Formula 1 Championship Prediction

Parv Bhargava[1], Tanmay Ambegaokar[2], Richik Ghosh[3,] Vishal Bakshi[4]

Data Science & The George Washington University, United States of America

Abstract— This data science project explores the fascinating realm of Formula 1 racing through the lens of comprehensive data analysis and visualization. Leveraging Python as the primary tool, we delve into an extensive Formula 1 dataset encompassing race results, driver information, team details, and various performance metrics. The project aims to uncover insights into the factors influencing race outcomes, driver performance, and team dynamics.

I. Introduction

Formula 1 (F1) racing, renowned for its blend of high-speed thrills and strategic intricacies, occupies a unique position in the world of motorsport. In the pursuit of victory, each team, comprised of two drivers, navigates through a grueling calendar of 23 races in 2023, vying for both individual driver glory and the coveted Constructors' Championship. At the heart of this riveting competition lies a meticulous points system that not only crowns the best driver but also determines the most adept engineering and strategy among the constructors.

The F1 points system is a crucial facet of the sport, with each of the 10 participating teams, known as Constructors, fielding two drivers, totaling 20 competitors on the grid. Points are awarded to drivers based on their finishing positions in each race, with the top 10 finishers earning 25, 18, 15, 12, 10, 8, 6, 4, 2, and 1 point(s) respectively. This hierarchy of points reflects the competitive nature of F1, where securing a podium finish is not only a matter of prestige but a strategic move in the pursuit of championship glory.

Beyond the core race results, additional points are up for grabs for drivers achieving the fastest lap, provided they finish within the top 10. This dynamic element injects an extra layer of excitement into the competition, emphasizing the need for speed and precision throughout the race duration.

The significance of these points extends beyond individual accolades, as they collectively contribute to both the Drivers' and Constructors' Championships. The former crowns the driver with the highest point total at the end of the season, while the latter honors the team that demonstrates excellence in engineering, strategy, and driver coordination. As we embark on this data science project, the F1 point system serves as a cornerstone in our analysis. By dissecting the points distribution, we aim to unravel patterns, trends, and insights that encapsulate the essence of Formula 1 racing. Our exploration will not only showcase the impact of the point system on individual and team standings but will also shed light on the strategic nuances that make Formula 1 a captivating spectacle for fans around the globe.

In subsequent sections, we delve into our methodologies, present key findings, and discuss the implications of our analysis on the competitive landscape of Formula 1.

II. Smart Questions

1. Can we do lap time analysis using EDA?
2. Utilizing supervised learning techniques, can we predict the final points earned by a driver?

III. Literature Review

The literature reviewed suggests a diverse range of analyses and methodologies applied to F1 datasets available on Kaggle. From exploratory data analysis to predictive modeling, network analysis, time-series analysis, sentiment

analysis, and feature importance studies, researchers have leveraged these datasets to gain insights into various aspects of Formula 1. As the F1 dataset on Kaggle continues to evolve, future research may delve deeper into these themes, exploring new perspectives and advancing our understanding of the intricate dynamics within the world of Formula 1 racing. The landscape of data science research has been greatly influenced by collaborative platforms that facilitate knowledge sharing, problem-solving, and innovation. Kaggle, a prominent platform in the data science community, has emerged as a key hub for researchers and practitioners alike. Kaggle not only hosts data science competitions but also provides a repository of diverse datasets and fosters a vibrant community where participants collaboratively tackle complex challenges.

IV. DATA DESCRIPTION

Our dataset is a comprehensive compilation of 14 CSV files, each capturing extensive information about Formula 1 races spanning a significant time frame from 1950 to 2023. These files collectively provide a wealth of key insights, offering a detailed perspective on various aspects of Formula 1 racing.

The dataset encompasses diverse information, including details from trial matches, constructor positions, the location (country) of each race, the total number of laps, the configuration of the racetrack (number of turns), lap times recorded, constructor points accumulated, championship points awarded, the nationality of drivers, pit stop statistics, and the number of victories achieved by each participant.

In essence, these 14 CSV files serve as a rich source of data, allowing for in-depth analysis and exploration of Formula 1 races over the years, covering a wide range of aspects that contribute to the dynamic and competitive nature of this iconic motorsport.

V. DATA PREPARATION

Steps followed in the entire process,

1. **Dealing with NA Values.**
   All the rows which had null values in the entire dataset of 14 CSV's were transformed to NaN (Not a Number).

```
circuits = pd.read_csv('../data/circuits.csv', na_values='\\N')
constructor_results = pd.read_csv('../data/constructor_results.csv', na_values='\\N')
constructor_standings = pd.read_csv('../data/constructor_standings.csv', na_values='\\N')
constructors = pd.read_csv('../data/constructors.csv', na_values='\\N')
driver_standings = pd.read_csv('../data/driver_standings.csv', na_values='\\N')
drivers = pd.read_csv('../data/drivers.csv', na_values='\\N')
lap_times = pd.read_csv('../data/lap_times.csv', na_values='\\N')
pit_stops = pd.read_csv('../data/pit_stops.csv', na_values='\\N')
qualifying = pd.read_csv('../data/qualifying.csv', na_values='\\N')
races = pd.read_csv('../data/races.csv', na_values='\\N')
results = pd.read_csv('../data/results.csv', na_values='\\N')
seasons = pd.read_csv('../data/seasons.csv', na_values='\\N')
sprint_results = pd.read_csv('../data/sprint_results.csv', na_values='\\N')
status = pd.read_csv('../data/status.csv', na_values='\\N')
```

Fig: Converting Null to NAN

2. **Data Wrangling.**
   We have taken a data wrangling approach that effectively extracts the number of turns from Wikipedia pages, providing a valuable dataset for Formula 1 circuit analysis. The use of `Beautiful Soup` for web scraping, coupled with pandas for data manipulation, demonstrates a robust methodology for acquiring the number of turns for each race circuit.

Formula 1 racing is a complex sport where driver performance is a crucial factor in team success. The introduction of specific performance metrics and date-based segmentation enhances our ability to analyze and compare driver achievements across different aspects of racing.

'winRate' quantifies a driver's success by dividing total wins by races, indicating consistent victory. 'fastestLapRate' assesses high-speed consistency by dividing total fastest laps by races. 'qualifyingWinRate' gauges a driver's qualifying prowess by dividing pole positions by races. 'firstHalf' facilitates temporal analysis, categorizing races in the first half of the year for season-level segmentation. These metrics collectively provide nuanced insights into a driver's overall performance, highlighting their ability to win races, maintain high-speed consistency, excel in qualifying, and adapt to varying phases of a racing season.

3. **Data Merging.**

The process of consolidating information from multiple CSV files was deemed imperative due to the strategic significance of the data contained within each file. These CSV files contained essential pieces of information vital to our analytical goals. To ensure a comprehensive and cohesive analysis, we engaged in a meticulous merging process.

The primary basis for the merging operation revolved around four key identifiers: raceId, constructorId, driverId, circuitId, and statusId. These identifiers served as the linchpins for merging data entries across the various files. By utilizing these specific identifiers, we aimed to integrate and align relevant data points accurately.

In essence, this merging endeavor was driven by the need to create a consolidated dataset that encompasses all pertinent information from disparate sources. The thoughtful application of merging criteria ensures that the resulting dataset is coherent, enabling a more insightful and comprehensive analysis of the underlying data.

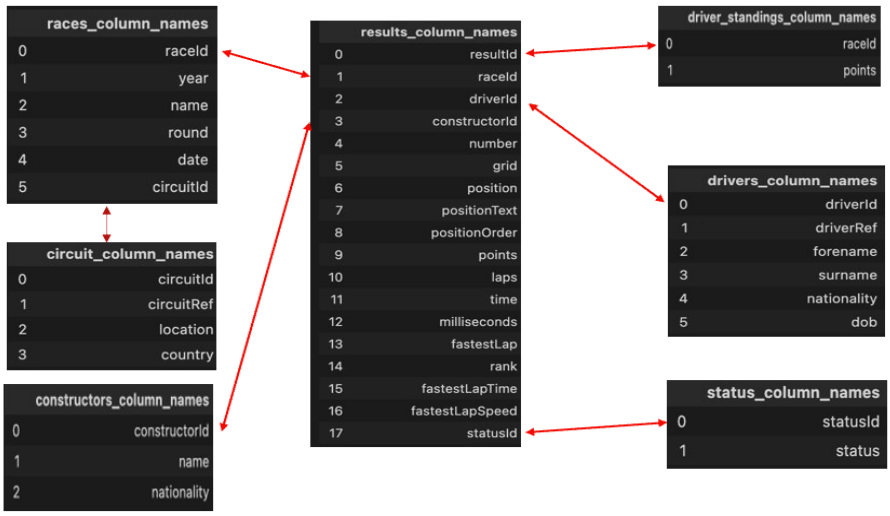

Fig: Data Frame Merging

VI. ANALYSIS

F1 RACING ANALYSIS & CHARACTERISTICS

EDA and Graphs

- The following analysis delineates the historical frequency of Formula 1 circuits utilized over the years. Italy stands out with the highest number of races in Formula 1 history, and Autodromo Nazionale di Monza emerges as the circuit within Italy hosting the most races.

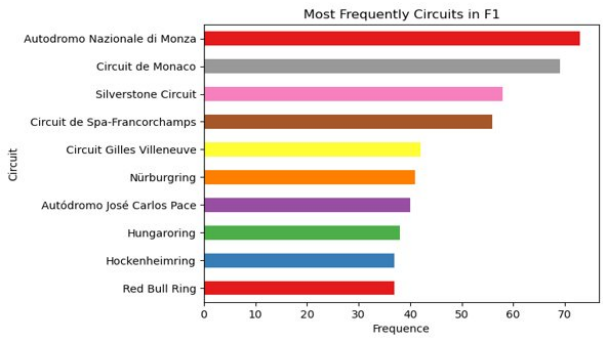

Fig: Most Frequently used Circuits

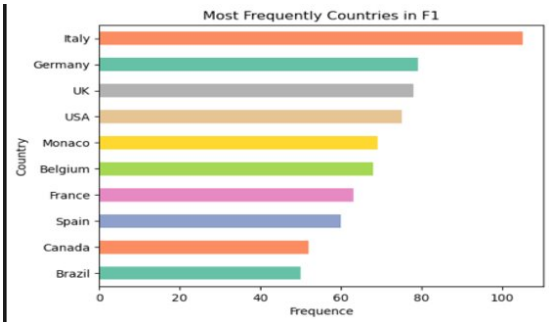

Fig: Countries with Most Frequent Races

- Observing the line plot, it is evident that the number of circuits has consistently risen each year. The most notable surge occurred between 1970 and 1980, with the count increasing from 12 to 16 circuits.
  We also see a steep decrease in 2020 because of the pandemic.

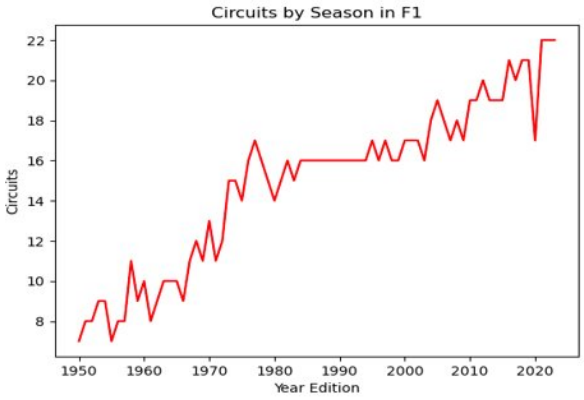

Fig: Number of Tracks from 1950-2022

- The geospatial plotting provided a comprehensive representation of the geographic locations of Formula 1 racetracks worldwide. It meticulously displayed the precise longitude and latitude coordinates of each racetrack situated across diverse continents. This detailed visualization contributes to a nuanced understanding of the global distribution of Formula 1 circuits and their geographical placements.

Fig: Geo Spatial Plot

- The pie chart illustrates the distribution of nationalities among participants in the Formula 1 championship. British and American F1 drivers collectively account for approximately 48.2% of the driver cohort, while the remaining portion comprises individuals from various other countries.

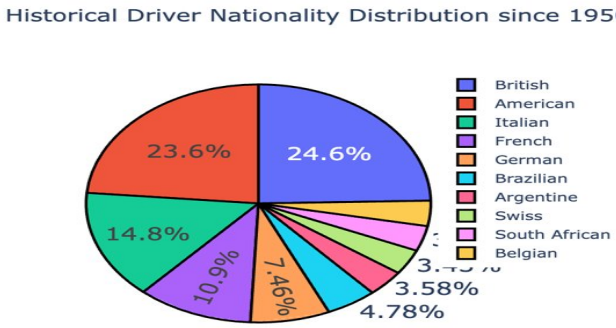

Fig: Distribution of Drivers of all Countries

VI. I. LAP TIME ANALYSIS

- The dynamic line chart provides a graphical representation of the lap progression and corresponding lap times for the top two drivers, namely Lewis Hamilton and Max Verstappen. Noteworthy events such as pit stops are discernible through spikes at lap 13 and lap 38, while a prominent spike post-lap 50 signifies the entry of the safety car onto the track.

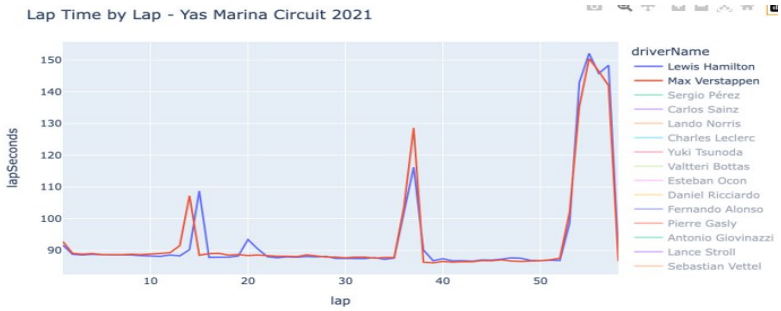

Fig: Dynamic Line Chart of Lap vs Lap Time

- In terms of lap times, this plot reveals that the conclusion of the race corresponds to the period when the car achieves its fastest lap. The analytical findings suggest that the swiftest lap is consistently recorded after 70-80% of the race duration has transpired.

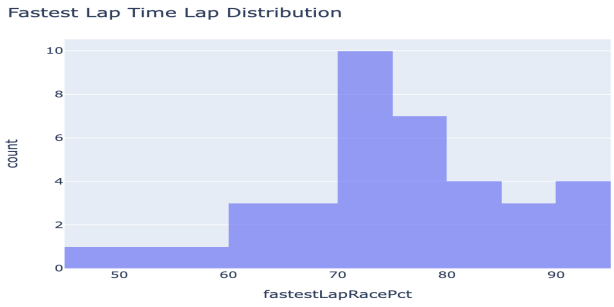

FIG: Fastest Lap Time Distribution

- Broadly observed, there is a discernible decline in the average lap time spanning the years 1990 to 2023. Periodic fluctuations are attributable to adjustments in regulations pertaining to the design and implementation of the cars.

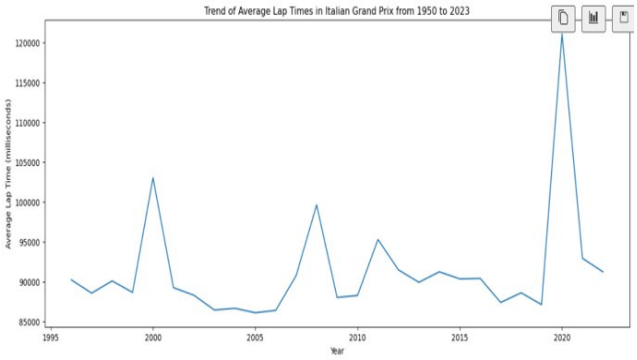

Fig: Dynamic Line Chart of Average Lap Time Analysis vs Year

VII. METHODOLOGY

In this analysis, four regression models—Linear Regression, Decision Tree Regressor, Multi-Layer Perceptron (MLP) Regressor, and K-Nearest Neighbors (KNN) Regressor—were employed to predict a driver's whole-year performance based on diverse features.

Supervised learning methods

Parametric Algorithms: -

Linear Regression: Linear regression is a supervised learning algorithm used for predicting a continuous outcome variable based on one or more predictor variables. It assumes a linear relationship between the independent and dependent variables. The model aims to find the best-fit line that minimizes the sum of squared differences between the observed and predicted values. In the context of predicting a driver's whole-year performance in motorsports, linear regression aims to quantify the influence of various factors, such as 'first_half_point,' 'winRate,' 'fastestLapRate,' and 'qualifyingWinRate,' on the target variable, 'whole_year_point.'

Non-Parametric Algorithms: -

K-nearest neighbor: k-Nearest Neighbors is a non-parametric, instance-based learning algorithm used for both classification and regression tasks. In KNN, the prediction for a new data point is determined by the majority class (for classification) or the average of k-nearest neighbors' values (for regression) in the feature space. The "k" represents the number of neighbors considered in making predictions. In our model we selected a k value of 5 which was selected by using the Elbow method.

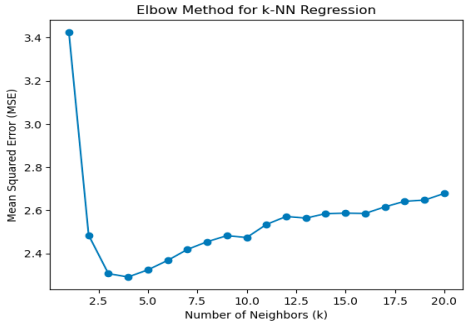

Fig. Elbow Method to Predict Best K Value.

Decision trees: Decision Trees are a versatile supervised learning algorithm used for both classification and regression tasks. The algorithm recursively splits the data based on the most informative features until a stopping criterion is met. We describe the modeling approach employed to predict the target variable, whole_year_point, using a Decision Tree Regressor. To enhance the model's generalization and mitigate overfitting, pruning is applied by optimizing the complexity parameter, ccp_alpha, through a grid search. The best ccp_alpha value was 5.0 which pruned the tree to remove overfitting.

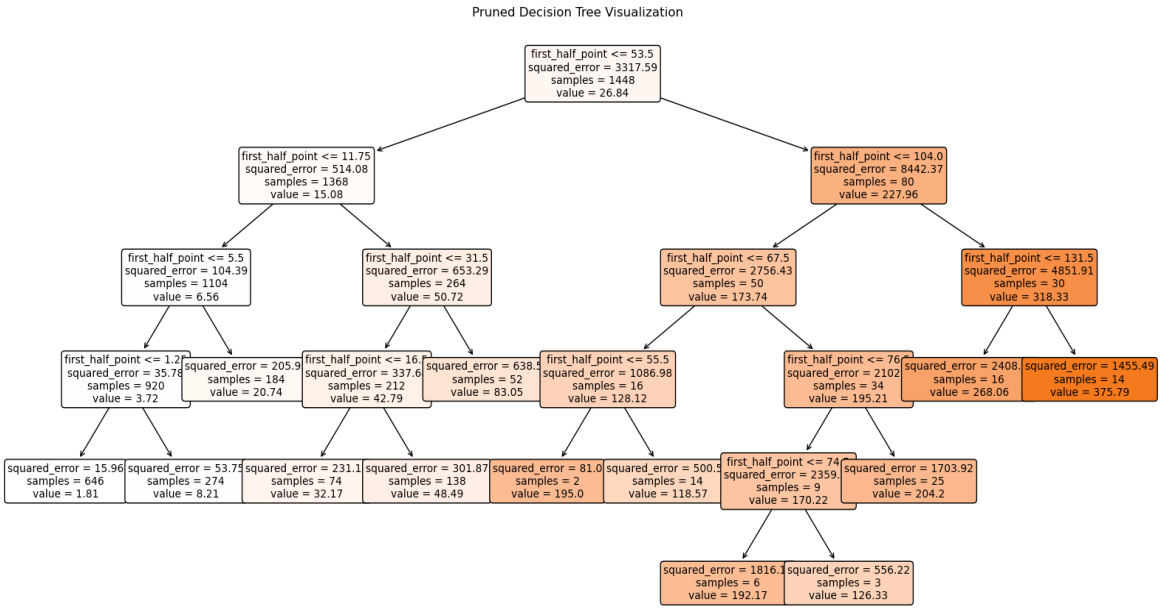

Fig. Decision Tree for our data

Unsupervised learning methods

MLP Regression: Multilayer Perceptron (MLP) Regression is a type of artificial neural network designed for regression tasks. Unlike traditional linear regression, MLP Regression can capture complex non-linear relationships between input features and the target variable. It belongs to the family of feedforward neural networks and is characterized by its multiple layers of nodes (neurons) organized into an input layer, one or more hidden layers, and an output layer. In our model we have used a hidden layer of 30, 20, 30,10,20,40 nodes in each layer. We trained the model for 300 epochs to get the best results.

VIII. RESULTS

In our comprehensive analysis of predictive models, we evaluated the performance of four distinct algorithms: Linear Regression, Pruned Decision Tree, MLP Regressor, and KNN Regressor. Each model was assessed based on key metrics, including Root Mean Squared Error (RMSE), R-squared (R^2), and Mean Absolute Percentage Error (MAPE). The results offer valuable insights into the strengths and weaknesses of each model.

	Model	RMSE	R^2	MAPE
0	Linear Regression	13.08	0.91	70.96
1	Pruned Decision Tree	13.08	0.88	61.17
2	MLP Regressor	1.32	0.82	N/A
3	KNN Regressor	14.97	0.88	70.96

Table: Results Analysis by Model

Linear Regression: The Linear Regression model demonstrated a robust fit with a low RMSE of 13.08 and a high R^2 of 0.91, indicating a strong ability to capture underlying patterns. However, the MAPE of 70.96 suggests some variability in predictions.

Pruned Decision Tree: Like Linear Regression, the Pruned Decision Tree exhibited a competitive performance with an RMSE of 13.08 and an R^2 of 0.88. The model showcased effective pattern capturing, although the MAPE slightly increased to 61.17.

MLP Regressor: The MLPRegressor stood out with exceptional accuracy, boasting a remarkably low RMSE of 1.32 and a respectable R^2 of 0.82. However, the lack of MAPE values raises interpretability concerns, emphasizing the trade-off between accuracy and transparency.

KNN Regressor: The KNN Regressor demonstrated competitive performance, with an RMSE of 14.97 and an R^2 of 0.88. Despite a higher RMSE, the model effectively captured underlying data patterns, as reflected in the MAPE of 70.96.

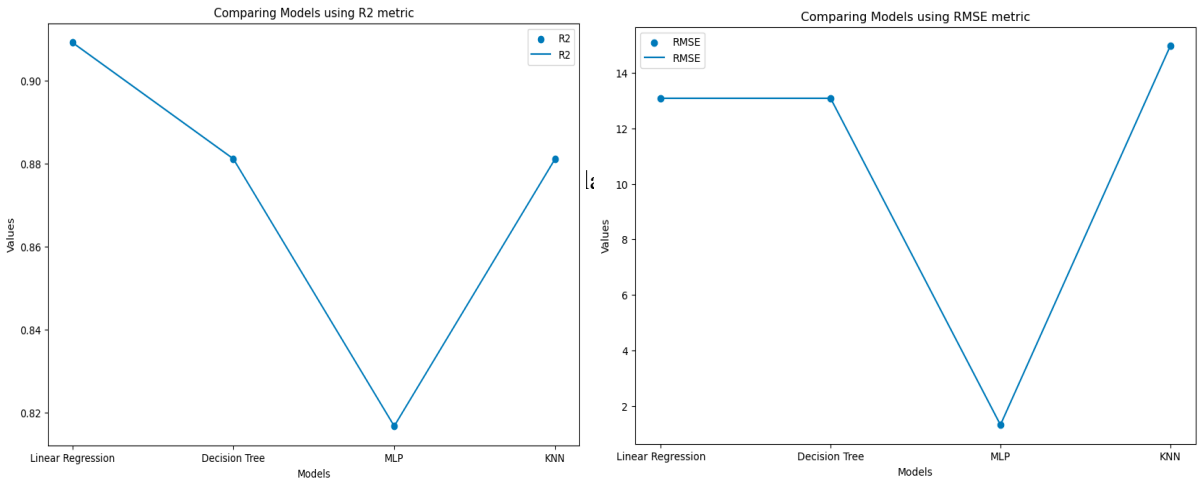

Fig: R2 and RMSE metric comparison

The graph shows that the MLP regressor has the least RMSE value and the Linear regression has the highest R2 value.

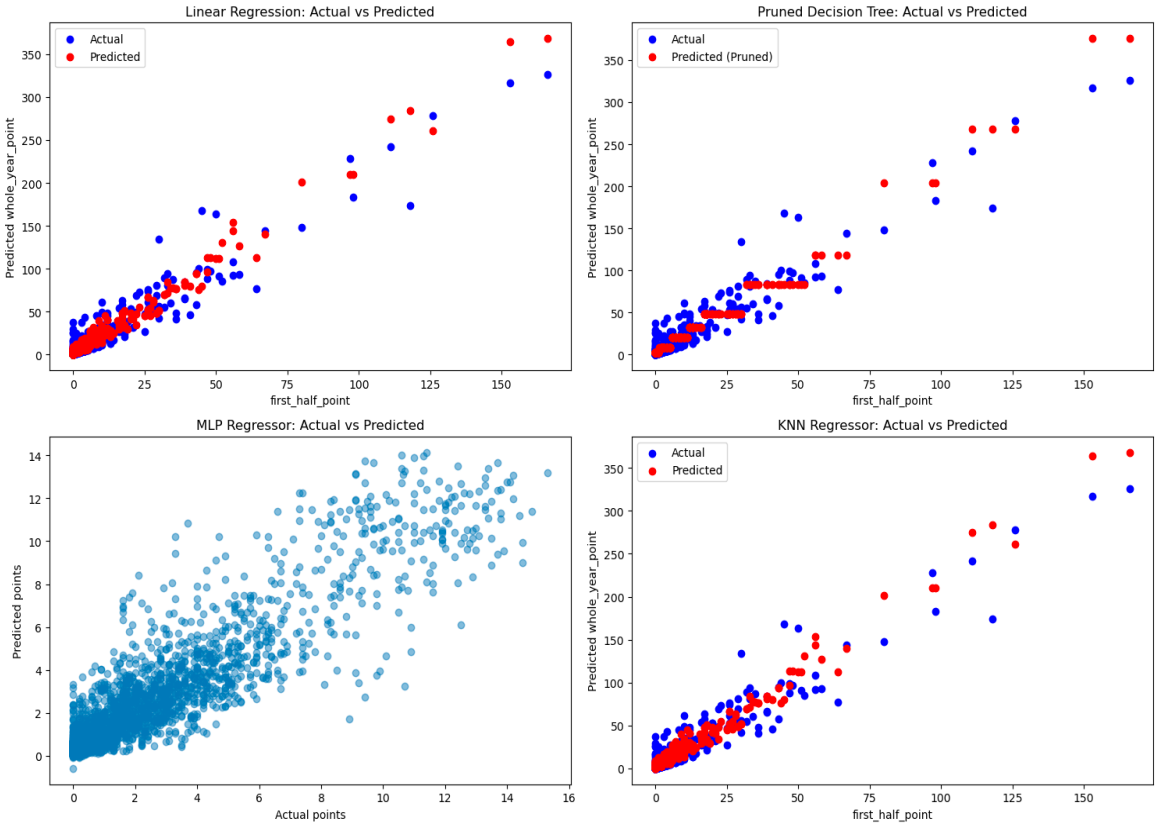

Fig: Plots of the four regression results

Our analysis provides a nuanced understanding of each model's strengths and trade-offs. Linear Regression and the Pruned Decision Tree offer robust and interpretable solutions. The MLP Regressor excels in accuracy but may pose challenges in interpretability. The KNN Regressor, with a competitive performance, strikes a balance between RMSE and R^2. Researchers should carefully consider these findings to select a model aligned with their specific use case and priorities.

IX. Conclusion

In this Formula 1 data science project, we conducted a thorough analysis spanning from 1950 to 2023, exploring race outcomes, driver performance, and team dynamics. Our examination covered diverse facets such as circuit frequencies, global track distribution, and driver nationalities.

To predict a driver's whole-year performance, we employed four regression models: Linear Regression, Pruned Decision Tree, MLP Regressor, and KNN Regressor. Each model's evaluation considered metrics like RMSE, R^2, and MAPE, revealing distinct strengths and trade-offs.

Linear Regression and the Pruned Decision Tree provided robust and interpretable insights, while the MLPRegressor excelled in accuracy. The KNN Regressor achieved a balance between RMSE and R^2.

This analysis contributes to Formula 1's understanding, showcasing data science's diverse methodologies. As Formula 1 evolves, future research can build on these insights, leveraging collaborative platforms like Kaggle to unravel the complexities of this iconic motorsport.

IX. Limitations and Future scope

The findings revealed that Linear Regression and Pruned Decision Tree models offered robust and interpretable insights into the intricate dynamics of Formula 1. On the other hand, the MLP Regressor demonstrated exceptional accuracy but introduced challenges in terms of interpretability. The KNN Regressor struck a balance between Root Mean Squared Error (RMSE) and R-squared (R^2).

While these models provide valuable contributions to our understanding of Formula 1, it's essential to acknowledge the limitations of the study. The scope and quality variations in historical data, potential inaccuracies, and the evolving nature of the sport are crucial considerations. The limited interpretability of the MLP Regressor also poses challenges in extracting actionable insights.

For future research, it is recommended to address these limitations by incorporating real-time data, considering external factors like team dynamics and technological advancements, and exploring alternative modeling approaches. Collaborative platforms like Kaggle can serve as invaluable resources for researchers to further unravel the complexities of this iconic motorsport. As Formula 1 continues to evolve, data science methodologies will play a pivotal role in uncovering new perspectives and advancing our understanding of this dynamic and captivating sport.

IX. References

- Formula One World Championship Limited. (n.d.). *F1 Analysis*. Retrieved from Formula 1: https://www.formula1.com/en/results.html/1950/races/94/great-britain/race-result.html