DEGREE PROJECT

Visualization of Golf Performance Data

Authors: Ahmed Raza & Shweta Singh

Supervisor: Anton Kalén

Master Programme in Data Science 2024

Luleå University of Technology
Department of Computer Science, Electrical and Space Engineering

Table of Contents

Table of Contents **2**
1. Abstract **3**
1.1 Keywords 3
2. Introduction **4**
3. Literature Review **7**
4.1 Objective 10
5. Methodology **12**
5.1 Tools and Libraries 16
5.2 Feature Importance Analysis 17
5.2.1. Random Forest Model 18
5.2.2. Decision Tree Model 19
5.2.3 SHAP Analysis 19
5.3 Ethical and Sustainability 21
6. Result **22**
6.1 Distribution of Players by Gender 22
6.2 Fairways Hit and Greens in Regulation (GIR) Analysis 22
6.3 Interactive Putting Performance Analysis Over Time 23
6.4 Interactive Player Performance Analysis by Shot Location Over Time 24
6.5 Interactive Scoring Trends Over Selected Rounds 25
6.6 Interactive Shot Outcome Analysis by Starting Location 26
6.7 Interactive Shot Performance Analysis Over Selected Rounds 27
6.8 Interactive Shot Performance Analysis by Location 28
6.9 Interactive Shot Performance Analysis by Starting Distance 29
6.10 Selected Player's Strokes Gained vs Average Strokes Gained 30
6.11 Best Player's Strokes Gained vs. Average Strokes Gained 31
6.12 Performance Factors in Golf 32
6.13 Shot Distance and Strokes Gained Net Distribution 33
6.14 Correlation Analysis of Strokes Gained Variables 34
6.15 Feature Importance Results 35
6.15.1 Random Forest Model 35
6.15.2 Decision Tree Model 35
6.16 SHAP Analysis Results 38
8. Conclusion **46**
9. Limitation & Future Work **47**
10. References **48**
11. Appendix **50**

1. Abstract

Golf is a sport where performance metrics play a crucial role in understanding player capabilities and success. This thesis explores various visualization and machine learning techniques to analyze and interpret golf performance data. By visualizing key metrics such as wins, strokes, putting accuracy, and tee-to-green performance, this study aims to uncover insights into player performance trends, correlations between different aspects of the game, and outliers. Additionally, the analysis will highlight the importance of specific features in predicting overall performance, helping to identify which aspects of the game most significantly impact success. The findings from this analysis can inform coaches, players, and analysts in optimizing training strategies, improving performance, and making data-driven decisions in golf.

The project will involve the creation of intuitive visualizations that offer actionable insights into various performance metrics. By leveraging these techniques, the project seeks to provide tailored benchmarks that account for the diverse demographics within the golfing community. This thesis endeavors to improve the accessibility and usability of golf performance data, thereby empowering elite golfers to make informed decisions regarding their training and competition strategies. Through a combination of innovative visualizations and robust benchmarks, this project aims to advance the field of golf performance analysis and contribute to the success of Swedish elite golfers on the international stage.

1.1 Keywords

Team sports, Data analytics, Visualization of Golf performance, Decision making, Python programming language

2. Introduction

In recent years, the intersection of sports and data analytics has revolutionized the way athletes and coaches approach performance improvement. With the advent of advanced tracking technologies and data visualization tools, athletes now have access to good information and insight into their performance, enabling them to make more informed decisions and optimize their training strategies. Exploring competitive sports data visualization represents a growing area of interest within the realm of information visualization, serving as a crucial foundation for examining human behavioral patterns and activity tendencies (Du & Yuan, 2021) [1]. Data visualization serves as a crucial tool for enabling individuals to comprehend the significance of data by summarizing and presenting it in a clear and concise manner (Qin et al., 2020) [2]. It entails the process of representing large datasets in an accessible, understandable, and straightforward format. By transforming data from various sources into visual elements such as length, location, shape, and color, data visualization provides a powerful means of presenting compelling data narratives, particularly for visually oriented individuals (Qin et al., 2020) [2]. In the sport of golf, data analysis plays a crucial role in understanding player performance, identifying strengths and weaknesses, and making informed decisions to enhance overall team performance. Golf visualization uses mental imagery to create a detailed picture of a desired outcome or scenario in the golfer’s mind. By visualizing different scenarios, shots, or holes, golfers can prepare for various situations on the course and develop a stronger mental game.

In the chapter, (Link D, 2024) [3] delineates the aims and techniques for visualizing information within the realm of sports. The primary purpose of sports visualizations lies in the identification of performance data patterns, trends, or anomalies, thereby facilitating the expeditious addressing of sport-specific inquiries.

A golf course comprises several distinct parts, each presenting its challenges and opportunities for players. Starting at the tee box, golfers aim to navigate the fairway, a clear path leading to the green. Along the way, they encounter rough terrain and strategically placed bunkers, requiring precision and skill to overcome. Water hazards add further complexity, testing golfers' accuracy, and decision-making. Finally, the green provides the ultimate challenge, with its finely manicured surface demanding mastery of

the putting stroke. By understanding these elements, golfers can strategically navigate the course, optimizing their performance and achieving success.

The primary aim of this thesis project is to empower golfers to comprehensively comprehend and elevate their game proficiency through an intuitive and accessible visual interface. The visual analysis will address the challenges that golfers face when trying to improve their performance by providing real-time feedback and performance analysis, assisting golfers in identifying areas for improvement and allowing coaches to design more effective training strategies. This project aspires to elevate the efficacy and precision of performance-enhancement strategies within the golfing community.

Fig. 1: Golf Course Layout

Golf course typically consists of several key parts:

1. **Tee Box:** This is where each hole begins. It's a designated area from which players hit their first shot. Tee boxes vary in size and may have different markers indicating the level of difficulty or the distance to the hole.
2. **Fairway:** The fairway is the close Short grass between tee and green that shows the actual lay-out of the hole. It provides a clear path to the hole and is usually the ideal place to land the ball after tee shot.
3. **Rough:** The rough is the longer grass bordering the fairway. It's more challenging to play from than the fairway because the grass is thicker and often uneven.
4. **Bunkers (or Sand Traps):** These are areas filled with sand, strategically placed along the fairway or near the green. They add challenge to the game, as hitting the ball into a bunker requires skill to get it back onto the green.
5. **Water Hazards:** These are areas of the course where water comes into play, such as ponds, lakes, or streams. Balls hit into water hazards usually incur a penalty.
6. **Putting Green:** This is the area around the hole where the grass is very short and smooth. It's where players use their putters to roll the ball into the hole.
7. **Hole:** Each hole on a golf course consists of a teeing ground, fairway, rough, hazards, and green, and ends with a hole where the flagstick is placed. Golfers aim to complete each hole in as few strokes as possible.

Moreover, data visualization plays a pivotal role in transforming complex datasets into intuitive visual representations. Through charts, graphs, and interactive dashboards, golfers can gain a comprehensive understanding of their performance and can get more accurate results of their game.

3. Literature Review

In recent years, the field of sports science has witnessed a growth in research interest focused on the development of innovative sports training systems tailored to enhance the skills and performance of beginner athletes. Recent studies have shown a growing interest in developing sports training systems aimed at assisting beginners in improving their skills.

In today's data-driven corporate environment, data visualization plays a critical role, aiding decision-making closely linked to key revenues in many industrial businesses (Muskan & Chander Prabha, 2022)[4]. The process of data visualization involves formulating a research question, gathering and cleaning data, selecting appropriate chart types and tools, and finally creating visual representations. As the demand for effective and accurate data processing grows, database specialists are increasingly sought after to handle large volumes of data efficiently (Muskan & Chander Prabha, 2022) [4]. Data visualization involves presenting data in graphical or pictorial form, making information easy to understand and aiding in the explanation of facts and determination of courses of action. Its application extends to various fields of study that require innovative ways of presenting large, complex information. The evolution of computer graphics has significantly influenced modern visualization techniques. This paper provides a concise introduction to the fundamentals of data visualization (Sadiku, M. N. O., Adediran, A. A., & Musa, I. A., 2016) [5]. This study explores various phases of data visualization, its types, and numerous applications, facilitating description, hypothesis generation, and idea development from large datasets, contributing to scientific research and prediction.

The extraction of vast amounts of data in every major sport has become commonplace, facilitating its utilization by various stakeholders including media, fans, athletes, and organizations. Collaboration with leading technology vendors has further emphasized the immense value of sports analytics. The pervasive nature, diverse content, and relative accessibility of sports data have made it an enticing domain for visualization researchers. Motivated by the exponential growth and widespread popularity, as well as the untapped potential of sports data visualization, this special issue compiles cutting-edge research in this emerging field Basole, (R. C., & Saupe, D., 2016)[6]. The chapter, (James, 2007)[7] highlights the limited research in golf performance analysis, emphasizing the need for advancements in performance indicators and statistical

techniques. The analysis encompasses various types of data utilized in golf performance assessment, including round scores and specific performance indicators deemed indicative of success. Moreover, it emphasizes the necessity for improvement in these indicators to more accurately reflect golfers' abilities and achievements.

This aligns with the proposal by (Liao et al., 2022)[8] for a golf swing analysis tool leveraging neural networks to aid in skill acquisition by identifying key differences between novice and professional motions. By utilizing unsupervised learning methods, the proposed system offers a promising approach for enhancing skill acquisition through targeted self-training. However, while neural networks show potential in this context, further research is needed to explore their effectiveness and potential limitations in sports skill acquisition.

In addition to individual skill development, sports data visualization has emerged as a crucial aspect of performance analysis. (Perin et al., 2018)[9] provide a comprehensive overview of recent advancements and challenges in sports data visualization, highlighting the importance of innovative visualization techniques and close collaboration with domain experts. By adopting these approaches, researchers have been able to address real-world challenges in sports analysis and decision-making, offering intuitive insights into complex sports data. However, despite these advancements, there remain critical research gaps that need to be addressed, such as the need for standardized visualization methods and tools that can be applied across different sports and skill levels. Exploratory data analysis (EDA) stands as a cornerstone of modern data science, offering indispensable insights into datasets through interactive exploration (Hullman & Gelman, 2021)[10]. Exploratory data analysis (EDA) is an essential component of data science, offering valuable insights into datasets through interactive exploration (Hullman & Gelman, 2021)[10]. By leveraging sophisticated software interfaces optimized for pattern discovery and data exposure, EDA enables analysts to uncover meaningful patterns and trends within complex datasets (Hullman & Gelman, 2021)[10]. This process emphasizes flexibility and discovery, yet it necessitates a grounding in theories of graphical inference to ensure valid interpretations and effective communication of findings. This approach, characterized by its emphasis on flexibility and discovery, is facilitated by sophisticated software interfaces optimized for pattern discovery and data exposure (Hullman & Gelman, 2021)[10]. However, effective EDA necessitates a solid

grounding in theories of graphical inference to ensure valid interpretations and effective communication of findings.

Data visualization techniques are utilized to present data graphically or pictorially, facilitating easier comprehension and interpretation. Various types of data visualization techniques exist, such as bar charts, line graphs, scatter plots, heat maps, and network diagrams, each suitable for representing different data types (Srivastava, 2023) [11]. Bar charts are effective for comparing data across categories, while line graphs visualize trends over time, and scatter plots depict relationships between two variables (Srivastava, 2023)[11]. The choice of data visualization technique depends on factors such as the data type, insights sought, and target audience, with effective visualization emphasizing clear presentation and visual appeal.

Overall, the literature reviewed highlights the importance of continued research and innovation in sports training systems, golf performance analysis, and sports data visualization. By addressing the limitations of existing approaches and leveraging emerging technologies, researchers can contribute to the advancement of sports science and enhance the performance and development of athletes at all levels.

4. Problem Statement

The problem addressed by this project lies in the inefficiency of current methods for analyzing and interpreting golf performance data, which often lack intuitive visualizations tailored to the specific needs of players, coaches, and analysts. Existing approaches fail to provide comprehensive insights into player performance trends, correlations between different aspects of the game, and actionable strategies for performance improvement. As a result, there is a pressing need to develop innovative visualization techniques that not only uncover hidden patterns and outliers within golf performance data but also empower stakeholders to make informed decisions regarding training strategies, performance optimization, and data-driven decision-making in the sport. With enhanced visualizations, golfers and coaches can dive deeper into the complexity of each game, gaining valuable insights into areas that require improvement. They can observe the precise movements, stroke techniques, and body positions throughout different stages of the game. By visualizing these aspects, this thesis empowers golfers and coaches with a deeper understanding of their strengths and weaknesses, enabling them to make more informed decisions on training strategies and golf game tactics.

4.1 Objective

To address this, the specific aim of this project is to develop advanced visualization methods that provide comprehensive insights into golf performance metrics, enabling stakeholders to gain a deeper understanding of player strengths and weaknesses, identify areas for improvement, and make data-driven decisions to enhance performance and strategic decision-making in golf.

The primary objective of this project is to develop visualizations and perform feature importance analysis of golf data that will empower golf coaches and players to gain valuable insights and make informed decisions regarding training strategies and performance optimization. This project will utilize feature importance analysis to specifically identify which performance metrics most significantly influence overall player success. These insights will allow coaches and players to focus their efforts on

the most critical areas that drive success, ensuring that training programs and strategic decisions are based on data-driven priorities.

Moreover, the visualizations can serve as powerful tools to showcase progress over time, compare performances with top athletes, and highlight areas of excellence. This aspect is especially crucial in encouraging golfers to strive for continuous improvement and set new personal goals.

5. Methodology

The methodology for this study follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which provides a structured approach to data analysis. This process is divided into six key phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

1. Business Understanding: The initial phase involves defining the objectives and goals of the analysis. The primary aim is to uncover insights into golf player performance metrics such as wins, strokes, putting accuracy, and tee-to-green performance. The insights gained will inform stakeholders on optimizing performance strategies and making data-driven decisions.

2. Data Understanding: In this phase, a comprehensive golf dataset containing various player performance metrics is acquired. The data is thoroughly explored to understand its structure, identify patterns, and detect any anomalies. Summary statistics and initial visualizations are utilized to gain an overview of the dataset.

- **Data Source :** The "**golf_strokes.csv**" dataset is sourced from the Swedish National Golf program, which collects and maintains detailed records of golf strokes performed by players across different tournaments and training sessions. This dataset is meticulously curated and updated regularly to ensure accuracy and reliability.
- **Data Description:** Upon loading the dataset into the analysis environment, a preliminary exploration reveals its structure and dimensions. The dataset comprises 4,441,144 rows and 17 columns, each representing a single golf stroke. The columns contain various attributes and metrics associated with each stroke, including:
    - **sex (object):** Indicates the sex or gender of the player who took the shot.
    - **shot_pk (int64):** A unique identifier for each golf shot.
    - **round_fk (int64):** An identifier for the round of golf during which the shot was taken.
    - **round_date (string):** Indicates the date when each golf shot was taken, enabling analysis of player performance trends.
    - **user_fk (int64):** An identifier for the player who took the shot.

- **hole_par (int64):** The par value for the hole on which the shot was taken (e.g., par 3, par 4, par 5).
- **hole_score (int64):** The score recorded for the hole on which the shot was taken.
- **shot_number_hole (int64):** The sequential number of the shot within the hole (e.g., 1st shot, 2nd shot, etc.).
- **shots_left (int64):** The number of shots remaining for the player to complete the hole.
- **from_location (object):** The starting location of the shot (e.g., tee box, fairway, rough).
- **result_location (object):** The ending location of the shot (e.g., fairway, green, bunker).
- **from_distance (int64):** The distance from which the shot was taken.
- **result_distance (int64):** The distance to the hole after the shot was taken.
- **shot_type (object):** The type of shot taken (e.g., drive, approach, putt).
- **sg_from (float64):** Strokes gained from the starting location of the shot, indicating the player's performance relative to the average.
- **sg_result (float64):** Strokes gained from the ending location of the shot, indicating the player's performance relative to the average.
- **sg_net (float64):** Net strokes gained, representing the overall performance gain or loss for the shot.

3. Data Preparation: Data preparation involves cleaning the dataset, handling missing values, and transforming the data to make it suitable for analysis. Features that are relevant to the study, such as wins, strokes, putting accuracy, and tee-to-green performance, are selected for further analysis.We addressed missing values in the dataset, particularly in the "From Location," "Result Location," and "Shot Type" columns, where significant numbers of missing values are observed.

- **From Location and Result Location:** The columns "From Location" and "Result Location" have 44,972 missing values each. These missing values may

indicate instances where the location from which the golf stroke was made or the resulting location of the stroke was not recorded or unavailable.

One approach to handling these missing values is to impute them with the most frequent location value or the mode. This ensures that the impact of missing values on the analysis is minimized while preserving the distribution of location data.

Alternatively, if the missing values represent a significant proportion of the data and cannot be reliably imputed, these rows may be removed from the dataset. However, this approach should be considered carefully to avoid significant data loss, especially if the missing values are scattered throughout the dataset.

- **Shot Type:** The "Shot Type" column has a large number of missing values, specifically 3,997,253. Missing values in this column may indicate instances where the type of golf shot performed (e.g., drive, putt, chip) was not recorded or available. Depending on the importance of shot type in the analysis, several approaches can be considered for handling these missing values.

  If shot type is crucial for the analysis, one option is to impute the missing values based on other attributes or characteristics of the golf stroke, such as distance or outcome. Alternatively, if the shot type is not critical for the analysis or if imputation is not feasible, these rows may be removed from the dataset. However, this should be done cautiously to avoid biasing the analysis or losing significant portions of the data.

4. Modeling: The core of the analysis involves the application of both visualization techniques and machine learning models:

- **Visualization Techniques:** Various techniques, including bar charts, histograms, scatter plots, line charts, box plots, heatmaps, stacked bar charts, and pie charts, are applied to the prepared data. These visualizations help in analyzing player performance trends, identifying correlations, and spotting outliers. Each visualization is carefully interpreted to extract meaningful insights.

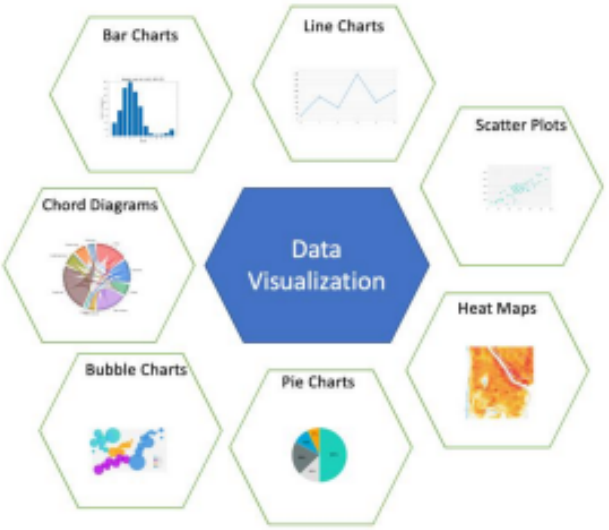


Fig. 2: Data Visualization Techniques

- **Machine Learning Models:** Random Forest and Decision Tree algorithms are employed to analyze the golf performance data, focusing on feature importance. The importance of each feature is assessed, providing targeted recommendations for prioritizing key areas for improvement.

5. Evaluation: The results from the visualizations and machine learning models are critically evaluated. Statistical analysis and domain knowledge are used to validate the findings, ensuring the reliability of the conclusions drawn. The evaluation phase ensures that the insights provided are both accurate and actionable.

6. Deployment: Finally, the insights and recommendations derived from the analysis are compiled into a report. This information serves as a foundation for optimizing training and strategic decision-making processes in golf performance management.

By adhering to the CRISP-DM framework and leveraging a combination of visualization techniques and machine learning models, this methodology aims to provide valuable, data-driven insights into golf performance metrics, enabling stakeholders to make informed decisions and optimize strategies effectively.

5.1 Tools and Libraries

For visualizing golf data, we are utilizing Python as the primary programming language. Python is an interpreted high-level programming language, meaning that it makes use of an interpreter to read the code and execute it line by line. Python is a dynamic programming language (variables can change in type) and allows different programming paradigms such as functional programming or object-oriented programming. The version of Python being used is Python 3.11.5. As for the visualization library, we have chosen Seaborn and Matplotlib due to their comprehensive functionality and ease of use. Matplotlib is a fundamental library for creating static, interactive, and publication-quality visualizations in Python. It provides a wide range of plotting functions, allowing for highly customizable plots and graphics.

The combination of Matplotlib and Seaborn offers a powerful toolkit for data visualization tasks. While Seaborn simplifies the process of creating complex statistical plots with its high-level interface and aesthetic enhancements, Matplotlib provides fine-grained control over plot elements and customization options. Matplotlib is a popular data visualization tool in Python, widely used by scientists and researchers. It helps create different types of charts, like line graphs and 3D plots, making data easier to understand. Developed by John Hunter and a team, Matplotlib is integrated with other Python libraries like NumPy and Pandas. Its user-friendly interface allows for customization, enabling the creation of high-quality visualizations for research and analysis. Overall, Matplotlib plays a key role in making complex data more accessible and insightful for users (Hafeez, Abdul & Sial, Ali. 2021)[12]. Seaborn offers a high-level interface for creating attractive and informative statistical graphics, making it a suitable choice for analyzing golf data. Seaborn, a Python library for statistical graphics, plays a pivotal role in the visualization process due to its seamless integration with Matplotlib and compatibility with pandas data structures. By automatically mapping data values to visual attributes like color, size, or style, Seaborn facilitates the generation of informative plots that effectively communicate key insights. Moreover, Seaborn handles statistical transformations internally and enhances plots with descriptive axis labels and legends.

Seaborn's versatility makes it a valuable asset throughout the lifecycle of a scientific project. Its streamlined syntax allows for rapid prototyping and initial data exploration, while its extensive customization options, coupled with seamless integration with

Matplotlib, enable the creation of polished, publication-quality figures (Waskom, M. L. 2021)[13].

It seamlessly integrates with Pandas data structures, allowing for efficient data manipulation and visualization. Additionally, Seaborn provides a wide range of statistical plotting functions, enabling the creation of complex visualizations with minimal code.

The advantages of using Matplotlib and Seaborn together include:

- **Comprehensive Functionality:** Matplotlib and Seaborn together cover a wide range of visualization needs, from simple line plots to complex statistical graphics.
- **Flexibility:** Matplotlib allows for extensive customization of plot elements, including colors, markers, line styles, and annotations. Seaborn enhances this flexibility with built-in themes and color palettes.
- **Integration with Pandas:** Both libraries seamlessly integrate with Pandas data structures, making it easy to work with DataFrame objects and perform data visualization directly from data stored in Pandas.
- **Publication Quality:** Matplotlib and Seaborn produce publication-quality visualizations suitable for inclusion in academic papers, reports, and presentations.
- **Community Support:** Both libraries have large and active communities, providing plenty of resources, tutorials, and examples for users to leverage in their data visualization projects.

5.2 Feature Importance Analysis

In this section, the methodology for evaluating the importance of features in the models is detailed. This analysis provides insights into the key factors influencing the outcome. Before conducting the feature importance analysis, it is crucial to split the dataset into training and testing sets to ensure that the model is evaluated on unseen data.

The dataset was divided into two distinct subsets: a training set and a test set, using the train_test_split function from the scikit-learn library. This function allows for a controlled and reproducible split, which is essential for evaluating model performance

effectively.

- **Training Set**: The training set consists of 80% of the total data. This portion of the dataset is used to train the machine learning model, enabling it to learn patterns and relationships between the features (X_train) and the target variable (y_train).
- **Test Set**: The remaining 20% of the data is reserved as the test set. This set is used exclusively to evaluate the model's performance after it has been trained. The purpose of the test set is to provide an unbiased estimate of how well the model generalizes to new, unseen data (X_test, y_test).

To ensure consistency in the results, a random state of 42 was used during the split. This ensures that the random splitting process is reproducible, meaning that the same training and test sets are generated each time the code is run.

The splitting process is critical in the context of feature importance analysis, as it ensures that the model's evaluation is based on data that was not seen during training. This method helps in accurately determining the contribution of each feature to the model's predictive power.

Both models were trained using the training dataset (X_train, y_train) and validated using the testing dataset (X_test, y_test) to ensure generalizability and performance evaluation.
For validation, the model's performance is evaluated on the X_test features and y_test target values after training. This helps in assessing how well the model generalizes to unseen data. The performance metrics used for evaluation include Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) score.

5.2.1. Random Forest Model

The Random Forest model was selected for its robustness in handling complex, non-linear relationships and its effectiveness in providing accurate predictions for regression tasks. This ensemble learning method combines multiple decision trees to enhance predictive performance and reduce overfitting.
The Random Forest model was implemented using the scikit-learn library, a popular Python library for machine learning. The model was configured with 100 trees

(n_estimators=100) which specifies the number of trees in the forest. The default settings for other parameters are used, which typically include max_depth (depth of each tree), min_samples_split (minimum samples required to split an internal node), and min_samples_leaf (minimum samples required to be at a leaf node). These defaults generally provide a good balance between performance and computation time.

5.2.2. Decision Tree Model

The Decision Tree model was selected for its ability to provide clear and interpretable results by splitting the data based on feature values, which can be useful for understanding the decision-making process. Although simpler compared to Random Forest, it can effectively capture non-linear relationships in the data.

Decision Trees create a flowchart-like structure where each node represents a decision based on a feature, making it easy to understand how predictions are made. The Decision Tree model was implemented using the scikit-learn library, which provides an efficient implementation of decision trees for regression tasks.

The Decision Tree model is initialized with default settings, which generally include parameters like max_depth (maximum depth of the tree), min_samples_split (minimum number of samples required to split an internal node), and min_samples_leaf (minimum number of samples required to be at a leaf node). These defaults allow the tree to grow until it perfectly fits the training data, which may lead to overfitting.

5.2.3 SHAP Analysis

To enhance the interpretability of our machine learning models, we utilize SHAP (SHapley Additive exPlanations) values. SHAP provides a robust framework for interpreting model predictions by attributing the contribution of each feature to the model's output. This approach is rooted in cooperative game theory, specifically Shapley values, which ensure a fair and consistent allocation of contributions among features.

According to (Lundberg and Lee, 2017) [21] SHAP values offer a unified perspective on model interpretability by aggregating various explainability techniques into a single coherent framework. They demonstrate how SHAP values can be used to explain the output of complex models in a way that is both accurate and interpretable. Their work

illustrates that SHAP values not only provide insight into individual predictions but also allow for a global understanding of feature importance across the dataset.

- **Rationale for SHAP Analysis**: SHAP analysis was chosen for its ability to provide a robust and consistent method for interpreting model predictions and quantifying the influence of features on player scores in golf. By leveraging Shapley values from cooperative game theory, SHAP ensures that the contributions of each feature to the model's predictions are accurately represented.
- **Implementation**: SHAP values were computed for both the Random Forest and Decision Tree models after training. The SHAP explainer, specifically shap.TreeExplainer, was utilized to calculate the SHAP values, providing detailed insights into how features such as sg_net_tee, sg_net_green, sg_net_ruff, sg_net_fairway, and sg_net_bunker contribute to predicting players' final scores. The SHAP summary plots were generated using the shap.summary_plot function, respectively, allowing for a thorough exploration of feature importance.
- **Interpretation**: The analysis focused on interpreting SHAP values through summary plots and bar plots. Summary plots revealed the overall distribution and impact of each feature across all predictions.
- **Visualization**: SHAP summary plots were utilized to visualize the overall impact of features on model predictions, clearly illustrating how each feature influences the output. The SHAP summary bar plot, which shows the mean absolute SHAP values, was also used to rank features by their overall importance.
- **Insights Gained**: The insights derived from the SHAP analysis highlighted critical performance metrics in golf, such as the significant impact of sg_net_tee and sg_net_green on player scores. These insights are instrumental in facilitating strategic decisions for enhancing player performance and optimizing training strategies, as they provide a clear understanding of which areas of the game contribute most to the overall score.

By applying SHAP to our Random Forest and Decision Tree models, we can quantitatively assess the impact of each feature on the predictions. This methodology helps in identifying which features are most influential and how they interact, thereby

providing a clearer understanding of the underlying decision-making processes of our models.

5.3 Ethical and Sustainability

The research and development work documented in this thesis keeps to ethical standards and guidelines. The data used in this study is anonymized and cannot be traced back to individuals, ethical approval for the collection or handling of personal data was not required. In our research, we're making sure our research is fair and good for the environment and society. Since the data we're using doesn't identify people, we don't need special permission to use it. We're also making sure our work doesn't harm the environment or communities. By following these guidelines, we're making sure our research helps without causing any harm.

6. Result

We are following various visualization techniques for our project. These various visualization techniques are suitable for representing golf performance data effectively.

6.1 Distribution of Players by Gender

The pie chart illustrates the distribution of players based on their gender. It presents the proportion of male and female players in the dataset. The chart provides a quick overview of the gender distribution among golf players, allowing for easy comparison between the two categories.

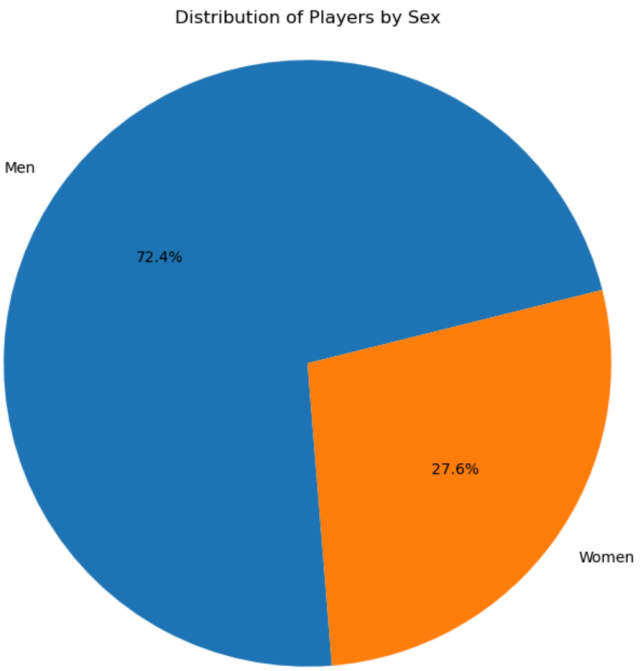


Fig. 3: Distribution of Players by Sex

6.2 Fairways Hit and Greens in Regulation (GIR) Analysis

Fairways hit and Greens in Regulation (GIR) are key metrics used to evaluate a player's ability to execute shots accurately and efficiently on the golf course. By analyzing these metrics over the dataset, we can gain insights into the consistency and effectiveness of players in achieving these fundamental objectives of the game. This evaluation is crucial for players, coaches, and analysts seeking to understand and improve player performance.

We created a graph that calculates the percentages of fairways hit and Greens in Regulation (GIR) for each player using the provided dataset. It then plots the trends of fairways hit and GIR percentages for each player over the dataset. The blue line represents the percentage of fairways hit, while the green line depicts the percentage of Greens in Regulation. This visualization offers insights into players' performance in hitting fairways and reaching greens in regulation, aiding in assessing their overall effectiveness in navigating the golf course.

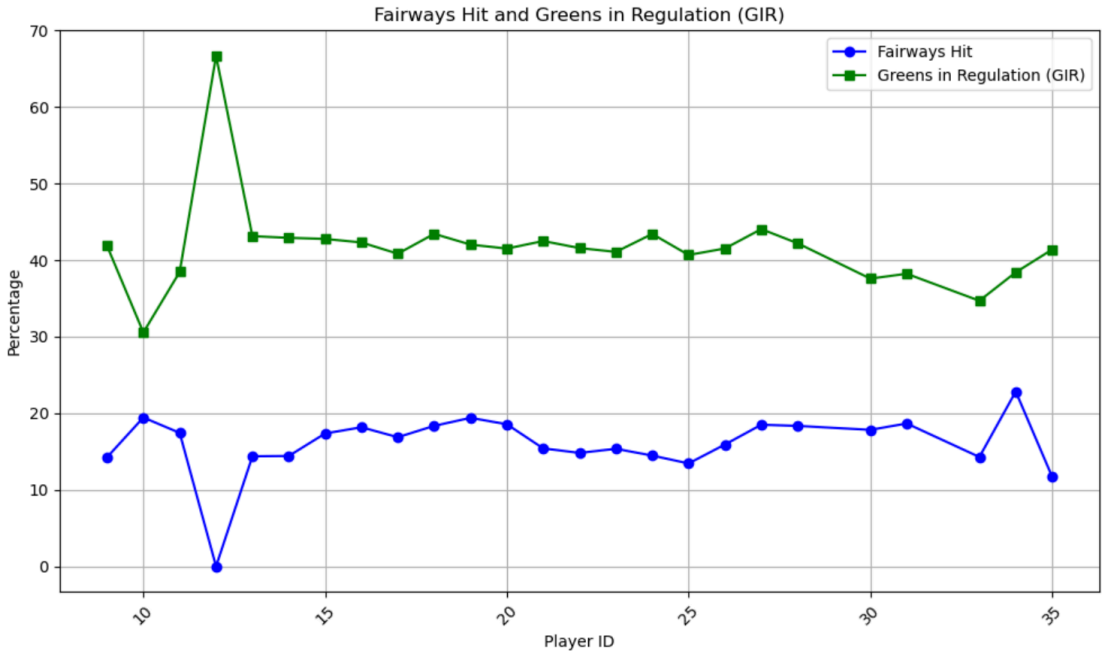


Fig. 4 : Fairways Hit & Greens in regulation (GIR)

6.3 Interactive Putting Performance Analysis Over Time

The interactive graph allows coaches to monitor how a player's putting performance evolves over time. By selecting a player ID from the dropdown menu, coaches can visualize the average number of putts made per round across different time intervals: weekly, monthly, and yearly.

The graph is divided into three subplots, each displaying the average number of putts per round over the specified time interval. The "Weekly Performance" subplot shows the weekly trend of putting performance, while the "Monthly Performance" subplot illustrates the monthly averages, and the "Yearly Performance" subplot displays the yearly averages.

The number of putts per round is calculated by subtracting the hole par from the hole score, indicating how many strokes a player took on the green to complete the hole. This

metric provides a measure of a player's proficiency in putting, with lower values indicating better performance.

Coaches can observe fluctuations in putting performance over time, identify periods of improvement or decline, and evaluate the effectiveness of training interventions. This granular analysis allows coaches to refine their coaching strategies, focus on specific aspects of putting performance, and optimize player development programs. The interactive nature of the tool provides coaches with flexibility in exploring putting performance data at different temporal resolutions, facilitating comprehensive performance assessment and strategic decision-making.

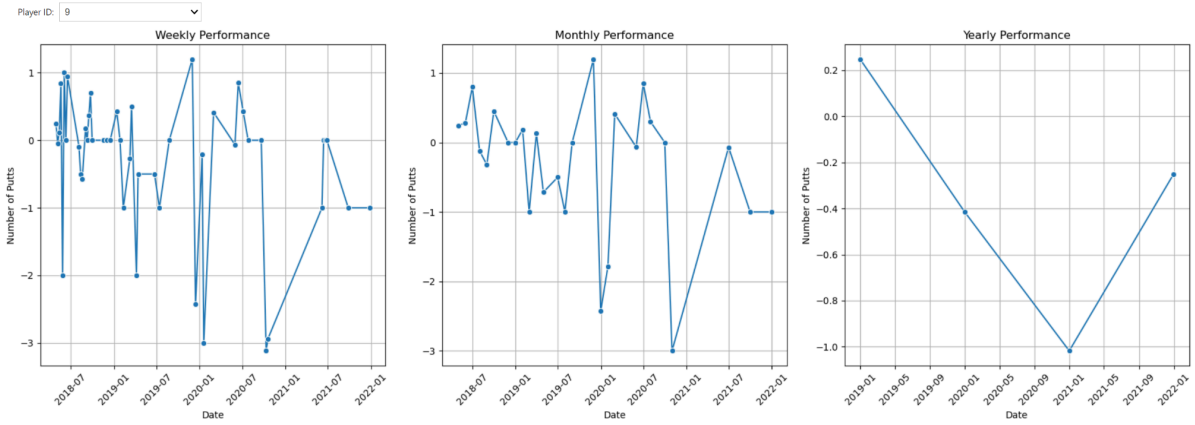


Fig.5: Weekly, Monthly, and Yearly Putting Performance Analysis for Player ID:9

6.4 Interactive Player Performance Analysis by Shot Location Over Time

The interactive graph allows coaches to monitor how a player's performance varies based on shot location over time. By selecting a player ID, from_location, and time interval, coaches can observe the trend of the rolling average sg_net, which indicates the player's performance relative to the average, for each shot location. The graph visualizes the performance trajectory, highlighting areas of strength and areas needing improvement. Coaches can use this information to adjust training regimens, focus on specific shot locations during practice sessions, and develop targeted strategies to enhance overall performance on the golf course. Additionally, the flexibility of choosing

different time intervals provides coaches with the ability to analyze performance trends at varying levels of granularity, from weekly fluctuations to longer-term trends over months or years.

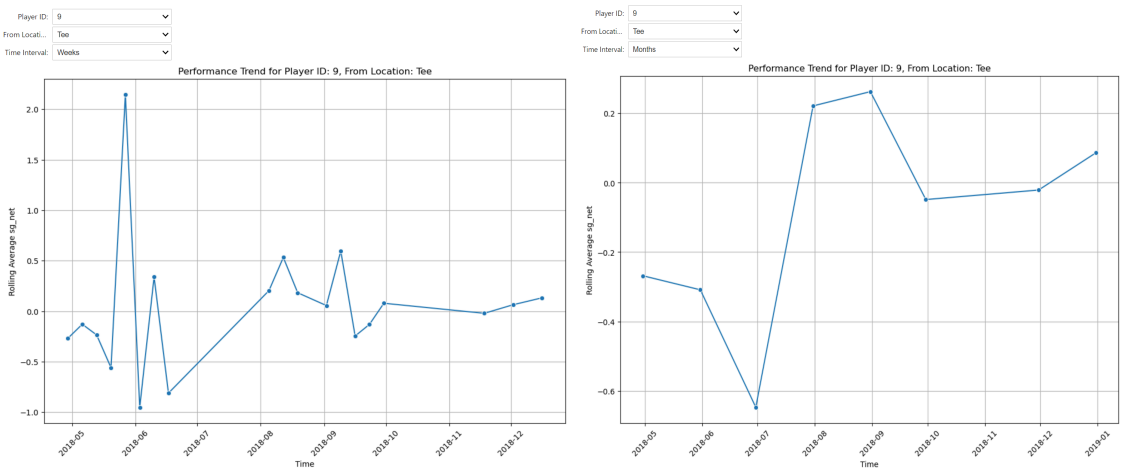


Fig.6: Weekly and Monthly Performance Analysis for Player ID:9

6.5 Interactive Scoring Trends Over Selected Rounds

Coaches can use this interactive tool to examine how a player's hole scores change across a selected range of rounds. They choose a player ID from the dropdown menu and adjust the slider to define the range of rounds for analysis.

The line chart displays the player's hole scores over the specified rounds, with the x-axis representing the round number and the y-axis representing the hole score. Coaches can observe trends in scoring consistency and identify any significant variations or patterns.

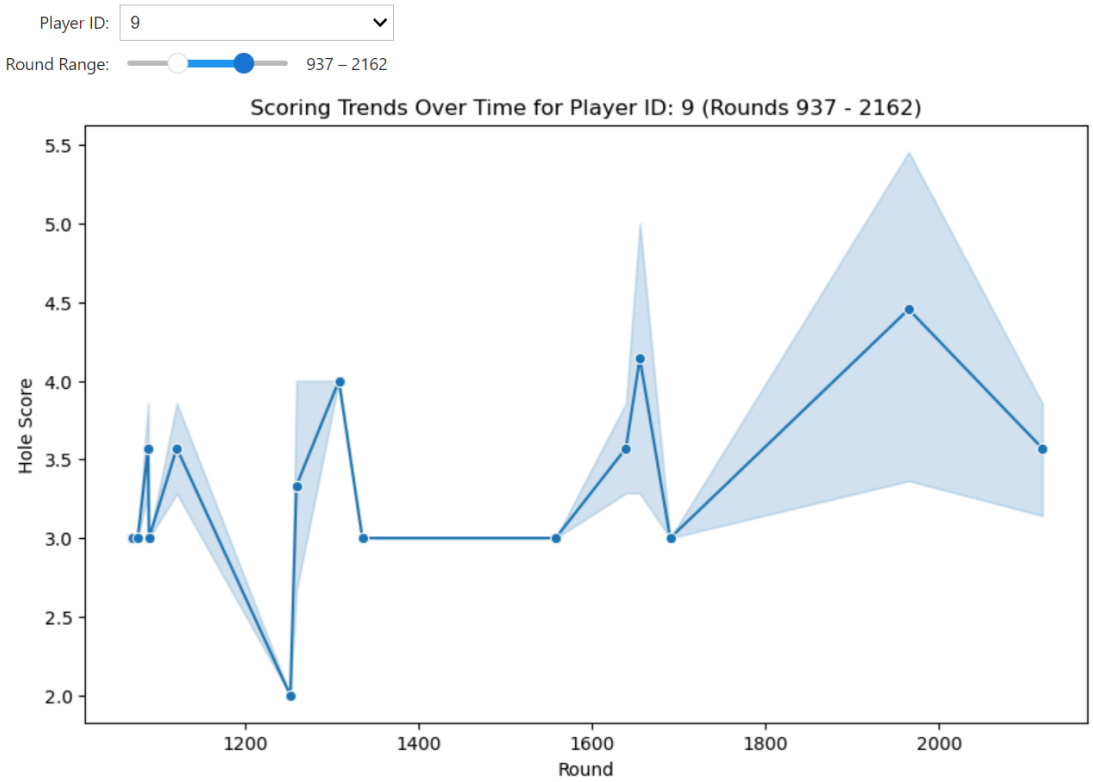


Fig.7: Scoring Trends over selected rounds for player ID:9

6.6 Interactive Shot Outcome Analysis by Starting Location

Coaches can use this interactive tool to analyze a player's shot outcomes from different starting locations (from_location). They select a player ID from the dropdown menu and choose the starting location from another dropdown menu.

The created bar chart displaying the distribution of shot outcomes, categorized by result location. Additionally, the average Strokes Gained Net (sg_net) value is annotated for each result location, providing insights into the player's performance relative to the average. By interactively exploring shot outcomes for different players and starting locations, coaches can tailor training strategies, address weaknesses, and optimize performance on the golf course.

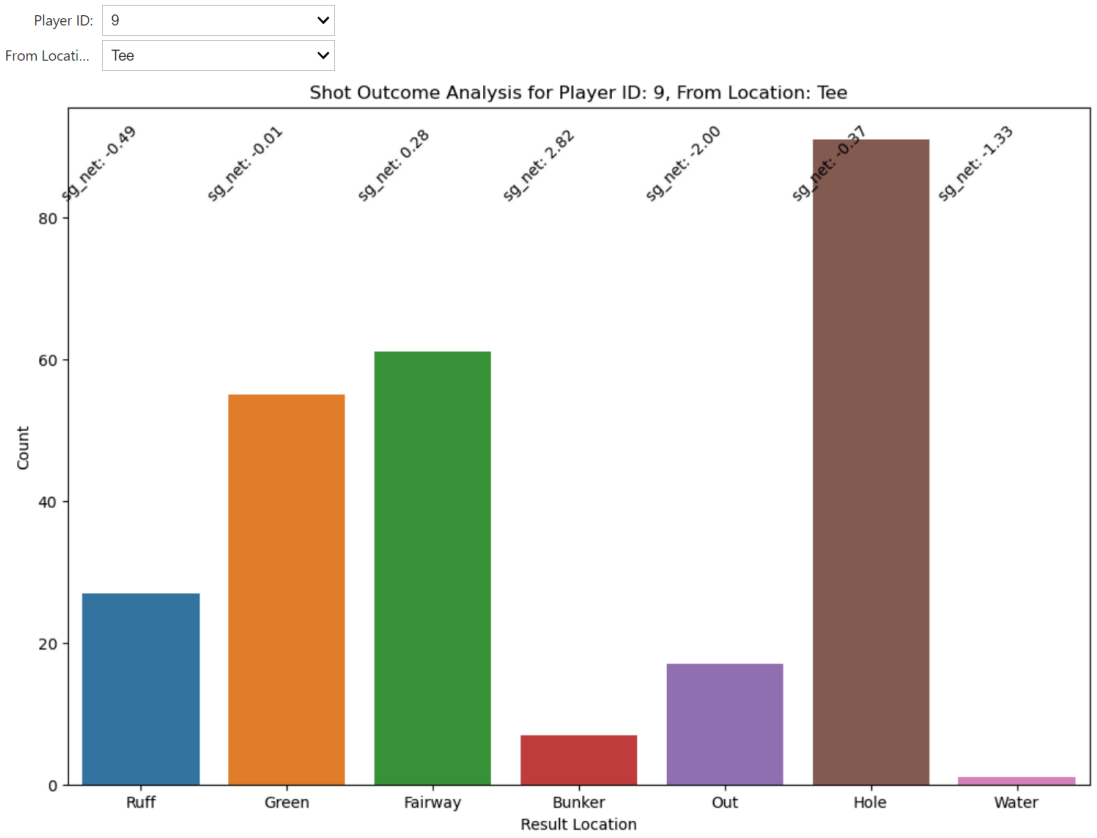


Fig.8: Shot Outcome Analysis for Player ID:9

6.7 Interactive Shot Performance Analysis Over Selected Rounds

Coaches can utilize this interactive tool to assess shot performance for a specific player across different rounds. They can select the player ID, starting location (from_location), result location, and range of round numbers using dropdown menus and a slider.

We created a line chart displaying the Strokes Gained Net (sg_net) values over the selected round number range. Each data point represents the sg_net value for a particular round, providing a visual representation of performance trends over time.

By interactively exploring shot performance based on various parameters, coaches can identify patterns, strengths, and weaknesses in a player's game. This facilitates targeted training strategies, skill development, and performance optimization, ultimately enhancing overall performance on the golf course.

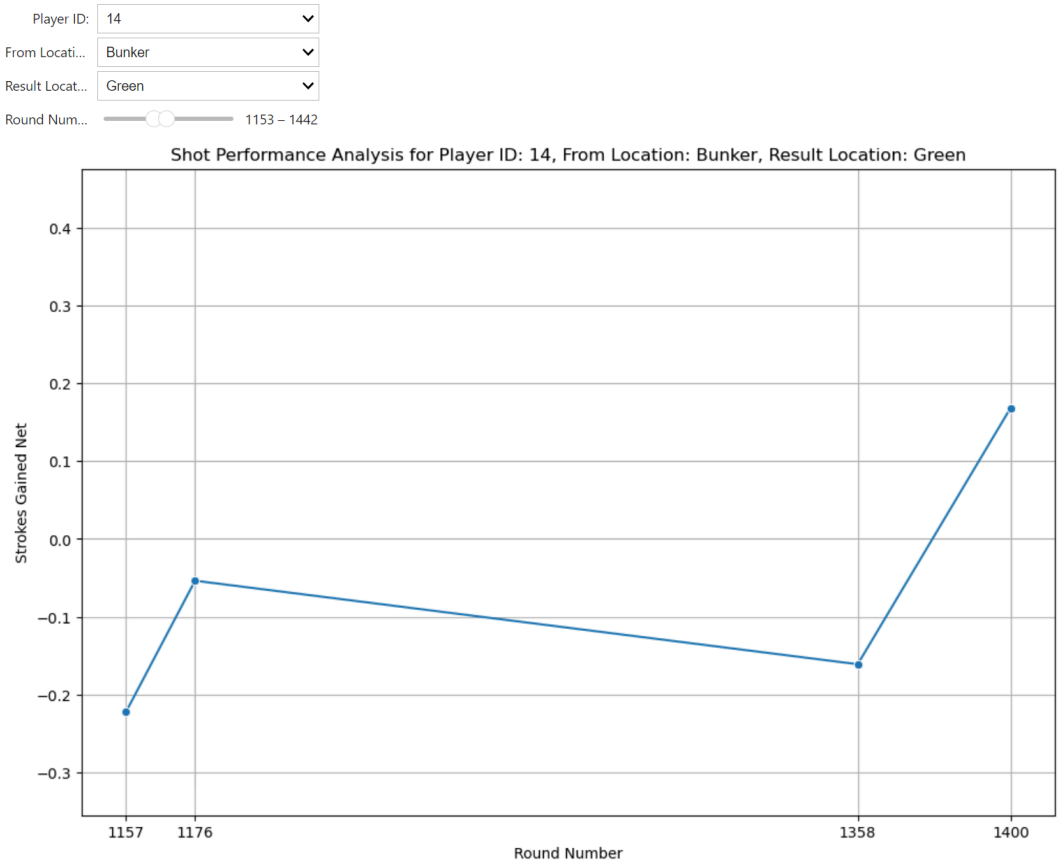


Fig.9: Short performance Analysis for player ID:14

6.8 Interactive Shot Performance Analysis by Location

Coaches can utilize this interactive tool to gain insights into a player's shot performance based on the starting location of the shot (from_location). Using the dropdown menu, coaches can select the player ID for whom they want to analyze shot performance.

We created a bar chart displaying the average sg_net for each from_location. The horizontal red dashed line represents the average sg_net across all locations, providing a reference point for comparison. This visualization allows coaches to identify which starting locations contribute positively or negatively to the player's overall performance.

The red dashed line represents the average Strokes Gained Net (sg_net) across all starting locations. It serves as a reference line to compare the performance of each starting location relative to the player's overall average performance. Locations above the red dashed line indicate that the player performs better than average from those starting positions, while locations below the line suggest areas where the player's performance may be below average and may require improvement.

By interactively exploring shot performance by location, coaches can tailor training strategies to address specific challenges, optimize shot selection, and improve overall performance on the golf course.

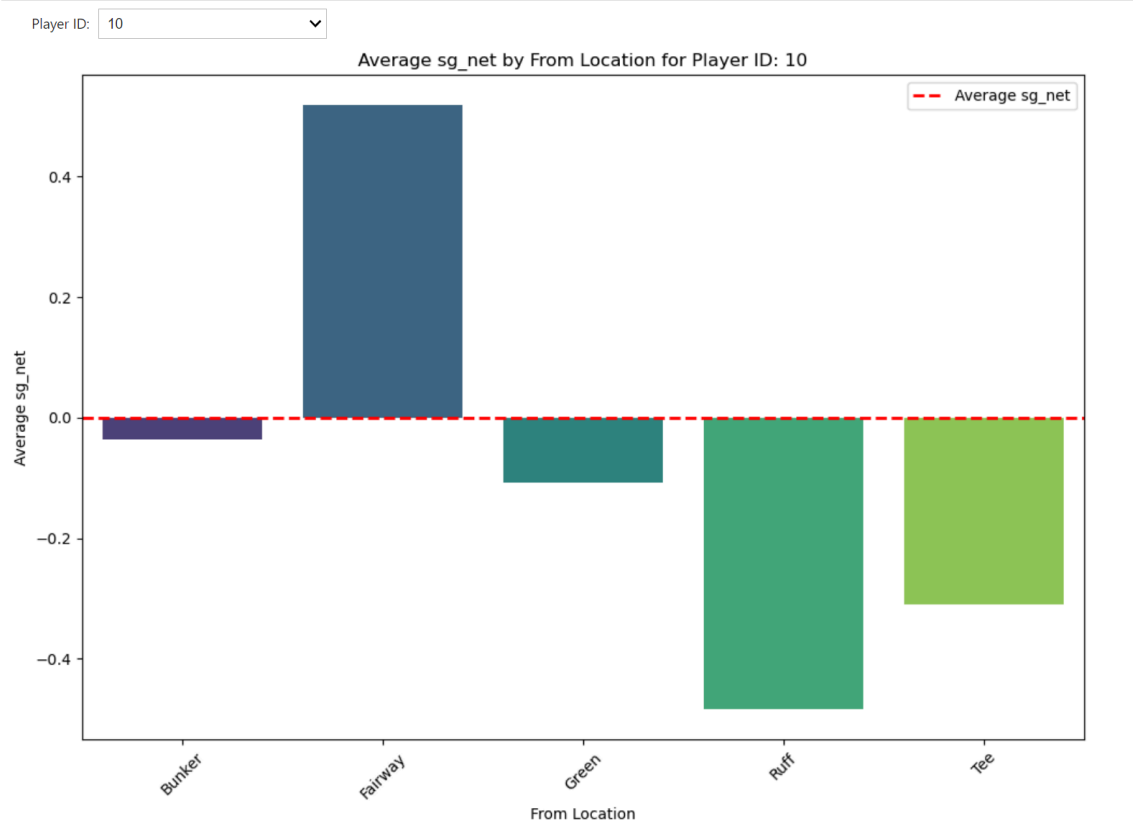


Fig.10: Shot Performance Analysis by Location for player ID:9

6.9 Interactive Shot Performance Analysis by Starting Distance

Coaches can utilize this interactive tool to gain insights into a player's shot performance based on the starting location of the shot (from_location). Using the dropdown menu, coaches can select the player ID for whom they want to analyze shot performance.

We created a bar chart displaying the average Strokes Gained Net (sg_net) for each from_location. The horizontal red dashed line represents the average sg_net across all locations, providing a reference point for comparison. This visualization allows coaches to identify which starting locations contribute positively or negatively to the player's overall performance.

By interactively exploring shot performance by location, coaches can tailor training strategies to address specific challenges, optimize shot selection, and improve overall performance on the golf course.

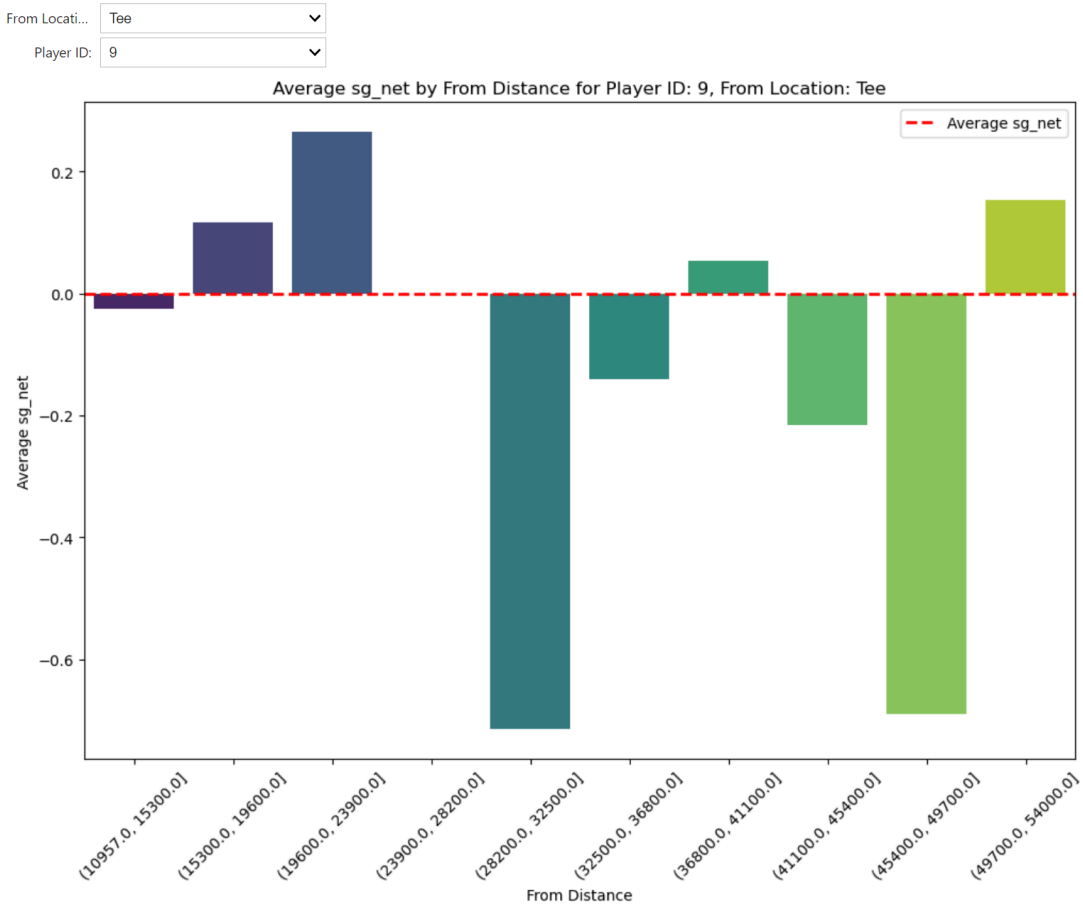


Fig.11: Shot Performance Analysis by starting distance for Player ID:9

6.10 Selected Player's Strokes Gained vs Average Strokes Gained

This interactive tool enables coaches to compare a selected player's shot performance (sg_net) to the average shot performance of all players across different starting locations (from_location). Coaches can use the dropdown menu to choose the player ID for whom they want to analyze shot performance.

We created a bar chart that displays two sets of bars for each from_location: one representing the average sg_net of all players (sky blue) and the other representing the selected player's sg_net (orange). This visualization allows coaches to identify areas where the selected player's performance deviates from the average performance of all players. By examining these deviations, coaches can pinpoint strengths and weaknesses in the selected player's game relative to their peers. This information can inform training strategies, helping coaches focus on areas where improvement is needed and capitalize on the player's strengths to enhance overall performance on the golf course.

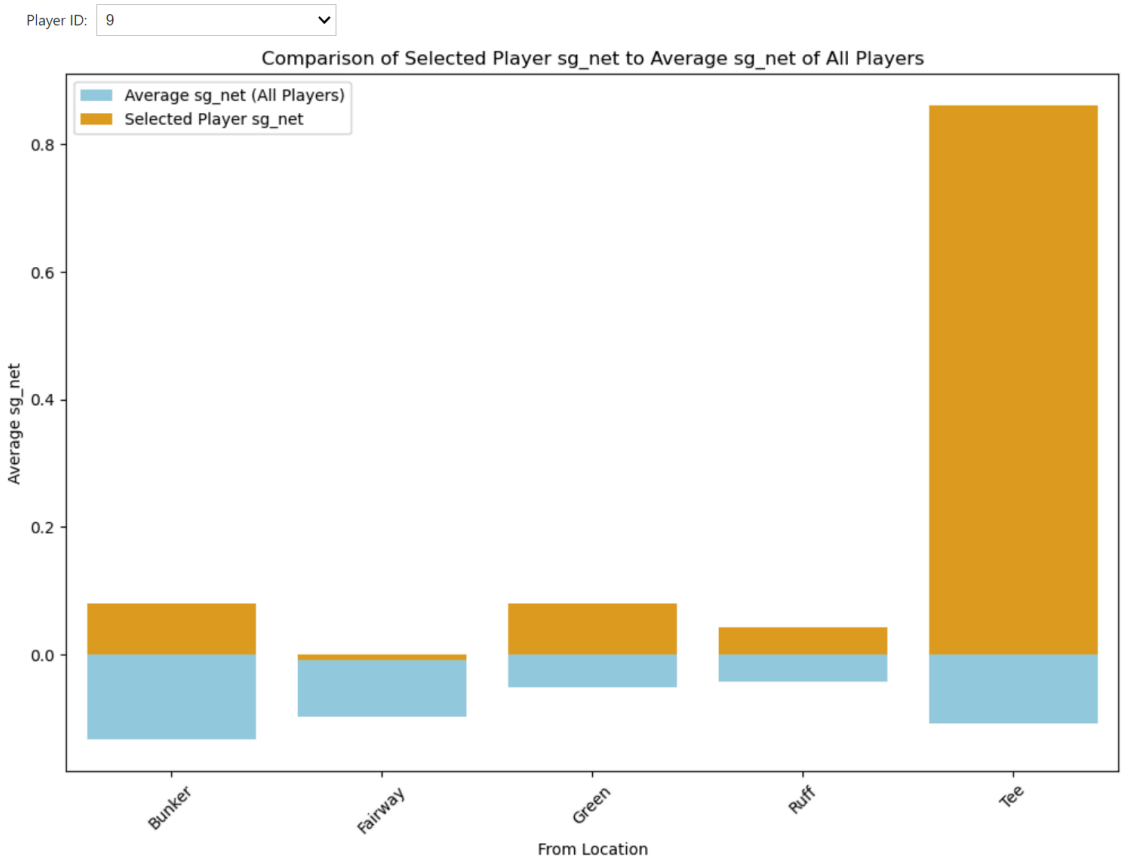


Fig.12: Comparison of Player ID 9 to Average Strokes Gained of All Players

6.11 Best Player's Strokes Gained vs. Average Strokes Gained

Coaches can use this interactive tool to analyze the performance of golf players based on their strokes gained (sg_net) for a specific year. The tool enables coaches to identify the best player for each year and compare their performance with the average strokes gained across all players.

By selecting a year from the dropdown menu, coaches can visualize the sg_net trend of the best player for that year along with the average sg_net line. This visualization allows coaches to gain insights into the performance dynamics of players over time, identify standout performers, and track overall performance trends. This chart provides valuable information for coaches to make data-driven decisions, tailor training programs, and optimize player strategies to enhance performance on the golf course.

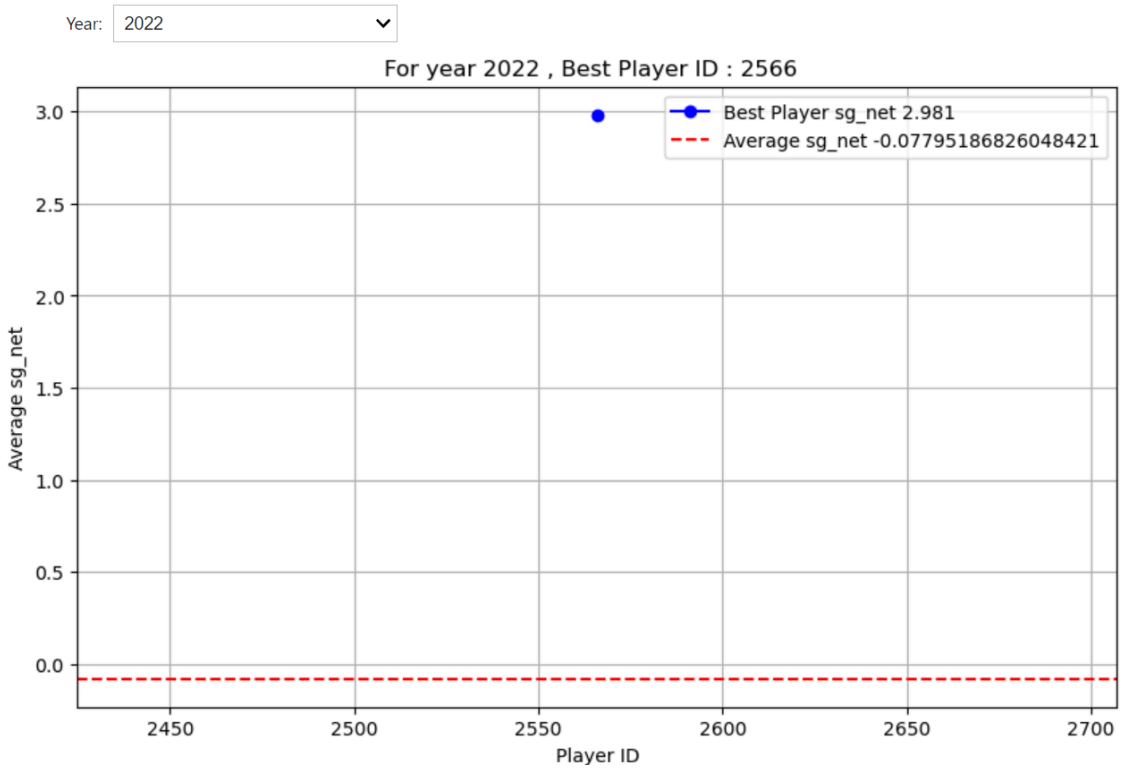


Fig. 13: Best Player vs Average Strokes Gained by All Players

6.12 Performance Factors in Golf

This combined analysis integrates multiple factors influencing golf performance, including performance by hole type, shot zone analysis, shot outcome analysis, and score distribution histogram. The first subplot provides insights into player performance across different types of golf holes, showcasing their scoring tendencies on par 3, par 4, and par 5 holes. The second subplot delves into shot zone analysis, illustrating the distribution of shot outcomes from various locations on the golf course. The third subplot focuses on shot outcome analysis, displaying the distribution of shot outcomes for each player. Finally, the fourth subplot presents the score distribution histogram, offering a visual representation of players' scoring patterns. By examining these factors collectively, coaches and analysts can comprehensively evaluate player performance and strategize for improvement.

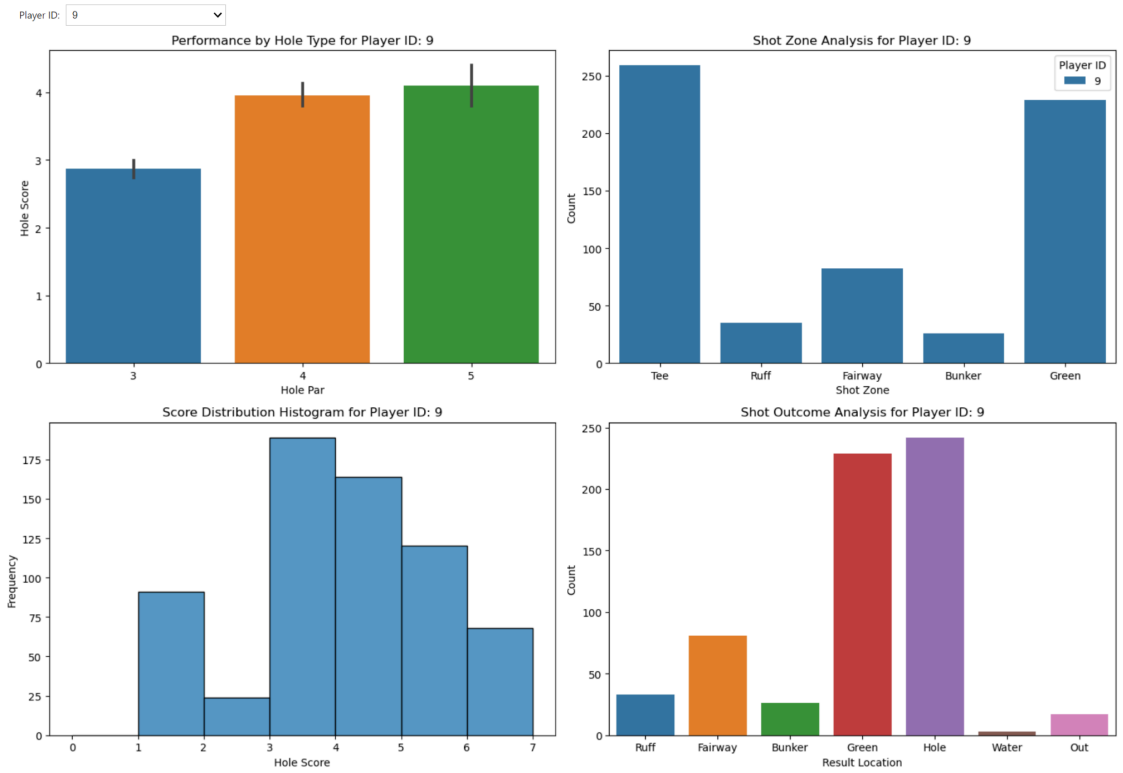


Fig.14: Performance by Hole type, Shot zone analysis, Shot outcome analysis, and Score distribution for Player ID 9

6.13 Shot Distance and Strokes Gained Net Distribution

This combined analysis presents insights into two crucial aspects of golf performance: shot distance and strokes gained net. The first subplot, the interactive shot distance histogram, illustrates the distribution of shot distances for each player, providing an understanding of their shot selection tendencies and distance coverage. The second subplot, the interactive strokes gained net histogram, showcases the distribution of strokes gained net for individual players, highlighting their performance relative to the average. By examining these histograms together, coaches and analysts can gain a comprehensive understanding of players' shot-making abilities and overall effectiveness on the golf course, facilitating targeted training and strategic planning.

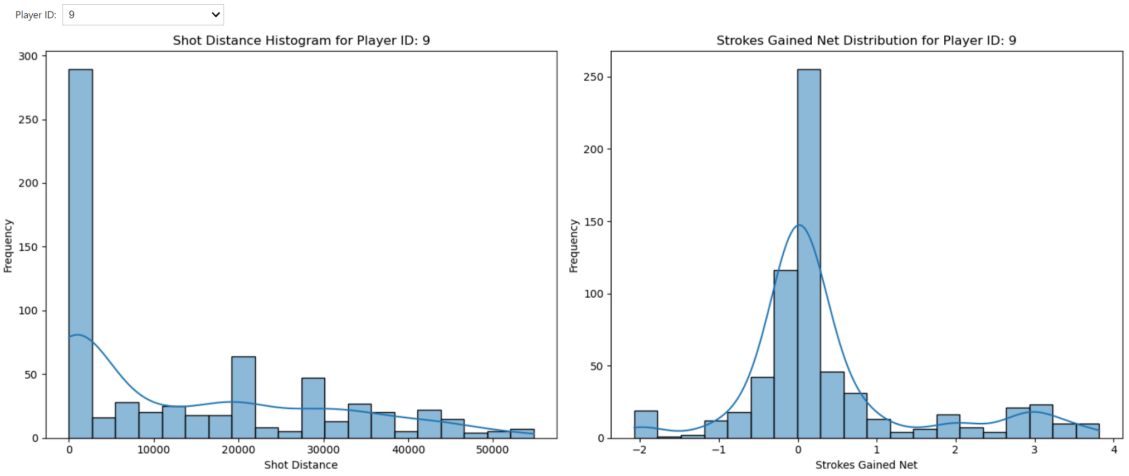


Fig.15: Shot distance and Strokes gained net for Player ID 9

6.14 Correlation Analysis of Strokes Gained Variables

The heatmap visualizes the correlation between strokes gained variables (sg_from, sg_result, sg_net) in the golf dataset. Each cell in the heatmap represents the correlation coefficient between two variables, ranging from -1 to 1. A positive correlation indicates that the variables move in the same direction, while a negative correlation indicates they move in opposite directions. This visualization helps identify relationships between different aspects of performance in golf shots.

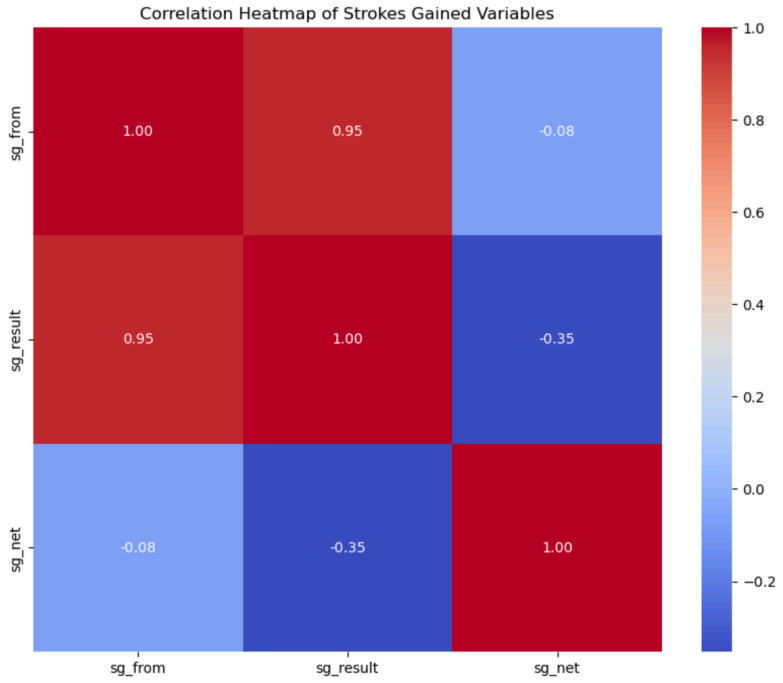


Fig.16: Heatmap for Strokes gained variables

6.15 Feature Importance Results

We are analyzing golf performance data using decision tree and random forest models to predict players' final scores based on their average Strokes Gained (SG) metrics from different locations on the course. The important features identified, such as `sg_net_ruff`,`sg_net_tee`, `sg_net_green`, `sg_net_fairway`, and `sg_net_bunker`, represent the player's performance from these specific areas. This analysis helps in understanding key factors that contribute to a player's overall performance and can inform coaching strategies.

6.15.1 Random Forest Model

To train the Random Forest model, we first split the dataset into training and testing sets, with 80% of the data allocated to the training set (X_train, y_train) and 20% to the testing set (X_test, y_test). We then initialized a RandomForestRegressor with 100 trees and trained the model using the training data. After training, the feature importance was determined using the feature_importances_ attribute of the model, which provides scores indicating the contribution of each feature to the prediction process. We visualized these importance scores through a bar plot, highlighting which features (such as sg_net_tee, sg_net_green, sg_net_ruff, sg_net_fairway, and sg_net_bunker) are most significant in predicting the players' final scores. This analysis helps in identifying key performance factors in golf.

6.15.2 Decision Tree Model

To train the Decision Tree model, the dataset was split into training and testing sets, with 80% of the data used for training (X_train, y_train) and 20% reserved for testing (X_test, y_test).Using the training data, a DecisionTreeRegressor was built to recursively partition the data based on feature importance. Post-training, feature_importances_ attribute was employed to quantify the impact of features like sg_net_tee, sg_net_green, sg_net_ruff, sg_net_fairway, and sg_net_bunker on predicting

player scores. Visualizing these scores via a bar plot highlighted key performance metrics in golf, aiding in strategic insights for player improvement and game tactics.

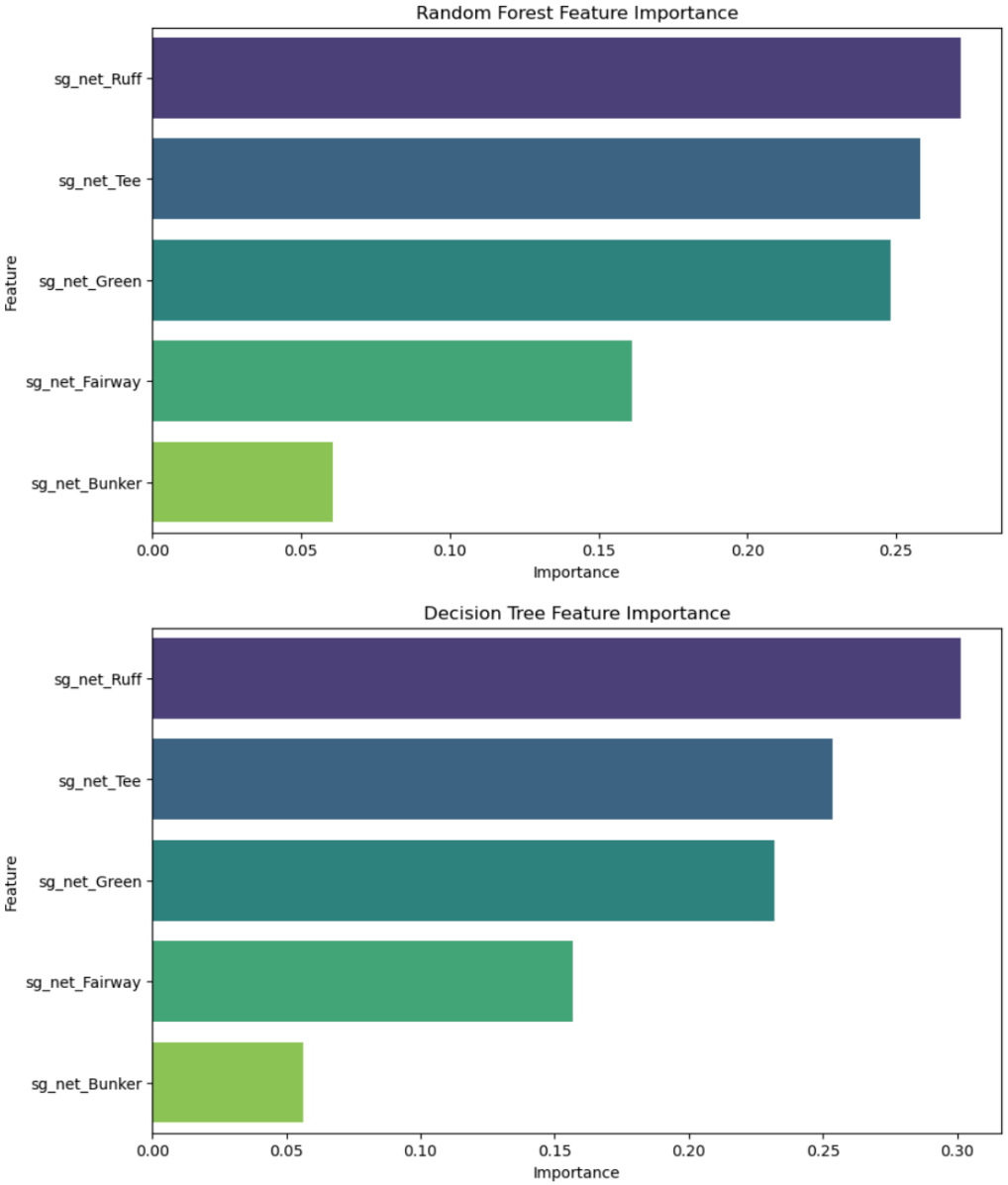


Fig.17: Feature Importance by Random Forest and Decision Tree Model

We included metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) for both the Random Forest and Decision Tree models. This allows us to evaluate and compare the performance of the models accurately.
Below is the result :

Model	Mean Squared Error	Mean Absolute Error	R2 Score
Random Forest	2711	30.1	0.53
Decision Tree	5337	40.8	0.082

The features identified as important in our models (`sg_net_ruff`,`sg_net_tee`, `sg_net_green`, `sg_net_fairway`, `sg_net_bunker`) represent the average Strokes Gained (SG) metric from various locations on the golf course. Strokes Gained (SG) is a performance metric used in golf to quantify a player's performance in various aspects of the game relative to the field. It is calculated as the difference between the actual number of strokes taken by a player to complete a task (e.g., tee shot, approach shot, putt) and the expected number of strokes taken by an average player in the same situation. Here's what each feature means:

1. **sg_net_ruff:** This represents the average Strokes Gained from the rough.It measures how well a player recovers from the rough. High positive values here indicate that the player can still perform well even when they miss the fairway, which can significantly impact their overall score. Effective play from the rough can mitigate the impact of missed fairways, leading to fewer dropped shots.
2. **sg_net_tee:** This represents the average Strokes Gained from the tee box. It indicates how well a player performs from the teeing area compared to the field. Higher positive values mean the player gains strokes on the field from their tee shots, contributing significantly to their overall score. Players who excel in driving the ball well from the tee can set themselves up for easier approach shots, leading to better scoring opportunities.
3. **sg_net_green:** This represents the average Strokes Gained on or around the green. It shows how effective a player is in putting and their short game around the green. Good performance in this area is crucial for scoring well, as it includes putting and short chips. Putting and short game are critical to scoring; players who gain strokes on the green can save pars and make more birdies.
4. **sg_net_fairway:** This represents the average Strokes Gained from the fairway.It shows how well a player performs their approach shots from the fairway.

Excellent performance from the fairway can lead to more birdie opportunities and better overall scoring. Strong approach play from the fairway directly influences the ability to hit greens in regulation, leading to more scoring chances.

5. **sg_net_bunker:** This represents the average Strokes Gained from bunkers.It indicates how well a player plays shots from sand bunkers. Being able to effectively get out of bunkers and place the ball close to the hole can prevent high scores on holes and maintain overall performance. Good bunker play ensures that players can recover from sand traps without losing strokes, maintaining their scoring potential.

6.16 SHAP Analysis Results

The next step involves using SHAP (SHapley Additive exPlanations) to explain the feature importance for different sets of feature values in Decision Tree and Random Forest models. SHAP is a method to explain individual predictions of complex models. It uses game theory to assign an important value to each feature, representing its contribution to the prediction. This helps in understanding how each feature influences the model's predictions for individual instances.

- **Y-Axis (Features)**: The y-axis lists the features used in the model: sg_net_Ruff, sg_net_Green, sg_net_Tee, sg_net_Fairway, and sg_net_Bunker.These features represent different aspects of a player's performance in golf.
- **X-Axis (SHAP Values)**: The x-axis shows the SHAP values, which represent the impact of each feature on the model's output. The SHAP value indicates how much a feature is pushing the prediction from the base value (average prediction) in either direction. Negative values push the prediction lower, and positive values push it higher.
- **Color (Feature Value)**: The color of the dots represents the value of the feature from low (blue) to high (red). This coloring helps to understand the relationship between the feature value and its impact on the prediction.

Random Forest SHAP Plot: The SHAP summary plots for the Random Forest model show the average impact of each feature on the model’s predictions. The features sg_net_tee, sg_net_green, sg_net_ruff, sg_net_fairway, and sg_net_bunker are

visualized, indicating their importance in predicting the final score. Higher values of sg_net_tee and sg_net_green have a significant positive impact on the final score, implying that strong performance from the tee and around the green are critical factors.

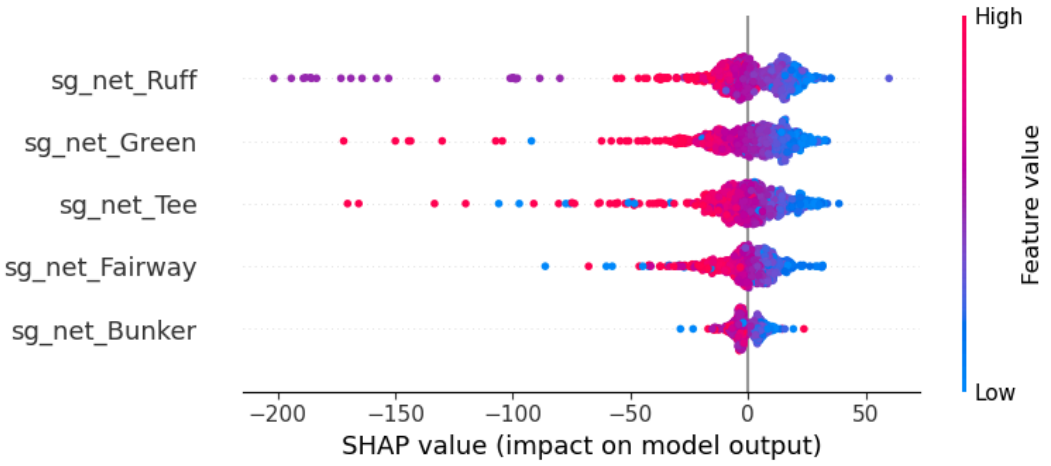


Fig.18: SHAP summary plot for Random Forest Model

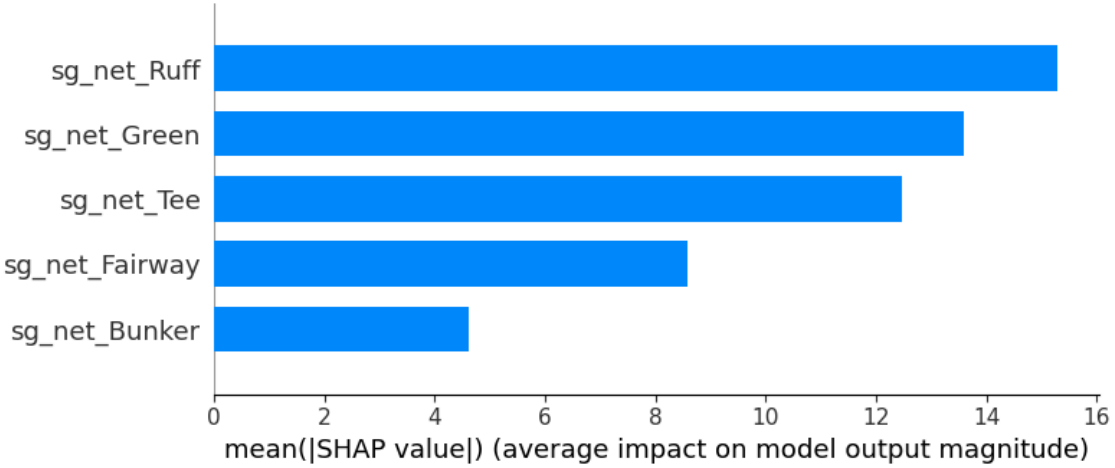


Fig.19: SHAP summary bar plot for Random Forest Model

Decision Tree SHAP Plot: The SHAP summary plots for the Decision Tree model show the average impact of each feature on the model's predictions similar to the Random Forest model.

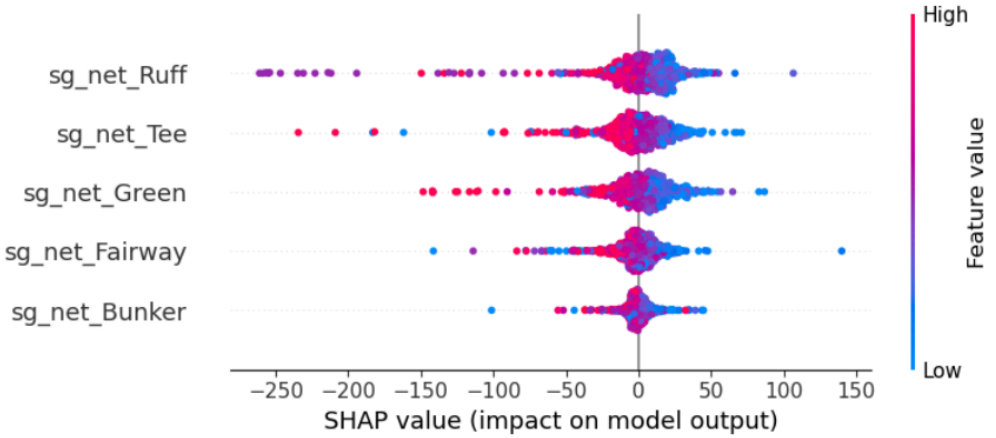


Fig.20: SHAP summary plot for Decision Tree Model

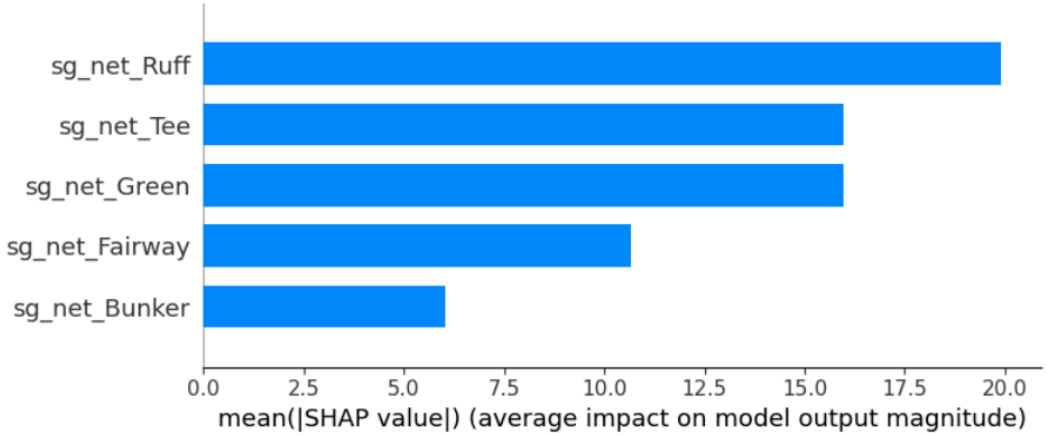


Fig.21: SHAP summary bar plot for Decision Tree Model

Here's what each feature means:

1. **sg_net_Ruff:** This feature has a mix of positive and negative SHAP values, indicating it can both increase and decrease the model's predictions.Higher values (red dots) tend to increase the model's output, while lower values (blue dots) tend to decrease it. This suggests that better performance from the rough (higher sg_net_Ruff) improves the player's predicted score.
2. **sg_net_Tee**: This feature shows a significant impact with SHAP values spread over a wide range.Higher values (red dots) tend to increase the prediction significantly, suggesting that strong performance from the tee is very important for a better score.
3. **sg_net_Green:** This feature has a narrower range of SHAP values, generally closer to zero. This suggests that sg_net_Green has a relatively moderate impact on the model's predictions.Similar to sg_net_Ruff, higher values (red dots)

increase the prediction, indicating that better performance on the green positively affects the predicted score.

4. **sg_net_Fairway:** This feature also has a wide spread of SHAP values.Higher values (red dots) tend to push the prediction higher, indicating that good performance from the fairway is crucial for a better score.
5. **sg_net_Bunker:** This feature mostly has SHAP values around zero but shows some strong negative SHAP values for lower feature values (blue dots).Poor performance from bunkers (lower sg_net_Bunker values) tends to significantly decrease the predicted score.

These plots indicate that sg_net_Tee and sg_net_Fairway have the most substantial impact on the model's predictions, with higher values significantly improving the predicted scores.

sg_net_Bunker has a notable negative impact when its values are low, showing the importance of good bunker play.sg_net_Ruff and sg_net_Green have moderate impacts, with higher values contributing positively to the score. The color gradient helps in understanding that better performance (higher feature values) in these areas generally leads to better-predicted scores.

These plots are useful for identifying which features are most influential and how they interact with the predictions, providing actionable insights into player performance.

7. Discussion

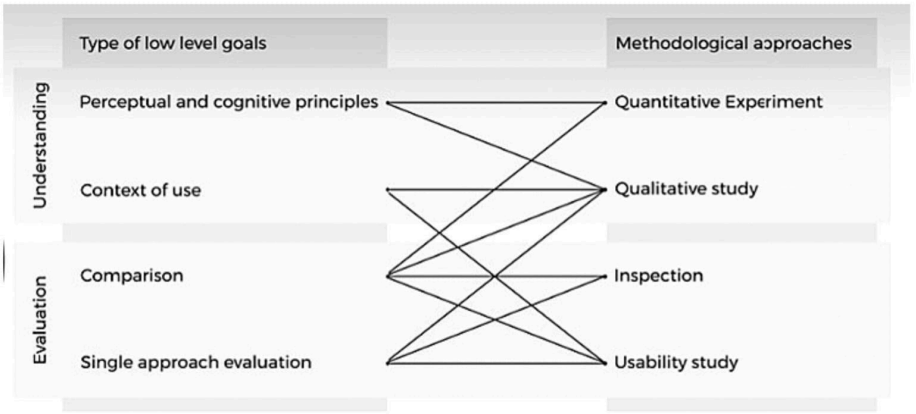
The findings from the visualization analysis provide valuable insights into golf performance metrics, player capabilities, and trends. These insights can inform coaching strategies, player development programs, and performance optimization techniques in golf. Additionally, the study demonstrates the effectiveness of visualization techniques in interpreting complex sports data and extracting actionable insights. The selected visualization techniques, including bar charts, histograms, and line charts, offer a comprehensive toolkit for exploring diverse aspects of golf performance data. These techniques provide flexibility and versatility, allowing for the representation of various data types and patterns. Each visualization technique serves a distinct purpose in revealing insights into player performance trends, correlations, and outliers. For instance, bar charts are effective for comparing categorical data, making them suitable for analyzing wins or performances across different tournaments. Line charts, on the other hand, excel in visualizing trends over time, such as the progression of performance metrics throughout a season or career. Moreover, the choice of visualization libraries, such as Seaborn and Matplotlib, is informed by their extensive functionality, ease of use, and compatibility with Python data structures.

In the literature review, various studies have emphasized the importance of exploratory data analysis (EDA) in uncovering meaningful patterns and trends within complex datasets, particularly in the field of sports data visualization. EDA serves as a cornerstone of modern data science, facilitating interactive exploration to reveal insights that contribute to scientific research and prediction (Hullman & Gelman, 2021)[10]. Reflecting on our methodology and results, we aimed to integrate principles of EDA into our visualization approach to effectively communicate insights from golf performance data. For instance, we proposed interactive bar charts to compare individual player performance against the average, aiming to provide clear and interpretable visualizations for players and coaches. While our visualizations succeeded in certain aspects, there is room for improvement in capturing the complexity of golf performance metrics comprehensively.

The choice of Python version 3.11.5 and the visualization libraries Seaborn and Matplotlib for our project was based on their comprehensive functionality, ease of use,

and suitability for our data visualization tasks. Matplotlib, being a fundamental library in Python, offers a wide range of plotting functions and Seaborn complements Matplotlib by simplifying the process of creating complex statistical plots with its high-level interface and elegant enhancements. These libraries served our project effectively by providing a powerful toolkit for data visualization, it's worth considering whether there could have been better alternatives available. For example, exploring newer libraries like Plotly or Altair may have offered additional interactive visualization capabilities or a more analytical approach to visualization, potentially enhancing our visualization outcomes. Overall, Seaborn and Matplotlib met our project requirements and offered a healthy set of features for data visualization.

To evaluate our visualizations, we could refer to methods outlined in Figueiras' paper titled "A Review of Visualization Assessment in Terms of User Performance and Experience" (Figueiras, Ana. ,2018)[14].



(a) Methodological approaches for each high and low level goal

Emphasis	Data collection methods
Performance (efficiency/errors)	Direct observation
Style and aesthetics	Performance measurements
Utility	Questionnaires
User experience (memorability/fun/engagement)	Self-reported preference questionnaires
Support for learning	Interviews
Support for insight	Eye tracking
Support for communication	Log analysis
Effect on the collaboration process	Experimenter-coded responses
Effect on the analytical reasoning process	Journaling
Literacy	
Usability	
Learnability	

(b) List of emphasis and data collection methods

Fig.22: Categorization in terms of goals, emphasis, methodological approaches, and data collection methods (Figueiras, Ana. , 2018)[14].

Specifically, we would focus on understanding user behavior and preferences through methods such as qualitative or quantitative experiments, usability studies, or inspection studies. These studies would help assess various aspects, including user performance, experience, and engagement with the visualizations. Overall, a combination of qualitative and quantitative approaches would provide comprehensive insights into the effectiveness of our visualizations in meeting user needs and preferences. User studies offer valuable insights into the effectiveness of visualizations and ensure they meet user needs.

The findings from the feature importance and SHAP analyses provide critical insights into the determinants of golf performance, shedding light on key metrics that significantly influence player scores. The feature importance analysis highlighted specific factors such as sg_net_tee, sg_net_green, sg_net_ruff, sg_net_fairway, and sg_net_bunker, which emerged as pivotal in predicting player outcomes. These insights are instrumental in identifying areas of strength and improvement for players, as well as guiding strategic decisions for coaches and performance analysts.

Moreover, the SHAP analysis further enhances our understanding by quantifying the impact of these features on individual predictions. By visualizing SHAP values through summary plots and dependency plots, we discerned how variations in each feature contribute to fluctuations in player scores. This granular level of insight not only validates the findings from the feature importance analysis but also provides a nuanced understanding of the interplay between different performance metrics in golf.

In conclusion, the careful selection of graphs, libraries, and techniques for the visualization analysis in this thesis was driven by the need for a comprehensive exploration of golf performance data and effective communication of findings. By leveraging a combination of versatile graphs and powerful libraries, the visualization analysis provides valuable insights into player capabilities, performance trends, and strategic considerations in golf. Furthermore, the comprehensive analysis of feature importance and SHAP in this study underscores their significance in explaining performance determinants in golf. These insights can inform coaching strategies, player development programs, and performance optimization techniques, ultimately enhancing the competitiveness and success of golfers at all levels. By adopting a systematic approach to data visualization and leveraging established methodologies, this project

contributes to advancing the field of sports performance analysis and informs decision-making in golf.

8. Conclusion

In conclusion, this thesis showcases the importance of visualization in analyzing golf performance data. By leveraging visualization techniques, valuable insights into player performance trends, correlations, and outliers are uncovered. Players and Coaches can check the performance of individuals to enhance better golf games. Furthermore, These insights have implications for player development, coaching strategies, and performance optimization in golf. Moving forward, further research can explore advanced visualization techniques and incorporate additional data sources to enhance our understanding of golf performance metrics.

9. Limitation & Future Work

One limitation of this study is the reliance on historical data, which may not capture the dynamic nature of player performance, such as the impact of a player's "day-form" (having a particularly good or bad day). Additionally, the visualization of golf performance data is influenced by numerous intangible factors that cannot be fully quantified through data alone.

In the future, there are several exciting possibilities for expanding on the work done in this study. One potential area of focus is the development of interactive web applications that allow players and coaches to log in and access personalized insights into their performance metrics. By integrating visualization tools and user-friendly interfaces, this application could help users analyze their strengths and weaknesses more effectively. Another thing for future exploration is the implementation of real-time performance monitoring systems. By integrating sensors and data collection devices, players and coaches could track performance metrics during training sessions and tournaments, allowing for immediate feedback and adjustment of strategies.

Future work could include adding predictive analysis to make the study more useful. There is potential to expand the application of visualization techniques to other sports beyond golf. By adapting these methods to sports like tennis, soccer, or basketball, valuable insights could be provided for athletes and coaches across a variety of disciplines.

In summary, the future holds exciting opportunities for further advancing the field of Golf performance analysis through visualization.

10. References

1. Du, M., Yuan, X. A survey of competitive sports data visualization and visual analysis. *J Vis* **24**, 47–67 (2021).
2. Qin, X., Luo, Y., Tang, N., & Li, G. (2020). Making data visualization more efficient and effective: a survey. The VLDB Journal, 29(1), 93-117.
3. Link, D. (2024). Visualization: Basics and Concepts. In: Memmert, D. (eds) Computer Science in Sport. Springer, Berlin, Heidelberg.
4. Muskan, G. Singh, J. Singh, and C. Prabha, "Data Visualization and its Key Fundamentals: A Comprehensive Survey," *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2022, pp. 1710-1714, doi: 10.1109/ICCES54183.2022.9835803.
5. Sadiku, M. N. O., Adediran, A. A., & Musa, I. A. (2016). Data Visualization. *International Journal of Engineering Research and Advanced Technology (IJERAT)*, 2(12), 11-16.
6. Basole, R. C., & Saupe, D. (2016). Sports Data Visualization [Guest editors' introduction]. IEEE Computer Graphics and Applications, 36(5), 24-26. doi: 10.1109/MCG.2016.85.
7. James, N. (2007). The Statistical Analysis of Golf Performance. International Journal of Sports Science & Coaching, 2(1_suppl), 231-249
8. -C. Liao, D. -H. Hwang and H. Koike, "AI Golf: Golf Swing Analysis Tool for Self-Training," in IEEE Access, vol. 10, pp. 106286-106295, 2022, doi: 10.1109/ACCESS.2022.3210261.
9. Perin, C., Vuillemot, R., Stolper, C. D., Stasko, J. T., Wood, J., & Carpendale, S. (2018, June). State of the art of sports data visualization. In Computer Graphics Forum (Vol. 37, No. 3, pp. 663-686).
10. Hullman, Jessica & Gelman, Andrew. (2021). Designing for Interactive Exploratory Data Analysis Requires Theories of Graphical Inference. Harvard Data Science Review. 10.1162/99608f92.3ab8a587.
11. Deepmala Srivastava, "An Introduction to Data Visualization Tools and Techniques in Various Domains," International Journal of Computer Trends and Technology, vol. 71, no. 4, pp. 125-130, 2023. Crossref, https://doi.org/10.14445/22312803/IJCTT-V71I4P116

12. Hafeez, Abdul & Sial, Ali. (2021). Comparative Analysis of Data Visualization Libraries Matplotlib and Seaborn in Python [HEC Y Cat]. International Journal of Advanced Trends in Computer Science and Engineering. 10. 2770-281. 10.30534/ijatcse/2021/391012021.
13. Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.
14. Figueiras, Ana. (2018). A Review of Visualization Assessment in Terms of User Performance and Experience. 10.1109/iV.2018.00035.
15. Data-driven decisions in Sports, Garcia de Baquedano, Gabriel, https://ltu.diva-portal.org/smash/record.jsf?pid=diva2%3A1770680&dswid=-595
16. An understandable visual analysis of swimming activity and performance, Vipin Primi, Thottiparambil Ravindran, Manjusha https://his.diva-portal.org/smash/record.jsf?pid=diva2:1776352
17. C. H. Chen, W. K. Härdle, & A. Unwin (Eds.). (2007). *Handbook of data visualization*. Springer Science & Business Media.
18. Y. Wu *et al*., "iTTVis: Interactive Visualization of Table Tennis Data," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 709-718, Jan. 2018, doi: 10.1109/TVCG.2017.2744218.
19. M. Kay, "ggdist: Visualizations of Distributions and Uncertainty in the Grammar of Graphics," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 1, pp. 414-424, Jan. 2024, doi: 10.1109/TVCG.2023.3327195.
20. Addepalli, Lavanya & Sindhuja, Sakinam & Gaurav, Lokhande. (2023). A Comprehensive Review of Data Visualization Tools: Features, Strengths, and Weaknesses. International Journal of Computer Engineering in Research Trends. 10. 10-20. 10.22362/ijcert/2023/v10/i01/v10i0102.
21. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

11. Appendix

This project was completed through the collaborative efforts of Ahmed and Shweta under the Supervision of Anton Kalén. Together, Ahmed and Shweta collaborated closely throughout the project, brainstorming ideas, conducting analyses, and refining the visualizations to ensure their relevance and effectiveness.

Ahmed, with expertise in data analysis and visualization techniques, led the project from conceptualization to implementation, Background, and Literature review. Shweta has done with methodology and her proficiency in Python programming and data manipulation was instrumental in processing the golf dataset and creating meaningful visualizations.

Throughout the project, Ahmed and Shweta relied on their collective skills and expertise to overcome challenges and achieve project objectives. In completing this project, we would like to acknowledge that while the primary efforts were made by our team members, Ahmed and Shweta, we also benefited from insights obtained through literature reviews, including resources provided by OpenAI and Google. These sources offered valuable guidance and inspiration throughout the research process, aiding in the development and implementation of our methodology and strategy.