*Data-Driven Exploration of Formula 1 Race Strategies Using Machine Learning*

Madhuri Muppa   
Data Analytics and Engineering  
George Mason UniversityFairfax, USA [mmuupa@gmu.edu](mailto:mmuupa@gmu.edu) G01414254   
Dhavani Avu  
Data Analytics and Engineering  
George Mason UniversityFairfax, USA  
davu@gmu.edu  
G01411610 Nivedita J  
Data Analytics and Engineering  
George Mason UniversityFairfax, USA [nj@gmu.edu](mailto:nj@gmu.edu) G01409066

***Abstract***

***This research delves into the intricate world of Formula 1, employing advanced machine learning techniques to analyze and optimize race strategies. With a comprehensive dataset spanning from 1950 to the present from the Ergast Motor Racing Data API, the study explores predictive modeling for race outcomes, the influence of distinct circuits, fluctuations in constructors' performance over specific years, and the impact of demographic factors on race results. This project contributes to the evolving field of F1 analysis, utilizing gradient boosting, logistic regression, and decision tree models. The research emphasizes the importance of meticulous data preprocessing, visualizations, and geographical mapping to enhance understanding. From the heart-pounding race dynamics to the precision of machine learning, this study explores the essence of Formula 1. Key questions, from forecasting race outcomes to understanding circuit influence, drive the research. A rich literature review sets the context, citing studies on correlation analysis and predictive models. The dataset, sourced from Ergast Motor Racing Data API, spans seven decades, providing a robust foundation for analysis. Geographical visualizations, driver demographics, and constructor performances unravel the sport's complexities. Machine learning takes the spotlight, with the Gradient Boosting algorithm boasting a stellar 96.6% accuracy in predicting race outcomes. In conclusion, this research not only decodes Formula 1 intricacies but also propels machine learning applications in motorsport analytics.***

# INTRODUCTION

In the heart-pounding world of Formula 1, where every turn, every roar of the engine, and every strategic pit stop is a heartbeat, the spirit of Mario Andretti's words reverberates: "If everything seems under control, you're just not going fast enough." It's more than a sport; it's a symphony of speed, precision, and adrenaline-pumping action.

Imagine sleek, supercharged cars tearing down tracks at mind-bending speeds, pushing the boundaries of what seems possible. These speed demons hit 220 to 230 miles per hour, a spectacle that eclipses the ordinary and transcends into the extraordinary. And why are we so enamored with this high-speed theater of sound and fury? Because Formula 1 isn't just about velocity; it's a collision of cutting-edge technology, split-second decision-making, and the pursuit of perfection.

Formula 1, the crown jewel of motorsport, isn't a mere race; it's a strategic dance on the asphalt. The racetrack is a battlefield, and the teams are akin to meticulous generals, mapping out their moves with precision. The cacophony of engines, the choreography of pit stops, and the strategy involved in every lap make it a riveting spectacle. And let's not forget the rules – a symphony of technical terms and regulations that add an intellectual layer to the visceral experience.

But Formula 1 is more than a sport; it's a global phenomenon, a fever that grips the world across a season of over 20 races. It's a theater where technology meets audacity, and every race is a chapter in an epic saga. The passion of the teams, the skill of the drivers, and the constant quest for innovation make it a storyline that transcends the racetrack.

So, why embark on a journey to dissect the nuances of Formula 1? For us, the team members, it's a passion project fueled by an unbridled love for the sport. Analyzing the strategies, understanding the dynamics, and peeling back the layers of Formula 1 is not just a project; it's an odyssey into the heart of something that stirs our souls. It's about unraveling the mysteries, decoding the tactics, and immersing ourselves in a world where milliseconds matter, and victories are earned through a fusion of man and machine.

This research isn't just an academic pursuit; it's a testament to the allure of Formula 1. It's about capturing the essence of a sport that transcends the ordinary, where every lap is a story, every strategy is a gamble, and every victory is a triumph of human ingenuity. Welcome to the world of Formula 1, where the passion for speed meets the precision of strategy, and the thrill is not just in the race but in the journey itself.

# Problem Statement:

Formula 1 teams encounter the formidable task of refining race strategies to elevate performance under ever-changing and dynamic conditions.

Research Questions:

* In what manner can machine learning models be leveraged to forecast Formula 1 race outcomes?
* How do distinct Formula 1 circuits exert influence on race results?
* Are there specific years marked by substantial fluctuations in constructors' performance?
* To what extent do demographic factors pertaining to drivers impact the overall outcome of races?

# Literature Review:

The exploration of Formula 1 race strategies in conjunction with machine learning applications is an emerging field with limited existing research. Patil et al. conducted a thorough data-driven analysis, employing correlation analysis and Principal Component Analysis (PCA) to discern influential race variables. Their study, which included a linear regression model, investigated the impact of various factors like tire types and starting positions on drivers' total points in a season [6].

Garcia Tejada delved into machine learning techniques such as decision trees, random forests, support vector machines, and neural networks to predict race outcomes and optimize strategies. The challenges in sports predictions, including weather impact and pit stop strategies, were underscored, emphasizing the crucial role of data preprocessing and visualization techniques in effective model training [2].

Jasper's article provided a practical demonstration of visualizing Formula 1 pit stop and tire strategies using Python. The author's approach involved importing necessary libraries, utilizing Fastf1 to load race data, and transforming data for insightful analysis. The visualization, achieved through horizontal stacked bar charts, offered a clear overview of strategic choices made during races [3].

Tobias Lampprecht et al.'s paper introduced an interactive web-based visualization tool tailored for Formula One races. This tool's unique features, such as a calendar-based overview, dynamic race position diagrams, and lap times line plots, provided comprehensive insights into race dynamics. The incorporation of diverse interaction techniques enhanced the user experience, offering a nuanced understanding of Formula One race data [7].

Naoki Saijo et al.'s study explored the relationship between pre-driving heart rate and driving performance in Formula Car Racing. Using a wearable monitor to track heart rate during real racing situations, the research shed light on the interplay of physical and mental stressors in a competitive racing environment [4].

P. Azzoni, D. Moro, and G. Rizzoni's paper applied time-frequency signal analysis methods to estimate engine performance parameters from the acoustic emission of Formula 1 engines. The study showcased the potential for extracting useful information from the acoustic emission of race engines, even in the absence of telemetry data [5].

The article titled "Dynamic Path Planning for Formula Autonomous Racing Cars" provided valuable insights into advanced path planning strategies for autonomous racing cars. The research enhanced understanding of adaptive path planning algorithms, contributing to the optimization of trajectory planning processes in dynamic and challenging racing environments [10].

Bekker et al.'s paper, "Planning Formula One race strategies using discrete-event simulation," presented a discrete-event simulation model mimicking on-track events during Formula One races. The model aimed to assist racing teams in planning and evaluating race strategies, showcasing the significance of simulation in aiding decision-making [1].

# Data Description and Data Cleaning:

The dataset employed for this research originates from the Ergast Motor Racing Data API, a comprehensive open-source repository tailored for Formula 1 enthusiasts. Functioning as a treasure trove of information, Ergast offers detailed insights into Formula 1 races, drivers, constructors, and associated events. Covering a vast expanse of time from 1950 to the present, the dataset acts as a pivotal resource for unraveling the intricate dynamics of Formula 1 over the years [8].

The Ergast dataset is a goldmine of information, featuring diverse columns that play a crucial role in our analytical pursuits. Among the key components are:

Circuit Information: Furnishing comprehensive details about race circuits, encompassing circuit ID, name, location, country, latitude, longitude, altitude, and URLs for additional references.

Race and Qualifying Information: Encompassing race-specific details such as race ID, year, round, circuit ID, name, date, and time. Qualifying information adds another layer of depth with details like qualifying ID, lap times, and positions.

Results and Standings: Serving as the backbone of our analysis, this segment contains exhaustive race results, constructor standings, and driver standings, providing a nuanced understanding of race outcomes, points earned, and positions achieved.

Constructor and Driver Details: Offering a detailed snapshot of constructor and driver entities, including their IDs, names, nationalities, and relevant URLs. Below is the elaborative description of the tables taken from the Ergast website.

Circuits: Information about racing circuits, including unique identifiers, names, locations, countries, latitude, longitude, altitude, and Wikipedia page links.

Constructor\_results: Records of constructor results for each race, indicating race and constructor IDs, points, and status.

Constructor\_standings: Standings of constructors for each race in a season, containing race and constructor IDs, points, positions, position texts, and wins.

Constructors: Details of constructors, including unique identifiers, names, nationalities, and Wikipedia page links.

Driver\_standings: Standings of drivers for each race in a season, with race and driver IDs, points, positions, position texts, and wins.

Drivers: Information about drivers, including unique identifiers, driver numbers, codes, forenames, surnames, dates of birth, nationalities, and Wikipedia page links.

Lap\_times: Lap-by-lap details for each driver in a race, specifying race and driver IDs, lap numbers, positions, lap times, and lap times in milliseconds.

Pit\_stops: Records of pit stops during races, containing race and driver IDs, stop numbers, lap numbers, stop times, durations, and durations in milliseconds.

Qualifying: Qualifying results for drivers, including unique identifiers, race and driver IDs, constructor IDs, numbers, positions, and lap times in Q1, Q2, and Q3.

Races: Information about races, such as unique identifiers, years, rounds, circuit IDs, names, dates, times, and Wikipedia page links. Additionally, it includes dates and times for Free Practice (FP1, FP2, FP3), Qualifying, and Sprint events.

Results: Detailed results for each driver in a race, including race and driver IDs, constructor IDs, numbers, grid positions, official positions, position texts, position orders, points, laps completed, finishing times, fastest laps, ranks, fastest lap times, fastest lap speeds, and status IDs.

Sprint\_results: Similar to the results table, but specific to Sprint events.

Seasons: Information about Formula One seasons, including unique identifiers (years) and Wikipedia page links.

Status: Finishing status options with unique identifiers and status descriptions.

The Ergast dataset achieves a commendable level of granularity, allowing us to delve into the finer details of Formula 1 races. With lap-by-lap information, qualifying results, pit stop details, and race-specific statistics, our research is empowered to construct robust models for comprehensive analysis.

Our reliance on this dataset is justified not only by its historical richness, spanning over seven decades, but also by its commitment to transparency and reliability. Being an open-source API, Ergast stands as a trustworthy resource, ensuring our research is supported by accurate and detailed data.

The integration of our dataset was achieved through the harmonious incorporation of 14 distinct datasets, intricately linked by foreign keys, meticulously aligning with the exacting specifications outlined in our project requirements. Focused on preserving the integrity of our data, we conscientiously dealt with null values denoted by "\N" within these datasets. Employing a meticulous approach, we opted for a strategy that entailed either the removal of these null values or their substitution with average values as deemed appropriate. This deliberate process was undertaken with precision, ensuring that our dataset retained its resilience and remained well-suited for the rigorous analyses central to our research objectives.

# Exploratory Data Analysis:

Formula 1, as a global spectacle, unfolds on racing circuits spanning the breadth of our planet. The annual calendar boasts approximately 20 races, each hosted at a distinct racetrack among the 77 scattered worldwide. This global expanse encapsulates iconic locations, from the historic twists of Monaco's city streets to the high-speed straights of Monza. The vivid tapestry of circuits paints a dynamic backdrop for the pinnacle of motorsport, where cutting-edge technology and driver skill converge in a thrilling showcase of speed, strategy, and precision. As the roaring engines echo across diverse landscapes, Formula 1 transcends borders, captivating audiences with its blend of luxury, glamour, and the relentless pursuit of excellence.

A screenshot of a map

Description automatically generated

Figure 1 Map of circuits

## Driver Demographics:

## Driver standings are a dynamic reflection of individual performances throughout the racing season in this sport. These standings, determined by cumulative points earned, showcase drivers' consistency and success. Points are awarded not only for race wins but also for top finishes, with the scoring system promoting competitiveness across the field. Achievements like setting the fastest lap further contribute to points. Driver standings are a pivotal metric in determining individual success, team contributions, and the battle for the prestigious driver's championship, adding excitement and narrative depth to each season.

A graph of a race

Description automatically generated with medium confidence

Figure 2 Fastest lap speed of players

A screenshot of a computer

Description automatically generated

Figure 3 Top players

United Kingdom securing the highest number of victories, followed closely by drivers from Germany. This observation aligns with the significant representation of drivers from these nations in the Formula 1 landscape. The prominence of UK and German drivers in the winner's circle suggests a compelling correlation between national representation and success on the track. This demographic insight adds a layer of understanding to the diverse and competitive nature of Formula 1, where drivers from specific countries play a substantial role in shaping the sport's outcomes.

A screenshot of a computer screen

Description automatically generated

Figure 4 Wins by Nationality

A graph of different colored bars

Description automatically generated

Figure 5 Driver count by Nationality

The findings from our analysis suggest that, at present, there is insufficient technical evidence to establish a direct correlation between the age of Formula 1 drivers and their average points accumulation. Despite conducting a comprehensive examination, no conclusive patterns or correlations have emerged, indicating that other factors might play more significant roles in influencing driver performance and point outcomes in the dynamic realm of Formula 1 racing. Further research and nuanced investigations may be necessary to unravel the complexities of age-related factors in the context of driver performance.

A graph of blue dots

Description automatically generated

Figure 6 Driver Age and points

## Constructors’ data analysis:

This graphical depiction encapsulates the historical performance trajectory of Formula 1 constructors over the years. The visualization unequivocally highlights the absence of a consistent upward trend among the majority of constructors. Notably, venerable teams such as McLaren and Williams have undergone substantial fluctuations, experiencing both declines and notable advancements in their overall performance. The dynamic nature of these performance shifts underscores the intricate challenges and competitive dynamics inherent in the Formula 1 arena, where teams navigate a complex landscape of technological advancements, strategic decisions, and evolving racing conditions.

A graph of different types of data

Description automatically generated with medium confidence

Figure 7 Constructors' performance over time

The intricacies of each racing circuit unveil a compelling truth, not every car is universally adept across all terrains. Some excel in navigating high-speed corners, others showcase superior performance in low-speed twists, and there are those finely tuned for the straight stretches of the track. This nuanced differentiation in car capabilities leads to a fascinating observation, i.e., certain constructors demonstrate peak performance on specific race circuits. The correlation between constructor strengths and diverse racing environments underscores the strategic complexity faced by teams in tailoring their machines to the unique demands posed by various tracks in the Formula 1 calendar.

A graph of different colored lines

Description automatically generated

Figure 8 Constructors' performance in Grand Prix

Correlation matrix

Positive correaltion:

The strongest positive correlation with positionOrder is with the variable points. This suggests that there is some relationship between the grid position and the points. The increase in points causes increase in posistionOrder

Negative Correlation:

There is a negative correlation between positionOrder and variables like rank, grid and position\_Laptimes. This suggests that higher of those varaiables might be associated with a less position order.

A screen shot of a graph

Description automatically generated

Figure correlation matrix

## Model Build and Evaluation

In the pursuit of predicting Formula One race outcomes, various machine learning algorithms were deployed, leveraging a meticulously curated dataset named resultDf. Below are the algorithms mainly used in this research.

Logistic Regression: it is a statistical method used for binary classification. It models the relationship between a binary dependent variable and one or more independent variables, providing probabilities for the occurrence of the binary outcome.

Gradient Boosting: It is an ensemble learning technique that combines the predictions of multiple weak learners, usually decision trees. It builds a strong predictive model by sequentially correcting errors of the previous models.

Naive Bayes: It is a probabilistic classification algorithm based on Bayes' theorem. It assumes independence between features, making calculations simpler. Despite its simplicity, it often performs well in classification tasks.

Decision Tree: It is a tree-like model where each internal node represents a decision based on a feature, each branch represents the outcome of the decision, and each leaf node represents the final outcome. It's used for both classification and regression tasks.

The selected features for these algorithms, encapsulating crucial race parameters, were thoughtfully chosen to encapsulate the essence of Formula One dynamics. The feature set included 'driverId', 'constructorId', 'grid', 'laps', 'fastestLap', 'rank', 'fastestLapTime', 'fastestLapSpeed', 'statusId', 'lap\_laptimes', 'lap\_duration', 'stop', 'lap\_pitstops', and 'pit\_stop\_duration.' To enhance model performance, a temporary dataset, tempDf, was created with scaled features. The initial foray into Logistic Regression, without feature scaling, yielded a commendable 94.9% accuracy. Subsequent refinement with feature scaling elevated accuracy to 95.2%.

A blue squares with white text

Description automatically generated

Figure 10 Confusion Matrix for Logistic regression

The exploration extended to Gradient Boosting, where the initial feature set and unscaled data achieved an impressive accuracy of 96.6%.

A blue squares with numbers and a blue square

Description automatically generated with medium confidence

Figure 11 Confusion Matrix for Gradient Boosting

Naive Bayes, applied to a different feature set without scaling, exhibited a solid accuracy of 93.6%.

A blue squares with white text

Description automatically generated

Figure 12 Confusion Matrix for Naive Bayes

A challenge surfaced with Decision Tree models, prompting the implementation of cross-validation and hyperparameter tuning to address overfitting. This iterative approach resulted in a Decision Tree accuracy of 92.6%.

A blue squares with numbers and a blue gradient

Description automatically generated with medium confidence

Figure 13 Confusion Matrix for Decision tree with Cross validation

Ultimately, the standout performer emerged as the Gradient Boosting algorithm, boasting the highest accuracy at 96.6%, substantiating its efficacy in predicting Formula One race outcomes. These methodical steps in model building and evaluation significantly contribute to the research's overarching goal of leveraging machine learning for insightful predictions in the high-speed realm of Formula One racing.

A screenshot of a phone

Description automatically generated

Figure 14 Model Evaluation

# CONCLUSION

# In our research, the application of machine learning, particularly the Gradient Boosting algorithm, has emerged as a standout performer in the realm of Formula 1 race predictions. The robust results achieved, with an impressive accuracy rate of 96.6%, underscore the algorithm's prowess. Its consistent and reliable performance positions it as a formidable tool for anticipating race outcomes in the fiercely competitive and unpredictable domain of Formula 1. This revelation not only showcases the potential of machine learning but also opens avenues for further exploration of advanced predictive models in motorsport analytics.

Delving into geographical visualizations has been a captivating journey, revealing the profound global impact of Formula 1 circuits on race dynamics. Through meticulously crafted maps, we've not only captured the essence of each circuit but also provided a unique perspective on the sport's diverse and dynamic landscape. These insights extend beyond the racetrack, shedding light on the intricate interplay between geographical elements and the nuanced strategies employed in Formula 1 races. The exploration of sport’s global influence adds a layer of depth to our understanding of the multifaceted nature of Formula 1.

The analysis of constructors' points over the years has uncovered pivotal moments that significantly shaped the competitive landscape of Formula 1. Traversing historical data has allowed us to pinpoint years marked by substantial shifts in the performance of constructors. This exploration not only enriches our historical understanding of Formula 1 but also lays the groundwork for delving deeper into the factors influencing these transformative periods. By identifying and analyzing these pivotal moments, we contribute to the ongoing narrative of Formula 1's evolution and competitiveness.

Our exploration of driver demographics through visualizations has provided nuanced insights into the diverse backgrounds of Formula 1 drivers and their potential impact on race outcomes. While we didn't identify clear linear trends, our scatter plots and bar charts have offered a unique lens through which to view the intricate relationships between demographic factors and performance metrics. This holistic understanding contributes to the broader narrative of Formula 1, shedding light on the multifaceted nature of driver influences on race dynamics. It invites further exploration into the intricate web of factors shaping the diverse and dynamic world of Formula 1 racing.

The outcomes of these cleaning and preprocessing efforts set the stage for robust analyses and model development. Subsequent proceedings in the final research paper will detail the application of machine learning models and analytical techniques to derive meaningful insights into Formula 1 race dynamics, strategies, and outcomes.

# LIMITATIONS

The study encounters several limitations that warrant consideration. Firstly, the reliance on data from the Ergast Motor Racing Data API introduces a potential limitation related to data completeness and accuracy. Any gaps or inaccuracies within this dataset could influence the reliability of the findings. Additionally, while the machine learning models, particularly Gradient Boosting, exhibited high accuracy, their generalizability beyond the current dataset needs validation. Formula 1's dynamic nature, characterized by ongoing rule changes, technological advancements, and team dynamics, poses a challenge as historical data may not fully capture current race intricacies. The demographic analysis, while insightful, might be constrained by the available data, and expanding it to include more socio-economic and psychological factors could enhance its depth. Lastly, the chosen features for model training might not encompass all relevant aspects influencing race outcomes, suggesting potential avenues for further feature exploration.

# FUTURE SCOPE

Looking ahead, there are promising avenues for expanding the research. Real-time predictions could be a valuable addition, incorporating live data feeds during races for dynamic adjustments based on evolving conditions. Enhancing demographic analysis by considering additional factors like driver experience, personal backgrounds, and psychological profiles could offer a more comprehensive understanding of their impact on race outcomes. Incorporating external factors such as weather conditions, audience sentiment, or geopolitical events during races might contribute to a more holistic model, accounting for influences beyond the immediate race environment. Exploring advanced machine learning architectures like neural networks or ensemble methods could further refine predictive accuracy and developing interactive tools for Formula 1 enthusiasts to explore and analyze race data in real-time could enhance engagement.

# ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Professor Can Nguyen for his exceptional guidance and insightful teachings throughout the course on Applications of Metadata in Complex Big Data Problems. Professor Nguyen's expertise and passion for the subject have been instrumental in shaping my understanding of the intricate challenges posed by big data in complex scenarios. Furthermore, I extend my appreciation for Professor Nguyen's efforts in introducing the fascinating world of data science through captivating guest lectures. His dedication to fostering a comprehensive understanding of these topics has been invaluable, and I am truly thankful for the knowledge and skills gained under his mentorship.

##### References

1. Bekker, J., Lotz, W. Planning Formula One race strategies using discrete-event simulation. J Oper Res Soc 60, 952–961 (2009). https://doi.org/10.1057/palgrave.jors.2602626
2. Garcia Tejada, L. (2023). Applying Machine Learning to Forecast Formula 1 Race Outcomes. Aaltodoc.aalto.fi. https://aaltodoc.aalto.fi/handle/123456789/122937
3. Jasper. “Visualizing Formula 1 Race Strategies in Python Using Fastf1, Pandas and Matplotlib.” Towards Formula 1 Analysis, 17 July 2022, medium.com/towards-formula-1-analysis/visualizing-formula1-race-strategies-in-python-using-fastf1-pandas-and-matplotlib-95fe6b3298fa. Accessed 19 Sept. 2023.
4. N. Saijo, R. Nishizono and M. Kashino, "Relationship between pre-driving heart rate and driving performance in formula car racing: a case study," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021, pp. 4957-4960, doi: 10.1109/EMBC46164.2021.9630288.
5. P. Azzoni, D. Moro and G. Rizzoni, "Time-frequency signal analysis of the acoustic emission of Formula 1 engines," Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis (Cat. No.98TH8380), Pittsburgh, PA, USA, 1998, pp. 441-444, doi: 10.1109/TFSA.1998.721456.
6. Patil, Ankur, et al. A Data-Driven Analysis of Formula 1 Car Races Outcome. Jan. 2023, pp. 134–46, <https://doi.org/10.1007/978-3-031-26438-2_11>.
7. T. Lampprecht, D. Salb, M. Mauser, H. Van De Wetering, M. Burch and U. Kloos, "Visual Analysis of Formula One Races," 2019 23rd International Conference Information Visualisation (IV), Paris, France, 2019, pp. 94-99, doi: 10.1109/IV.2019.00025.
8. Terms & Conditions – Ergast Developer API. (n.d.). Retrieved October 16, 2023, from http://ergast.com/mrd/terms/
9. What is Formula 1? (n.d.). Www.rookieroad.com. <https://www.rookieroad.com/formula-1/what-is/>
10. Y. Liu et al., "Dynamic Path Planning for Formula Automous Racing Cars," 2021 40th Chinese Control Conference (CCC), Shanghai, China, 2021, pp. 6087-6093, doi: 10.23919/CCC52363.2021.9550038.

**.**