance to team strategy, fan engagement to business model has been transformed into multiple folds (Srivastav, 2023) . According to a report by Fortune Business Insights, global sports analytics market has forecasted to reach form $3.78 billion in 2023 to $22.13 billion by 2030 (Srivastav, 2023), showing a substantial growth of 28.7% CARG. One prime example for this is the sports of F1, which at current time is so data driven backed up by AWS, providing top notch viewing experience for the viewers by showing real time stats during race like tire deterioration, traction ratio, all the sector times with coloured mini sectors, too see where exactly a driver has improved his lap time and DRS status.

**Project Timeline:**

****

**Executive Summary**

This dissertation aims to provide a comprehensive analysis of the factors influencing outcomes in professional tennis matches, focusing on ATP match dataset. By examining key performance metrics—such as break point saving percentage, ace rate, and first serve win percentage—across multiple ATP seasons, this study seeks to uncover patterns and insights that can inform both strategic decisions and performance optimization in tennis.

The analysis includes nearly 4,000 data points from ATP matches performed between 2019 and 2023 at various tournament levels. A detailed Exploratory Data Analysis (EDA) was performed to uncover trends, correlations, and anomalies in the data, providing the framework for further statistical investigations.

To investigate the impact of various playing conditions, notably the court surface, the study used statistical techniques such as Analysis of Variance (ANOVA) and Tukey's Honestly Significant Difference (HSD) test. These methodologies revealed considerable differences in player performance indicators across different surfaces, providing useful insights into how players adapt to changing conditions and the strategic implications of surface type on match outcomes.

In addition to surface analysis, the dissertation examines the dynamic aspect of tennis matches using Hidden Markov Models (HMM). By analysing match patterns and extracting hidden states from both winners' and losers' performance data, the work provides a nuanced perspective on match dynamics, including as momentum shifts and important turning points.

Building on these core insights, the dissertation proposes a machine learning approach for forecasting the outcome of the important third set in a match. A classification model was created to predict whether a player will win the third set using previous match data and feature engineering approaches.

Overall, this dissertation combines statistical rigor with advanced modelling techniques to enhance our understanding of professional tennis match outcomes. The insights gained from this study not only contribute to the academic discourse on sports analytics but also have practical implications for players, coaches, and other stakeholders in the tennis community.

**Acknowledgment**

I would not have been able to complete this dissertation without the help of others, I would like to take this opportunity to thank them. Firstly, I would like to thank University of Strathclyde for the opportunity, it’s been a great learning experience. I would also like to thank my supervisor Ruaraidh McPike who answered all my questions and supported me throughout the dissertation, helping me understand the projects objectives and setting up feedback meetings ultimately improved the quality of my research work.

**Ethics**

Except where explicitly stated, all the work in this dissertation – including any appendices – is my own and was carried out by me during my MSc course. It has not been submitted for assessment in any other context.

Signed

# Table of Contents

[1 Table of Contents 6](#_Toc175522811)

[2 Table of Figures 7](#_Toc175522812)

[3 Introduction: 9](#_Toc175522813)

[4 Literature Review 10](#_Toc175522814)

[4.1 Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments 10](#_Toc175522815)

[4.2 Visualizing Professional Tennis Upsets: ATP 2012-2014 Men's Singles Matches 10](#_Toc175522816)

[4.3 The logit function and its applications in sports modelling 11](#_Toc175522817)

[4.4 Tracking the Evolution of a Tennis Match Using Hidden Markov Models 11](#_Toc175522818)

[4.5 Data-driven analysis of point-by-point performance for male tennis player in Grand Slams 12](#_Toc175522819)

[4.6 Capturing Momentum in Tennis (Manuel, March 4 2022) 13](#_Toc175522820)

[4.7 Quantification of momentum in tennis matches and its impact: a study based on AHP-EWM method and data analysis (Wu, May 2024) 13](#_Toc175522821)

[4.8 Applications of Machine Learning to Optimize Tennis Performance: A Systematic Review 14](#_Toc175522822)

[4.9 A Statistical Model to Predict the Results of Novak Djokovic's Matches in the Australian Open Tennis Event Using Binary Logistic Regression 14](#_Toc175522823)

[4.10 Momentum Analysis of Tennis Matches Based on Logistic Regression 15](#_Toc175522824)

[5 Data and Methodology 16](#_Toc175522825)

[5.1 Overview 16](#_Toc175522826)

[5.2 Description of the ATP tennis match dataset 17](#_Toc175522827)

[5.3 Primary key for each match and spitting “scores” into distinctive columns 18](#_Toc175522828)

[5.4 New metrics derived from dataset: 19](#_Toc175522829)

[6 Data Visualizations 20](#_Toc175522830)

[6.1 Combined data frame which includes data from 2020 – 2023: 20](#_Toc175522831)

[6.2 Age vs Rank Correlation: 22](#_Toc175522832)

[6.3 Top 30 players for year 2023 based on Win Ratios: 23](#_Toc175522833)

[6.4 Visualizing winning pattern for top 30 players in a tennis match 23](#_Toc175522834)

[6.5 Distribution of Wins Across Different Tennis Surfaces 24](#_Toc175522835)

[6.6 Boxplot representations of derived metrics 25](#_Toc175522836)

[7 Statistical tests 26](#_Toc175522837)

[7.1 Surface Impact Analysis Using ANOVA and Tukey HSD 26](#_Toc175522838)

[7.2 One-Way ANOVA for bp\_saved% and ace% 26](#_Toc175522839)

[7.3 Tukey's Honest Significant Difference (HSD) Test 26](#_Toc175522840)

[7.4 Analysis of Variance (ANOVA) for Break Points Saved 27](#_Toc175522841)

[7.4.1 Data Preparation 27](#_Toc175522842)

[7.4.2 ANOVA Model for Main Effects 27](#_Toc175522843)

[7.4.3 ANOVA Model with Interaction Term 28](#_Toc175522844)

[7.4.4 Subset Analysis by Era 28](#_Toc175522845)

[7.5 Implementation of Hidden Markov Models (HMM) for Analysing Tennis Matches 28](#_Toc175522846)

[7.5.1 Data Preparation 29](#_Toc175522847)

[7.5.2 Implementation of Hidden Markov Model (HMM) 29](#_Toc175522848)

[7.5.3 Interpretation of Hidden States: 29](#_Toc175522849)

[8 Machine learning implementation 30](#_Toc175522850)

[8.1 About the dataset 30](#_Toc175522851)

[8.2 Feature Engineering and Model building 31](#_Toc175522852)

[9 Analysis and Inferences 33](#_Toc175522853)

[9.1 Average minutes per match for 2020-2023 and 2000-2003 33](#_Toc175522854)

[9.2 Average Age of winner per tournament across different Levels (2000-2003 vs 2020-2023) 33](#_Toc175522855)

[9.3 Winner Height and Rank Analysis in ATP matches 34](#_Toc175522856)

[9.4 Effects of Age parameter over players ranking 35](#_Toc175522857)

[9.5 Winning trend demonstrated by top seeded ATP players for year 2023 36](#_Toc175522858)

[9.6 Comprehensive Analysis of winners and losers’ performance metrics 38](#_Toc175522859)

[9.7 Surface-Level impact analysis 38](#_Toc175522860)

[9.8 Two-Way ANOVA on Surface and Era Interaction 40](#_Toc175522861)

[9.9 Summary of Hidden Markov Model (HMM) Analysis and Implementation 42](#_Toc175522862)

[9.10 Interpretation of Machine Learning model results 43](#_Toc175522863)

[10 Conclusion and Future work 45](#_Toc175522864)

[11 References 46](#_Toc175522865)

[12 Appendices 47](#_Toc175522866)

# Table of Figures

[Figure 1: Graphical model for awarding a point in a tennis match 12](#_Toc175522867)

[Figure 2 Momentum formula 14](#_Toc175522868)

[Figure 3: Methodology flowchart 16](#_Toc175522869)

[Figure 4 : Meta data for column match\_id for dataframe [atp] 18](#_Toc175522870)

[Figure 5: Scores of each set for matches played in ATP 2023 19](#_Toc175522871)

[Figure 6 : Average Minutes per Tournament Level for 2020-2023 20](#_Toc175522872)

[Figure 7 : Average Minutes per Tournament Level for 2000-2003 20](#_Toc175522873)

[Figure 8 : Average winner age per tournament level for years 2020-2023 21](#_Toc175522874)

[Figure 9 Average Winner Age Per Tournament level for 2000-2003 21](#_Toc175522875)

[Figure 10 : Winner Rank vs Height Distribution for players in 2023 21](#_Toc175522876)

[Figure 11 : Age vs Rank Correlation for year 2023 22](#_Toc175522877)

[Figure 12 : Age vs Rank Correlation for year 2001 22](#_Toc175522878)

[Figure 13 : Bar Chart representation of top 30 players according to win ratios 23](#_Toc175522879)

[Figure 14 : Winning trend for top 30 players 23](#_Toc175522880)

[Figure 15 : 100% stacked bar viz for top 25 players across different surfaces 24](#_Toc175522881)

[Figure 16: Serve efficiency and Ace% Boxplot viz for winner and loser for year 2023 25](#_Toc175522882)

[Figure 17 : Boxplot viz : Breakpoints Saved, & 1st serve point won% for winner and loser in 2023 25](#_Toc175522883)

[Figure 18: Target column set3\_won metadata 30](#_Toc175522884)

[Figure 19: List of independent and dependent variable for our model 31](#_Toc175522885)

[Figure 20: Feature Importance ranked in descending order 31](#_Toc175522886)

[Figure 21: Anova test results for break point win rate 39](#_Toc175522887)

[Figure 22: Tukey HSD pairwise comparison for Break point saved 39](#_Toc175522888)

[Figure 23: ANOVA test results for Ace% for winner 40](#_Toc175522889)

[Figure 24: Tukey HSD pairwise comparison for Ace point 40](#_Toc175522890)

[Figure 25: bp\_saved\_pct ~ C(surface) + C(era) 41](#_Toc175522891)

[Figure 26: bp\_saved\_pct ~ C(surface) + C(era) 41](#_Toc175522892)

[Figure 27: Boxplot viz for breakpoints saved over different surfaces in 2 eras 41](#_Toc175522893)

[Figure 28: Boxplot viz for acepoints saved over different surfaces in 2 eras 42](#_Toc175522894)

[Figure 29: Classification Report for XGboost + PCA model 44](#_Toc175522895)

[Figure 30: Classification report after applying SMOTE on Class[0] 44](#_Toc175522896)

# Introduction:

Given the extensive scope of analysing ATP tennis match data, the main objective of this dissertation is to study the dataset and define particular research questions that can provide significant insights into the factors that influence match outcomes in professional tennis.

The primary purpose is to undertake a thorough Exploratory Data Analysis (EDA) on the most recent complete ATP season, with the goal of understanding significant statistics and visualising them. This preliminary research will help to find interesting features that can be examined further, possibly through the analysis of historical data from prior seasons.

This dissertation will investigate the following research directions:

1. **Evolution of Match Statistics Across Eras:** This research will analyse the evolution of match statistics in professional tennis over time by examining data from multiple ATP seasons. A well-devised Exploratory Data Analysis (EDA) will be a significant component of this study, exploring various metrics and factors to uncover trends in the sport’s dynamics, tactics, and competitiveness. Advanced statistical techniques such as ANOVA and Tukey HSD will be employed to gain deeper insights into the shifts observed across different eras.
2. **Predicting Third Set Outcomes in Matches:** The dissertation will develop a machine learning model aimed at predicting the outcome of the crucial third set in tennis matches. Recognizing the third set's significant impact on the overall match result, the model will leverage player metadata and performance metrics from the earlier sets to estimate the likelihood of a player winning this decisive set.
3. **Surface Impact on Player Performance:** This research will investigate how different playing surfaces (e.g., clay, grass, hard court) influence key player performance metrics, such as break point saving percentage and ace rate. By conducting a surface impact analysis using statistical methods like ANOVA and Tukey HSD, the study aims to provide insights into how surface types affect match outcomes and player strategies.

With possible applications in player development, coaching tactics, and tournament preparation, the investigation of these study aims will deepen our understanding of the elements that impact professional tennis success.

# Literature Review

The following literature review explores the intersection of tennis and data analytics, highlighting key research and developments in this field. This section examines various studies that have applied data-driven approaches to analyse different aspects of tennis, from player performance to match outcome prediction. By reviewing these works, we aim to provide a comprehensive overview of the current state of tennis analytics, identify common themes and methodologies, and contextualize our own research within the broader landscape of tennis data science. Each subsection focuses on a specific study or group of related studies, discussing their methodologies, findings, and contributions to the field.

## Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments

In the field of tennis match prediction, a groundbreaking study (Saaty & Gu, June 2019) emerged that combined empirical data with expert judgments, achieving a remarkable 85.1% accuracy rate. This research utilized an extensive dataset from the ATP World Tour, spanning from 1968 to 2015, encompassing 165,974 data points across 44 features. The study employed an Analytic Network Process (ANP) model to predict match outcomes, focusing specifically on contests between Djokovic and Federer. To assess the significance of player performance metrics, the researchers conducted a Wilcoxon rank sum test on 20 key features. Linear regression was then applied to determine the relative importance of each feature through coefficient analysis.

In preparing data for the ANP model, the researchers innovatively grouped similar performance factors into clusters, effectively reducing the number of factors from 20 to 12. A key aspect of this study was its integration of expert opinions and pre-match judgments. These judgments were carefully constrained to be based solely on the features incorporated into the ANP model. Experts made paired comparisons drawing on their knowledge and experience, while also considering correlation coefficients from historical data. The researchers also factored in the current form of players and additional judgment opinions to create a prioritized list of influencing factors.

The ANP model's super matrix was then used to synthesize all these priorities, producing a final prediction. This methodology represents a significant advancement in tennis match prediction, combining the strengths of data-driven analysis with the nuanced insights of expert judgment. This approach not only demonstrated high predictive accuracy but also provided a framework for understanding the complex interplay of factors that influence tennis match outcomes.

## Visualizing Professional Tennis Upsets: ATP 2012-2014 Men's Singles Matches

This blogs by Tyler Knutson (Knutson, 2016) walks though the analysis of performance metrics and Odds of upset matches; where an underdog player defeats a most likely a favourite / higher seed player. So, identify the winner of the match before the match, the concepts of “Odds” is leveraged, odds greater than 2.0 denotes an underdog while odds less than 2.0 represents the favourite. A density distribution of break points in a upset match is studied to analyse the impact on break point metric. This blog mainly highlighted 2 main findings:

Firstly, studies have observed a tendency among bookmakers to allocate their resources disproportionately when setting odds. Bookmakers focus more intently on matches with less predictable outcomes, dedicating less time to precisely calibrating odds for matches with clear favourites. For instance, in matches with heavy favourites, the difference between setting underdog odds at 20:1 versus 18:1 was often neglected.

Secondly, analyses of betting patterns across multiple sportsbooks have revealed that odds are frequently adjusted not just based on perceived probabilities, but also in response to the distribution of wagers. In a notable example, when two sportsbooks initially offered identical odds (1.3 for the favourite, 3.5 for the underdog), the book receiving heavier action on the favourite adjusted their odds to 1.2 and 4.5 respectively. This adjustment serves to encourage more balanced betting, allowing sportsbooks to manage their risk exposure more effectively.

## The logit function and its applications in sports modelling

This (Hyatt-Twynam, 2013) blog post provides in-depth exploration of logit function and its applications in sports application. Logit function can be used as transformation while working with probabilities. Linear shifts in logit space are the fundamentals behind logistic regression. The author presents a compelling case for the utility of logit function in transforming probabilities to improve modelling accuracy and interpret uncertainties. The author argues that linear shifts in probability space can lead to unrealistic outcomes, especially in extreme probability scenarios.

The blog post effectively demonstrates the versatility of the logit function in sports modelling, from basic probability transformations to more complex applications in Bayesian inference and market pricing. The author emphasizes the critical importance of quantifying uncertainties in probability estimates for both bookmakers and sports traders. Instead of relying on a single probability value, the blog suggests developing a probability density function that describes the likelihood of each possible true probability value.

The blog introduces the logit-normal distribution as an ideal choice for representing this uncertainty. This distribution is created by applying a normal distribution in logit space. The author states: "Typically in science, for physical measurements, the logical choice would be a normal distribution, however if this is used for p\_A, you're claiming there's a chance p\_A could be outside the range 0 to 1”

## Tracking the Evolution of a Tennis Match Using Hidden Markov Models (Kolonias, et al., August 2004)

The author (Kolonias, et al., August 2004) have addressed the problem of extracting higher level semantic information from low level feature from multimedia content. A sequence of video is given as a input to the model A graphical model is developed to represent the usual progression of the game, which helps to identify and examine elementary events within the tennis sequence. Using this fundamental sequence of events, reasoning for higher events is performed such as awarding the current point. Computer Vision models is used to detect the event and HMM is used to perform higher level reason.

This (Kolonias, et al., August 2004) works aims to replace the original scene evolution model with a series of smaller models, each of which aims to accurately depict a specific moment in the match's history. The most important thing to remember during this process is that when we combine each model in this set, we will end up with a model that is identical to the original.

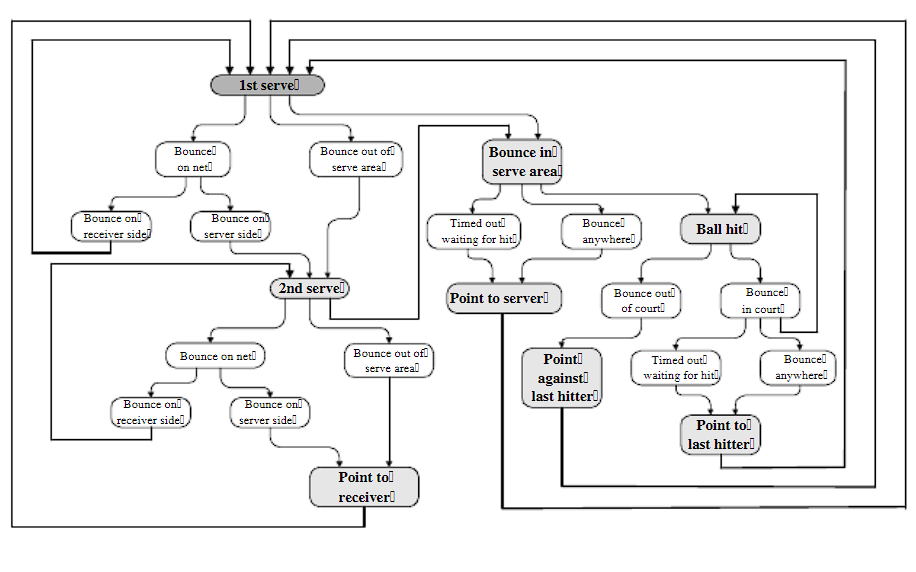


Figure 1: Graphical model for awarding a point in a tennis match

## Data-driven analysis of point-by-point performance for male tennis player in Grand Slams

This (Yixiong Cui, 2019) article illustrates the work on analysing the performance metrics of tennis by incorporating 29675 data points through Classification Tree analysis. The author leveraged the point level data of 145 Grand slams main draw men’s Singles matches played by top ranked players from year 2011 to 2016. A two-step cluster analysis was performed with Euclidean distance as measure and Schwart’s Bayesian criterion. The classifying metrics were (i) Serve speed (ii) Rally length.

This article projects the point level analysis of players using classification tree model. Humidity and Ball types were 2 unique features which could be found being used by the author. Another key finding was due to the slow court surface of Roland Garros, it was easier to break to the server’s game and players tend to player more aggressively. The influence of Rally points was noticeable in certain ways: Short and longer rally helped player to win more matches than medium rally. The serving efficiency of a player is not only affected by the surface type and the level of the tournament but also by the skill level of the opposition

In a thorough examination of tennis performance throughout Grand Slam tournaments, the author discovers that a classifying tree model portrays some critical insights concerning a player's effectiveness. The model shows that the analysed player had higher overall point-winning percentages in the Australian Open (AO) and US Open (US) than Roland Garros (RG) and Wimbledon (W).

Specifically, the tree model shows that serve speed has a considerable impact on point-winning chance. Slower serves, on the other hand, resulted in decreased point-winning percentages, which were most noticeable in the AO and US. Interestingly, it demonstrates consistent return efficacy throughout all Grand Slams, with the player winning roughly one-third of points against first serves and 40% against second serves, regardless of event. This classification approach offers a comprehensive understanding of how surface qualities and serve dynamics influence match outcomes, which is useful for strategic planning and performance analysis in professional tennis.

## Capturing Momentum in Tennis (Manuel, March 4 2022)

This (Manuel, March 4 2022) study introduces the concept of a "momentum swing," defined as a shift in momentum from one player to another by a magnitude of 3% or more. This threshold provides a concrete metric for identifying significant turning points within a match. The model's ability to consider a wide range of factors and provide point-by-point analysis of match dynamics offers potential applications for players, coaches, and analysts in developing strategies and understanding match progression.

In a comprehensive study of tennis match dynamics, researchers have developed a sophisticated chain of predictive models to analyse and quantify the concept of momentum in professional tennis. The study, which utilizes data from 1.5 million points played on the WTA tour between 2012 and 2020, presents a novel approach to understanding the ebb and flow of tennis matches.

The authors introduce a multi-layered model that considers various factors including court type, current match state, in-match statistics, and pre-game odds. This model chain begins by predicting the probability of winning the next point, which then feeds into predictions for game, set, and ultimately match outcomes. This hierarchical approach allows for a granular analysis of how individual points impact the overall match probability.

Central to the study is the concept of "leverage," which quantifies the importance of each point in terms of its potential to change the match outcome. Building on this, the researchers define "momentum" as an exponentially weighted moving average of the leverage gained by a player. This definition takes into account both the recency and the importance of points, with more recent and higher-leverage points having greater influence on the momentum score.

## Quantification of momentum in tennis matches and its impact: a study based on AHP-EWM method and data analysis (Wu, May 2024)

In this report (Wu, May 2024) , the author deeply analyses the concept of momentum formation in tennis match and its extent of influence on winning the match. Using Analytic hierarchy process (AHP) and the entropy weight method (EMP) methodology, Welch’s Test and Run test, non-randomness of momentum metric is investigated. One interesting metric considered by the author during data pre-processing stage is p1\_d : the longer the distance traversed by an athlete, the greater their level of exhaustion consequently. Evaluating Criterion Matrix, final AHP and EWM weights are computed which are then fed into a concise formula to calculate momentum

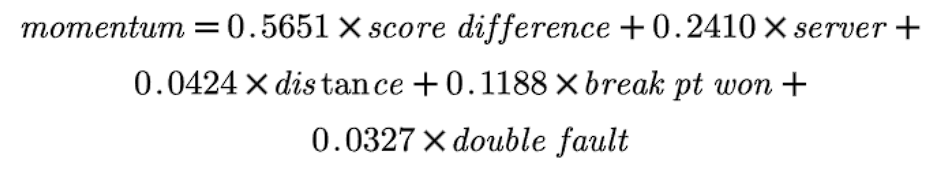
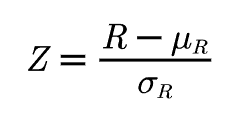


Figure 2 Momentum formula

To establish the non-random nature of momentum and likelihood of winning the match, Z test run is performed using below given formula:



Hypothesis test was undertaken on Wimbledon 2023 tournament – 1301 as reference for player Carlos Alcaraz . Z value obtained was -7.402, which indicate that w that the data are non-random data based on the variable Momentum\_p1, and the level is significant

## Applications of Machine Learning to Optimize Tennis Performance: A Systematic Review

This (Sampaio, et al., June, 2024) report sheds light over various approaches adopted to assess the performance stats of players. Notable reviewed mythologies include Psychological and affective states where a sensor based approach was employed to leverage inertial measurement units (IMUS) (Havlucu Hayati, 2022) worn by small group of elite coaches (n=2) and 4 professional players. BY using Long short-term Memory Recurrent Neural Network, an accuracy of 85% was achieved. He further dwelled over the scope of AI for predicating players optimal performance “zones”.

This report also displays how various ML approaches yields different results for instance studies include Backpropagation and Neural Network and Convolutional Neural Networks being utilized to evaluate tactical performances and classifying groundstroke stances (Yu Zhang, 2023), (Yang, January 2023) .While Random Forest was utilized to perform Classification of moment patters and predicting point winners for a set. (Jernej & Ziva, 2021) (Panjan, et al., June 2010). (Yu Zhang, 2023) Designed a model for assessing player performance using decision trees and a common classification technique while deployed convolutional neural network to analyse batting strength and angles based on video footage (Yang, January 2023). CNNs excel at image recognition tasks, making them ideal for this application.

## A Statistical Model to Predict the Results of Novak Djokovic's Matches in the Australian Open Tennis Event Using Binary Logistic Regression

This (Choudhary, et al., 2023) provides a detailed analysis with the purpose of predicting Novak Djokovic's match outcomes at the Australian Open. The study uses binary logistic regression to examine 147 matches from the 2013 to 2021 Australian Open tournaments. The model examines several predictor variables, such as aces, double faults, first serve points won, and breakpoints converted, among others, to determine their impact on the match outcome, which is defined as a win or a loss. The findings show that, while many factors were considered, the breakpoint conversion was the most significant predictor, resulting in a model that correctly classified 88.9% of Djokovic's matches. The study emphasises the significance of specific performance metrics in predicting tennis match outcomes, particularly in the case of elite players such as Djokovic. The study's methodology included rigorous statistical testing, such as tests for multicollinearity and model fit, to ensure the reliability of the findings.

The logistic regression model explained 58.8% of the variance in match outcomes (Nagelkerke R²=0.588). This finding emphasises the complexities of tennis, where subtle differences in performance metrics can have a significant impact on the likelihood of winning, particularly in high-stakes tournaments like the Australian Open. However, the study acknowledges its limitations, particularly the exclusion of Djokovic's 2022 performance due to his absence from the tournament, which may have influenced the generalisability of the findings. Despite this, the study lays the groundwork for future research on predictive modelling in sports.

## Momentum Analysis of Tennis Matches Based on Logistic Regression

This paper (Liu, 2024) introduces an innovative approach to predicting tennis match outcomes by integrating logistic regression with the entropy weight method. The study focuses on identifying key performance indicators (KPIs) that significantly impact match results and uses these indicators to build a predictive model. By applying the entropy weight method, Liu ensures that the most influential KPIs are weighted appropriately, enhancing the model's accuracy. The logistic regression model developed in this study achieved an impressive 94.3% accuracy, underscoring the effectiveness of this combined approach in predicting match outcomes based on selected match statistics such as first-serve scoring rate and ace rate.

A notable contribution of this paper is the introduction of the concept of momentum as a predictive factor in tennis matches. Momentum is quantified as the slope of the pointing probability, which is calculated through the logistic regression model. This novel approach to defining and measuring momentum provides a dynamic perspective on how real-time changes in match statistics can influence outcomes. The strong correlation between momentum and match results, validated through a chi-square test with a significance level of less than 0.001, highlights the importance of momentum in determining match outcomes. This finding aligns with and extends the existing literature by not only focusing on static predictors but also incorporating the temporal dynamics of performance during a match.

# Data and Methodology

## Overview

The flowchart above provides a overview of the methodology employed in this research, detailing the step-by-step process undertaken to analyse player performance and build a classification model. The methodology is divided into distinct stages, each contributing to the overall objective of understanding and predicting player outcomes based on historical and current data.

This methodology provides a robust framework for analysing player performance over time, comparing different eras, and building predictive models based on comprehensive and normalized datasets. The use of statistical tests ensures the reliability of the findings, while the classification model serves as a practical tool for forecasting player success

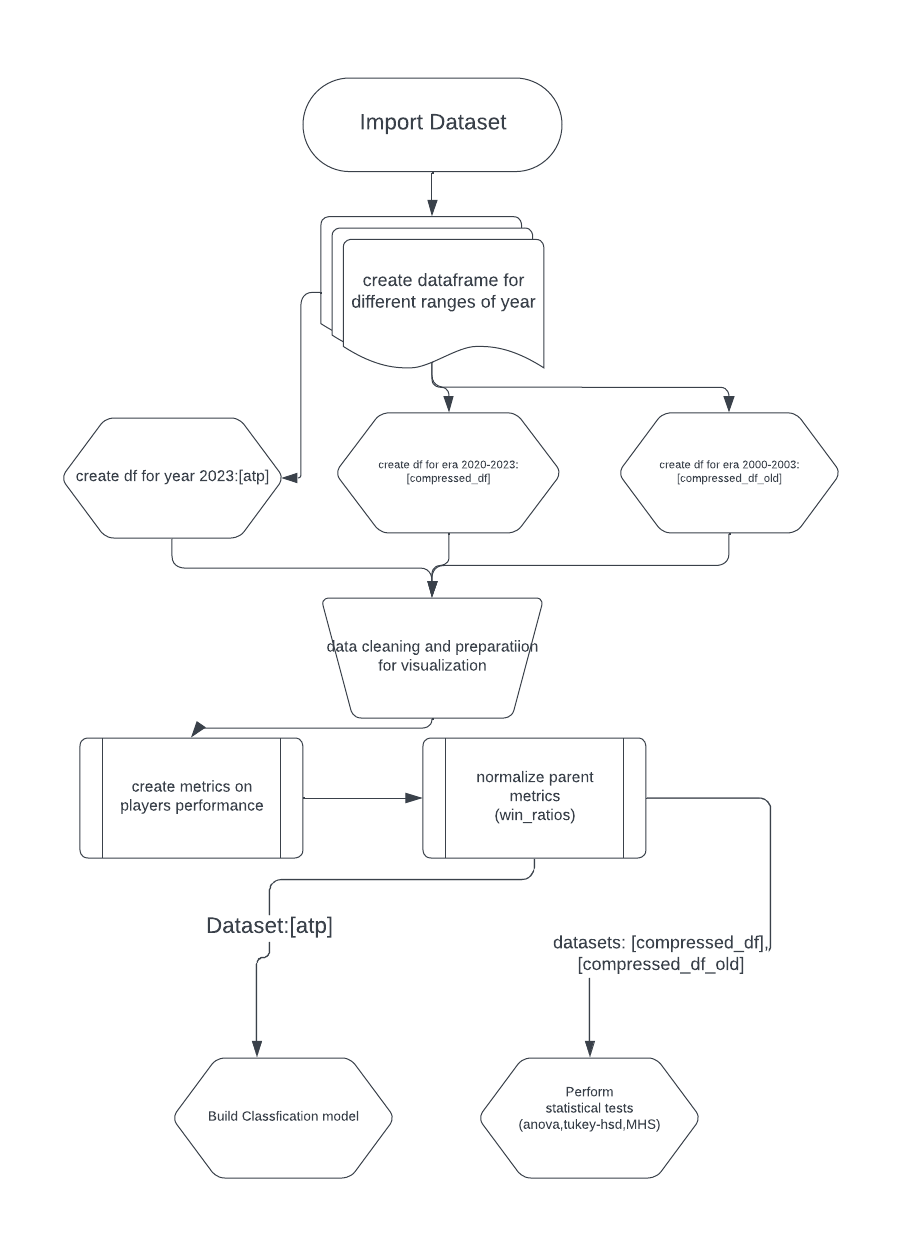


Figure 3: Methodology flowchart

## Description of the ATP tennis match dataset

The dataset (JeffSackmann, 2016) used for this report is from GitHub repository by JeffSackmann .

This dataset contains comprehensive information about tennis tournaments and matches, including player statistics and match outcomes. The data is organized into several categories:

1. Tournament Information:

|  |  |
| --- | --- |
| tourney\_id | A unique identifier for each tournament |
| tourney\_name | The name of the tournament, as various tourneys are played at different levels are listed in this: 470 different tourneys held from 2020-2023 |
| surface | The type of court surface, possible values are [Hard, Grass, clay] for year 2000-2003 era Carpet as a surface was also found |
| draw\_size | The number of players in the tournament |
| tourney\_level: | Indicates the level of the tournament (e.g., Grand Slam, Masters 1000, other level tourney, Challenger’s, Satellites/ITFs, finals, Davis cup) => Distinctive tourney levels values are [G, M, A, C, S, F, D] |
| tourney\_date | The start date of the tournament |

Table 1 Tournament Information

2. Match Information:

|  |  |
| --- | --- |
| match\_num | A match-specific identifier |
| score | The final score of the match |
| best\_of | Indicates whether the match is best of 3 or 5 sets |
| round | The tournament round of the match |
| minutes | The duration of the match |

Table 2: Match Information

3. Player Metadata:

For both winner and loser:

|  |  |
| --- | --- |
| player\_id: | Unique identifier for each player |
| name | Player's full name |
| Hand | Player’s dominant hand |
| ht | Player’s height in cms |
| ioc | Player’s country code |
| age | Players age at the time of the tournament |
| rank | Players ATP rank at the time of the tournament |
| rank\_points | Player’s ranking points |

Table 3: Players Metadata

5. Match Statistics:

For both winner and loser:

|  |  |
| --- | --- |
| ace | Number of aces served |
| df | Number of double faults |
| svpt | Total serve points |
| 1stIn | Number of first serves made |
| 1stWon | Number of first-serve points won |
| 2ndWon | Number of second-serve points won |
| SvGms | Number of service games |
| bpSaved | Number of break points saved |
| bpFaced | Number of break points faced |
| seed | Player's seeding in the tournament |
| entry | Type of entry into the tournament (e.g., wild card, qualifier) |

Table 4: Match data

This dataset provides a rich source of information for analysing tennis matches, player performance, and tournament characteristics. It allows for in-depth analysis of various aspects of the game, including serving performance, player rankings, and match outcomes across different tournaments and surfaces.

## Primary key for each match and spitting “scores” into distinctive columns

We concatenated the match\_num column and tourney\_id column to uniquely identify each match . The new column created was match\_id which acted like a primary key .

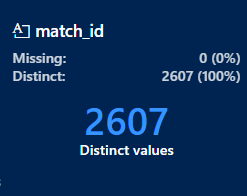


Figure 4 : Meta data for column match\_id for dataframe [atp]

To study in depth about the set progression for each match, we split the set scores into individual columns [set1, set2,set3,set4,set5]. The split columns will be re-framed into original dataframe using the “match\_id” column as our primary key. For winners and losers: 'set1\_gm\_diff', 'set2\_gm\_diff', 'set3\_gm\_diff', 'set4\_gm\_diff', 'set5\_gm\_diff: these columns hold the game differences for each set. Columns 'set1\_ot\_diff', 'set2\_ot\_diff', 'set3\_ot\_diff', 'set4\_ot\_diff', 'set5\_ot\_diff' hold the tiebreaker differences for each set.

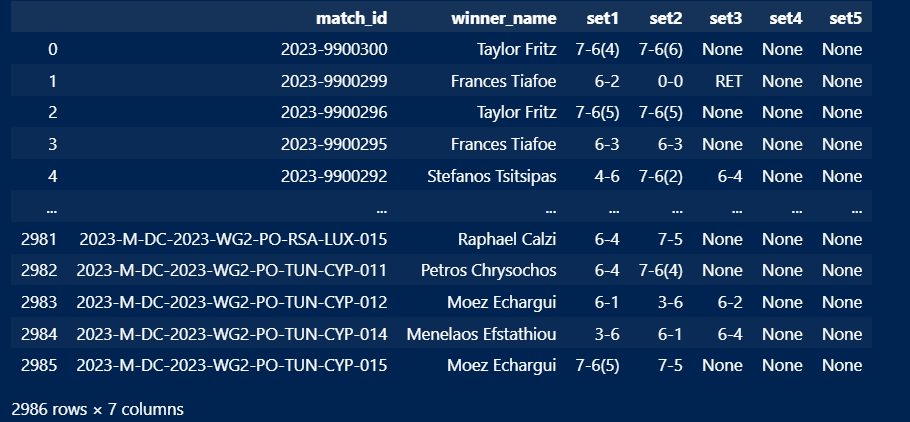


Figure 5: Scores of each set for matches played in ATP 2023

## New metrics derived from dataset:

Following are the new metrics that were derived from the existing columns from the dataset:

* 1st\_srv\_w%: First serve win percentage: Denotes the percentage of 1st serves won by total number of 1st serves played.
* 2nd\_srv\_w%: Second serve win percentage - Denotes the percentage of 2nd serves won over total number of 1st serves played.
* bp\_saved%: Break points saved percentage - Denotes the percentage of break points saved by player in the match over total number of break points faced.
* ace%: Ace percentage: denoted the ace conversion of a player over total number of 1st serve points
* serve\_efficiency% : Serve efficiency percentage - denotes the serving efficiency of a player in a match . Formula used to derive this metric is:

# Data Visualizations

## Combined data frame which includes data from 2020 – 2023:

Exploratory data analysis over ATP data from year 2020 to 2023 for average duration in minutes per Tournament level over years was performed. A grouped bar plot was created to visualization who the match duration has varied over years for each tournament. The subset of a dataframe was created using aggregate function over minutes column. An interactive python data visualization library called Plotly was used to implement this. Similarly, a pair of grouped bar plots for year 2020-2023 and 2000-2003 was implemented to visualize average age of winners ever tournament level

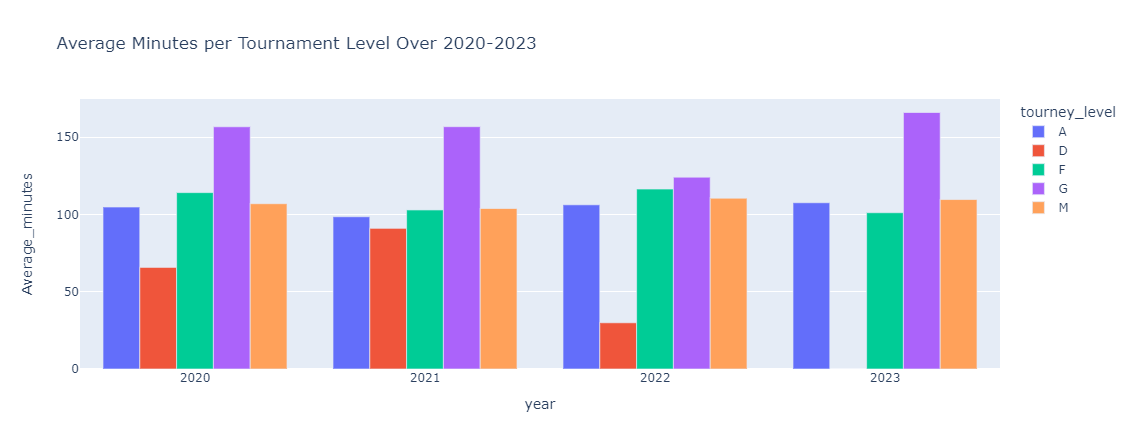


Figure 6 : Average Minutes per Tournament Level for 2020-2023

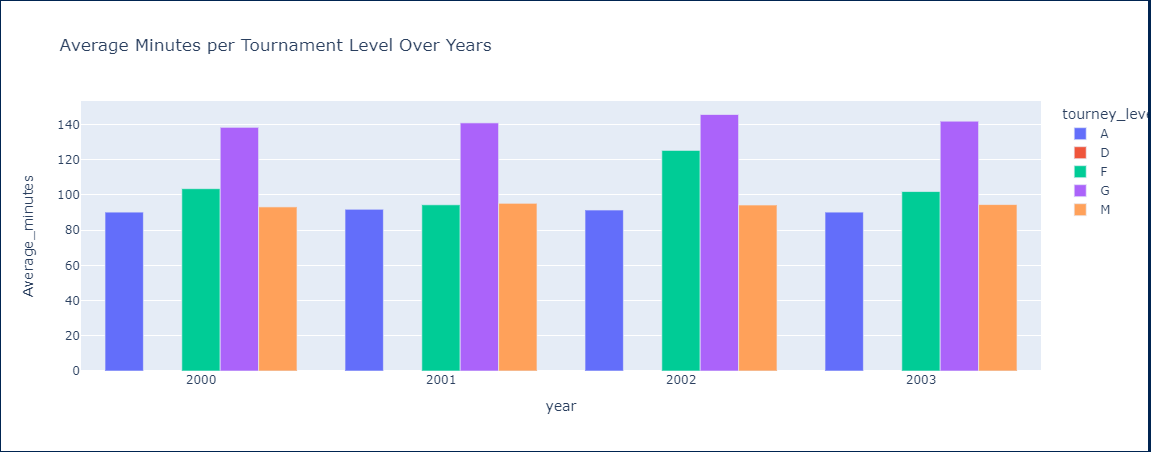


Figure 7 : Average Minutes per Tournament Level for 2000-2003

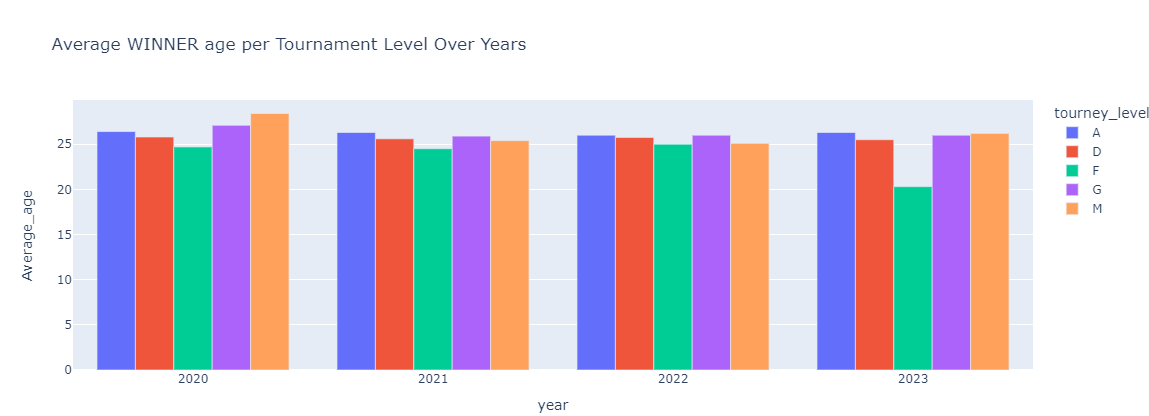


Figure 8 : Average winner age per tournament level for years 2020-2023

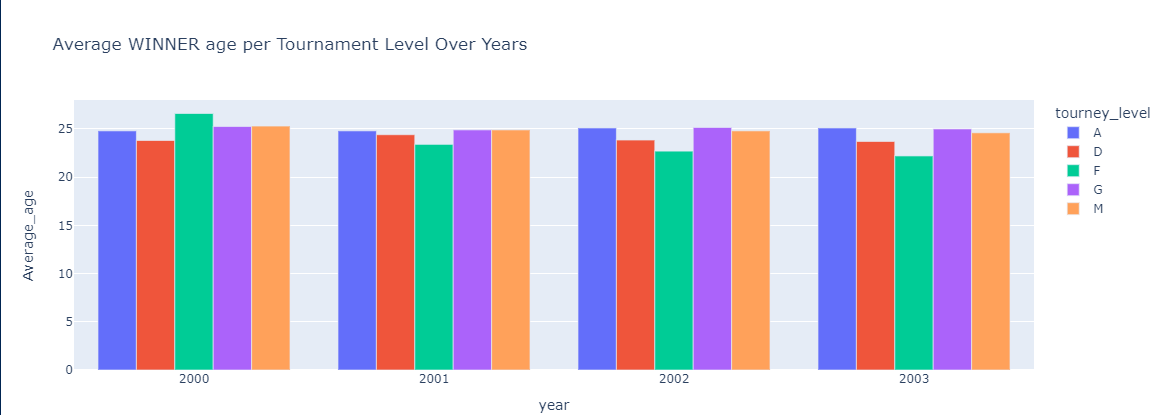


Figure 9 Average Winner Age Per Tournament level for 2000-2003

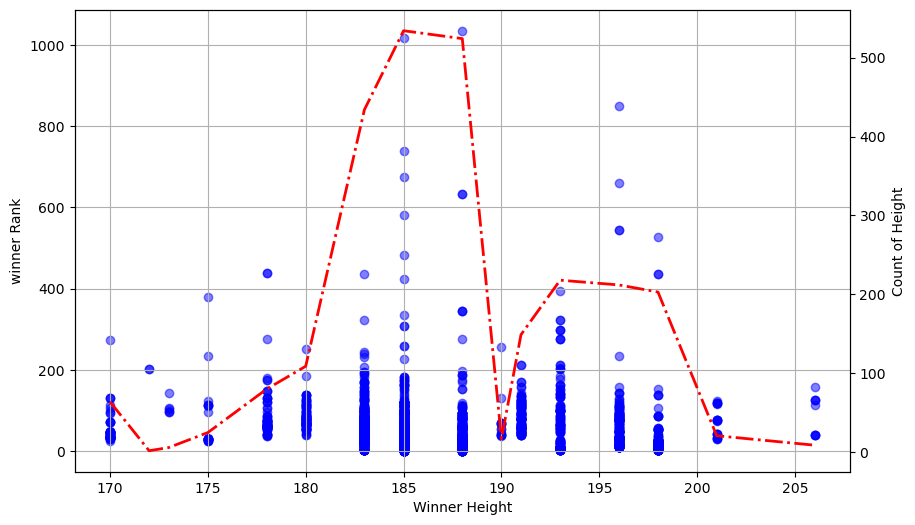


Figure 10 : Winner Rank vs Height Distribution for players in 2023

The provided visualization in Figure 10 presents a comprehensive exploration of the relationship between winner height and winner rank in a tennis dataset. The following EDA techniques have been utilized:

* **Scatter Plot:** A scatter plot was employed to visualize the relationship between winner height and winner rank. This technique allows for a direct comparison of the two variables, revealing any potential trends .
* **Histogram:** A histogram was incorporated on a secondary y-axis to depict the distribution of winner heights. This provides insights into the frequency of different height ranges among the winners.
* **Overlayed Density Plot:** A density plot was superimposed on the histogram to offer a smoother representation of the height distribution. This technique helps in identifying the underlying probability density function of the data.

## Age vs Rank Correlation:

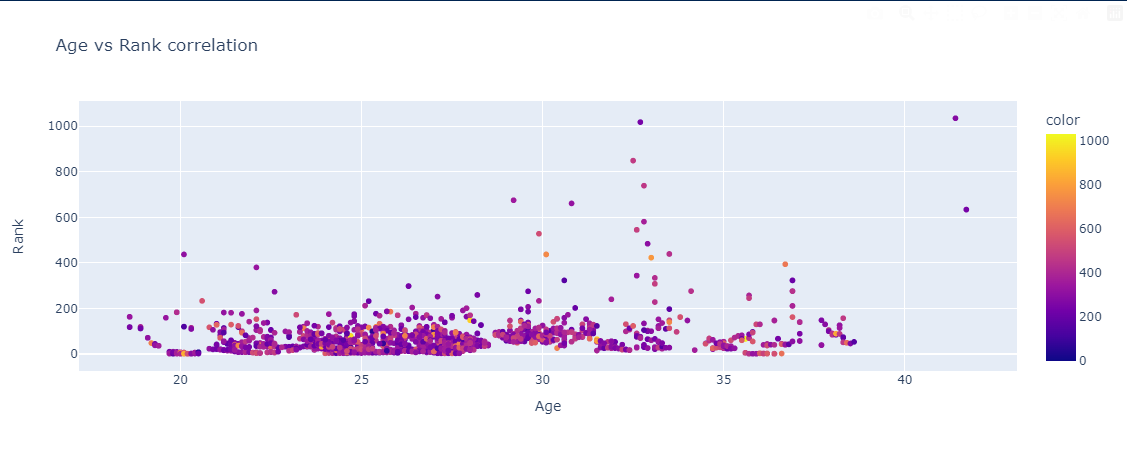


Figure 11 : Age vs Rank Correlation for year 2023

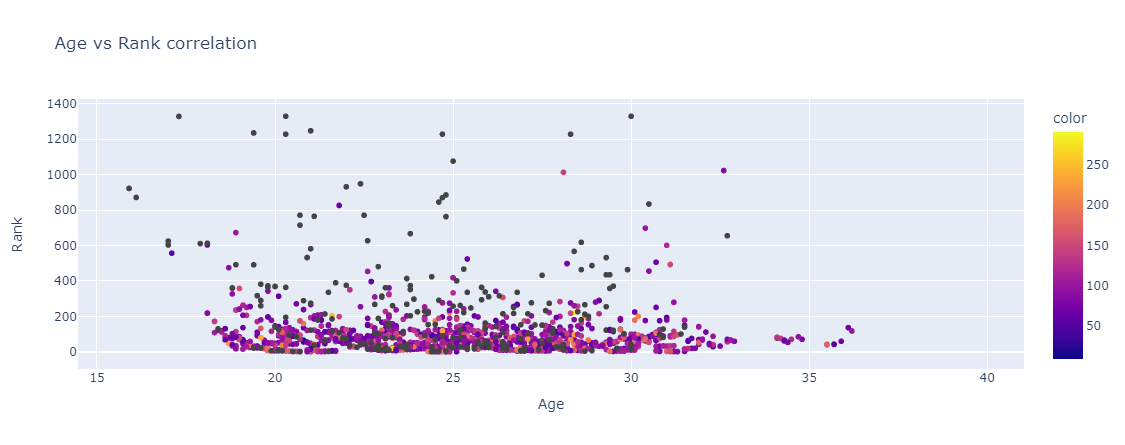


Figure 12 : Age vs Rank Correlation for year 2001

This visualization employs a scatter plot to explore the relationship between winner age and winner rank in a tennis dataset, along with an additional dimension represented by colour. The code utilizes Plotly Express (pt.scatter), a high-level library for creating interactive visualizations in Python.

* Data Source: The data for the scatter plot is extracted from two columns in the “atp” DataFrame: winner\_age (winner's age) and winner\_rank (winner's rank).
* Colour Dimension: An additional data point, minutes (presumably representing match duration), is used to colour-code the data points in the scatter plot. This allows for a preliminary exploration of how match duration might relate to both winner age and rank.
* Axis Labels and Title: Descriptive labels are provided for the x-axis ('Age'), y-axis ('Rank'), and the overall plot title ('Age vs Rank correlation').

By using a scatter plot with color-coding, this visualization offers a multifaceted exploration of the data, aiding in the initial investigation of potential relationships between winner age, winner rank, and match duration.

## Top 30 players for year 2023 based on Win Ratios:

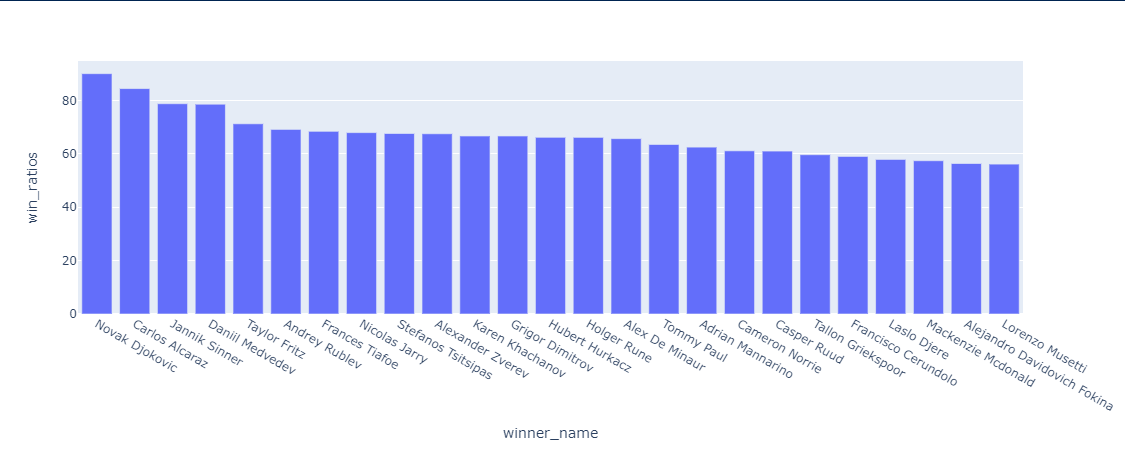


Figure 13 : Bar Chart representation of top 30 players according to win ratios

The provided visualization presents a bar chart illustrating the win ratios of various tennis players for year 2023. The following EDA techniques have been utilized:

* Bar Chart: A bar chart was chosen to visually represent the win ratios of each player. This format allows for easy comparison and identification of the highest and lowest performers.
* Sorting: The data was sorted in descending order of win ratios, ensuring that the players are ranked from highest to lowest win percentages. This facilitates a clear understanding of the performance hierarchy.
* Data Normalization: The decision to create a separate DataFrame with win ratios as a column suggests that the data underwent normalization. This process likely standardized the win ratios, ensuring that they are comparable across different players or periods, potentially accounting for variations in playing frequency or opponent strength.

## Visualizing winning pattern for top 30 players in a tennis match

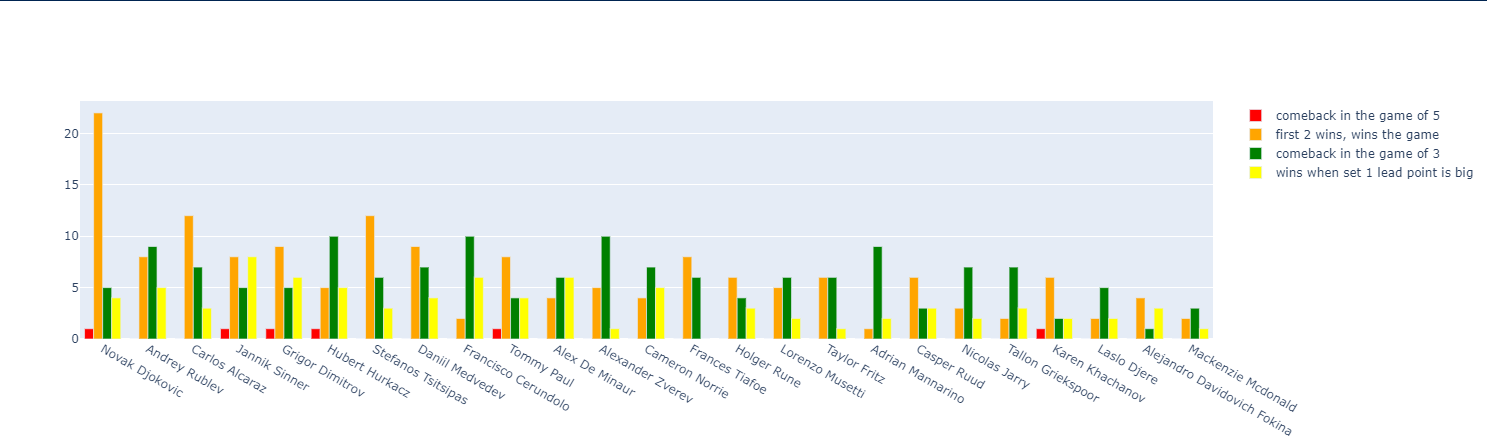


Figure 14 : Winning trend for top 30 players

In this section, a visualization of the winning patterns of various tennis players has been presented using a grouped bar plot. The objective was to categorize and illustrate the different methods through which players secure victory in a match. Four main winning patterns or labels were identified:

* First 2 Wins, Wins the Game (Yellow): This label represents scenarios where a player wins the first two sets and subsequently wins the match.
* Wins When Set 1 Lead Point is Big (Orange): This label captures matches where a significant lead in the first set contributes to an overall match victory.
* Comeback in the Game of 5 (Red): This category indicates matches where a player makes a comeback in the fifth set to win the match.
* Comeback in the Game of 3 (Green): This label represents matches where a player recovers in the third set and goes on to win the match.

This plot provides a clear comparison of the different winning patterns across various players, offering insights into their match-winning tendencies based on these predefined categories

## Distribution of Wins Across Different Tennis Surfaces

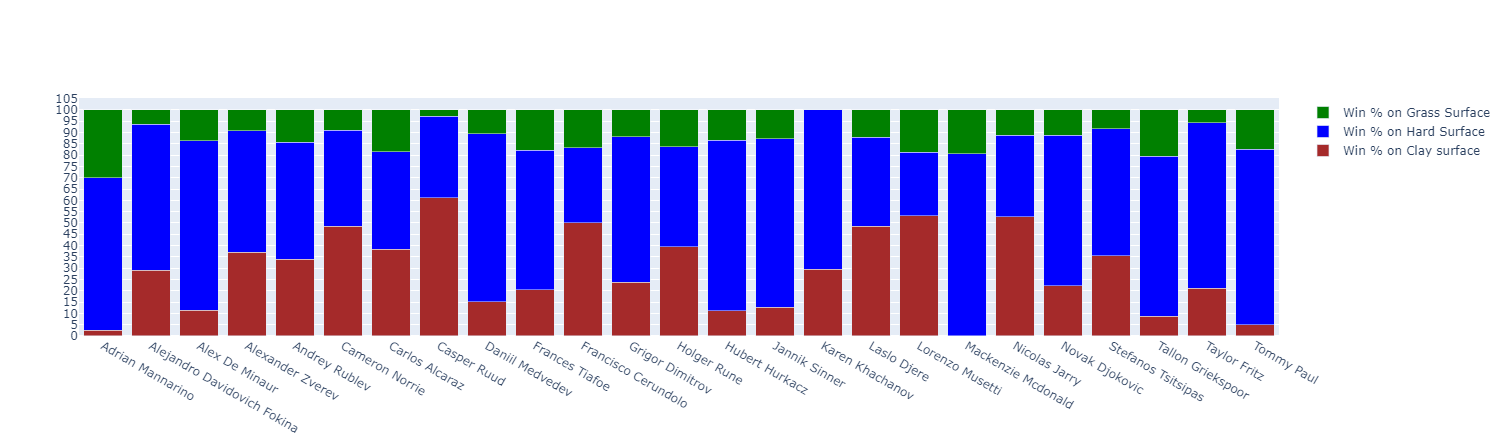


Figure 15 : 100% stacked bar viz for top 25 players across different surfaces

Each bar in the plot corresponds to a player, with the stacked segments indicating the relative success across the three surfaces. The use of a 100% stacked bar plot ensures that the total height of each bar is consistent, facilitating an easy comparison of surface performance distributions among the players.

The above representation utilizes a 100% stacked bar plot to analyse and understand the distribution of wins by the top 25 tennis players across different types of surfaces. The aim is to visualize how player success varies depending on the surface—whether it be grass, hard, or clay courts.

The win percentages on each surface are categorized into three segments:

1. **Win % on Grass Surface (Green)**: This segment shows the proportion of wins each player has achieved on grass courts.
2. **Win % on Hard Surface (Blue)**: This segment represents the proportion of wins on hard courts, reflecting player performance on this common surface type.
3. **Win % on Clay Surface (Red)**: This segment illustrates the percentage of wins on clay courts, a surface known for its unique playing conditions.

Each bar in the plot corresponds to a player, with the stacked segments indicating the relative success across the three surfaces. The use of a 100% stacked bar plot ensures that the total height of each bar is consistent, facilitating an easy comparison of surface performance distributions among the players.

## Boxplot representations of derived metrics

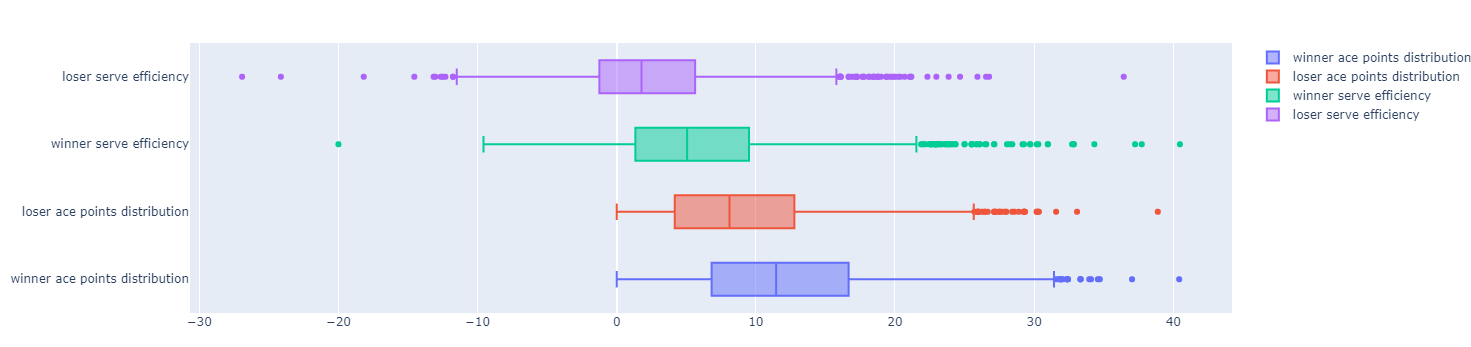


Figure 16: Serve efficiency and Ace% Boxplot viz for winner and loser for year 2023

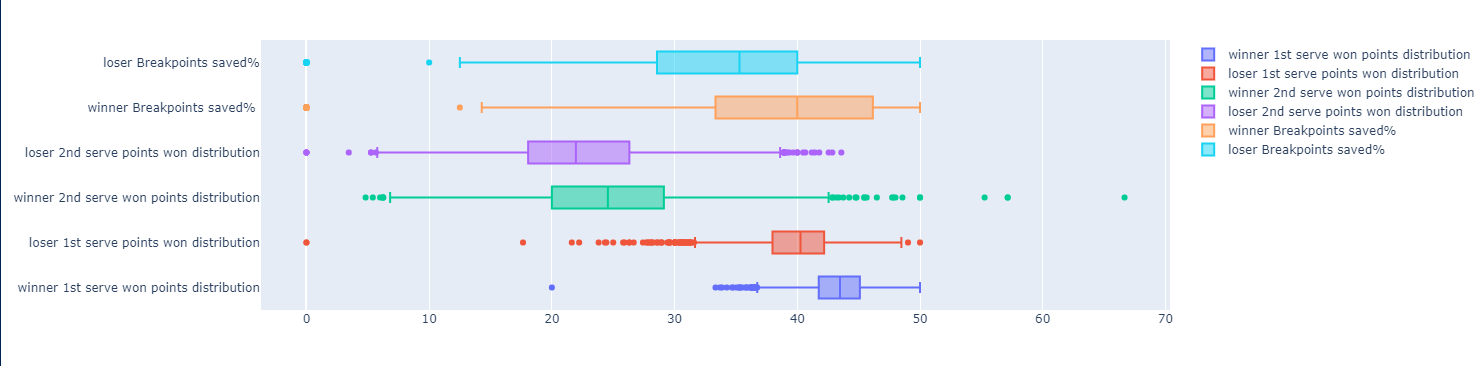


Figure 17 : Boxplot viz : Breakpoints Saved, & 1st serve point won% for winner and loser in 2023

The provided visualizations consist of two boxplots, each presenting different aspects of tennis match statistics for winners and losers.

Figure 16 focuses on service efficiency and ace points distribution for both winners and losers. It highlights the variation in serve efficiency and the number of aces delivered by each player. Again, the boxplot structure allows for an easy comparison between winners and losers across these metrics, showcasing the spread of data points and identifying any significant outliers that may exist within the dataset.

Figure 17 boxplot displays the distribution of several critical match statistics related to service games. The variables represented include the percentage of breakpoints saved by both winners and losers, as well as the distribution of points won on first and second serves by both winners and losers. The boxplot clearly shows the range, interquartile range, and any outliers for each of these variables, providing a comparative view of how winners and losers perform in these key areas of the game.

These visualizations offer a detailed snapshot of the distribution and variability in key performance metrics, effectively illustrating the differences in how winners and losers perform across critical aspects of tennis matches.

# Statistical tests

## Surface Impact Analysis Using ANOVA and Tukey HSD

In this section, statistical techniques, particularly Analysis of Variance (ANOVA) and Tukey's Honestly Significant Difference (HSD) test, are used to examine the impact of the three different tennis surfaces—Clay, Grass, and Hard—on key performance metrics: the percentage of break points saved (bp\_saved%) and the percentage of aces (ace%). This analysis emphasises on player data from 2023 to gain insight into how surface type influences these critical aspects of match performance.

Tennis is unique in that it is played on a range of surfaces, each with a substantial impact on gameplay and player performance. The primary surfaces—Clay, Grass, and Hard—provide different playing conditions that influence ball speed, bounce height, and player movement. Given these variances, it is critical to investigate how players' abilities to save break points and hit aces differ among surfaces.

To carefully evaluate these variations, ANOVA was chosen as the proper statistical procedure. ANOVA is used to see if there are statistically significant differences in the means of a dependent variable—here, bp\_saved% and ace%—across many groups, in this case, the various surfaces. Following ANOVA, Tukey's HSD test is used to conduct pairwise comparisons of the surfaces, determining which surfaces differ from one another.

## One-Way ANOVA for bp\_saved% and ace%

A one-way ANOVA was conducted to examine whether the type of surface—Clay, Hard, or Grass—significantly affects the performance metrics of players.

1. **Break Points Saved (bp\_saved%):** The objective is to determine if there are significant differences in the bp\_saved% among the three surfaces. Players were grouped based on the surface they played on, and their bp\_saved% values were compared across these groups. The F-value and p-value from the ANOVA test were used to assess the overall effect of surface type on bp\_saved%.
2. **Aces (ace%):** The objective is to evaluate whether the percentage of aces served (ace%) differs significantly across Clay, Hard, and Grass surfaces. Similar to the analysis for bp\_saved%, the ace% values were compared across surface types using one-way ANOVA. The F-value and p-value provided insights into the influence of surface type on ace%.

## Tukey's Honest Significant Difference (HSD) Test

Following the one-way ANOVA, Tukey's HSD test was applied to conduct pairwise comparisons between the surfaces. This test helps identify which specific pairs of surfaces show significant differences in the performance metrics.

1. **Break Points Saved (bp\_saved%):**
   1. **Clay vs. Grass:** A significant difference was found between Clay and Grass surfaces, indicating that players saved a higher percentage of break points on Clay compared to Grass.
   2. **Clay vs. Hard:** No significant difference was observed between Clay and Hard surfaces, suggesting similar bp\_saved% on these surfaces.
   3. **Grass vs. Hard:** A significant difference was detected between Grass and Hard surfaces, with a higher percentage of break points saved on Hard courts compared to Grass.
2. **Aces (ace%):**
   1. **Clay vs. Grass:** The analysis revealed a significant difference between Clay and Grass surfaces, with players serving a higher percentage of aces on Grass.
   2. **Clay vs. Hard:** A significant difference was also noted between Clay and Hard surfaces, with more aces served on Hard courts.
   3. **Grass vs. Hard:** A smaller yet significant difference was found between Grass and Hard surfaces, with Grass courts yielding slightly more aces.

This methodological approach, combining ANOVA and Tukey HSD, provides a robust framework for understanding the impact of surface type on player performance metrics, with detailed results to be discussed in the subsequent inference and conclusion sections.

## Analysis of Variance (ANOVA) for Break Points Saved

This section describes the use of Analysis of Variance (ANOVA) to examine the effects of surface type and era on the percentage of break points saved (bp\_saved\_pct) by tennis players. The analysis was conducted using different models to explore both the main effects and interaction effects of these factors.

### Data Preparation

Before conducting the ANOVA tests, certain columns in the dataset were renamed to maintain clarity and consistency: The column serve\_efficiency% was renamed to serve\_efficiency\_pct. The column ace% was renamed to ace\_pct. The column bp\_saved% was renamed to bp\_saved\_pct.

These changes ensure that the column names are consistent and appropriate for statistical modeling.

### ANOVA Model for Main Effects

The first ANOVA model was designed to analyze the main effects of two categorical variables—surface type (surface) and era (era)—on the dependent variable, bp\_saved\_pct.

**Model Specification:** This model assesses the individual contributions of surface type and era to variations in the percentage of break points saved by players. An ordinary least squares (OLS) regression model was fit using the formula:

1. **ANOVA Table:** The ANOVA table was generated to examine the statistical significance of each factor. The typ=2 argument was used to produce a Type II ANOVA, which tests each main effect after accounting for the others.

### ANOVA Model with Interaction Term

To explore potential interaction effects between surface type and era, a second ANOVA model was implemented with an interaction term:

1. **Model Specification:** This model evaluates whether the effect of surface type on bp\_saved\_pct varies across different eras. It helps in understanding if the relationship between surface and break points saved is consistent over time or if it has changed across different tennis eras. The formula for this model included the interaction between surface and era:
2. **ANOVA Table:** Again, a Type II ANOVA was conducted to determine the significance of both the main effects and the interaction term.

### Subset Analysis by Era

To gain deeper insights, the data was further divided into two subsets representing distinct eras: 2000-2003 and 2020-2023. Separate ANOVA tests were then performed for each era to assess the effect of surface type on bp\_saved\_pct within these time periods. The dataset was split into two subsets: data\_2000\_2003: Contains data for the era 2000-2003 & data\_2020\_2023: Contains data for the era 2020-2023.

1. **Model Specification:** For each subset, an OLS regression model was fit using the formula:
2. **Purpose:** These models were intended to compare the impact of surface type on break points saved within each era, providing a clearer understanding of how this relationship might have evolved over time.
3. **ANOVA Tables:** Separate ANOVA tables were generated for each era to evaluate the significance of surface type in each specific context.

This comprehensive approach using ANOVA allows for a nuanced analysis of how surface type and era influence player performance in terms of saving break points. The findings from these tests will be discussed in the inference and conclusion sections of the dissertation.

## Implementation of Hidden Markov Models (HMM) for Analysing Tennis Matches

In this section, a Hidden Markov Model (HMM) is utilized to analyse and uncover the latent states during tennis matches for both winners and losers. The use of HMMs is particularly fitting in this context due to the sequential and temporal nature of tennis match data, where a player's performance evolves over the course of the match, influenced by various factors such as fatigue, momentum, and psychological states.

Hidden Markov Models are a type of statistical model that is well suited for evaluating time-series data. They are based on the assumption that the system being modelled is a Markov process with unobserved states. In tennis, observable variables like points won, break points saved, and serve efficiency are influenced by underlying states that represent a player's form or condition at any given time during the match. These states cannot be directly observed, but can be inferred from the sequence of match statistics.

### Data Preparation

To apply the HMM, specific match-related features were first extracted and organized into a suitable format . Feature Selection: The dataset was filtered to include relevant columns that capture critical aspects of match performance:

* match\_id, ply\_id, and results for identifying the match and player.
* bp\_saved%: The percentage of break points saved by the player.
* ace%: The percentage of aces served by the player.
* serve\_efficiency%: An overall metric of serving efficiency.
* setX\_ot\_diff (where X = 1 to 5): The overtime point differential for each set.
* setX\_gm\_diff (where X = 1 to 5): The game differential for each set.

### Implementation of Hidden Markov Model (HMM)

The Hidden Markov Model was applied to the processed data to identify hidden states within the match dynamics:

**Model Selection:** A Gaussian HMM was chosen for the analysis, given its suitability for continuous data and its ability to model the probabilistic transitions between states.

The features used for fitting the model were:

* ace%: Percentage of aces served.
* bp\_saved%: Percentage of break points saved.
* serve\_efficiency%: Serve efficiency percentage.
* set\_point\_diff: Game differential across sets.
* overtime\_set\_point\_diff: Overtime point differential across sets.

### Interpretation of Hidden States:

The hidden states identified by the HMM can be interpreted as representing different momentum phases within the match, from both the winning and losing player’s perspective.

1. **State 0:** This could represent a phase of the match where momentum is neutral or slightly against the player. For example, it might correlate with scenarios where the player is slightly behind in game or set points but still within reach of shifting momentum.
2. **State 1:** This state might represent a phase where the player is under significant pressure, possibly losing momentum. Characteristics might include lower serve efficiency, fewer aces, or a significant negative point differential.
3. **State 2:** This state likely represents a positive momentum phase, where the player is either in control or has just gained an advantage. High serve efficiency, a higher percentage of aces, and positive set or game point differences would characterize this state.

# Machine learning implementation

## About the dataset

In this section, we present the implementation of a classification machine-learning model utilizing the ATP tennis dataset spanning from the years 2020 to 2023. The dataset underwent well defined preprocessing as described in the preceding sections, ensuring that all necessary transformations and data cleaning were applied.

The primary objective of this model is to predict the winner of the third set in a tennis match. In this context, the target variable, set3\_won, is a binary classification outcome where ‘0’ indicates the player who lost the third set, and ‘1’ signifies the player who won it.

The focus on predicting the winner of the third set stems from the crucial role this set plays in determining the outcome of a match. In best-of-three (Bo3) formats, the third set acts as the decisive moment where the match is won or lost, making it critical for both players. In best-of-five (Bo5) matches, the third set often serves as a momentum-shifting point, setting the stage for the remainder of the match. Understanding and predicting the dynamics of the third set not only highlights its importance but also provides strategic insights that can significantly influence match outcomes.

Alongside existing features, three additional binary metrics: set1\_won, set2\_won, and set3\_won—were derived specifically for this analysis, where 0 represents a set lost and 1 a set won.

The final dataset utilized for model training comprised 7,474 data points. It is important to note the inherent class imbalance within the target variable: 623 instances of set3\_won = 0 (indicating a loss) and 6,851 instances of set3\_won = 1 (indicating a win). This imbalance represents a typical challenge in classification tasks, where the model might become biased toward the majority class, potentially leading to suboptimal predictions for the minority class.

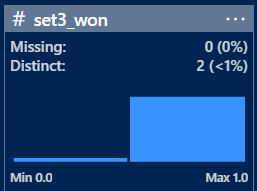


Figure 18: Target column set3\_won metadata

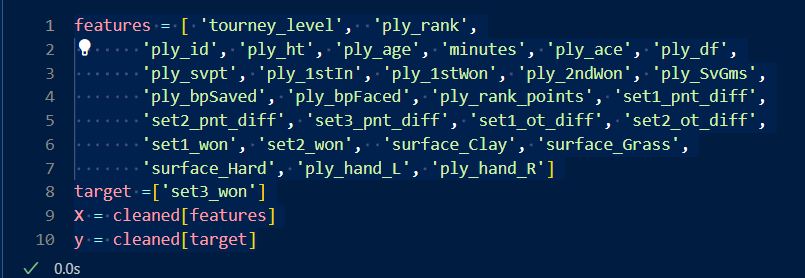


Figure 19: List of independent and dependent variable for our model

Given the challenges posed by the dataset, particularly the class imbalance, the XGBoost (Extreme Gradient Boosting) classification algorithm was selected for this study. XGBoost is a robust and highly efficient gradient boosting framework known for its ability to handle various challenges in predictive modelling.

One of its primary strengths lies in its ability to handle class imbalance through the incorporation of weighted loss functions. By adjusting the scale\_pos\_weight parameter, XGBoost balances the significance of both positive and negative classes, effectively mitigating the risk of bias toward the majority class. This feature is crucial in our context, where the target variable exhibits a significant imbalance, ensuring that the model remains sensitive to the minority class. It also provides robust insights into feature importance, allowing us to identify and focus on the most impactful features within the dataset.

## Feature Engineering and Model building

Feature Engineering is an important step in machine learning to convert raw data and features relevant metrics for model training. This procedure employs techniques such as normalisation, scaling, and dimensionality reduction.

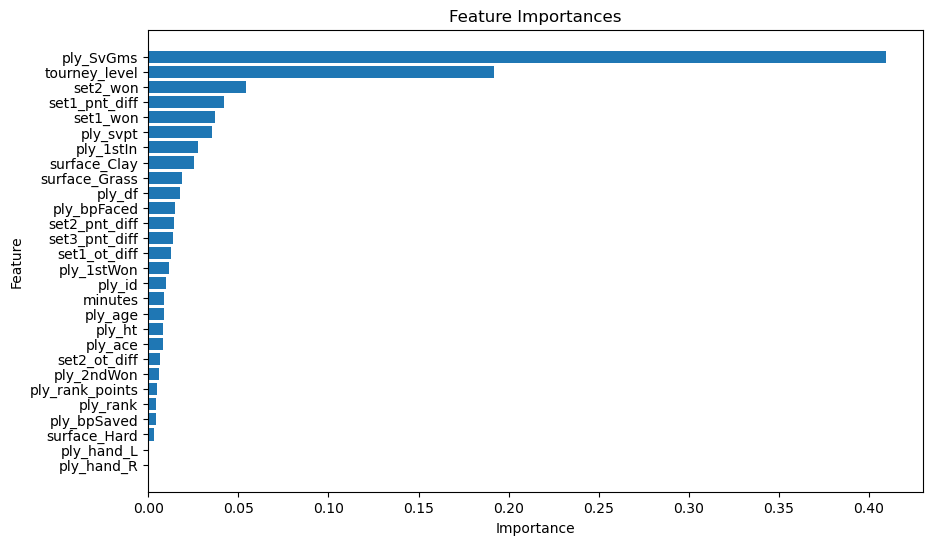


Figure 20: Feature Importance ranked in descending order

StandardScaler was used on all features to ensure normalisation within a given range. This is a frequent strategy when dealing with characteristics of different scales, as greater magnitudes can induce bias into the model. Dimensionality reduction was achieved using Principal Component Analysis (PCA). PCA determines linear combinations of the original variables that represents the most variability in the data. By setting n\_components to 7, we reduced the dimensionality of the dataset from 28 to 7 .

The Train-Test Split method is widely used to evaluate model performance. The dataset was divided into two sections: training (60%), and testing (40%). The training set is used to train the model, and the testing set is used to evaluate its generalisation ability. On the testing set, the PCA model achieved a 93.14% accuracy rate. Cross-validation with 10 folds was used to ensure that the model did not overfit. The mean cross-validation accuracy of 92.55% indicates that the model performs consistently across different data subsets. A more detailed analysis of the model's performance and insights into its behaviour will be provided in subsequent sections.

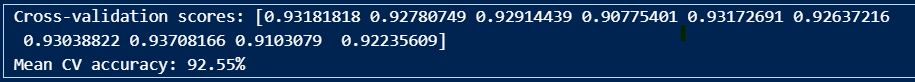


Figure 20: CV scores for XGboost + PCA model

To address potential class imbalance issues in the dataset, SMOTE was applied before training the XGBoost model. SMOTE is an technique that is used to create synthetic instances of the minority class to balance the dataset.This approach combines the strengths of SMOTE in handling class imbalance, PCA for dimensionality reduction, and XGBoost for its powerful classification capabilities. By applying SMOTE before model training, we ensure that the XGBoost classifier learns from a balanced representation of all classes, potentially improving its performance on minority classes.

SMOTE was imported from the imbalanced-learn library and instantiated with a fixed random state to ensure reproducibility resulting in a balanced dataset (X\_res, y\_res). The dataset was split into 70-30 ratio, with a fixed random state = 42 . An XGBoost classifier was trained using the resampled data. Finally, the trained model predicted on the resampled test set (X\_test\_res), completing the implementation of this combined approach to address the imbalance in third-set outcomes and improve classification performance. This model gave is accuracy of 95.5% with majority and minority classes both more balanced. A detailed analysis of this model's performance, including its effectiveness in handling class imbalance and its predictive accuracy for third set outcomes, is presented in the subsequent sections of this report.

The implementation of SMOTE in this context aims to deal with bias towards the “set3\_won” = 1 class and improve the classification of instances from all classes, particularly in scenarios where class distribution is skewed.

# Analysis and Inferences

## Average minutes per match for 2020-2023 and 2000-2003

In analysing the average match duration (in minutes) from Figure 6 : Average Minutes per Tournament Level for 2020-2023 and Figure 7 in Section 6.1 across different tournament levels in the ATP Tour, the data is segmented into two distinct periods: 2000-2003 and 2020-2023. This temporal comparison allows for an examination of trends and potential shifts in match duration over the last two decades. Both periods exhibit a relatively consistent range of average match durations across the years, with a slight increase in the latter period (2020-2023). This could be indicative of evolving playing styles, technological advancements, or changes in tournament structures.

**Variation in Tournament Levels:**

1. **2000-2003:** Tournament levels such as A and F generally feature higher average match durations, often exceeding 120 minutes in some years, particularly in 2002. Lower-tier tournaments (D, G, M) show more variability, with match durations typically ranging between 80 and 120 minutes.
2. **2020-2023:** A similar pattern is observed in the later period, with the top-tier tournaments (A, F) continuing to have the longest matches. Notably, the year 2023 sees a pronounced increase in match duration for tournament level A, peaking around 140 minutes. The duration for lower-tier tournaments has shown a more stable trend, with matches averaging between 90 and 110 minutes.
3. Increased match duration at higher tier tournaments could suggest a greater level of competition with more matches extending to longer duration due to closely contested sets.

## Average Age of winner per tournament across different Levels (2000-2003 vs 2020-2023)

The analysis focuses on the average age of winners in ATP tournaments across different levels (A, D, F, G, M) during two distinct periods: 2000-2003 and 2020-2023. The grouped bar plots provide a visual comparison of how the average age of winners has evolved over these two periods.

From Figure 8 and Figure 9 when we perform comparative analysis:

1. **Winner's Age Stability:** Data from 2000 to 2003 show a consistent trend in winner age, implying a relatively uniform age distribution among top performers across tournament levels.
2. **Emerging Generational Shifts**: In contrast, the 2020-2023 period indicates a generational shift, particularly at the G level, where the average age of winners appears to be decreasing by 2023. This shift may indicate the rise of younger players who are beginning to challenge the dominance of more experienced players at certain tournament levels.
3. **M Level Consistency**: Across both periods, the M level tournaments, which are usually the most prestigious events, have a consistently higher average age. This implies that winning at the highest levels of competition still necessitates extensive experience and maturity.

The evolution of competition from 2000-2003 to 2020-2023 demonstrates how the ATP tour may be in a transition phase, with established players continuing to dominate the most prestigious events while younger players are breaking through, particularly in lower-level tournaments. The analysis of the average age of winners across different tournament levels over these two periods indicates that, while the core of ATP tennis remains dominated by experienced players, particularly in higher-tier tournaments, there is a clear trend of younger players emerging as serious competitors. This shift could be attributed to broader trends in the sport, such as advances in training, recovery, and playing style, which could allow younger players to compete at high levels earlier in their careers. Further analysis over a longer time period could help to determine whether this trend will continue or is a one-time occurrence.

## Winner Height and Rank Analysis in ATP matches

The plot in Figure 10 for winner rank vs Height distribution for year 2024 contains a superimposed scatter plot with red dashes lines indicating the count of players for each height interval, which provides the insights into the distribution of winner heights to their ranking

**Concentration of Heights Around 180-185 cm:**

The plot shows a significant concentration of winners with heights ranging from 180 to 185 cm. This height range appears to be associated with a wide range of rankings, both lower (closer to 0, indicating a better rank) and higher. However, the density of data points indicates that many successful players fall within this height range, making it a common characteristic.

**Higher Ranks and Heights:**

An interesting pattern emerges around 185-190 cm. The plot shows a peak in the number of players within this height interval, but these players tend to have relatively higher rankings (in the 400-1000 range), implying that, while this height is common, players within this range may not always be the best-ranked, possibly due to other factors such as agility, speed, or technique.

**Trends at extreme heights:**

As player height exceeds 190 cm, the winner rank decreases significantly (closer to zero) in Figure 10 , implying that taller players are ranked higher. This trend continues until around 200 cm, when it becomes clear that many top-ranked players fall into this height range. This could imply that a combination of reach, power, and serve effectiveness frequently helps taller players achieve high rankings. There are fewer data points at the extreme ends of the height spectrum (170 cm and >200 cm), suggesting that fewer ATP tournament winners fall into these height categories. For the tallest players (>200 cm), the distribution shows that while there are fewer of them, those who do achieve success frequently have very high ranks.

The analysis of the relationship between height and winner rank in ATP tournaments reveals several important trends that could have significant implications for the sport:

* **Serve Dominance:** The growing success of taller players, particularly those over 190 cm, suggests that serve power and effectiveness are becoming more important in modern tennis. Taller players often benefit from a natural advantage in serving due to their higher trajectory and longer reach, which allows them to generate more power and achieve a greater angle on their serves. This trend may be pushing the sport towards a style of play that emphasises serves and quick points, potentially decreasing the frequency of long rallies.
* **Baseline Play and Reach:** Taller players not only benefit from a powerful serve but also from an extended reach, which can make it easier to cover the court and retrieve difficult shots. This can lead to a more aggressive baseline game, where taller players can dictate play with powerful groundstrokes while also being able to defend against passing shots more effectively.
* **Physical Demands:** While taller players might enjoy certain advantages, the physical demands on their bodies are often greater. The stress on joints, particularly in the knees and lower back, can be significant due to the biomechanics of taller bodies. As a result, these players might be more prone to injuries, which could impact their career longevity. This trend could lead to a higher turnover rate among taller players at the top levels, with fewer of them enjoying long, consistent careers compared to their shorter counterparts.
* **Adaptations in Training and Conditioning:** To counteract these risks, there might be an increased focus on specialized training and conditioning for taller players. This could include targeted exercises to strengthen vulnerable areas, as well as tailored recovery protocols to ensure that these athletes can maintain peak performance while minimizing injury risks.

## Effects of Age parameter over players ranking

In Figure 11 and Figure 12 the scatter plots between age and rank for ATP players for year 2023 and 2021:

**Extended Peak Performance:** Section 6.2 shows a broad distribution of ages among top-ranked players, with a notable number of players in the 28-35 age range still achieving high rankings. Figure 11 indicates that the peak performance age for tennis players has extended, allowing players to remain competitive well into their 30s.

**Densely Populated Younger Age Group:** Similar to previous years, the 18-25 age group remains densely populated with players, many of whom hold high rankings. However, this group now faces more competition from older players, suggesting a more competitive landscape where age is less of a determining factor for success than in the past.

**Presence of Veterans:** The plot highlights the presence of several players over the age of 35 who still maintain competitive rankings. This trend suggests improved longevity in tennis careers, potentially due to advancements in sports science, better injury management, and modern training regimens.

The 2001 scatter plot shows a concentration of younger players, particularly in the 18-25 age range, dominating the top ranks (closer to 0). This indicates that during this period, younger players were more likely to achieve high rankings early in their careers, with fewer older players competing at the highest levels. Players in the 26-30 age range are still present in the rankings but tend to occupy lower ranks compared to their younger counterparts. This suggests that in 2001, players were more likely to experience a decline in performance or ranking as they aged. Following Trends were observed while performing comparative analysis for both the years:

**Shift in Peak Age for Performance:** The comparison between 2001 and 2023 reveals a significant shift in the peak age for performance in ATP tennis. While 2001 was dominated by younger players, the 2023 data indicates that players now maintain high rankings well into their 30s. This shift likely reflects changes in the sport, including advancements in player conditioning, recovery technologies, and strategic adaptations that allow for longer careers.

**Greater Competition Across Age Groups in 2023:** Figure 11 shows a more even distribution of rankings across a broader age range, suggesting that competition has intensified across all age groups. Younger players are no longer the sole occupants of top rankings, as older players have adapted and continue to compete at a high level, leading to a more age-diverse competitive field.

**Impact of Sports Science and Technology:** The increased presence of older players in the top ranks in 2023 may be attributed to advancements in sports science, including improved training methods, nutrition, and injury prevention. These factors likely contribute to the extended careers of players, enabling them to maintain high performance levels even as they age.

**Challenges for Younger Players:** While younger players still populate the top rankings, the increased competition from older players in 2023 suggests that breaking into the top ranks has become more challenging. Younger players may need to develop more comprehensive skill sets earlier in their careers to compete effectively against seasoned veterans.

## Winning trend demonstrated by top seeded ATP players for year 2023

The analysis of Figure 14 and Figure 15 provides a comprehensive view of how the top ATP players achieve their victories and how their success varies across different surfaces. The data highlights the importance of both physical and mental attributes in determining match outcomes. Players like Djokovic, who excel in various winning patterns and across all surfaces, exemplify the modern tennis player’s need for versatility and adaptability. The correlation between a player’s winning pattern and their surface specialization also underscores the strategic differences required to succeed on different courts. Understanding these dynamics can provide deeper insights into player performance and help predict outcomes in different tournament settings.

Figure 14 illustrates the winning patterns of the top 30 ATP players, categorized into four distinct types of victories:

1. **Comeback in a Game of 5 Sets (Red)** represents instances where a player was trailing but managed to stage a comeback to win in a five-set match. Novak Djokovic stands out with the highest count in this category, indicating his resilience and ability to turn around challenging matches.
2. **First Two Sets Win, Wins the Game (Orange)** depicts players who secured victory after winning the first two sets. The data shows that many players, such as Daniil Medvedev, Stefanos Tsitsipas, and Jannik Sinner, frequently win their matches after securing a strong start by winning the first two sets. This suggests that these players are particularly strong front-runners who can maintain their lead effectively.
3. **Comeback in a Game of 3 Sets (Green)** shows the number of times a player has come back to win a match after being down in a three-set match. Players like Alexander Zverev and Carlos Alcaraz have a notable number of wins in this category, indicating their mental toughness and ability to perform under pressure, even in shorter match formats.
4. **Wins When Set 1 Lead Point is Big (Yellow)** This pattern represents victories where a player won the first set by a significant margin and went on to win the match. This type of victory is common across many players, including Andrey Rublev, Hubert Hurkacz, and others, suggesting that a strong start often leads to match success, emphasizing the importance of momentum in tennis.

The diversity in winning patterns across the top players indicates a variety of strategies and psychological resilience in high-pressure situations. For instance, Djokovic's high count in comebacks illustrates his exceptional mental fortitude, while others like Medvedev and Tsitsipas dominate when they establish an early lead.

Figure 15 provides a breakdown of the win percentages of the top 25 ATP players across three different surfaces: grass, hard, and clay.

Grass surfaces typically favour players with strong serves and quick points. The data shows that players like Novak Djokovic and Matteo Berrettini have high win percentages on grass, reflecting their proficiency on this faster surface. The relatively lower percentage of matches played on grass is due to the limited grass-court season, which is concentrated around Wimbledon. Hard courts are the most common surface on the ATP tour, and most players show a strong performance on this surface. Players such as Daniil Medvedev and Alexander Zverev have a high win percentage on hard courts, underscoring their ability to adapt to the surface’s unique characteristics, which balance speed and bounce Clay courts favour players with strong baseline play and endurance. The dominance of players like Rafael Nadal on clay is well-known, although he is not featured in this specific dataset. Among the players shown, Carlos Alcaraz and Casper Ruud exhibit a high win percentage on clay, highlighting their comfort and skill on slower surfaces that require strategic point construction and patience.

The winning patterns observed in Figure 14 can be correlated with the surface specializations in Figure 15. For example, Novak Djokovic’s ability to stage comebacks is well-suited to the varied demands of different surfaces. His high win percentages across grass, hard, and clay in Figure 15 underscore his versatility and adaptability, making him a formidable opponent on any surface. Similarly, players like Carlos Alcaraz, who have high win percentages on clay, are likely to perform well in scenarios that demand endurance and tactical play, as suggested by their winning patterns in Figure 14

## Comprehensive Analysis of winners and losers’ performance metrics

In this section we will analyse key performance metrics and its influence in perspective of a winner and a loser of the matches played in 2023. Figure 16 represents boxplot viz for Serve Efficiency and Ace Points Distributions for winner and loser.

The boxplot for winner serves efficiency (green) shows a higher median value with a more compact distribution compared to the loser serve efficiency (purple). This indicates that winners tend to have more consistent and higher serve efficiency, which likely contributes to their success in matches. The spread of the loser serve efficiency is wider with more outliers, suggesting that losers exhibit greater variability in their serve performance, which might lead to their defeats. The ace points distribution for winners (blue) is skewed towards higher values with some outliers, indicating that winners generally score more aces, which is a crucial factor in gaining the upper hand in matches. In contrast, the ace points distribution for losers (red) is lower and more spread out, with a few positive outliers but generally less impact compared to the winners. This suggests that the ability to score aces effectively is a distinguishing factor between winning and losing.

Figure 17 offers a detailed boxplot analysis of breakpoints saved and serves points won distribution for both winners and losers, focusing on both first and second serves.

The distribution of first serve points won by winners (purple) is higher with a tighter interquartile range compared to losers (red). This demonstrates that winners are generally more effective in winning points off their first serve, which is crucial in setting the tone for the match. Losers show a wider spread and lower median in first serve points won, indicating inconsistency in converting first serves into winning points. The winner's second serve points distribution (green) is also higher compared to the losers (blue). This suggests that even on their second serve, winners are better at winning points, which indicates a robust all-around serving game. Losers have a lower median and wider spread, showing more variability and less effectiveness on their second serves, which could be a significant disadvantage in match play.

The ability to save breakpoints is a key differentiator between winners and losers. Players who can consistently save breakpoints not only maintain their service games but also exert psychological pressure on their opponents, leading to better overall match control. Efficient serving, a strong ace game, and the ability to manage critical points such as breakpoints are the key factors that distinguish winners from losers in competitive matches.

## Surface-Level impact analysis

In this analysis, we aimed to determine whether the mean break point win rate for winners differs significantly across various tennis court surfaces—Clay, Grass, and Hard. The analysis was conducted using a one-way ANOVA, followed by a post-hoc pairwise comparison using Tukey's HSD test to identify specific differences between the surfaces.

The ANOVA test was performed to compare the mean break point win rates across the three surfaces. The results are as follows:

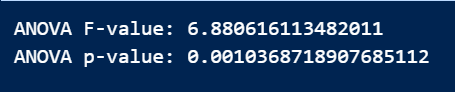


Figure 21: Anova test results for break point win rate

The p-value is less than the alpha level of 0.05, indicating that there is a statistically significant difference in the mean break point win rates across at least one pair of surfaces. Therefore, we reject the null hypothesis (H0) that the mean break point win rates for winners are the same across all surfaces.

To pinpoint which specific surface pairs exhibit significant differences, a Tukey's HSD test was conducted. The results of the Tukey test are as follows:

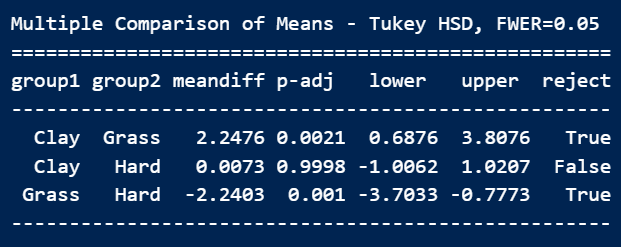


Figure 22: Tukey HSD pairwise comparison for Break point saved

**Clay vs. Grass:** The significant difference suggests that the slower nature of Clay courts, which often leads to longer rallies and more baseline play, allows players more opportunities to defend break points effectively, compared to the faster Grass courts.

**Grass vs. Hard:** The significant difference here highlights that Hard courts, with their balanced speed, provide players with more consistent bounce and predictability, possibly aiding in better break point defence compared to Grass, which is faster and has a lower bounce.

**Clay vs. Hard:** The lack of significant difference between Clay and Hard courts suggests that while these surfaces differ in speed, their impact on break point defence may be mitigated by other factors such as player adaptability and strategy.

Hard courts are typically seen as the most balanced surface in tennis. They offer a medium pace with consistent bounce, which can aid in breakpoint defence due to the predictability of the surface. Grass courts, on the other hand, are faster and lower-bouncing, which can make it harder to defend break points as points are often shorter, and the serve becomes a more dominant factor. Clay courts are indeed slower and typically lead to longer rallies, favouring players who excel in baseline play and defence. This generally provides more opportunities to defend break points effectively, as players have more time to set up their shots and recover between points.

Similarly ANOVA tests were conducted for Ace% for winners across all surfaces and here are the results :



Figure 23: ANOVA test results for Ace% for winner

To identify where these differences lie, we conducted a pairwise comparison using the Tukey HSD test. The results are as follows:

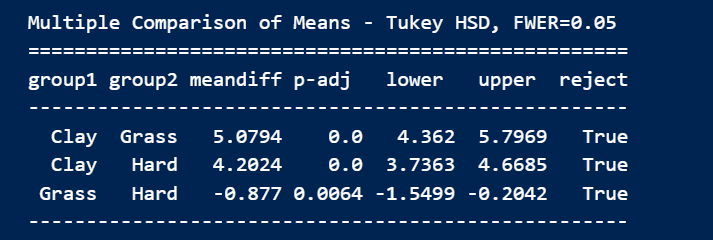


Figure 24: Tukey HSD pairwise comparison for Ace point

There is a significant difference between Clay and Grass surfaces (**Mean Difference:** 5.0794), with Grass courts showing a higher ace percentage. This makes sense given the faster nature of Grass courts, which favour strong serves and often lead to more aces. A significant difference was also observed between Clay and Hard courts (**Mean Difference:** 4.2024), with Hard courts yielding more aces than Clay. This is consistent with the fact that Hard courts are faster and more conducive to powerful serves compared to the slower Clay courts.

The analysis confirms that the type of court surface significantly impacts the percentage of aces served during tennis matches. Specifically, **Grass courts** allow for the highest percentage of aces due to their fast nature and low bounce, making it easier for players to win points directly off their serve. **Hard courts** also support a high ace percentage, though slightly less than Grass, due to their balanced speed and consistent bounce. **Clay courts** have the lowest ace percentage, consistent with their slower pace and higher bounce, which gives returners more time to react to serves. These findings are consistent with the known characteristics of each surface and provide insight into how players might adjust their serving strategies based on the surface they are playing on.

## Two-Way ANOVA on Surface and Era Interaction

This analysis is to investigate whether the effect of different tennis surfaces on the break point save percentage (bp\_saved\_pct) has changed across two distinct eras. Specifically, the analysis aims to determine if surfaces like Grass and Clay, which traditionally have different characteristics, have become more similar in their impact on break point save percentages over time.

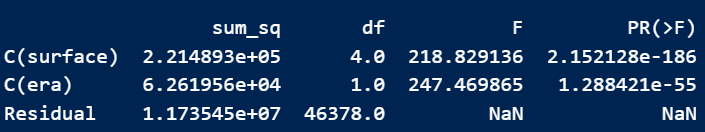
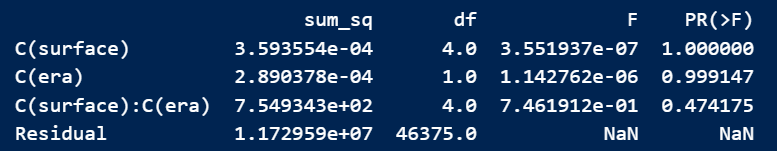


Figure 25: bp\_saved\_pct ~ C(surface) + C(era)

Figure 26: bp\_saved\_pct ~ C(surface) + C(era)

In Figure 26 , the main effects of surface and era on the break point save percentage were analysed without considering their interaction we obtained a F-value 218.8194 and P-value of 2.15x 10-186. The very low p-value indicates that the type of surface significantly affects the break point save percentage. This suggests that the inherent characteristics of each surface (e.g., speed, bounce) continue to play a crucial role in determining how effectively players can save break points. Similarly, the era has a significant effect, indicating that there have been changes over time that impact break point saves percentages. These changes could be attributed to shifts in playing styles, advancements in equipment, or improvements in player conditioning and strategy.

In Figure 25, second model introduced an interaction term between surface and era to explore whether the impact of surfaces like Grass and Clay on break point save percentages has changed over time. Surface\* Era interactions give us a result of F-value of 0.7461 and p\_value of 0.4742. The interaction effect between surface and era is not significant, as indicated by the high p-value. This suggests that the effect of different surfaces on break point save percentages has remained consistent across the two eras. In other words, surfaces like Grass and Clay have not become more similar in their impact over time. The traditional differences between these surfaces in terms of how they affect play (e.g., the fast, low-bouncing nature of Grass vs. the slow, high-bouncing nature of Clay) appear to persist regardless of the era.

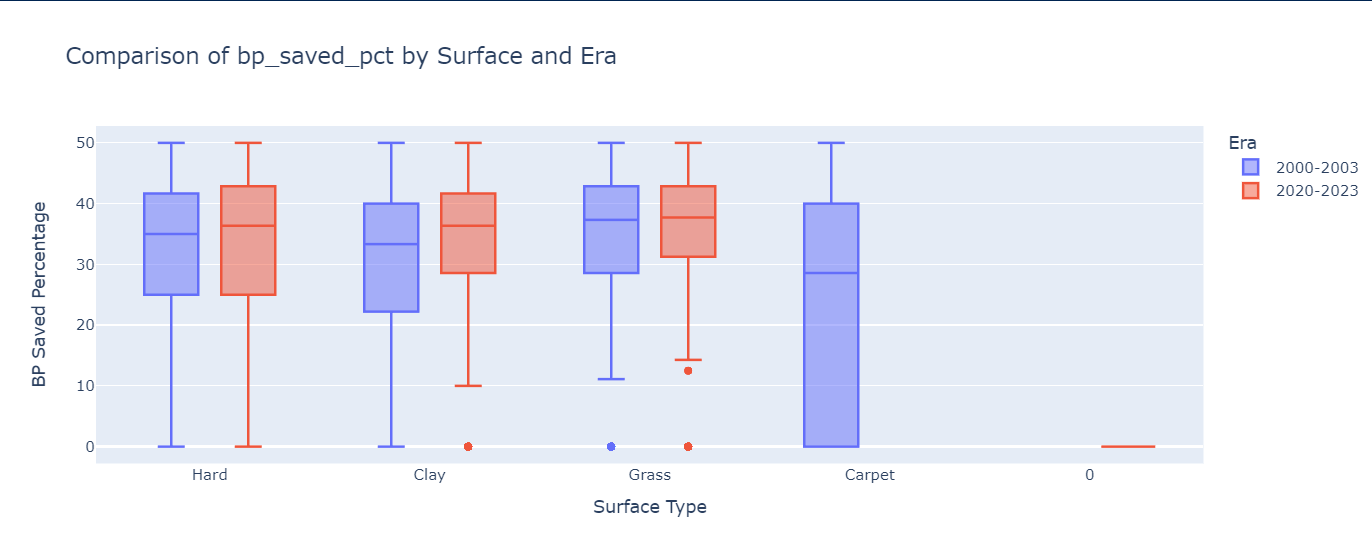


Figure 27: Boxplot viz for breakpoints saved over different surfaces in 2 eras

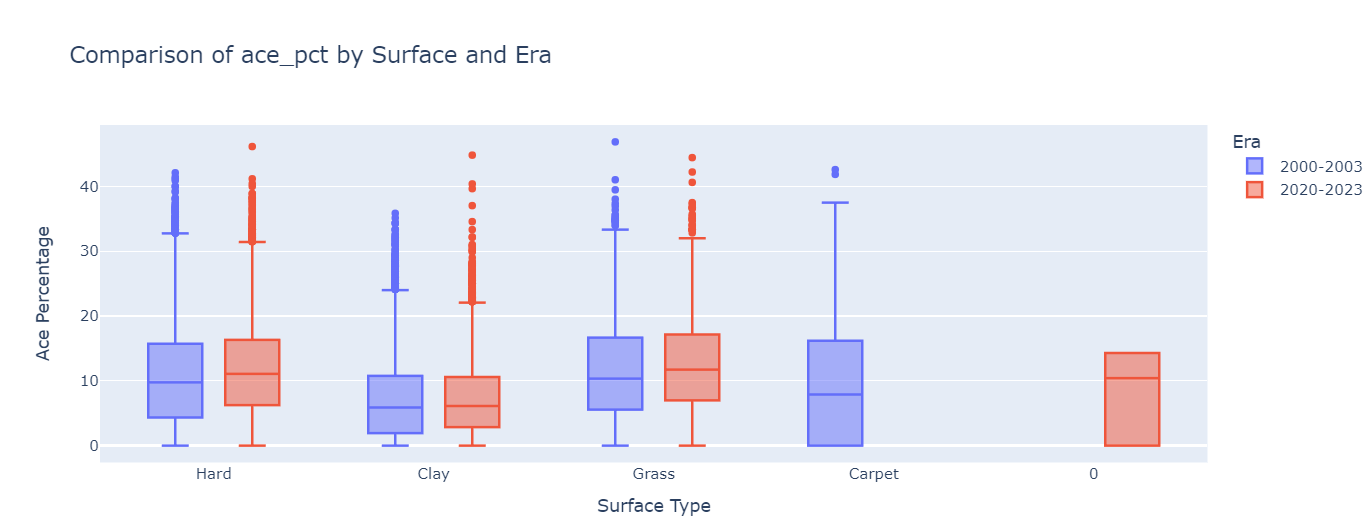


Figure 28: Boxplot viz for acepoints saved over different surfaces in 2 eras

The two boxplots provided in Figure 27 & Figure 28 offer a visual representation of the break point save percentage (bp\_saved\_pct) and ace percentage (ace\_pct) across different surfaces (Hard, Clay, Grass, and Carpet) and across two distinct eras (2000-2003 and 2020-2023). These visualizations help to further support and validate the findings from the two-way ANOVA regarding the influence of surface and era on tennis performance metrics.

## Summary of Hidden Markov Model (HMM) Analysis and Implementation

The primary objective of implementing a Hidden Markov Model (HMM) in this context was to analyse momentum shifts during tennis matches from both the winning and losing players' perspectives. The HMM was designed to identify underlying "hidden" states that correspond to different phases of momentum within the match. By mapping these states to observable performance metrics (e.g., ace percentage, break points saved, serve efficiency), the goal was to gain deeper insights into the dynamic nature of tennis matches, beyond the surface-level match scores. Interpretation of States in our Hidden Markov Model are as follows:

1. **State [0]**: Likely represents a neutral or transitional phase where the player is neither fully in control nor losing momentum.
2. **State [1]**: Generally, reflects a challenging phase where the player might be under pressure or struggling to maintain momentum.
3. **State [2]**: Indicates a phase of positive momentum or control, where the player is likely performing well and is in a dominant position.

Let's analyse the Hidden Markov Model (HMM) states for each player in the given matches. The states will help us understand the momentum shifts experienced by both winning and losing players during the matches. Lets us consider a Match ID: 2023-0301296

**Player 104755 (Winning Player): Hidden Markov States:** [1]

**Explanation:** Player 104755 is in state [1] throughout the match. State [1] likely represents a challenging phase where the player was under pressure, possibly struggling with momentum. Despite being in this state, the player managed to win the match with a score of 1-6, 6-1, 6-1. This suggests that although the player faced difficulties early on (losing the first set 1-6), they managed to regain control and secure the victory, even while remaining in a state that might typically indicate challenges.

**Player 105676 (Losing Player): Hidden Markov States:** [2]

**Explanation:** Player 105676 is in state [2], which often indicates a phase of control or positive momentum. However, despite being in this state, the player lost the match. The scoreline (1-6, 6-1, 6-1) suggests a complete reversal of momentum after the first set, where the losing player dominated the first set but then lost control. This paradoxical situation might indicate that the state [2] was more reflective of the initial set's momentum, but the player couldn't maintain it across the remaining sets.

**Shortcomings of the Present Model:**

With only 3 hidden states, the model might oversimplify the complex dynamics of a tennis match. The model assumes that transitions between states are consistent throughout the match. However, the early stages of a match might have different momentum dynamics compared to later stages (e.g., match point situations). A more advanced model might incorporate varying transition probabilities depending on the match phase. There were instances where a player was in state [2] (interpreted as positive momentum) but still lost the match. This indicates that the state may not fully capture the complexity of momentum, especially when a player is performing well but still gets outplayed in critical moments.

The Hidden Markov Model provided valuable insights into the momentum dynamics of tennis matches, revealing how players transition through different phases of control and struggle. However, the current implementation has several limitations, including a potentially oversimplified state model, limited feature selection, and a lack of context sensitivity

## Interpretation of Machine Learning model results

In this section we will discuss about the results obtained from our XGboost classification model from Section 8.2 predicting the winner of the third set. The initial XGBoost model was built using all 28 features without any feature engineering or dimensionality reduction. While XGBoost is known for its ability to handle complex datasets, the model showed signs of overfitting, as indicated by a discrepancy between training and validation performance.

To address the overfitting, PCA was applied, reducing the feature set from 28 to 7 principal components. This step aimed to retain the most significant information in the data while removing redundant features that could contribute to overfitting. After applying PCA, the model's accuracy improved to 93.14%, with a mean cross-validation accuracy of 92.55%. This suggests that PCA successfully reduced the model's complexity while retaining the most relevant information.

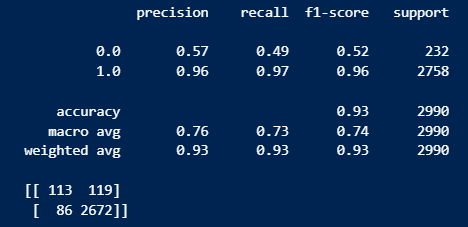


Figure 29: Classification Report for XGboost + PCA model

The classification report as shown in Figure 29 , however, showed a disparity in performance between the majority class (label 1) and the minority class (label 0). While the model performed well for the majority class, the minority class had significantly lower precision, recall, and F1-score, indicating that the model was biased towards predicting the majority class. To counter the imbalance in the target classes, SMOTE was applied to generate synthetic samples for the minority class. This balanced the dataset and allowed the model to learn more effectively from the minority class instances. Post-SMOTE, the model's accuracy further improved to 95.50%, with a balanced classification report showing high precision, recall, and F1-scores for both classes

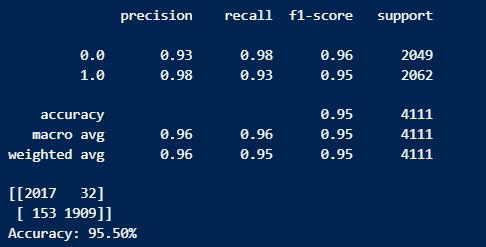


Figure 30: Classification report after applying SMOTE on Class[0]

**Shortcomings and Potential Improvements**

While PCA helped reduce dimensionality, further feature engineering could be explored to derive new features that capture more nuanced aspects of tennis matches, potentially improving model performance further. Although the model achieved high accuracy, XGBoost complexity means that it can still be prone to overfitting, especially if the dataset were to grow or change. Regularization techniques or model pruning could be considered to further mitigate this risk. Introducing contextual features such as player fatigue, historical performance in similar match conditions, or psychological factors could enhance the model’s ability to predict outcomes in real-world scenarios where such factors play a critical.

# Conclusion and Future work

This dissertation set out to explore and predict the winner of the third set in ATP tennis matches, recognizing the critical role this set plays in determining the outcome of a match. Through comprehensive analysis and the application of machine learning techniques, several key insights were uncovered. The research demonstrated significant differences in the age distribution of players, highlighting how player demographics can influence match outcomes. Additionally, the analysis provided strong evidence of the impact of the era and surface type on crucial match statistics, such as break points saved, underscoring how the playing conditions and time period can shape performance.

Moreover, the interaction between surface and era on match statistics was explored, revealing complex dynamics that influence player success. This interaction was initially examined by comparing two distinct time periods, with future research potentially extending this analysis to cover a continuous timeline, offering a more detailed understanding of how these factors evolve over time.

A Hidden Markov Model (HMM) was also implemented to capture the hidden states of momentum shifts during a match. While the model showed promise in identifying key momentum phases, further refinement and the inclusion of additional data could enhance its predictive power, making it a valuable tool for real-time match analysis.

Looking forward, future work could delve deeper into the interaction between surface and era, and explore the HMM in greater detail, potentially integrating more contextual variables to improve the model's accuracy and applicability. The findings of this dissertation not only contribute to the field of sports analytics but also provide practical insights that could influence strategic decision-making in professional tennis.

# References

Choudhary, P., Dubey, S., Brijwal, D. & Paswan, R., 2023. A statistical model to predict the results of Novak Djokovic's matches in the Australian open tennis event using the binary logistic regression. *International Journal of Statistics and Applied Mathematics.*

Drive-analytics, n.d. *The Gradual yet steady rise of Tennis Analytics.* [Online]   
Available at: https://www.drive-analytics.com/tennis.php

Havlucu Hayati, A. B. ,. E. T. ,. C. A. ,. O. O., 2022. Toward Detecting the Zone of Elite Tennis Players Through Wearable Technology. *Frontiers in Sports and Active Living,* Volume 4.

Hyatt-Twynam, S., 2013. *The logit function and its applications in sports modelling : Sports Trading Network.* [Online]   
Available at: https://www.sportstradingnetwork.com/article/logit-function-applications-sports-modelling/

JeffSackmann, 2016. *Github JeffSackmann.* [Online]   
Available at: https://github.com/JeffSackmann/tennis\_atp

Jernej, R. & Ziva, M. R., 2021. Skill Level in Tennis Serve Return Is Related to Adaptability in Visual Search Behavior. *Frontiers in Psychology.*

Knutson, T., 2016. *Visualizing Professional Tennis Upsets: ATP 2012-2014 Men's Singles Matches.* [Online]   
Available at: https://nycdatascience.com/blog/student-works/visualizing-professional-tennis-upsets-atp-2012-2014-matches/

Kolonias, I., Christman, W. J. & Kittler, J., August 2004. Tracking the Evolution of a Tennis Match Using Hidden Markov Models. *Lecture Notes in Computer Science.*

Liu, H., 2024. Momentum Analysis of Tennis Matches Based on Logistic. *Highlights in Science, Engineering and Technology,* Volume 101.

Manuel, J., March 4 2022. *Capturing Momentum in Tennis.* [Online]   
Available at: https://theanalyst.com/eu/2022/03/capturing-momentum-in-tennis

Panjan, A., Sarabon, N. & Filipcic, A., June 2010. Prediction of the successfulness of tennis players with machine learning methods. *Kinesiology.*

Saaty, T. L. & Gu, W., June 2019. Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments. *Journal of Systems Science and Systems Engineering 28(3):317-343,* Volume 28, pp. 317-343.

Sampaio, T. et al., June, 2024. Applications of Machine Learning to Optimize Tennis Performance: A Systematic Review. *Applied Sciences.*

Srivastava, S., 2023. *appinventiv.* [Online]   
Available at: https://appinventiv.com/blog/data-analytics-in-sports-industry/

Srivastav, S., 2023. *The Business Opportunities and Challenges of Implementing Analytics in Sports.* [Online]   
Available at: https://appinventiv.com/blog/data-analytics-in-sports-industry/

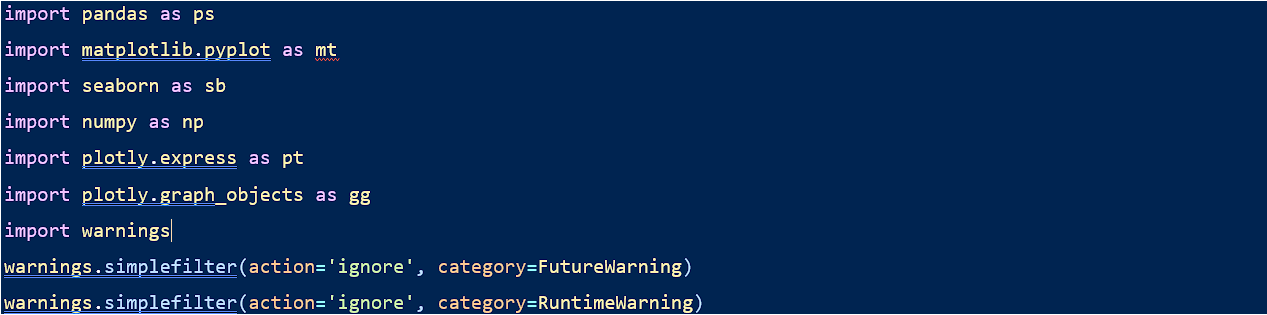
Wu, W., May 2024. Quantification of momentum in tennis matches and its impact: a study based on AHP-EWM method and data analysis. *Highlights in Science Engineering and Technology.*

Yang, J. L. Z., January 2023. The Biomechanical Analysis on the Tennis Batting Angle Selection Under Deep Learning. *IEEE Access,* Volume 11.

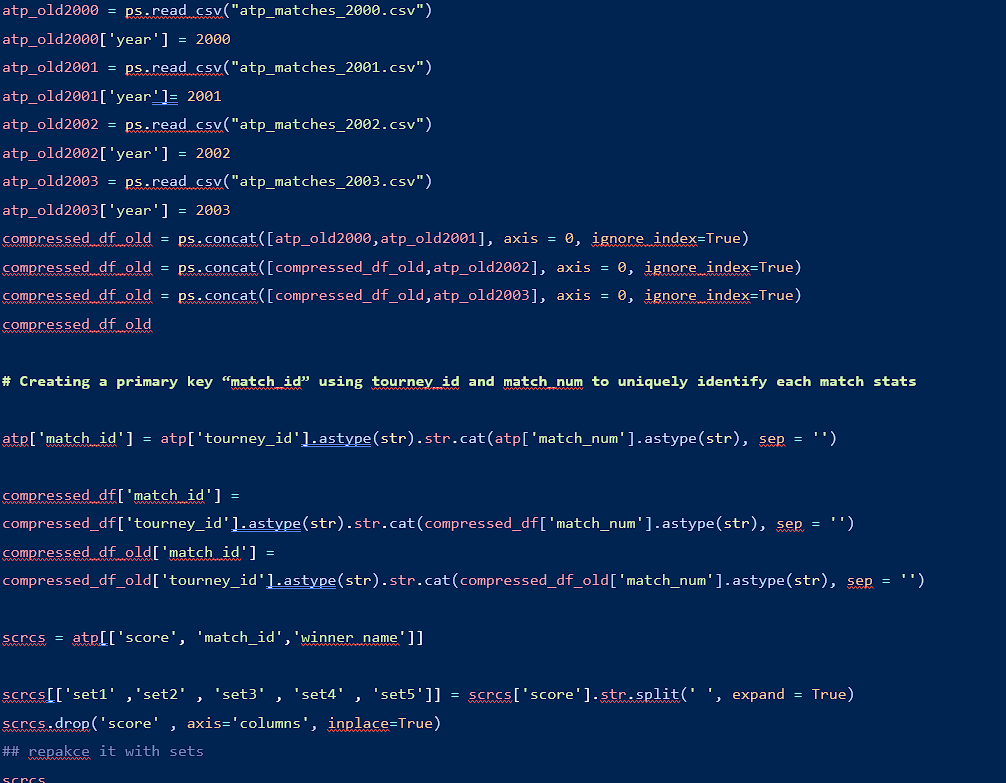
Yixiong Cui, H. L.-Á. G., 2019. Data-driven analysis of point-by-point performance for male tennis player in Grand Slams. *revistas,* Volume 15.

Yu Zhang, 2023. The optimization of college tennis training and teaching under deep learning,. *Heliyon,* 10(4).

# Appendices



Reading dataset from csv



Function to create set game difference for winners

def set1\_gm\_diff(scrcs):

    for i in range(scrcs.shape[0]):

        scores = scrcs['set1'][i]

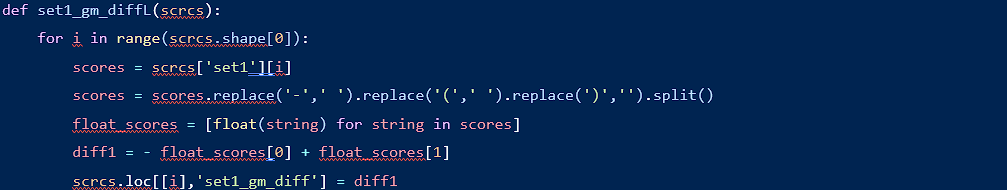
        scores = scores.replace('-',' ').replace('(',' ').replace(')','').split()

        float\_scores = [float(string) for string in scores]

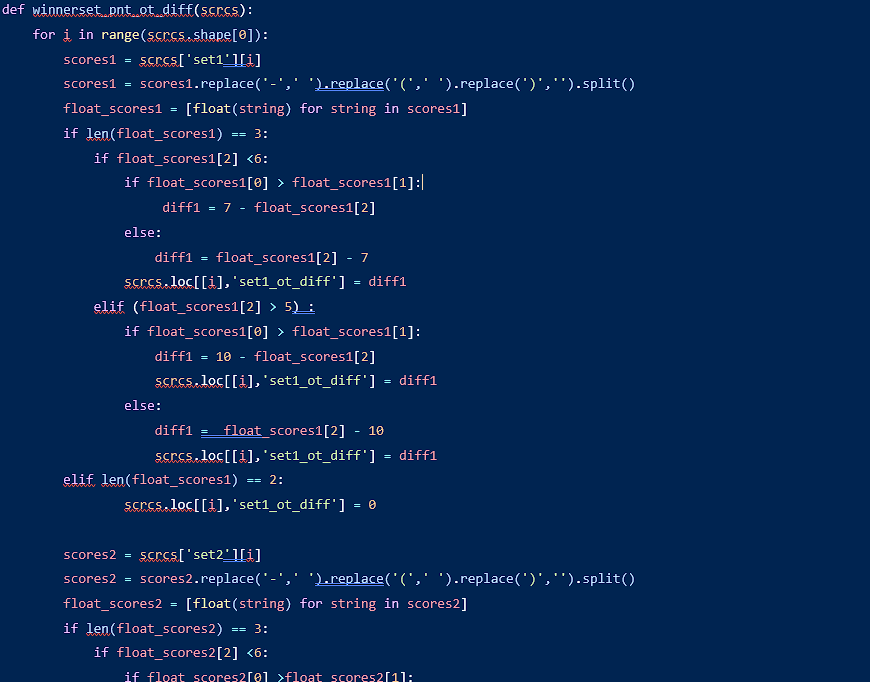
        diff1 = float\_scores[0] - float\_scores[1]

        scrcs.loc[[i],'set1\_gm\_diff'] = diff1

Function to create set game difference for losers



Function for Winner overtime set game difference



diff2 = 7 - float\_scores2[2]

                else:

                     diff2 =  float\_scores2[2] -7

                scrcs.loc[[i],'set2\_ot\_diff'] = diff2

            elif (float\_scores2[2] > 5):

                if float\_scores2[0] >float\_scores2[1]:

                    diff2 = 10 - float\_scores2[2]

                    scrcs.loc[[i],'set2\_ot\_diff'] = diff2

                else:

                    diff2 = float\_scores2[2] - 10

                    scrcs.loc[[i],'set2\_ot\_diff'] = diff2

        elif len(float\_scores2) == 2:

                scrcs.loc[[i],'set2\_ot\_diff'] = 0

        scores3 = scrcs['set3'][i]

        scores3 = scores3.replace('-',' ').replace('(',' ').replace(')','').split()

        float\_scores3 = [float(string) for string in scores3]

        if len(float\_scores3) == 3:

            if float\_scores3[2] <6:

                if float\_scores3[0] >float\_scores3[1]:

                    diff3 = 7 - float\_scores3[2]

                else:

                     diff3 =  float\_scores3[2] -7

                scrcs.loc[[i],'set3\_ot\_diff'] = diff3

            elif (float\_scores3[2] > 5):

                if float\_scores3[0] >float\_scores3[1]:

                    diff3 = 10 - float\_scores3[2]

                    scrcs.loc[[i],'set3\_ot\_diff'] = diff3

                else:

                    diff3 = float\_scores3[2] - 10

                    scrcs.loc[[i],'set3\_ot\_diff'] = diff3

        elif len(float\_scores3) == 2:

                scrcs.loc[[i],'set3\_ot\_diff'] = 0

        scores4 = scrcs['set4'][i]

        scores4 = scores4.replace('-',' ').replace('(',' ').replace(')','').split()

        float\_scores4 = [float(string) for string in scores4]

        if len(float\_scores4) == 3:

            if float\_scores4[2] <6:

                if float\_scores4[0] >float\_scores4[1]:

                    diff4 = 7 - float\_scores4[2]

                else:

                     diff4 =  float\_scores4[2] -7

                scrcs.loc[[i],'set4\_ot\_diff'] = diff4

            elif (float\_scores4[2] > 5):

                if float\_scores4[0] >float\_scores4[1]:

                    diff4 = 10 - float\_scores4[2]

                    scrcs.loc[[i],'set4\_ot\_diff'] = diff4

                else:

                    diff4 = float\_scores4[2] - 10

                    scrcs.loc[[i],'set4\_ot\_diff'] = diff4

        elif len(float\_scores4) == 2:

                scrcs.loc[[i],'set4\_ot\_diff'] = 0

        scores5 = scrcs['set5'][i]

        scores5 = scores5.replace('-',' ').replace('(',' ').replace(')','').split()

        float\_scores5 = [float(string) for string in scores5]

        if len(float\_scores5) == 3:

            if float\_scores5[2] <6:

                if float\_scores5[0] >float\_scores5[1]:

                    diff4 = 7 - float\_scores5[2]

                else:

                     diff4 =  float\_scores5[2] -7

                scrcs.loc[[i],'set5\_ot\_diff'] = diff4

            elif (float\_scores5[2] > 5):

                if float\_scores5[0] >float\_scores5[1]:

                    diff4 = 10 - float\_scores5[2]

                    scrcs.loc[[i],'set5\_ot\_diff'] = diff4

                else:

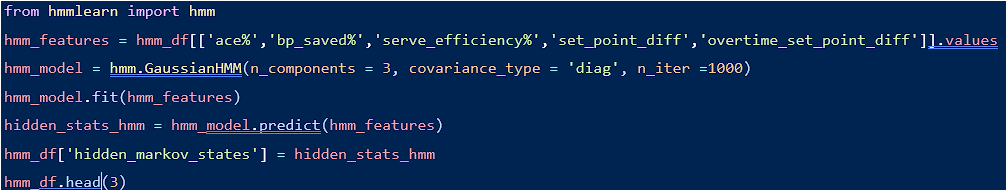
                    diff4 = float\_scores5[2] - 10

                    scrcs.loc[[i],'set5\_ot\_diff'] = diff4

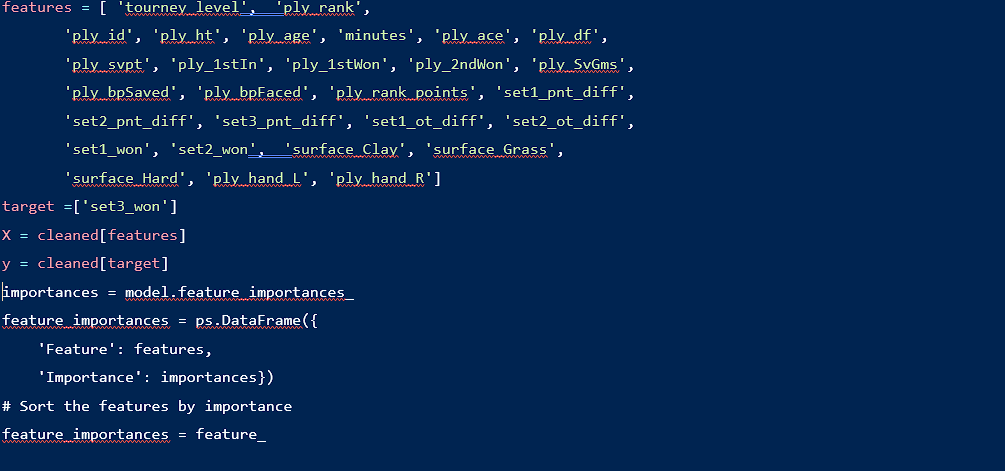
        elif len(float\_scores5) == 2:

                scrcs.loc[[i],'set5\_ot\_diff'] = 0

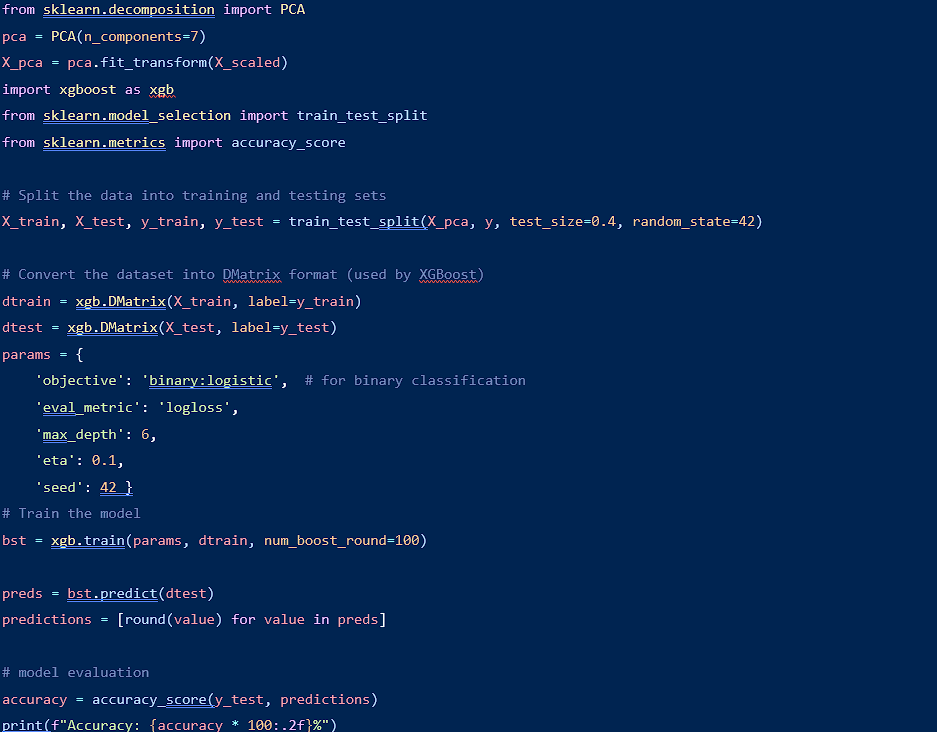
Hidden Markov Model code



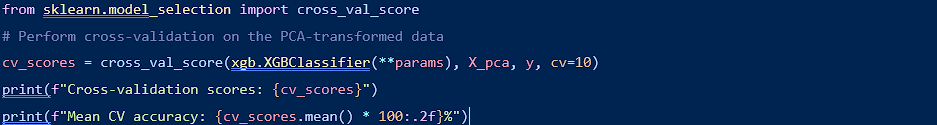
Feature Selection and Target variable for ML model



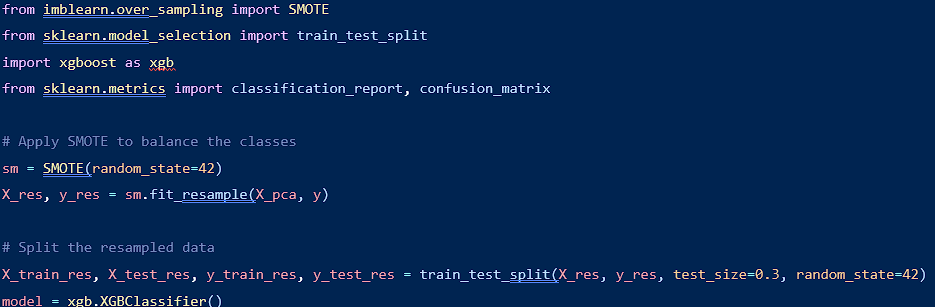
XGboost with PCA



Cross Validation



SMOTE implementation over XGBoost + PCA

model.fit(X\_train\_res, y\_train\_res)

y\_pred\_res = model.predict(X\_test\_res)

print(classification\_report(y\_test\_res, y\_pred\_res))

print(confusion\_matrix(y\_test\_res, y\_pred\_res))

accuracy = accuracy\_score(y\_test\_res, y\_pred\_res)

print(f"Accuracy: {accuracy \* 100:.2f}%")