A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

FROM DATA TO PODIUM: A MACHINE LEARNING MODEL FOR PREDICTING FORMULA 1 PIT STOP TIMING

VALENTIN HETTMANN

Work project carried out under the supervision of:

Michail Batikas

Abstract

This thesis explores Formula 1 pit stop strategies through advanced analytics, with a focus on driver clustering in relation to performance, tactical, and behavioural aspects. Our approach led to the identification of four distinct driver categories, providing a framework to investigate various pit stop strategies. By integrating these driver profiles into predictive models, the study delves into the impact of driver characteristics on team strategy and pit stop efficiency. We introduce a novel dimension by developing a binary prediction model for pit stop timing, thoroughly evaluated within a simulation environment. This research contributes to a more refined understanding of strategic elements in Formula 1, demonstrating the role of tailored analytic methods in optimizing racing tactics and decision-making processes.

Keywords

Data, Data Analytics, Sports Analytics, Formula 1, Strategy, Prediction Model, Machine Learning, Simulation

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Contents

Tables of Figures

List of Tables

1 Introduction

In the competitive world of sports, the strategic use of analytics has revolutionised how competition is understood and won. This transformation is particularly evident in Formula 1, a sport where the smallest decisions can have significant impacts; leveraging data effectively is not merely an advantage but a cornerstone of success. Central to these strategies are the critical decisions made during pit stops, which can significantly influence race outcomes. In this context, the role of advanced analytics has become increasingly prominent, offering new avenues to optimize these crucial decisions.

While research has been conducted on various aspects of Formula 1 strategy, a notable gap exists in exploring driver clustering and its potential impact on strategic decisions as well as the evaluation of analytical approaches. This thesis is motivated by the prospect of harnessing advanced analytics and driver clustering to add additional insights to pit-stop strategies. The aim is to explore how a refined understanding of driver behaviour can enhance race outcomes and the interpretation of analytical outcomes when integrated with data-driven approaches.

This study is guided by the research question: "*How can the application of advanced analytics and driver clustering in Formula 1 enhance the accuracy and effectiveness of strategic decision-making, particularly in the context of optimising pit stop strategies?*"

The objectives are twofold: firstly, to develop a robust clustering model based on performance, tactical, and behavioural metrics; and secondly, to use the results of this model within analytical frameworks, namely prediction models, to assess its impact especially on pit stop timing and tire selection.

Our methodology centres around leveraging data mainly from the open-source FastF1 API, focusing on recent seasons to construct a clustering model, providing a multifaceted perspective

on driver profiles. The methodology extends to applying this clustering within analytical prediction models, aiming to shed light on the intricacies of pit-stop strategies and potentially add optimisations or additional insights.

This research aims to contribute significantly to the fields of sports analytics and Formula 1 strategic planning. By offering a novel perspective on driver behaviour and its implications for race strategy, the findings of this study could potentially guide teams and drivers towards more informed and effective decision-making. The insights gained could not only enrich academic discourse but also provide practical tools for strategic optimisation in the high-pressure environment of Formula 1 racing.

To lay the foundation for our analysis, we begin with a comprehensive literature review that covers the broader spectrum of sports analytics, with a specific focus on data analytics in Formula 1. This review will also discuss the application of clustering in other sports, noting its relatively nascent presence in Formula 1 literature. Following this, we detail our methodology, including our full data collection process and the clustering of Formula 1 drivers based on their performance metrics. In subsequent chapters, we examine various aspects of pit stop strategy—such as pit timing and explore how these elements interact with the identified driver behaviour clusters.

2 Literature Review

2.1 Evolution and Scope of Sports Analytics

Sports analytics refers to the application of scientific techniques for investigating and modelling sports performance. It entails organising historical data in a structured manner, applying predictive analytical models to the data, and utilising information systems to inform decision-makers (Morgulev, Azar, and Lidor 2018). This practice enables sports organisations to secure a competitive advantage through enhanced player performance analysis, game strategy formulation,

health and injury prevention, fan engagement, and operational strategy optimisation (Nadikattu 2020; Tan 2023). Originating in the 1960s with notational analysis in sports like American Football and Basketball (Hughes and MFranks 2004), sports analytics has evolved significantly with technological advancements and the rise of Big Data. Today, its application extends across various sports domains, from individual sports like Tennis to team sports such as Football and has become central in the data-driven world of motorsports like Formula 1. This wide adoption underscores the broad reach and applicability of sports analytics across diverse sporting disciplines (Bai and Bai 2021).

As technology evolved, sports analytics underwent a significant transformation, unveiling new avenues for analysing and improving sports performance. A recent comprehensive review by Ghosh et al. (2023) classifies the technological advancements in sports analytics into three primary research fields: sensors, computer vision, and wireless and mobile-based applications. These areas form the foundation of modern sports analytics, providing essential methods to collect, analyse, and interpret data. Such technological advancements prove particularly pertinent in Formula 1, a sport renowned for its data-driven approach. Incorporating hundreds of sensors in racing cars facilitates real-time data collection, covering various parameters such as speed, temperature, and throttle percentage (Shapiro 2023). Effective analysis of this data offers invaluable insights for making informed decisions on pit stop strategies, car setups, and race strategies. The inclusion of artificial intelligence (AI) and machine learning (ML) algorithms amplifies the potential of these technological advancements, enabling more sophisticated analysis and predictive modelling (Ghosh et al. 2023; Dindorf et al. 2023). Integrating these novel technologies with sports analytics methodologies has opened and revolutionised research opportunities in Formula 1 racing.

2.2 Analytical Approaches in Formula 1 Racing

As established in the preceding section, sports analytics is central to enhancing competitive strategies across various sports disciplines. This becomes particularly evident in Formula 1, a sport where even the minutest decision can significantly alter race outcomes. In F1, where the stakes are high and the financial implications are vast, the synergy between data analytics and elite sporting performance exemplifies the comprehensive capabilities of sports analytics. The body of literature directly engaging with sports analytics in the context of Formula 1 is limited both in terms of the quantity of research and the range of thematic focus. Broadening the search parameters to include 'Circuit Racing Motorsports' - thereby encompassing NASCAR, Formula E, and similar series - yields an expanded body of work. Preliminary examination suggests that while there has been increasing interest in this field, it appears that the field has not yet reached a point of saturation, indicating sufficient opportunity for further research and contribution.

The existing literature can be divided into several overarching categories, however, two are predominant: lap time simulations and race simulations. As Heilmeier (2018) highlights, it is crucial to differentiate between race simulations and the more prevalent lap time simulations. The latter predominates in the literature and typically focuses on the physical or engineering aspects rather than on the holistic view of an entire race. Siegler (2000) identifies three distinct approaches to lap time simulations: Steady State, Quasi-Static, and Transient. Heilmeier (2019) published a study on quasi-static lap time simulation, applying it to both Formula 1 and Formula E to illustrate its utility. Colunga (2014) examined the modelling of transient cornering and suspension dynamics, along with the investigation of control strategies for an ideal driver within a lap time simulation framework. In a similar vein, Timings (2014) contributed to this body of work by aiming to develop a robust lap time simulation, referring to its comprehensive nature and its resilience in varying conditions.

However, for this thesis, race simulations and their components, specifically those that prioritize pit stop strategies as a key component in modelling or predicting race outcomes, are of greater relevance. Such simulations are instrumental in forecasting final standings by accounting for various factors, including driver interactions, empirical fuel consumption models, tire wear, and probabilistic effects. Bekker (2009) developed one of the earlier holistic race simulations to replicate key on-track activities in Formula 1, such as mechanical failures, overtaking manoeuvres, and pit stops. This model facilitates strategy planning by simulating the mechanical and physical dynamics of a race, thereby offering a team a potential advantage. More recently, Heilmeier (2018) outlined a simulation methodology for circuit motorsport racing strategies, which considers variables such as pit stops, tire choices, and tire degradation. The tool is designed to rapidly simulate races based on discrete lap data and adjustable strategy inputs. Building on this previous work, Heilmeier (2020) introduces a new further improving the simulation. This advanced version surpasses simple optimisation models by providing a comprehensive, automated simulation that responds in real-time to race dynamics. Heilmeier (2020) suggests that the current methodology for optimising pit stop decisions and associated tire compound selections might benefit from additional exploration in the future. Furthermore, he acknowledges the omission of complex strategies, such as the undercut - a tactic where a driver pits and switches to faster tires to gain time on rivals who pit later - from current models. He advocates for the integration of such tactics into subsequent models to enrich the decision-making process.

In addition to comprehensive racing simulations, focused research activities are directed at optimising specific components of the racing domain and simulations themselves, such as pit stops. These efforts aim to refine these aspects to their utmost efficiency. One research exemplifies this by employing machine learning algorithms to aid tire strategy decisions in the NASCAR series.

This work utilises predictive analytics, employing historical race data to forecast positional shifts consequent to variables such as tire change frequency and tire lifespan. Extensive feature testing has revealed that support vector regression and LASSO regression yield the highest accuracy in results (Tulabandhula and Rudin 2014). Additionally, Bell (2016) advances the analysis of Formula 1 by conducting a comprehensive analysis of performance determinants in Formula 1, examining the evolving contributions of team dynamics and driver skills over time. The study results in a systematic ranking of drivers, offering insights into the qualifications of the 'potential best' based on a quantifiable set of criteria. Furthermore, Monte Carlo methods and analysis of probabilistic factors play a significant role in simulating the inherent variability present in lap times, pit stops, race incidents, and potential degradation of vehicle parts. The study of Heilmeier (2020) builds upon this foundation, providing a comparative analysis of the seminal works of Bekker (2009), Phillips (2014), and Salminen (2019), thereby extending the understanding of these stochastic elements in race simulations.

The existent body of research on Formula 1 is mainly based on the scope of publicly available data, which has been limited to lap times, and race outcomes. Such data has predominantly been sourced from the Ergast API, a privately maintained database for Formula 1 statistics (Ergast 2009). However, the introduction of the Fast F1 API marks a significant progression in data availability, offering not only the information provided by the Ergast API but also a more comprehensive set of F1 data. This includes telemetry data, official weather statistics, and track information (FastF1 2020). The introduction of the Fast F1 API thus promises to broaden the scope of current literature and models by facilitating the incorporation of mechanical details of the vehicle (such as current gear, RPM, and speed) and more granular data like positional coordinates, distances between drivers, and time specific air and track temperatures. The newly available data points, particularly

telemetry data, which have received little attention in the scientific literature to date, present potential opportunities to enhance and broaden current analytical frameworks. The richer and more granular nature of this data allows for a more detailed examination of specific drivers' behaviours based on positional and telemetry information. A prospective methodological approach might include clustering drivers according to their behavioural patterns in various racing scenarios. Such clustering could yield valuable insights into performance differentiators and decision-making processes throughout a race and its strategic decisions.

2.3 Clustering Techniques in Sports Analytics and Formula 1

Clustering, a core technique in unsupervised machine learning, groups data points into distinct categories based on common attributes without predefined labels (Pedregosa et al. 2011). In sports analytics, clustering is a key tool, enabling teams and coaches to decode complex patterns and detect subtle correlations. This method can help identify performance trends and effective team compositions, which, in turn, can be leveraged to improve predictive models that forecast future sports outcomes based on the grouping of player or play characteristics. Clustering may uncover non-intuitive groupings or strategies, providing a strategic advantage in the competitive world of professional sports. Its utility and adaptability across various sports disciplines are well-documented, with a significant body of literature.

In Basketball analytics, player assessment and categorisation have evolved well beyond the confines of traditional position labels. One study by Duman, Sennaroğlu, and Tuzkaya (2021) applies hierarchical cluster analysis to a rich dataset of game-related statistics from 15 NBA seasons, uncovering four to six distinct playing styles within each traditional position. The clusters, characterised by unique attributes and skill sets, offer a multifaceted perspective on player

capabilities, yielding strategic insights for player placement and team composition. Muniz and Flamand (2022) introduce an advanced clustering technique based on weighted networks. The methodology starts with k-means clustering to form preliminary groupings, which then inform a network where players are interconnected by weighted edges reflecting performance similarities. Employing the Louvain method for community detection, the study identifies eight player archetypes, surpassing the insight offered by the five traditional positions. They further enhance this method by using tracking data, adding a layer of depth to the analysis with precise measurements of player movements and interactions. For Football, a fuzzy clustering model that can handle mixed data types has been introduced. This model assigns objective weights to attributes such as player performance metrics, positional data, and physical characteristics, thereby uncovering clusters that the complexity of mixed-attribute data might hide. Such detailed analysis facilitates the identification of player clusters according to their on-field roles, skill sets, and physical profiles, which is instrumental in tactical team structuring and player market valuation (D'Urso, De Giovanni, and Vitale 2023). A study in Tennis clustered 1188 Grand Slam players by analysing their physical and play style data, revealing four distinct profiles. Through two-step cluster analysis and further MANOVA and discriminant analysis, the research identified how factors like height and handedness correlated with performance, notably in serving and net play (Cui et al. 2019). Clustering in sports analytics reaches beyond mainstream sports, extending its application to disciplines like Badminton. Sinadia and Murwantara (2022) apply k-means and hierarchical agglomerative clustering to analyse Badminton athletes' performances, identifying clusters based on game results and consecutive scoring. K-Means clustering discerns four distinct performance clusters, with one particularly strong cluster associated with high scores and consecutive points, supported by similar results from hierarchical clustering.

Collectively, these studies showcase the breadth and depth that clustering techniques bring to sports analytics. They enable a sophisticated segmentation of athletes, offering a granular understanding that can guide coaching decisions, athlete development, and competitive strategy formulation. In the data-rich context of Formula 1, the application of clustering techniques to group drivers by various data-driven similarities could potentially yield insights that extend beyond traditional performance metrics. While such methods have provided detailed understandings in other sports disciplines, the extent to which they have been applied to driver analysis in Formula 1 remains limited. Most existing studies on driver performance within this sport have focused on race results without a significant exploration of specific clustering techniques that have been beneficial in broader sports analytics.

For instance, one pioneering study by Phillips (2014) presents a statistical model that quantifies the contributions of drivers and teams to performance, using championship points as the metric. In another notable work by Rockerbie and Easton (2021), they employ an econometric method to dissect the relative impact of driver skill versus car technology on race results, discussing the "80-20 rule" but primarily concentrating on race positions and outcomes without delving into driver-specific skill and performance metrics. Similarly, a recent study utilises a Bayesian multilevel rank-ordered logit model to analyse historical ranking data, aiming to distinguish between driver skill and constructor efficiency (Van Kesteren and Bergkamp 2023). However, like its predecessors, it does not extensively investigate drivers' characteristics or behaviour.

While providing valuable insights, the discussed studies often centre around broader performance metrics and race outcomes without specifically exploring the clustering of drivers based on their attributes or behaviours. This gap indicates an opportunity for further research into the application of driver clustering techniques within Formula 1. Exploring driver clustering, especially by using

detailed telemetry data, expands the analytical horizon and could establish a foundation for in-depth examination of strategic components, such as pit stop tactics and other decision-making processes in the sport. Furthermore, the comprehensive use of telemetry data remains underutilised in existing models. Incorporating telemetry data into clustering analyses could reveal more profound insights into driver behaviour, driving styles, and performance metrics, which go beyond the conventional race outcomes and rankings.

3 Data

In the upcoming section, we comprehensively cover all pertinent aspects related to our data acquisition and processing, that were relevant for the further course of the work.

3.1 F1DataFetcher

The publicly accessible API from FastF1 and the associated Python library can be used to acquire the required data. In contrast to the previously common ERGAST API, the FastF1 API also includes interesting aspects such as weather, car position, and telemetry information in addition to the usual Formula 1 data. Telemetry information includes data such as speed, revolutions per minute etc. Position data not only informs about the exact position of a car on the racetrack but also provides information about direct duels by showing which driver is driving in front of a driver and how far away he is. This data can be a great asset and a clear advantage over existing literature. That's why the decision on FastF1 as the main source of data was made. Using this API is relatively simple and works without any problems. Interested readers are referred to the detailed documentation provided by the operators (FastF1 3.1.6). However, the API has one limitation for efficient data collecting purposes: the user can only ever load one event of a session, be it qualifying, race, free

practice, etc., from the API. This is not due to the user's authorisation but merely to the logic of the API. To get around this, we created a new logic.

The Python class we developed is called F1DataFetcher. This allows us to pull data for multiple races simultaneously from the API into a notebook, automatically creates the required data frames in the required structure and outputs them as Panda's data frames. In addition to the challenge of being able to receive multiple data at once, it was also important to be able to save the data so that local access without the need of reloading data every time was granted. Therefore, we created a relational database in PostgreSQL. This allows to save the individual data frames that are collected by the API via the F1DataFetcher in a meaningful, coherent, and comprehensible way and to keep them ready for further use. The database was created manually in pgAdmin4, and a connection was established via the Python library ‘sqlalchemy’. This package allows users to send SQL commands in Python to SQL-based databases and receive pandas’ data frames. The methods of the FastF1 Data Fetcher were enhanced by the capability of sending collected data automatically to a specified PostgreSQL data base. The final relational database contains a total of 10 tables. The structure and relations are composed as shown in Figure 1.

The final F1DataFetcher class therefore contains ten methods plus one more to collect different data frames from FastF1 (and ERGAST API) and optionally store them in a database with the above schema. The data frames are on different levels which will be explained in the following along with the methods used to collect this data.

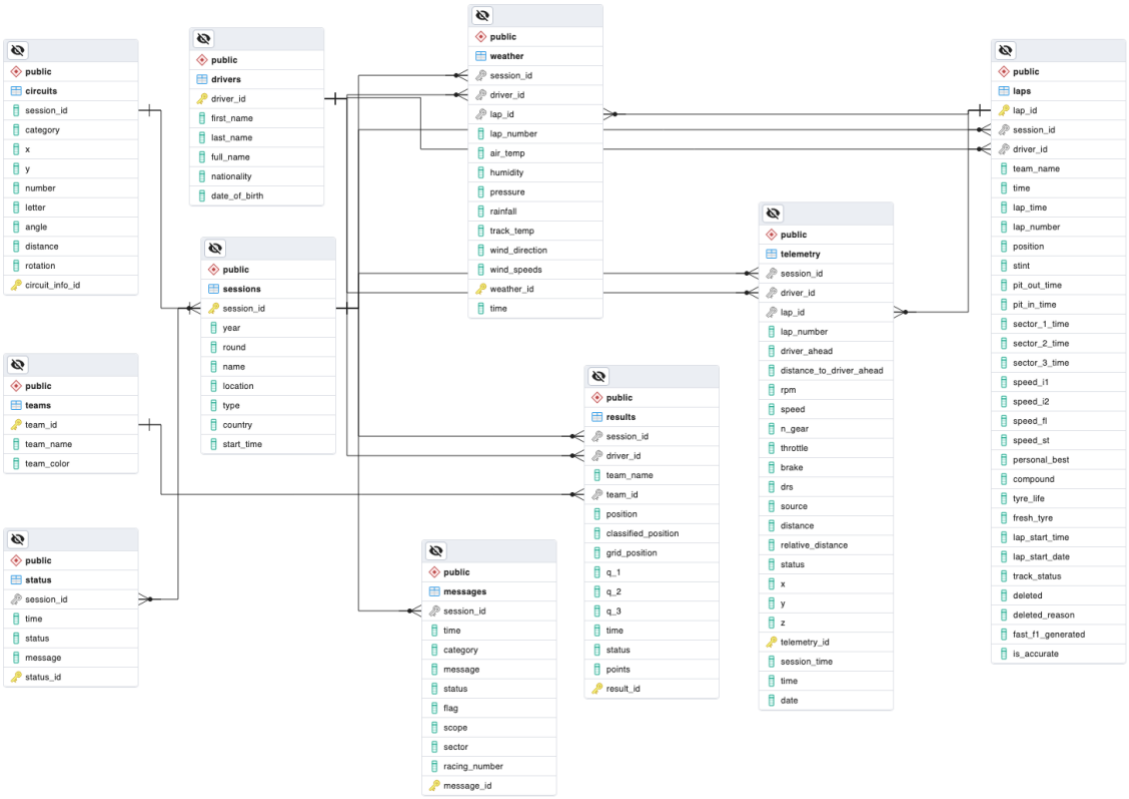

Figure 1: Simplified Schema of Our Own Data Base

The first group of methods collect general information, such as drivers or teams. Although this data is collected on session level, the information is overarching and not tied to individual sessions. A comprehensive list of the methods used to collect this data is shown in Table 1.

Table 1: F1DataFetcher Methods - Group 1

Method	Description
get_multiple_sessions()	Load detailed information on grand prix events, optionally store to database
get_unique_drivers()	Load detailed information on unique drivers, optionally store to database
get_unique_teams()	Load detailed information on unique teams, optionally store to database

The second data group pertains to session-level data. Table 2 enumerates the methods employed to gather this specific data type.

Table 2: F1DataFetcher Methods - Group 2

Method	Description
get_session_results()	Load detailed information on grand prix results, optionally store to database
get_race_control_messages()	Load detailed information on race control messages, optionally store to database
get_track_status()	Load detailed information on flags and statuses, optionally store to database
get_circuit_info()	Load detailed information on track characteristics, optionally store to database
get_weather()	Load detailed information on weather characteristics, optionally store to database

The concluding set of methods gathers highly granular data, collected individually on a lap-by-lap basis for each driver. Refer to Table 3 for a comprehensive overview.

Table 3 F1DataFetcher Methods - Group 3

Method	Description
get_laps()	Load detailed information on laps, optionally store to database
get_telemetry()	Load detailed information on telemetry, highly granular
store_to_postgresql()	store highly granular telemetry data to database

The methods within the F1DataFetcher offer a comprehensive and user-friendly approach to efficiently collect extensive data from the FastF1 API, automatically storing it in PostgreSQL. These methods were subsequently employed to establish a complete database, serving as the input data for the entire project.

3.2 Data Limitations

For feature engineering, data pre-processing and normalization, data was limited for better performance and other reasons. The data filtering process involved the following steps:

- **Exclusion of races during rain**: Grand Prix events occurring in rainy weather conditions were excluded from the analysis. The rationale behind this decision stems from the incomplete data for races held in the rain, as not all drivers participated in each event. Given the substantial differences in metrics on wet tracks and the varying participation rates, it was deemed appropriate to remove these data points from the input dataset used for feature engineering.
- **Removal from retired drivers**: Exclusion of data from retired drivers was implemented due to ambiguity in determining the reasons behind retirements—whether it was the driver's fault, another driver's influence, a strategic decision, or a mechanical issue. To ensure data reliability, the decision was made to retain only those data points where drivers successfully completed all laps of a race, omitting instances of retirement on a specific track.
- **Removal of interrupted laps**: Eliminating interrupted laps was crucial in handling noisy data caused by race interruptions due to accidents, weather conditions, or other factors. Instances where a red flag was issued, signifying a race interruption, were excluded from the dataset. This approach aimed to mitigate distortions in lap times and pit stop times, ensuring the integrity of the engineered features.

4 Clustering

4.1 Introduction to Clustering in Racing Analysis

Transitioning to clustering, a principal method in unsupervised machine learning, we leverage the refined dataset to explore and interpret the complex dynamics of Formula 1 racing.

Clustering is essential in the analysis and interpretation of complex data sets. This technique, as elucidated by Jain (2010), involves the organization of a collection of patterns into clusters, based on similarity. Patterns within a cluster exhibit a higher degree of similarity to each other than to those in different clusters, thereby revealing the intrinsic structure of the data. The significance of clustering extends beyond mere data categorization. Jain (2010) emphasizes its role in exploratory data analysis, summarization, and compression, as well as its utility as a tool for hypothesis generation and testing. In diverse fields such as image and pattern recognition, bioinformatics, and information retrieval, clustering is instrumental in uncovering hidden structures, identifying anomalies, and simplifying complex data landscapes.

In the realm of sports analytics, particularly in Formula 1 racing the application of clustering takes on a more defined and critical role. In the high-octane world of Formula 1 racing, understanding the nuances of driver behaviour is essential for gaining a competitive edge. This study employs a clustering methodology to delve into the intricacies of driving behaviour, aiming to achieve a set of specific objectives that enhance our comprehension of what differentiates top performers on the track. A central goal of this research is the characterization of driver profiles through clustering. This involves categorizing drivers into distinct groups based on a comprehensive array of performance metrics and observed behaviours on the track. The aim is to construct detailed profiles that capture the essence of each driver's approach to racing, including their performance, behaviour, and tactics. These profiles are expected to provide a holistic view of each driver's strengths and areas for improvement, offering insights into the diverse strategies and techniques employed in

Formula 1 racing. Furthermore, the study engages in a comparative analysis across the derived clusters. This analysis is crucial for identifying the unique characteristics that distinguish each cluster, thereby shedding light on what sets apart the most successful drivers. This comparative approach is anticipated to reveal valuable insights into the factors that contribute to effective driving strategies and overall race performance.

Finally, the clustering analysis aims to provide practical strategic implications for teams and drivers. By understanding the distinct behavioural clusters, teams can tailor their training programs, strategy development, and in-race decision-making to better align with the identified strengths and weaknesses of their drivers. This objective is predicated on the belief that data-driven, personalized strategies can significantly enhance performance and competitiveness in Formula 1 racing. Aligning training and strategies with the insights gained from clustering analysis, teams can optimize their approach to each race, maximizing their potential for success in this dynamic and challenging sport.

4.2 Feature Engineering

Before being able to cluster driver behaviour and performance, it is vital to define metrics and input features that capture the essence of what is tried to analyse. Since Formula 1 is a complex competition with various influential factors, it was decided to limit it down to three main aspects: The performance metrics employed to evaluate a driver's performance encompass behavioural metrics, which illustrate drivers' conduct during driving, and tactical metrics, reflecting the strategic choices made by a driver or team throughout a race.

In the following section, an overview of the features created for the clustering of Formula 1 drivers is represented, detailing the data utilized for their calculation.

4.2.1 Performance Metrics

Despite the significant dependency of performance on the car's capabilities, for which we have insufficient data, our objective was to quantify specific dimensions of driver performance. We defined three main aspects of driver performance: lap time, position gain, and consistency in driving. The first two metrics are self-explanatory and do not require additional clarification. The third metric, consistency, aims to quantify a driver's uniformity on the track, specifically regarding lap time variations. These three metrics effectively represent a comprehensive assessment of a driver's performance.

Table 4: Clustering - Performance Metrics

Variable	Name	Description
Lap time	avg_lap_time	Average lap time for each driver on all races
Position gain	position_gain	Average number of positions gained (or lost)
Consistency	consistency	Standard deviation in average lap time

4.2.2 Behavioural Metrics

More complicated than defining metrics to see a driver's performance is mapping various aspects of driving behaviour to input features. The focus was primarily on aggressiveness and a driver's behaviour towards other drivers, especially the driver ahead of him. To analyse behaviour in direct competitive scenarios, we quantified various gaps and distances relative to the leading driver, thereby attempting to delineate patterns of conduct in these head-to-head duels. After all, the five new input features for capturing driver behaviour were defined as the *distance to the driver ahead*, the *time within range*, *persistence*, *overtakes*, and *defensive actions*.

Table 5: Clustering - Behavioural Metrics

Variable	Name	Description
Distance to driver ahead	`avg_distance`	Average distance to the opponent ahead
Time within range	`time_within_range`	Average time a driver spends in a range of 50 meters behind his opponent
Persistence	`persistence`	Standard deviation of the distance to the driver ahead
Overtakes	`overtakes`	Average number of successful overtakes
Defensive actions	`position_maintained`	Average number of laps without losing a position

4.2.3 Tactical Metrics

As tactical decisions that are made by team leads and the driver himself during constant communication play an essential role in Formula 1 and can impact a driver's final position and overall performance significantly, capturing differences in strategic decisions was the goal that was aimed for. Two main aspects of Formula 1 tactics are pit stop timing and compound choice. Deciding when to pit stop, which tires to wear and how many pit stops a driver does can drastically change the outcome of a race. To capture the strategic decisions for each driver, a new set of features was created from the data to capture the pit stop timing and the compound strategies.

Table 6: Clustering - Tactical Metrics

Variable	Name	Description
Pit time	`pit_time`	Average time a pit stop takes
Pit window: Early	`early`	Percentage of pit stops in first 25% of the laps
Pit window: Mid	`mid`	Percentage of pit stops in mid 50% of the laps
Pit window: Late	`late`	Percentage of pit stops in last 25% of the laps
Laps on SOFT	`laps_on_soft_percentage`	Percentage of laps driven on SOFT compound
Laps on MEDIUM	`laps_on_medium_percentage`	Percentage of laps driven on MEDIUM compound
Laps on HARD	`laps_on_hard_percentage`	Percentage of laps driven on HARD compound

4.2.4 Normalization

Normalization is a common technique to scale data to a standard range, eliminate redundant or noisy data**,** and improve algorithm performance significantly (Patel and Mehta 2011). Literature shows that it is common practice in machine learning to apply normalization techniques to data before using it as input data (Gopal, Patro, and Kumar Sahu 2015).

Since most clustering algorithms, such as K-means and others, rely on scaled data, we needed to normalize our data points. Another reason to follow that approach is like motorsports in general and Formula 1 in particular: Information on the actual car build is not publicly accessible and kept top secret in the individual teams. They have different characteristics not shared by the teams participating in Formula 1. There might be cars performing better in general than other cars. Since the focus was merely on the driving capabilities of the individual drivers only, a need to reduce the potential bias that could influence our clustering algorithms was urged.

Besides that, some input features can be significantly influenced by track characteristics, such as the number of corners on a track and their angle. Since data from various Grand Prixes was used, results achieved on multiple tracks were compared. To make them comparable, again, there is a need for data normalization.

There are several ways of scaling and normalizing data, the most common being z-score scaling and min-max normalization. For this purpose, the decision was made to apply the latter to our input data. Min-max normalization is a technique that keeps the original relationship among the data (Gopal, Patro, and Kumar Sahu 2015). It follows a simple calculation:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

The Python library 'scikit-learn' (Pedregosa et al. 2011) provides a standard pre-processing algorithm for this kind of normalization that we applied to our data.

4.3 Selection of Clustering Algorithm

After normalizing our data to ensure comparability and reduce potential biases, as detailed previously, we move to the next crucial step in our analysis: the selection of an appropriate clustering algorithm. The choice of an algorithm is crucial, as it must align with the nature of our normalized data and the specific objectives of our study. The selection of the K-Means clustering algorithm for the analysis of driver behaviour in Formula 1 racing is underpinned by several compelling reasons that align with the specific nature of the data involved. A primary factor is the simplicity and computational efficiency of K-Means, as highlighted in Jain (2010) study. This characteristic is particularly advantageous given the large volume and complexity of data generated in Formula 1. K-Means offers an efficient and straightforward approach, which is essential for the

initial exploratory analysis where a fundamental understanding of data grouping is sought. Additionally, the suitability of K-Means for clustering numerical data, as noted by Arthur and Vassilvitskii (2007), aligns well with the mostly numerical nature of our engineered data. This alignment ensures that K-Means is well-equipped to handle the specific types of data encountered in this research. In addition, the efficiency of K-Means in processing large datasets is a significant advantage, as it allows for the swift and effective handling of the extensive data typical in Formula 1 racing. This efficiency, coupled with the ease of interpretation of its results, as Jain (2010) notes, makes K-Means a practical choice for researchers and analysts who require clear and actionable insights from their data.

K-Means has certain limitations that must be considered in its application. One challenge is that it operates under specific assumptions about the nature of the clusters it forms. Typically, the algorithm assumes that clusters are spherical and of similar size. However, as Arthur and Vassilvitskii (2007) point out, these assumptions may not always hold true in real-world data scenarios, including those encountered in Formula 1 racing. This limitation necessitates an understanding that the clusters formed may not perfectly represent the complex and varied nature of the data.

Additionally, K-Means is known to be sensitive to the initial choice of centroids. This sensitivity can lead to variability in results across different runs of the algorithm, as observed by Celebi, Kingravi, and Vela (2013). This characteristic requires careful consideration and potentially multiple iterations to ensure robustness in the clustering outcomes. This need for precise selection of starting points leads directly into the next crucial step of our analysis: determining the optimal number of clusters.

4.4 Algorithm Parameters and Tuning - Determining the Number of Clusters

In defining the number of clusters for our k-means analysis, we employed a diverse approach to ensure methodological rigor. Initially, we utilized the elbow method, widely recognized for its effectiveness in cluster analysis (Kodinariya and Makwana, 2013). This method involved plotting the explained variance against the number of clusters and identifying an 'elbow' point where the rate of decrease in variance sharply changes. Our analysis indicated an elbow point between three and four clusters, suggesting that further increasing the number of clusters would result in marginal improvements in explained variance.

To complement this quantitative method, domain knowledge regarding the performance of drivers was also incorporated. This additional layer of analysis provided valuable context in guiding the decision-making process. Specifically, the decision to opt for four clusters was influenced by the distinct characteristics of drivers Russel and Latifi, who formed a separate fourth cluster, in contrast to the other three clusters. Given their different tire choices compared to the other drivers and their overall race positioning at the lower end of the grid, it was deemed appropriate to assign them to a separate cluster. This decision underscores the importance of integrating quantitative methods with domain-specific knowledge for more insightful and informed analyses in complex situations.

While the silhouette score, a measure of cluster cohesion and separation, was also calculated as part of our evaluation, it was the combination of all three together that ultimately led us to choose four clusters. This number of clusters appeared to offer a reasonable balance, ensuring that the clusters were distinct and meaningful without introducing overfitting or unnecessary complexity. The decision to opt for four clusters was made with the goal of achieving a segmentation that would allow for an insightful analysis of driver behaviour.

4.5 Results

Applying the outlined clustering methodology, drivers in the Formula 1 season of 2019, 2020, and 2021 were effectively segregated into four distinct clusters, which can be seen in the Scatterplot in Figure 2. This is based on the presented combination of performance, tactical, and behavioural metrics.

The allocation of drivers across these clusters is as follows:

- **Cluster 0:** Albon, Raikkönen, Perez, and Magnussen.
- **Cluster 1:** Hamilton, Verstappen, and Bottas.
- **Cluster 2:** Grosjean, Kvyat, Leclerc, Norris, Ocon, Gasly, Ricciardo, Sainz, Stroll, Giovinazzi, and Vettel.
- **Cluster 3:** Latifi and Russell.

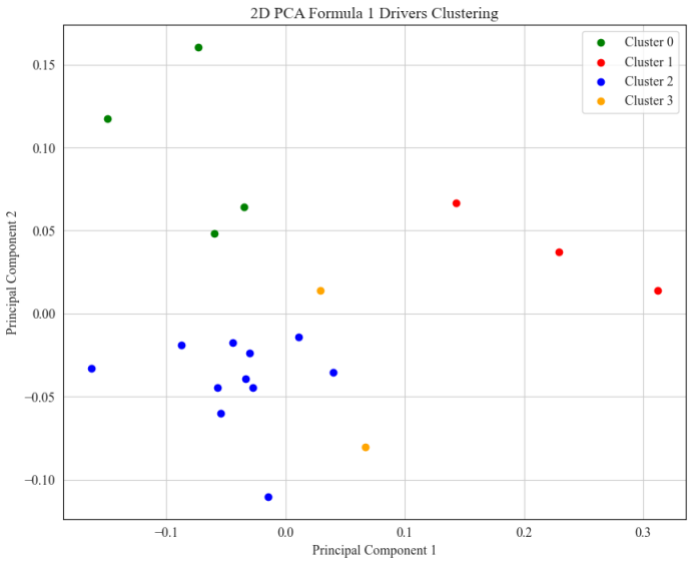

Figure 2: Scatterplot Illustrating Driver Clusters in 2D PCA Space

An initial analysis reveals a reasonable distribution reflective of known performance tiers within the sport. The top performers, typically dominating the front of the pack, are distinctly grouped in Cluster 1, while those frequently at the rear form Cluster 3. The midfield drivers, recognized

through domain knowledge as a diverse group in terms of tactics and performance, are split between Clusters 0 and 2.

To describe the clusters in more detail and to interpret the classification, the underlying performance, tactical and behavioural metrics, which were defined as input features, need to be put into perspective. The different degrees of these features can also be seen in Figure 3. Here, the averages of the features are shown in the form of a parallel plot enabling a visual comparison between the clusters.

4.6 Interpretation

When interpreting the clustering results, it is important to recall that in Formula 1 the competition involves the top 20 drivers in the world. This elite group ensures a high degree of skill parity, as illustrated by our parallel plot analysis. The analysis underscores the close competition and high level of competence inherent in this sport. However, upon closer examination, subtle but meaningful variations in performance can still be observed. It's crucial to emphasize that while these differences are partly due to the varying budgets of the teams, and consequently the performance of the cars, the vehicle is not the sole determinant of a driver's overall performance. Teams with larger financial resources often have an advantage in terms of car performance, but the skills and decisions of the driver are also key to success. Teams with limited budgets may face challenges in vehicle performance, yet the individual performance and strategy of the driver can mitigate these disadvantages to some extent. This interplay between car performance and driver skill is a key element in analysing the competitive dynamics of Formula 1.

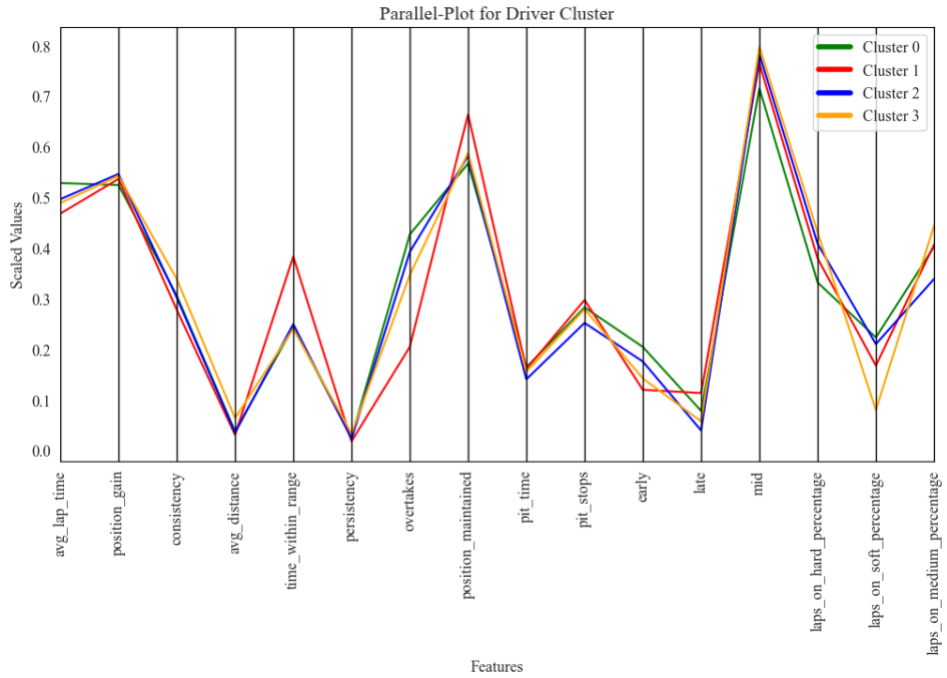

Figure 3: Parallel Plot Showing Mean Features of Driver Cluster

In our subsequent analysis, we will explore the specific characteristics of the clusters, bearing in mind the influence of vehicle differences, to provide a comprehensive understanding of the factors driving performance in Formula 1. This detailed examination starts with Cluster 0 and is based on the mean of each feature for each cluster.

Drivers in Cluster 0 are characterized by above-average lap times suggesting a focus on endurance and tactical positioning rather than outright speed. Their notable position gains reflect effective racing behaviour, with a high number of overtakes suggesting skill in dynamic racing scenarios. The moderate consistency of their performance can potentially be attributed to the positioning of the drivers in the midfield. Interactions between the drivers are much more frequent here, which has a greater impact on lap times. These interactions are also reflected in the frequent overtaking manoeuvres. The cluster's pit strategy, marked by average pit times and a preference for early stops, points to a tactical edge in race management. This is complemented by their significant usage of

hard tires, implying a strategy centred on durability and long-term race planning over the immediate speed offered by soft tires.

In essence, drivers in Cluster 0 excel in strategic manoeuvring and consistency, prioritizing race longevity and tactical positioning over peak lap speed. This approach aligns with drivers who may not have the fastest cars but utilize strategic race management to optimize their overall standings.

Moving on to Cluster 1, which is characterised by the lowest average lap times, indicating superior performance. Their excellent position maintenance skills, combined with the highest average amount of pit stops, point to strategic control of both race pace and pit strategy. The frequent pit stops in this strategy are likely a tactical decision to optimize tire performance, considering there is usually sufficient time for an additional pit stop towards the end of the race without losing positions. This approach aligns with the balanced tire usage observed, where a high percentage of laps are run on both medium and hard compounds. Their exceptional consistency underscores a notable ability to deliver stable, reliable performances across various races. Additionally, their highest time within range of the driver in front, paired with a lower number of overtakes, suggests they often lead races and if not remain close to the driver in front, efficiently managing their positions and indicating an aggressive approach to overtaking manoeuvres. In essence, Cluster 1's drivers demonstrate a blend of speed, strategic expertise, and consistent performance, marking them as elite competitors in the Formula 1 field.

Next is Cluster 2 which presents a profile of competent race management with a focus on strategic tire usage. This cluster's average lap times and position gains, along with a consistency level akin to Cluster 0, indicate a balanced approach to race performance, matching speed with strategic

positioning. The drivers in this cluster exhibit a lower frequency of overtakes compared to Cluster 0, yet they maintain their positions effectively during races. This suggests a focus on consistent lap performance and strategic placement rather than aggressive overtaking manoeuvres. Their ability to hold positions is indicative of skilful ability, particularly in defending against competitors and capitalizing on track opportunities. A notable aspect of Cluster 2 is their pit strategy. These drivers tend to pit the earliest among all clusters and have the lowest average number of pit stops. The fewer pit stops suggest a preference for longer stints on track. This aligns with their unique tire usage pattern, as Cluster 2 is the only group to perform more laps on hard tires compared to medium ones. This preference indicates a strategy leaning towards tire durability and longer race stints, possibly to maintain consistent lap times and reduce the time lost in pit stops. The hard tire's longevity would be particularly advantageous in races with high tire degradation or where preserving tire life is crucial for race strategy.

In summary Cluster 2 drivers excel in strategic racing behaviour and tire management, combining average lap times with effective position holding and a distinctive early pitting strategy, highlighting a focus on endurance and tactical adaptability in the Formula 1.

Lastly, Cluster 3 presents a unique set of characteristics. Their lap times and position gains are comparable to those in Cluster 2, suggesting a similar approach in terms of speed and race advancement. However, the distinctive feature of this cluster is the lowest consistency among all groups, as indicated by the highest standard deviation in average lap times. This variability could be attributed to a range of factors such as adapting to different race conditions, varying strategic choices, or fluctuations in both driver and car performance. Another notable aspect of Cluster 3 is their racing style, which seems to focus on endurance and stability. This is inferred from their

longest average distance to the driver ahead. Such characteristics may imply a tendency towards a more calculated and defensive racing approach, possibly prioritizing maintaining a steady pace and avoiding risks. The tire usage pattern in this cluster further supports this interpretation. The drivers in Cluster 3 have the lowest percentage of laps on soft tires, which are typically used for aggressive, short-term speed advantages. Their reluctance to use soft tires might indicate a preference for strategies that emphasize tire longevity and sustained performance over short bursts of speed. This approach is often seen in races where managing tire degradation and fuel consumption is critical for achieving a favourable outcome.

In conclusion, Cluster 3's drivers appear to adopt a strategic approach focused on endurance and stability, with less emphasis on aggressive manoeuvres and high-speed performance. Their lower consistency might reflect a more adaptive or variable strategy, responding to the specific demands of each race. This approach, combined with a cautious tire strategy, suggests a focus on long-term race management rather than immediate position gains.

In summary, the clustering analysis of Formula 1 drivers reveals distinct strategic profiles across the four groups. Cluster 0 and 2 both emphasize endurance and tactical positioning, but Cluster 2 distinguishes itself with an earlier pitting strategy and a greater emphasis on hard tires. Cluster 1 stands out for its combination of high speed, strategic pit stops, and consistent performance, showcasing the traits of front-running drivers. In contrast, Cluster 3 is defined by its variability in performance and a more conservative tire strategy, indicating a focus on stability and endurance. This comparative analysis highlights the diverse approaches in Formula 1, from aggressive speed and tactical prowess in Cluster 1 to the more calculated and endurance-focused strategies in Clusters 0, 2, and 3.

5 A Machine Learning Model for Predicting Formula 1 Pit Stop Timing

In this part, we examine the strategic element of pit stop timing in Formula 1, a key determinant in the outcome of races. Our approach combines the creation of a predictive model for pit stop decisions with an analysis based on previously defined driver clusters and their characteristics. This integration allows us to not only develop a sophisticated prediction model but also to critically evaluate its effectiveness and applicability across diverse driver profiles. By doing so, we aim to provide insights that are tailored to the distinct styles and strategies inherent in Formula 1 racing, thereby enhancing the strategic decision-making process in this highly competitive sport.

5.1 Methodology

5.1.1 Data

The following analysis predominantly utilizes data from completed laps in various Formula 1 races spanning the 2014 to 2021 seasons. We selected this timeframe due to the consistency in regulations affecting engine requirements and other vehicle-specific components. Specifically, this era in Formula 1 is notable for the implementation and continued use of 1.6-litre V6 hybrid turbocharged engines (FIA 2011). However, it is crucial to acknowledge the ongoing vehicular advancements that occurred within the regulatory parameters during this period. While there were continuous developments, the regulations provided a stable framework, ensuring a level of comparability. This consistent regulatory environment enables a foundational assumption for the analysis: the data from these seasons are comparable and can be used as a basis for predictive modelling. Such comparability is crucial as it ensures that any conclusions drawn are grounded in a period of relative stability in terms of vehicle specifications and regulations.

5.1.2 Data Acquisition

The data acquisition involved two primary sources. For the 2020 season onward, we extracted data directly from our database, as outlined in Section 3.1. For the 2014-2018 seasons, data was obtained from an external SQLite database provided by Heilmeier (2020), which was compiled using the Ergast API, a precursor to the FastF1 API currently in use. To integrate these datasets, we needed to make structural modifications to our data architecture, ensuring the seamless merging of data from 2014 to 2021. This integration created a comprehensive dataset for analysis.

5.1.3 Data Pre-processing

In the data pre-processing stage, an examination of the collected data was conducted. We applied various constraints to ensure a refined dataset without significant noise. The constraints imposed were as follows:

- **Exclusion of Races with Missing Lap Times**: Certain laps lacked complete time data, which could not be accurately reconstructed using sector times or timestamps from adjacent laps. Given the fine margins in lap times (often a matter of seconds or milliseconds) crucial in differentiating drivers, imputation techniques and statistical methods were tested but deemed inadequate, leading to errors such as negative intervals or inconsistencies between positions and cumulative lap times.
- **Exclusion of Races Involving Wet Tire Usage**: The analysis operates under the fundamental premise of excluding rain-affected races. This is in line with the data limitation of our clustering approach in Section 3.2. Therefore, races with pit stops for wet tire fittings were disregarded.
- **Disregard of Pit Stops After 90% Race Progression**: Based on the explanation by Heilmeier et. al (2020) pit stops occurring after 90% of the race completion were excluded, as they are

generally not aligned with standard strategic decisions, often influenced by the regulations for additional championship points for the fastest lap.

- **Removal of Atypical Lap and Pit Stop Times**: Laps and pit stops with significantly deviated driving and standing times, were eliminated. The average of regular, non-interrupted races, across the database multiplied by two was used as the respective threshold.
- **Exclusion of First and Second Lap Data**: Data from the initial two laps were excluded, as pit stops during this phase are typically non-strategic, often provoked by damage or unforeseen incidents during the race start.
- **Removal of Non-Tire Change Pit Stop Data**: Due to the overall goal of analysing and predicting pit stops, which are part of a tire strategy, Pit stops not involving tire changes were excluded. These are usually carried out due to technical problems such as the replacement of the front wing, etc.
- **Exclusion of Data for Drivers Finishing Outside Top 10**: As already defined and applied by Heilmeier et. al (2020) lap data for drivers finishing beyond the 10th position, which does not contribute to world championship points, were omitted. This is intended to exclude potentially ineffective strategies.

The implementation of the pre-processing criteria resulted in a refined dataset comprising 77,552 rows. As expected, the resulting dataset showed a strong imbalance in the decision to perform a pit stop, which was later defined as a target variable. This is plausible, as drivers usually perform no more than two to three pit stops on an average of 60.83 laps. The final data distribution showed 75.537 rows for no pitstop and 2.015 for the pitstop decision.

5.1.4 Race Simulation

We evaluated our predictive model using a race simulation methodology, simulating Formula 1 races from 2014 to 2019 under various conditions. The established simulation framework by Heilmeier (2020), as mentioned in our literature review, was employed. This framework accounts for long-term dynamics, including mass reduction due to fuel consumption and tire degradation, and the interactions among all race participants. The simulation's lap-by-lap discretization facilitates rapid computation, an essential feature for processing complex race scenarios efficiently. Moreover, it includes probabilistic elements to simulate varying race conditions, which can be evaluated using Monte Carlo simulations.

Integrating our model into this simulation required structural and parameter adjustments to align with the training phase's features. This was crucial for preserving input integrity and ensuring accurate output interpretation in further analyses.

5.1.5 Feature Selection

5.1.5.1 Output Variable

In our predictive model for Formula 1 pit stop strategy, the output variable *decide pitstop* was defined as a binary decision: whether a driver should execute a pit stop on a given lap. This binary classification framework is particularly suitable for several reasons. First, it directly aligns with the fundamental decision criteria in pit stop strategy, which is essentially dichotomous – to pit or not to pit. Additionally, the binary nature of the output simplifies the modelling process. It narrows down the focus to this key decision point, effectively reducing the complexity inherent in the complex nature of race strategy. This aids in maintaining clarity in the decision-making process, making it easier to interpret and apply the model's predictions in practical racing situations.

5.1.5.2 Feature Engineering

In the initial phase of feature engineering, we analysed a wide range of features. This review encompassed features identified from existing literature in the field and additional features developed for this analysis. To integrate the prediction model into the existing simulation framework, we identified features already utilized in the simulation. We then adapted and refined these features for our specific dataset. The initial set of features included *Race Progress*, *Tire Age Progress*, *Position*, *Relative Compound*, *Racetrack Category, Safety Car Status*, *Remaining Pitstops*, *Tire Change of Pursuer*, and *Close Ahead*, as identified by Heilmeier et al. (2020). In the pursuit of identifying variables directly pertinent to the pit stop decision, we engineered various additional ones. These, combined with the previously mentioned existing ones, formed the basis for the following feature selection process.

5.1.5.3 Feature Selection

The feature selection process was a vital step in developing our prediction model, as it involved identifying and selecting a subset of relevant features from a larger pool. This stage is essential not only for creating a model that is both efficient and effective but also for enhancing its performance (Guyon and Elisseeff 2003).

The process was executed using a combination of two methods, tailored to suit both linear and non-linear models in our later model comparison. Initially, we employed Variance Inflation Factors (VIF) to evaluate multicollinearity among the features. This step was particularly relevant for linear models like logistic regression in our comparison, as multicollinearity can obscure the importance of individual predictors and due to this undermine model stability (Daoud 2018). Identifying and addressing multicollinearity was crucial to ensure the distinct and substantial contribution of each feature in linear models, especially in scenarios predicting binary outcomes like pit stops.

Following this, we utilized a Random Forest classifier to assess feature importance. Random Forest, an ensemble learning method known for its robustness to multicollinearity and capacity to handle complex, non-linear relationships, provided valuable insights into the relative significance of different features (R. C. Chen et al. 2020; Li et al. 2017). This approach was particularly relevant for our non-linear models, such as Random Forest and XGBoost, in the model comparison. The insights gathered from the Random Forest analysis were instrumental in identifying the most influential variables for predicting pit stops. This informed our feature selection process, guiding us towards features with the highest predictive power across both linear and non-linear models in our comparative analysis.

5.1.5.4 Final Input Features

Based on the previously defined feature selection process, we identified a final feature set, which forms the basis for the rest of the process. A total of 5 of the newly created features were recognised as important and added to the feature set. The definition and relevance of these is as follows:

- *Interval:* Defines the time gap in seconds to the preceding competitor. This feature is central in strategic decision-making, as smaller intervals may prompt a pit stop to prevent time loss in slower traffic.
- *Gap rolling average:* Calculates the gap of the last two rolling averages over the respective last three completed laps. This metric helps assess the driver's current performance trend, which can be a determinant in deciding pit timings.
- *Lap time standard deviation:* Measures the variation of the most recent lap time compared to the last three laps. A higher deviation may indicate inconsistency or changing conditions, influencing pit stop decisions.

- *Position change:* Reflects the position change from the previous lap, normalized by the number of active drivers on track. This feature is relevant for understanding the driver's relative performance, which can impact strategic decisions.
- *Compound interaction:* This feature integrates the type of tire used with its relative age, providing a detailed view of tire performance over time.

Among the newly introduced features, two have replaced existing ones in the set, resulting in a net increase of three features in total. The first change was the replacement of *Relative Tire Age* with *Compound Interaction*. This substitution provided a more comprehensive view of tire performance by combining the type of tire with its wear over time. Additionally, we replaced the *Close Ahead* feature, which was a Boolean indicating if a driver was within 1.5 seconds of another, with the *Interval* feature. It offers a more precise measurement of the time gap to the preceding driver, enhancing the accuracy of strategic decision-making inputs.

Table 7: Feature Details - Name, Classification, Type, and Value Range

Classification	Feature	Categorical or Numerical	Value Range
Existing Features	Race progress	Numerical	[0.0, 1.0]
	Position	Numerical	[0, 20]
	FCY status	Categorical	{0, 1, 2, 3, 4}
	Remaining pit stops	Categorical	{0, 1, 2, 3}
	Tire change of pursuer	Categorical	{true, false}
	Race track category	Categorical	{1, 2, 3}
New Features	Compound Interaction	Numerical	[0.0, 1.0]
	Interval	Numerical	Variable
	Gap rolling average	Numerical	Variable
	Lap time standard deviation	Numerical	Variable
	Position change	Numerical	Variable

5.1.6 Pre-processing Pipeline

Our machine-learning pipeline was designed for optimal data pre-processing, starting with generating the *decide pitstop* target variable and excluding variables outside the defined feature set. The features were categorized by data type and transformed accordingly: one-hot encoding for categorical variables to enable accurate model interpretation without implied ordering by creating a binary column for each category of a variable, and Standard Scaler for numerical variables to normalize each feature by removing the mean and scaling it to unit variance ensuring model training efficiency and feature scale uniformity (Pedregosa et al. 2011).

Missing data handling was also crucial. We employed the Simple Imputer for categorical data, replacing missing values with the most common category, and median replacement for numerical data to maintain data integrity and minimize bias. These pre-processing steps were consolidated using a Column Transformer, ensuring a uniform and efficient data transformation process for the modelling phase.

5.1.7 Model Training

For this analysis, we employed and compared four distinct machine learning algorithms: Logistic Regression, Random Forest Classifier, XGBoost, and a Neural Network. This diverse selection of models allowed for a robust comparison across different algorithmic approaches.

Logistic Regression is a foundational model for binary classification tasks like the pitstop decision. It's particularly useful for its interpretability and efficiency in scenarios with linear relationships (Zaidi and Al Luhayb 2023).

The Random Forest Classifier is an ensemble learning method known for its high accuracy and, effectiveness in handling complex datasets with multiple features (Breiman 2001). XGBoost is a

highly efficient and scalable implementation of gradient boosting known for its performance in a wide range of classification tasks (T. Chen and Guestrin 2016). And finally, a Neural Network offers a deep learning approach, suitable for capturing non-linear relationships in the data.

Before preparing the data for further use, we excluded specific previously defined races from the data set so that they could be used as final, previously unknown data in the later simulation.

The remaining pre-processed data was then divided into training and test sets with an 80/20 split. This split facilitates independent evaluation of unseen test data, crucial for assessing generalization and detecting overfitting during training. Given the imbalance in the target variable *decide pitstop*, with it being the minority class, stratified sampling was essential. With this approach we ensured that both the training and test sets have a proportionate representation of the target variable, mitigating the risk of bias and increasing the models' ability to generalize.

We selected the F1-Score as the primary metric for model evaluation, which is the harmonic mean of precision and recall, providing a more balanced measure of a model's performance than accuracy. Especially in scenarios where class distribution is uneven, it effectively captures the trade-off between the model's precision (defined as the proportion of true positives among all positive predictions) and recall (the proportion of true positives correctly identified by the model) (Zhao 2023; Pedregosa et al. 2011). With this we established a comprehensive assessment of the model's ability to predict pit stops.

Finally, we applied hyperparameter tuning to each model to optimize their performance. For Logistic Regression, Random Forest, and XGBoost, RandomizedSearchCV with 10-fold cross-validation and 100 iterations was employed, totalling 1000 fits per model. This method balances the thoroughness of the search with computational efficiency. The Neural Network underwent tuning using the Keras Tuner, an approach tailored for deep learning models.

To tackle the class imbalance in the *decide pitstop* target variable, we evaluated oversampling methods like various forms of SMOTE but finally decided to experiment with a class weights strategy in our model training, assigning higher importance to the minority class. With this method we tried to effectively balance the class distribution without altering the dataset's integrity, thereby enhancing the models in predicting minority class instances. To further safeguard against overfitting, we evaluated all models using both training and test scores.

5.1.8 Model Selection

After implementing the methodology as outlined, we identified and evaluated four distinct machine learning models. Table 8 provides a detailed overview of these models, including their optimized parameters determined through our earlier hyperparameter tuning process. Our selection criteria prioritized the F1 score, as the key metric for model evaluation.

Table 8: Final Pit Decision Models Post-Hyperparameter Optimization

Model	Best Parameters	Test F1 Score
Logistic Regression	C: 2.19, Penalty: L2, Solver: newton-cg	0.4963
Neural Network	Number Layers: 4, Activation: relu, sigmoid Dropout Rate: 0.5 Optimizer: Adam Learning Rate: 0.001, Batch Size: 256, Loss Function: Binary Cross Entropy	0.6886
Random Forest	Max Depth: 17, Max Features: sqrt, Min Samples Leaf: 2, Min Samples Split: 6, N Estimators: 60	0.7031
XGBoost	Max Depth: 6, Colsample bytree: 0.7744, Learning Rate: 0.225, Subsample: 0.9236, N Estimators: 295	0.7270

Among the models, the XGBoost model demonstrated superior performance based on the F1 score, indicating its effectiveness in handling the class imbalance while maintaining a balance between precision and recall. Beyond this, the XGBoost model's inherent features such as handling various types of data, scalability, and efficiency in processing also contributed to its selection (T. Chen and Guestrin 2016). Therefore, it has been chosen for further development and analysis. For clarity and simplicity in our subsequent discussions, this optimized XGBoost model will be referred to as 'the model'.

5.2 Results

5.2.1 Simulation Process

To evaluate the model and its predictions, as well as to integrate them with the previously identified driver clusters, we employed a simulation as described before. For this, we selected one driver from each cluster and used the test races that were excluded from the training process, namely Abu Dhabi 2019, Spielberg 2019, Japan 2019, and Canada 2019. These races were chosen based on criteria including the fewest safety car deployments and retirements. Additionally, our clustering is limited to the 2019-2021 seasons, with 2019 serving as the essential year that intersects both our Simulation and Clustering analyses. It's important to recognize that while driver clustering offers a structured approach, it doesn't guarantee completely identical results within each cluster.

To achieve robust and reliable results, we established six distinct scenarios. These scenarios were designed to facilitate the assessment of the model both individually for each cluster and collectively across all drivers. Each scenario was replicated 5,000 times per race.

- **Scenario I (Real Strategy):** In this scenario, the simulation incorporates the actual pit stop strategies executed in the respective race. This approach serves as a benchmark, allowing for a comparison of the simulation with real-world race outcomes and as a later reference point.

- **Scenarios II-V (Cluster Specific):** These scenarios involve the selection of one driver from a specific cluster. In each scenario, only this driver employs the decision model for pit stops.
- **Scenario VI (Complete Grid):** This scenario entails the use of the decision model by all drivers participating in the race.

In our analysis, we compared the outcomes of Scenario I, which mirrors real race strategies, with actual race results to ensure a baseline for accuracy. For Scenarios II through VI, we used the results of Scenario I as a reference to maintain normalized comparability and mitigate simulation biases. To assess the effectiveness of our model, we calculated the average change in final race positions over 5,000 simulation runs across the four selected test races. In our analysis, a positive deviation indicates a positional improvement, while a negative deviation suggests a decrease. The detailed results of this analysis, including the impact on driver positions, are outlined in Figure 4.

5.2.2 Simulation Results

Looking at the results in Table 9, it is initially noticeable that the outcomes of Scenario I, which simulates real-world strategies without model interaction, show deviations from actual race results. This variance is particularly pronounced for Magnussen and Ricciardo. As Heilmeier et al. (2020) highlight, such discrepancies can arise in the simulation due to possible inaccuracies or parameterization errors.

Table 9: Comparative Analysis of Average Outcomes: Actual Results vs. Scenario I Simulation

Driver	Cluster	Real Race	Scenario I	Position Change
MAG	0	16.25	14.15	2.10
HAM	1	2.50	2.10	0.40
RIC	2	8.75	10.93	-2.18
RUS	3	16.75	15.85	0.90

In the scenarios tailored to specific clusters (Scenarios II-V) which are ultimately compared to Scenario I, visualised in Figure 4, a noteworthy trend emerges where the model appears to enhance the performance of drivers from Clusters 0 and 3. Notably, Magnussen, a representative of Cluster 0, demonstrates an average improvement of 0.65 positions when utilizing the model exclusively. This suggests that the model's strategies may be particularly advantageous for drivers in this cluster, who typically focus on endurance and tactical positioning rather than outright speed. Similarly, Russel from Cluster 3 experiences a more subtle yet positive change, with an average improvement of 0.25 positions compared to Scenario I. In stark contrast, drivers from Clusters 1 and 2, exemplified by Ricciardo and Hamilton, show a noticeable decline in performance when adopting the model's strategies. Ricciardo sees a substantial reduction in average positions, dropping by 1.33 places. This outcome could reflect a misalignment between the model's recommendations and the inherent strategies of drivers in these higher-tier clusters, known for their sophisticated racing skills and strategic agility.

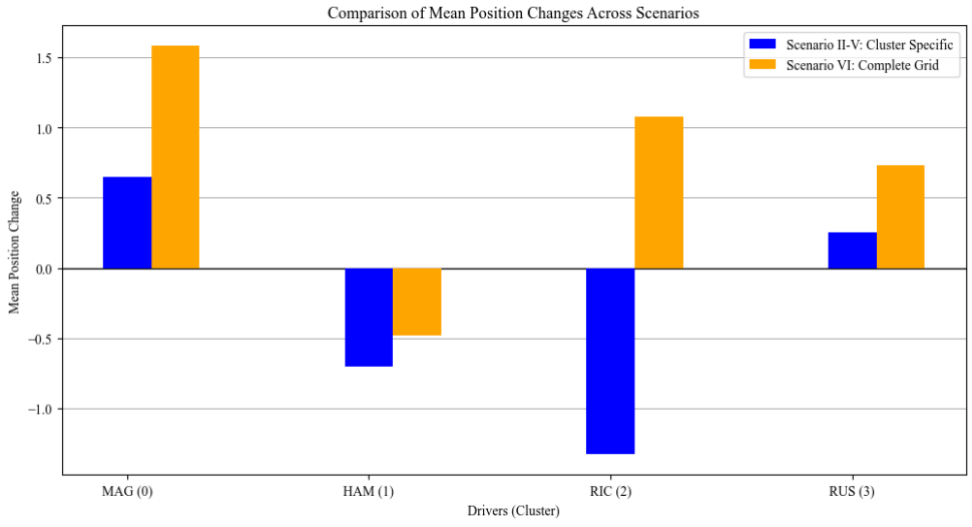

Figure 4: Variation in Positions: Highlighted Scenarios vs. Baseline from Scenario I

Hamilton's scenario warrants special consideration. As a top-tier driver with one of the strongest cars during the period under review, his limited improvement or deterioration under the model's

guidance can be attributed to the already optimized nature of his racing strategies and the minimal margin for enhancement. His case highlights the challenges in improving performance for drivers at the pinnacle of the sport, where gains are often marginal and harder to achieve.

A critical observation from Scenario VI, where the model is universally applied to all drivers, reveals an overall improvement in performance for three of the drivers, with the notable exception of Hamilton. Ricciardo's case, however, stands out uniquely. Intriguingly, when the model is applied solely to him, his performance deteriorates. Yet, in the scenario where all drivers use the model, Ricciardo's results show an improvement. This contrast indicates that while the model's strategies may not favour Ricciardo in isolation, they yield benefits in a context where all competitors are equally subject to the same model-driven approach.

This scenario seemingly establishes a more uniform competitive environment, potentially neutralizing the distinct advantages held by individual team strategies. Consequently, the model emerges as one of the key differentiators in strategic approach across the grid. The uniform application of the model levels the playing field, altering the race dynamics. Drivers like Ricciardo, who might not benefit from the model individually, gain relative advantages when all competitors adopt it, suggesting that the universal adoption could also worsen the position for some drivers, favouring those like Ricciardo who might capitalize on the altered strategic landscape.

However, the decline in Hamilton's performance remains a significant observation. It suggests that Hamilton, who often leveraged superior strategies from his high-tier team, finds these advantages diminished in a scenario where every driver follows a standardized, model-driven strategy. This implies that some of Hamilton's prior successes were partly due to strategic superiority, now less impactful in a uniformly strategic environment with a less advanced model.

5.2.3 Cluster-specific Interpretation

To validate the variation between drivers from different clusters, we simulated the Suzuka 2019 race as a test case multiple times and extracted the strategy suggestions from one of the simulation runs as an example. The examination of the strategies of Kevin Magnussen and Lewis Hamilton resulting from the model being applied to the drivers according to Scenarios II-V showcases the exemplary reasoning behind the changes in positioning connected to the defined cluster characteristics.

As illustrated in Figure 5, the model's strategic recommendation for Magnussen significantly differed from his actual race strategy. It proposed a delayed first pit stop from lap 17 to lap 20, removing the need for a second stop. This strategy, involving a longer first stint on MEDIUM tires and foregoing the switch to SOFT tires, is in line with Cluster 0's emphasis on endurance and tactical positioning. By reducing the number of pit stops and opting for the more durable HARD tires in the second stint, Magnussen's average final position improved from 13.4 to 11.9. This approach capitalizes on Cluster 0's preference for consistency over short-term speed gains, which likely contributed to Magnussen's improved standing.

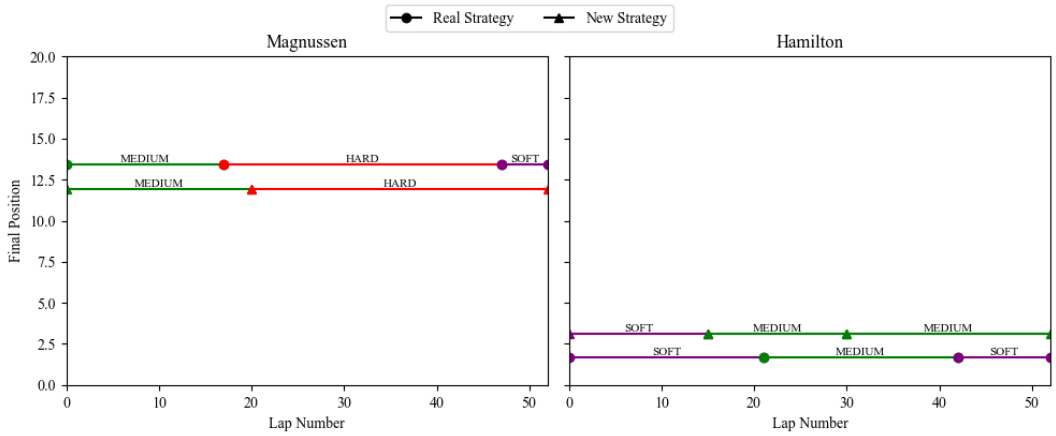

Figure 5: Updated Pitstop Strategies for Drivers in Scenarios II-V Post Model Implementation

In contrast, the model suggested an earlier pit strategy for Hamilton, with stops on laps 15 and 30 instead of his actual strategy of stops on laps 21 and 42. This led to a strategy change from an

aggressive SOFT-MEDIUM-SOFT sequence to a more conservative SOFT-MEDIUM-MEDIUM approach. Hamilton's final average position declined from 1.7 to 3.1. The model's recommendation for earlier stops and a prolonged stint on MEDIUM tires deviates from the high-speed, aggressive strategies typical of drivers in Cluster 1. This shift might have conflicted with Hamilton's cluster style, characterised by low lap times and consistent, strategic race pace control, explaining the decline in his performance.

5.3 Implications

The implications of these observations are meaningful. It indicates that while the model can enhance performance for some drivers, it may not universally benefit those who already operate at an optimized strategic level. This leads to an imperative need for a deeper exploration of how varied strategies impact drivers within their respective clusters. Specifically, it calls for examining the model's adaptability to the unique strengths and tactical preferences inherent in each cluster showing the current limitations of the model.

6 Discussion

In the initial phase of our study, we explored the application of advanced analytics in Formula 1, focusing on how driver clustering might inform pit-stop strategies. Our methodical approach to clustering revealed four distinct driver categories within the dataset. This classification emerged from a comprehensive analysis incorporating three key dimensions: performance, tactical, and behavioural patterns. These dimensions encompassed a variety of metrics, each contributing to a deeper understanding of driver profiles.

Despite the high-performance level uniformly exhibited by the drivers, which led to some similarity in cluster characteristics, our analysis succeeded in distinguishing between nuanced aspects of driver behaviour and strategy. The resulting profiles offered four differentiated interpretations of driver profiles and decision-making styles.

Subsequently, these clusters were instrumental in underpinning predictive models related to pit stop strategies. Our investigation spanned critical areas, especially including optimal pit timing, and tire selection. Through this analysis, we established associations between the divergent driver profiles and the outcomes of our analytical models. These connections allowed us to attribute specific strategic decisions and outcomes to identifiable cluster characteristics, thereby providing a deeper insight into the interplay between driver behaviour and pit stop strategy.

The research findings underscore the central role of driver clustering in enhancing the efficacy of analytical models within Formula 1. Notably, the application of driver clusters in our Pitstop Prediction Models provided a more detailed understanding, enabling us to trace and explain strategic decisions in relation to specific driver characteristics. This approach yielded insights that would likely remain obscured under a more generalized analytical framework.

These results from our study imply the critical importance of tailoring analytical models to reflect the distinct characteristics of drivers and teams in Formula 1. This customization is vital for capturing essential features and characteristics that might otherwise be overlooked in a broader, more generalized approach. Our findings suggest that the success of strategic models in Formula 1 hinges on their ability to account for the unique attributes and behaviours of individual drivers and teams.

This study, while thorough, encounters several limitations that should be considered when interpreting the results. Primarily, the availability of telemetry data through the FastF1 API, limited to the 2019 season onwards, restricts the temporal scope of our analysis of driver clusters and its evolution. Although future seasons will incrementally enrich our dataset, the absence of pre-2019 data constrains our historical analysis.

Additionally, a significant unknown in our study is the detailed information on car characteristics, which likely influences driver performance and tactics and ultimately the resulting clusters. The unavailability of these specifics, typically kept confidential by teams, may limit our understanding of the technical factors impacting driver behaviour. We attempted to mitigate this through feature engineering and normalization techniques but acknowledge that this is an approximation.

Another deliberate exclusion for our initial cluster analysis was races affected by rain. This decision was based on the disproportionate representation of wet races in our dataset, which could potentially skew the results. While this aids in maintaining data consistency, it omits the distinct strategic dynamics and behaviours prevalent in wet conditions. Furthermore, the challenge of quantifying track difficulty, despite having data on corners and marshal lights since 2019, presented another limitation. Our approach to normalising this data was a method to address this issue, but it remains uncertain.

These limitations underscore the need for cautious interpretation of our findings, particularly in their application to strategic decision-making in Formula 1. The clustering methodology applied for the 2019 to 2021 seasons and the resulting application on our advanced analytical models demonstrates promise for future applicability, yet it's important to consider these constraints and assumptions. Despite these challenges, we believe our methodology is robust and adaptable for future seasons.

Considering the dynamic nature of motorsport in general and Formula 1 in particular, which is subject to constantly evolving strategies, technologies and rules, our work is not intended to be a final solution, but rather a stimulus for thought for future work and analysis. As the scope of the data grows over time, there is the potential to deepen and develop our methodology. This promises even deeper and more differentiated insights into the strategic complexity of Formula 1 and motorsports. One key area is the expansion of the dataset. With the FastF1 API continually updating, incorporating data from subsequent seasons would not only improve the robustness of our clustering model but also facilitate longitudinal analyses. Such studies could track the evolution of driver strategies and team tactics over time. If telemetry data for seasons before 2019 becomes accessible, it would offer valuable historical insights, enriching our understanding of the sport's strategic development in response to technological and regulatory shifts.

Rain-affected races, excluded from our current analysis, represent a distinct strategic element in Formula 1. Future research could delve into these scenarios, perhaps through specialized models or by incorporating new features into existing frameworks. This focus could unveil how teams and drivers adapt to the challenges posed by variable weather conditions.

Another promising opportunity is the integration of track characteristics into the analysis. Understanding the influence of different track designs and surface conditions on driver

performance and pit stop strategies would add considerable depth to our model. Additionally, while data on specific car builds is largely confidential, future collaborations with Formula 1 teams or utilization of public data could shed light on the relationship between car technology, driver skills, and strategic choices. Also, the field of data analytics and machine learning is rapidly advancing, offering the potential for refining our clustering model with more sophisticated algorithms. These advancements could handle larger and more complex datasets, providing finer-grained insights into driver performance and pit strategies.

As Formula 1 continues to evolve, so does the opportunity for deeper analytical exploration. The advancements in data collection and analysis present fertile ground for enriching our understanding of this complex sport. By adapting to these changes and refining our methodologies, we can contribute not only to academic discourse but also to the practical strategic toolkit of teams and drivers. This research represents a step towards an even more data-driven and detailed comprehension of Formula 1 racing, setting the stage for future scholarly and practical advancements in the field.

7 Conclusion

This thesis has delved into the realm of Formula 1 pit stop strategies, employing advanced analytics to understand how driver clustering informs strategic decisions. We discovered four distinct driver categories based on performance, tactical, and behavioural dimensions, providing a nuanced view of driver profiles within the sport. This classification served as a foundation for examining various facets of pit stop strategy, including optimal pit timing, tire selection and the influence of different pit strategies. The research illustrates the impact of driver characteristics on pit stop strategies in Formula 1.

This study enhances our comprehension of how individual behaviours and decisions intertwine with team strategies by linking divergent driver profiles to strategic outcomes. Using driver clusters in predictive models has shed light on the complexities of strategic decision-making in this high-speed sport. This thesis contributes to the understanding of Formula 1 racing by offering a data-informed perspective on the strategic elements of the sport. It underscores the value of bespoke strategies that consider the distinct qualities of drivers and teams, highlighting the potential of analytics in refining racing tactics.

In summary, this thesis provides an insightful exploration of the strategic dimensions of Formula 1 pit stops. It offers a clearer picture of how data analytics can be applied to decode the intricacies of racing strategies, enriching our understanding of this dynamic and technologically advanced sport.

References

Anıl Duman, Eyüp, Bahar Sennaroğlu, and Gülfem Tuzkaya. 2021. "A Cluster Analysis of Basketball Players for Each of the Five Traditionally Defined Positions." *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*. https://doi.org/10.1177/17543371211062064.

Arthur, David, and Sergei Vassilvitskii. 2007a. *K-Means++: The Advantages of Careful Seeding*. *Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms*. Vol. 8. https://doi.org/10.1145/1283383.1283494.

———. 2007b. *K-Means++: The Advantages of Careful Seeding*. *Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms*. Vol. 8. https://doi.org/10.1145/1283383.1283494.

Bai, Zhongbo, and Xiaomei Bai. 2021. "Sports Big Data: Management, Analysis, Applications, and Challenges." *Complexity* 2021: 1–11. https://doi.org/10.1155/2021/6676297.

Bekker, J., and W. Lotz. 2009. "Planning Formula One Race Strategies Using Discrete-Event Simulation." *Journal of the Operational Research Society* 60 (7): 952–61. https://doi.org/10.1057/palgrave.jors.2602626.

Bell, Andrew, James Smith, Clive E. Sabel, and Kelvyn Jones. 2016. "Formula for Success: Multilevel Modelling of Formula One Driver and Constructor Performance, 1950-2014." *Journal of Quantitative Analysis in Sports*. https://doi.org/10.1515/jqas-2015-0050.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45: 5–32. https://doi.org/http://dx.doi.org/10.1023/A:1010933404324.

Celebi, M. Emre, Hassan A. Kingravi, and Patricio A. Vela. 2013. "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm." *Expert Systems with Applications* 40 (1): 200–210. https://doi.org/10.1016/J.ESWA.2012.07.021.

Chen, Rung Ching, Christine Dewi, Su Wen Huang, and Rezzy Eko Caraka. 2020. "Selecting Critical Features for Data Classification Based on Machine Learning Methods." *Journal of Big Data* 7 (1): 1–26. https://doi.org/10.1186/s40537-020-00327-4.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016:785–94. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.

Colunga, Ivan Fernandez, and Andrew Bradley. 2014. "Modelling of Transient Cornering and Suspension Dynamics, and Investigation into the Control Strategies for an Ideal Driver in a Lap Time Simulator." *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 228 (10): 1185–99. https://doi.org/10.1177/0954407014525362.

Cui, Yixiong, Miguel Ángel Gómez, Bruno Gonçalves, and Jaime Sampaio. 2019. "Clustering Tennis Players' Anthropometric and Individual Features Helps to Reveal Performance Fingerprints." *European Journal of Sport Science* 19 (8): 1032–44. https://doi.org/10.1080/17461391.2019.1577494.

Daoud, Jamal I. 2018. "Multicollinearity and Regression Analysis." In *Journal of Physics: Conference Series*. Vol. 949. Institute of Physics Publishing. https://doi.org/10.1088/1742-6596/949/1/012009.

Dindorf, Carlo, Eva Bartaguiz, Freya Gassmann, and Michael Fröhlich. 2023. "Conceptual Structure and Current Trends in Artificial Intelligence, Machine Learning, and Deep Learning Research in Sports: A Bibliometric Review." *International Journal of Environmental Research and Public Health* 20 (1). https://doi.org/10.3390/ijerph20010173.

D'Urso, Pierpaolo, Livia De Giovanni, and Vincenzina Vitale. 2023. "A Robust Method for Clustering Football Players with Mixed Attributes." *Annals of Operations Research* 325 (1): 9–36. https://doi.org/10.1007/s10479-022-04558-x.

Ergast. 2009. "Ergast Developer API." 2009. http://ergast.com/mrd/.

"FastF1 3.1.6." n.d. Accessed December 18, 2023. https://docs.fastf1.dev/.

FIA. 2011. "FIA Formula One World Championship Power Unit Regulations." FEDERATION INTERNATIONAL DE L'AUTOMOBILE.

Ghosh, Indrajeet, Sreenivasan Ramasamy Ramamurthy, Avijoy Chakma, and Nirmalya Roy. 2023a. "Sports Analytics Review: Artificial Intelligence Applications, Emerging Technologies, and Algorithmic Perspective." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. John Wiley and Sons Inc. https://doi.org/10.1002/widm.1496.

———. 2023b. "Sports Analytics Review: Artificial Intelligence Applications, Emerging Technologies, and Algorithmic Perspective." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 13 (5). https://doi.org/10.1002/widm.1496.

Gopal, S, Krishna Patro, and Kishore Kumar Sahu. 2015. "Normalization: A Preprocessing Stage."

Guyon, Isabelle, and André Elisseeff. 2003. "An Introduction to Variable and Feature Selection André Elisseeff." *Journal of Machine Learning Research* 3: 1157–82.

Heilmeier, Alexander. 2020. "F1-Timing-Database: SQLite Database Containing Formula 1 Lap and Race Timing Information for the Seasons 2014 - 2019." 2020. https://github.com/TUMFTM/f1-timing-database.

Heilmeier, Alexander, Maximilian Geisslinger, and Johannes Betz. 2019. "A Quasi-Steady-State Lap Time Simulation for Electrified Race Cars." In *2019 Fourteenth International Conference*

on Ecological Vehicles and Renewable Energies (EVER). IEEE. https://doi.org/10.1109/EVER.2019.8813646.

Heilmeier, Alexander, Michael Graf, Johannes Betz, and Markus Lienkamp. 2020. "Application of Monte Carlo Methods to Consider Probabilistic Effects in a Race Simulation for Circuit Motorsport." *Applied Sciences (Switzerland)* 10 (12). https://doi.org/10.3390/app10124229.

Heilmeier, Alexander, Michael Graf, and Markus Lienkamp. 2018. "A Race Simulation for Strategy Decisions in Circuit Motorsports." In *2018 21st International Conference on Intelligent Transportation Systems (ITSC) Maui, Hawaii, USA, November 4-7, 2018*. Maui.

Heilmeier, Alexander, André Thomaser, Michael Graf, and Johannes Betz. 2020. "Virtual Strategy Engineer: Using Artificial Neural Networks for Making Race Strategy Decisions in Circuit Motorsport." *Applied Sciences (Switzerland)* 10 (21): 1–32. https://doi.org/10.3390/app10217805.

Hughes, Mike, and Ian MFranks. 2004. "Notational Analysis of Sport Second Edition: Systems for Better Coaching and Performance in Sport."

Jain, Anil K. 2010. "Data Clustering: 50 Years beyond K-Means." *Pattern Recognition Letters* 31 (8): 651–66. https://doi.org/10.1016/J.PATREC.2009.09.011.

Kesteren, Erik Jan Van, and Tom Bergkamp. 2023. "Bayesian Analysis of Formula One Race Results: Disentangling Driver Skill and Constructor Advantage." *Journal of Quantitative Analysis in Sports*. https://doi.org/10.1515/jqas-2022-0021.

Li, Jundong, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017. "Feature Selection: A Data Perspective." *ACM Comput. Surv* 50 (6): 94–139. https://doi.org/10.1145/3136625.

Morgulev, Elia, Ofer H. Azar, and Ronnie Lidor. 2018. “Sports Analytics and the Big-Data Era.” *International Journal of Data Science and Analytics* 5 (4): 213–22. https://doi.org/10.1007/s41060-017-0093-7.

Muniz, Megan, and Tulay Flamand. 2022. “A Weighted Network Clustering Approach in the NBA.” *Journal of Sports Analytics* 8 (4): 251–75. https://doi.org/10.3233/jsa-220584.

Nadikattu, Rahul Reddy. 2020. “IMPLEMENTATION OF NEW WAYS OF ARTIFICIAL INTELLIGENCE IN SPORTS.” *Journal of Xidian University* 14 (5). https://doi.org/10.37896/jxu14.5/649.

Patel, Vaishali R., and Rupa G. Mehta. 2011. “Impact of Outlier Removal and Normalization Approach in Modified K-Means Clustering Algorithm.” *IJCSI International Journal of Computer Science Issues* 8 (5): 331–36.

Pedregosa, Fabian, Vincent Michel, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Jake Vanderplas, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–30. http://scikit-learn.sourceforge.net.

Phillips, A. 2014. “Building a Race Simulator.” https://f1metrics.wordpress.com/2014/10/03/building-a-race-simulator/.

Phillips, Andrew J.K. 2014. “Uncovering Formula One Driver Performances from 1950 to 2013 by Adjusting for Team and Competition Effects.” *Journal of Quantitative Analysis in Sports* 10 (2): 261–78. https://doi.org/10.1515/jqas-2013-0031.

Rockerbie Duane, and Easton Stephen. 2021. “Race to the Podium: Separating and Conjoining the Car and Driver in F1 Racing.”

Salminen, Tomi. 2019. “Race Simulator: Downloadable R Program Code.” 2019. https://f1strategyblog.wordpress.com/2019/05/10/race-simulator-downloadable-r-program-code/.

Shapiro, Joel. 2023. “Data Driven at 200 MPH: How Transforms Formula One Racing.” January 26, 2023. https://www.forbes.com/sites/joelshapiro/2023/01/26/data-driven-at-200-mph-how-analytics-transforms-formula-one-racing/.

Siegler, Blake, Andrew Deakin, and David Crolla. 2000. “Lap Time Simulation: Comparison of Steady State, Quasi-Static and Transient Racing Car Cornering Strategies.” In *SAE Motorsports Engineering Conference & Exposition*. https://doi.org/10.4271/2000-01-3563.

Sinadia, Herbie Ewaldo, and I. Made Murwantara. 2022. “Sports Analytics: A Comparison of Machine Learning Perfomance for Profiling Badminton Athlete.” In *Proceedings - 2022 1st International Conference on Technology Innovation and Its Applications, ICTIIA 2022*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICTIIA54654.2022.9935852.

Tan, Xiaomeng. 2023. “Enhanced Sports Predictions: A Comprehensive Analysis of the Role and Performance of Predictive Analytics in the Sports Sector.” *Wireless Personal Communications*, October. https://doi.org/10.1007/s11277-023-10585-z.

theOehrly. 2020. “FastF1.” 2020. https://docs.fastf1.dev/index.html.

Timings, Julian, and David Cole. 2014. “Robust Lap-Time Simulation.” In *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 228:1200–1216. SAGE Publications Ltd. https://doi.org/10.1177/0954407013516102.

Tulabandhula, Theja, and Cynthia Rudin. 2014. “Tire Changes, Fresh Air, And Yellow Flags: Challenges in Predictive Analytics For Professional Racing.” *Big Data* 2 (2): 97–112. https://doi.org/10.1089/big.2014.0018.

Zaidi, Abdelhamid, and Asamh Saleh M. Al Luhayb. 2023. “Two Statistical Approaches to Justify the Use of the Logistic Function in Binary Logistic Regression.” *Mathematical Problems in Engineering* 2023 (April): 1–11. https://doi.org/10.1155/2023/5525675.

Zhao, Zibin. 2023. “Transforming ECG Diagnosis:An In-Depth Review of Transformer-Based DeepLearning Models in Cardiovascular Disease Detection.” *Department of Chemical and Biological Engineering*, June. http://arxiv.org/abs/2306.01249.