National College of Ireland

Predictive Model for Pitstop Strategy in Formula 1 using Ensemble Learning

MSc Research Project
Masters of Science in Data Analytics

Aniketh Mahesh Rao
Student ID: X22166343

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Aniketh Mahesh Rao
Student ID:	X22166343
Programme:	Masters of Science in Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Noel Cosgrave
Submission Due Date:	14/12/2023
Project Title:	Predictive Model for Pitstop Strategy in Formula 1 using Ensemble Learning
Word Count:	6228
Page Count:	15

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	14th December 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Model for Pitstop Strategy in Formula 1 using Ensemble Learning

Aniketh Mahesh Rao
X22166343

Abstract

The research explores the complex world of Formula 1 racing, where winning or losing might depend on split-second decisions made during pit stops. This project investigates how tire deterioration, weather, and sensor data affect pit-stop tactics and race results using advanced analytics and supervised machine learning. The research aims to optimize pit-stop decision-making by building historical data-driven models. One distinctive feature of the model is its adaptability to user input, which enables people to enter their own data or preferences and receive corresponding results from the predictive model. This opens the door to the development of an optimized pit-stop race strategy model. The research uses concepts from ensemble learning to rethink race strategy, driven by the strategic significance of pit stops and the dynamic nature of Formula 1 racing. The all-encompassing model combines a variety of data sources, including as historical race results, telemetry data, weather patterns, and tire performance factors, to offer a flexible tool for optimizing pit stop tactics in a range of racing contexts. The model for predicting pit stops correctly determines the best laps to stop, following the patterns of Formula 1 races. The aforementioned findings demonstrate the model's resilience and capacity for generalization, highlighting its versatility across various datasets and its possible practical uses in scenarios where comprehensive data may be scarce.

1 Introduction

Formula 1 racing is a symphony of speed, accuracy, and strategy; in many cases, the strategic decisions made for a pit stop will determine the outcome of the race. The research is an ambitious attempt to re-imagine Formula 1 race strategy through modern understanding of the data and code machine learning concepts such as Ensemble Learning and acknowledging the importance it plays in pit-stop strategy making in Formula 1 motorsport. The complex dynamics of Formula 1 racing, where each and every team is developing their cars to be the best and making continuous improvements to change the rivals and to win the championship, served as the inspiration for this research. Pit stops are more than simply a standard halt for maintenance or tyre change in this fast-paced environment; they are a tactical advantage that may have a big influence on the result of the race. The selection of tires, the timing of pit stops, and the adjustment to shifting track conditions are all critical choices requiring careful consideration and planning.

This research is based on factors such as tire degradation, which can have a significant impact on race strategy. Tire performance is a complicated variable that depends on several factors, including the track's layout, the weather conditions, and the performance

traits of the car. Over the course of a race, tire deterioration affects not just the vehicle's speed but also opens an window to determine the best time to make pit stops.[1] The research's main goal is to create an ensemble learning model that forecasts pit-stop strategy in a range of racing scenarios, including past race results, telemetry data, weather trends, and tire performance parameters, that requires a approach that can learn from a broad range of data sources, in contrast to traditional models that could concentrate on certain races or situations Stoppels (2017). To provide a flexible and reliable tool for pit stop strategy improvement, the model integrates data from various circuits, weather scenarios, and race circumstances to build a predictive model. In conclusion, the research is an investigation into the future of motorsport pit-stop strategy rather than merely a research project. With passion for Formula 1 racing and with the accuracy of ensemble learning, this project seeks to improve knowledge of pit-stop tactics while also completely changing the way teams tackle one of the most important parts of the competition.

Research Question Designing a predictive model that integrates tire degradation, weather conditions, and telemetry data effectively to determine the optimal lap for pit stops in Formula 1 racing to maximize the probability of winning the race?

2 Related Work

Since there hasn't been much scholarly study done on Formula 1, this literature review covers a number of years to provide a deep and comprehensive knowledge of Formula 1 motorsport. The chosen research articles that are being reviewed focus on as well as explaining the rules and operations of Formula 1 as a sport. This literature analysis aims to provide a thorough knowledge of the interaction of machine learning methodologies, predictive analytics, and Formula 1 by carefully examining these technological contributions in race prediction and race strategies.

2.1 Pit Stop Race Strategy Models in Motorsport

In the field of motorsport racing, strategic planning and execution has made great strides with the use of machine learning algorithms and predictive analysis of data, especially with relation to pit stop strategies. Numerous articles have explored the field of Formula 1, with each research offering distinctive viewpoints and approaches Choo (2015). The research summarizes and adapts the key findings from these publications, research work and places them in the context of a novel project that integrates various data points from different types of circuits to create an ensemble-learning-based pit stop racing strategy model. Research conducted by Choo (2015) addresses a machine learning software created at MIT for forecasting changes in track position in professional car race i.e., NASCAR. The creative application of machine learning and thorough data analysis has provided useful implications for decision-making in real time for race racing companies. However, its concentration on certain racing series and seasons, limits its usefulness to some extent as extensive track knowledge has not been adapted using the machine learning software which the authors have created, hence the lack of real-time telemetry data may restrict its forecasting power for all the tracks. Piccolomini et al. (n.d.) work stands out for its creative use of neural network designs and using neural network algorithms, specifically Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU), to anticipate tyre

[1] Understanding Tire Degradation `https://www.youtube.com/watch?v=wdj7uzla5pE`

strategy during Formula 1 races. The model's inability to precisely estimate race timings and integrate circuit variables and different race scenarios causes need for improvement and plan a strategized model which can overcome these setbacks.

The "Virtual Strategy Engineer" is a model which demonstrates the potential of machine learning in sports strategy by using Artificial Neural Networks to make racing strategy choicesHeilmeier et al. (2020). Although it makes use of a lot of data from Formula One seasons, various limitations are brought to light by the data imbalance and dependence on historical data. On unknown circuits this model has not been tested which creates a downside for this model. Finally, the research done by Tulabandhula and Rudin (2014) exemplifies the innovative application of analytical methods for within-race prediction and decision-making. Despite its strengths in data-driven insights and comprehensive feature construction, the complexity of racing data and reliance on historical data present substantial challenges in race prediction as inclusion of weather is important in within-race prediction. Tyre degradation is one of the reason a racing car has to make a pitstop, so that the tyres can be changed and a fresh new set of tyres can be put on, having said that the authorsSulsters and Bekker (2018) have investigated the application of machine learning models for forecasting tire replacement choices in Formula 1 racing and makes a noteworthy contribution in the direction of tyre degradation strategy. The study has done thorough investigation of tyre replacement decision-making and the applicability of its methodology. Its exclusive focus on certain races and drivers, however, limits its applicability to a variety of racing scenarios Piccinotti (2021).

2.2 Pit Stop Race Strategy in Formula 1 Using Advanced Data Analytics and Machine Learning

Several research publications that examine different aspects of race strategy reflect a distinct perspective, they all emphasize the need for more thorough and flexible analytical models. The research done by O'Hanlon (2022) focuses on forecasting the results of the 2021 Grand Prix Championship using machine learning, specifically through Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR) models. This paper's strengths lie in its comprehensive data analysis and utilization of advanced ML techniques, but it faces limitations as it only predicts the final ranking of the championship and not for individual races conducted throughout the year. Having said that, the research conducted FRANSSEN (2021) use Deep Neural Networks (DNN) and Radial Basis Function Neural Networks (RBF) for predicting Formula 1 outcomes. This study stands out for its innovative approach in using advanced neural networks and its data handling. However, both the papers try different approach to predict the championship outcomes. In research Stoppels (2017) explores the use of ANNs for predicting Formula One finish results as done by O'Hanlon (2022). It offers a comprehensive theoretical foundation and practical application in race prediction but is limited by lack of data for far newer circuits. Similarly, in the research paper by SICOIE (n.d.) uses historical data as well but it limits the research as there is lack of new track data. Lastly, Haghighat et al. (2013) offers an extensive review of different data mining techniques used in sports analytics. The paper's strengths are its comprehensive analysis and in-depth methodological review, but it lacks practical case studies and can be complex for readers without a background in data science.

2.3 Framework in Formula 1

In the dynamic and technologically driven world of Formula 1 racing, strategic decision-making, especially pit stops, has increasingly shifted its reliability towards data-driven methodologies. One such significant study is "Strategy for Optimizing an F1 Car's Performance based on FIA Regulations" by Bopaiah and Samuel (2020). This research optimizes Formula 1 car performance through numerical simulations under the 2019 FIA Formula One regulations Bopaiah and Samuel (2020) [2]. Employing GT-Suite software, the study stands out for its detailed modelling of the car's power unit and the innovative use of real-world circuit data, particularly using Optical Character Recognition (OCR) to extract data from onboard F1 videos. However, the study's focus on specific FIA regulations and its reliance on simulation data restrict its applicability as Formula 1 races are involve various unpredictable factors such as weather conditions and accidents. In parallel, this research Bunker and Thabtah (2019) offers a critical analysis of Artificial Neural Networks (ANNs) in sports prediction, proposing a novel framework, SRP-CRISP-DM. This framework is a structured approach to sports result prediction using machine learning, tailored to the unique requirements of sports analytics. Despite its comprehensive review of ANN applications in sports prediction, the paper's complexity and focus on team sports may limit its utility for non-experts and individual sports applications.

In contrast, the proposed research to build a pit stop race strategy model represents a significant leap forward in this domain of motorsport. By employing ensemble learning and incorporating a wide array of data types, including telemetry, tyre degradation and weather data, the model aims to address the limitations identified in previous studies. The inclusion of data from various circuit types promises a more generalized and adaptable approach, potentially overcoming the constraints of models focused on specific races or conditions. This comprehensive approach, coupled with the understanding of the dynamic nature of motorsport gained from the reviewed papers, positions the proposed model as a more robust and versatile tool in the which the Formula 1 teams can make strategic decision-making. In summary, although the reviewed papers provide a solid basis for applying machine learning and predictive analytics to motorsport, the proposed research can provide a more comprehensive and flexible pit stop race strategy, increasing the likelihood of success in the highly competitive and unpredictable world of motorsport racing that is formula 1.

3 Methodology

To develop and execute machine learning research project, it is essential that an industry framework is chosen as a project guide. The chosen framework for this research project is Knowledge Discovery in Databases (KDD) process model. This will include understanding the Formula 1 Motorsport business, understanding various track and weather data, which will require data preparation before modelling, and at the end evaluating the developed model and understanding its impact on Formula 1 community.

[2]Formula 1 Regulations `https://www.fia.com/regulation/category/110`

3.1 Business Model of Formula 1

The exploration of the research begins with an insightful review of the related work, offering valuable insights into the environment of Formula 1, spanning its regulations, rules, and statistical landscape. However, understanding the problem statement and objectives demands a broader adaptation of the sport. This enthusiasm and willingness to understand the sport will be cultivated through user experience into the sport with diverse channels, including real-time engagement by watching live Grand Prix events, tapping into the motorsport documentaries available online, and immersing oneself in the historical narrative of Formula 1. A new perspective can also be acquired from the vibrant community of Formula 1 enthusiasts and experts [3] [4] who express themselves by writing blogs dedicated to F1, moreover, the official F1 website stands as a comprehensive repository housing both historical and contemporary race results and race details, along with a roadmap of future Grand Prix events [5] [6]. It is crucial to acknowledge the unique circumstances surrounding the 2023 F1 season, the addition of new circuits creates an intriguing layer to the context within the research which therefore changes the dynamics and challenges faced during this specific season.

3.2 Data Understanding

The dataset utilized in this research originates from two open-source platforms, namely Tracing Insight [7] and Pitwall [8]. These websites are collaborative initiatives within the Formula 1 community, serving as valuable resources for both supporting the sport and facilitating learning opportunities for enthusiasts who are eager to understand the nature of the sport. The dataset consists of historical race data spanning from the inaugural year of Formula 1 in 1950 to the present year, 2023. The primary data source is the Formula 1 website, where information is systematically gathered.

Tracing Insight contributes trackside data essential for a comprehensive analysis. This dataset involves critical details such as tire degradation metrics, track temperature variations, diverse speed tracking metrics for drivers, and an array of weather-related data encompassing air pressure, wind speed, among other sensor-generated information.

Supplementary data is obtained from the Pitwall website, another community-driven platform dedicated to Formula 1. This source specifically provides valuable insights into driver positions at each lap interval, as well as the final ranking positions at the end of races. The collaborative nature of these open-source platforms gives a collective effort of the Formula 1 community in knowledge sharing and analysis within motorsport.

3.3 Data Preparation - EDA

In the initial phases of data preparation for this research project, three distinct datasets were obtained from two Formula 1 websites: Tracing Insights and Pitwall. Tracing Insights contributed two CSV files which includes both race-related and weather data. A comprehensive examination for null values was done, and adjustments were made in

[3] `https://f1tv.formula1.com/page/5351/legends-of-f1`

[4] `https://www.formula1.com/`

[5] Drive To Survive, online film recording, Netflix, `https://www.netflix.com/ie/title/80204890`

[6] `https://www.youtube.com/@WeAreTheRace`

[7] Tracing Insight `https://huggingface.co/spaces/tracinginsights/F1-analysis`

[8] Pitwall `https://pitwall.app/`

columns where null entries were identified to ensure the overall cleanliness and integrity of the dataset.

The dataset had a substantial presence of categorical data, with fields such as driver names and team affiliations. To address these categorical values for analysis, a systematic transformation into numerical representations was done. Another significant aspect of the dataset was a column denoting data for fresh tires, initially presented in a binary true/false format. This information was converted into a binary numerical format (0 and 1) to facilitate meaningful analysis.

A critical phase of data refinement involved handling columns with significant null values, particularly those related to pit stop times (pitin and pitout). Given that pit stops are integral part of Formula 1 races, the null values were addressed. These columns played an important role in subsequent feature engineering. By making pitin and pitout times, a 'pitflag' column was created, representing instances where pit stops occurred (coded as 1) and others where they did not (coded as 0). This 'pitflag' column emerged as a crucial target variable for the subsequent analysis. After performing the initial data cleaning all the three files were combined into one file which consisted of almost 25000 rows with 40 columns and before the dataset was utilized in model building, a standardization process was applied to ensure consistency and comparability of the data across various features. This comprehensive data preparation approach laid to robust and insightful analysis in the later phases of the research project.

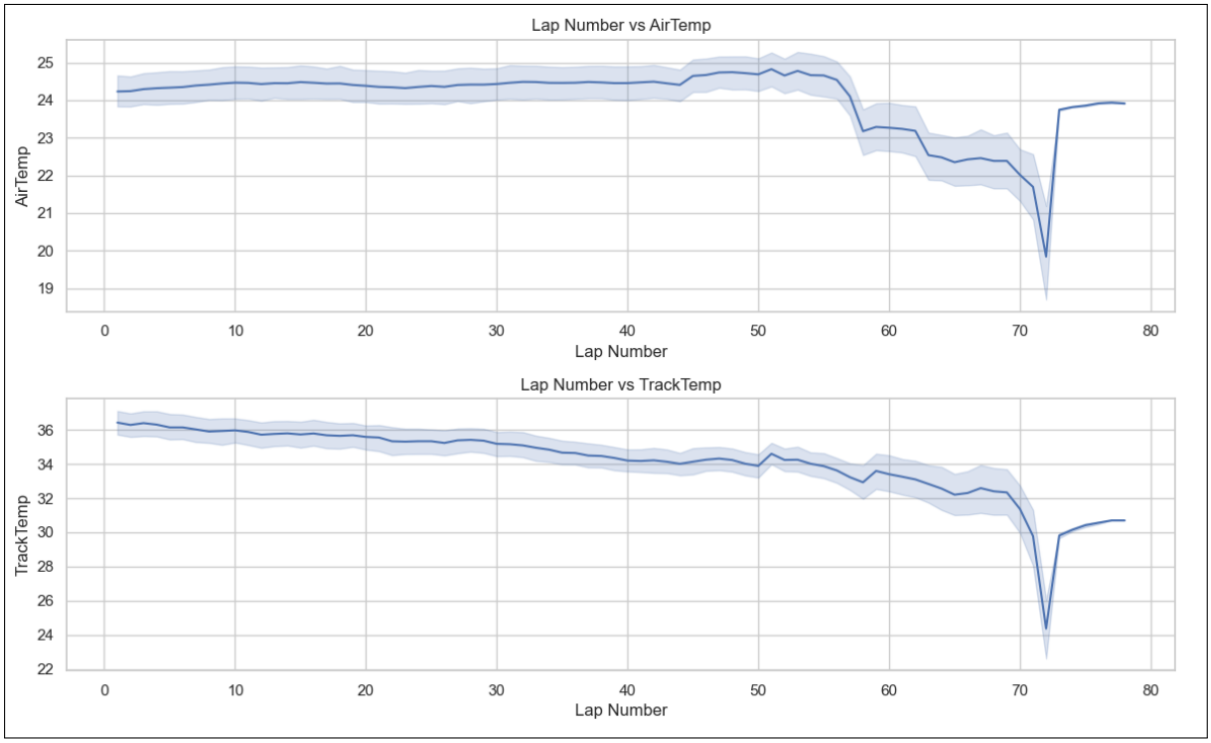

Figure 1: Lap Number Vs Air Temperature

The graphic displays in Figure 1 two plotted line graphs that indicate the connection between the track temperature and air temperature, two separate temperature data, and the lap number. Since the lap number is shown by the same x-axis on both graphs,

it is likely that the data originates from a racing event, such a Formula 1 race, where temperature may have a big impact on performance.

Air Temperature versus Lap Number: The air temperature and lap number are plotted on the top graph. The y-axis displays the air temperature, expressed in degrees. A range band or confidence interval that surrounds the line on the graph denotes the measurement uncertainty or variability. As the number of laps increases, there seems to be a modest drop in air temperature, with a noticeable dip towards the later laps.

Track Temperature versus Lap Number: The track temperature is shown on the bottom graph along with the lap count. Because of the heat produced by the tires' friction on the track, the track temperature is often greater than the air temperature. This graph displays a slow drop in temperature with a more noticeable dip at the conclusion. It also has a range band.

Both figures might be included in a more comprehensive research on race strategy and vehicle performance, or they could be used to examine how temperature changes impact racing conditions. A particular incident that occurred during the race, such shifting weather, might be the cause of the abrupt reductions in temperature that occurred in the final laps.

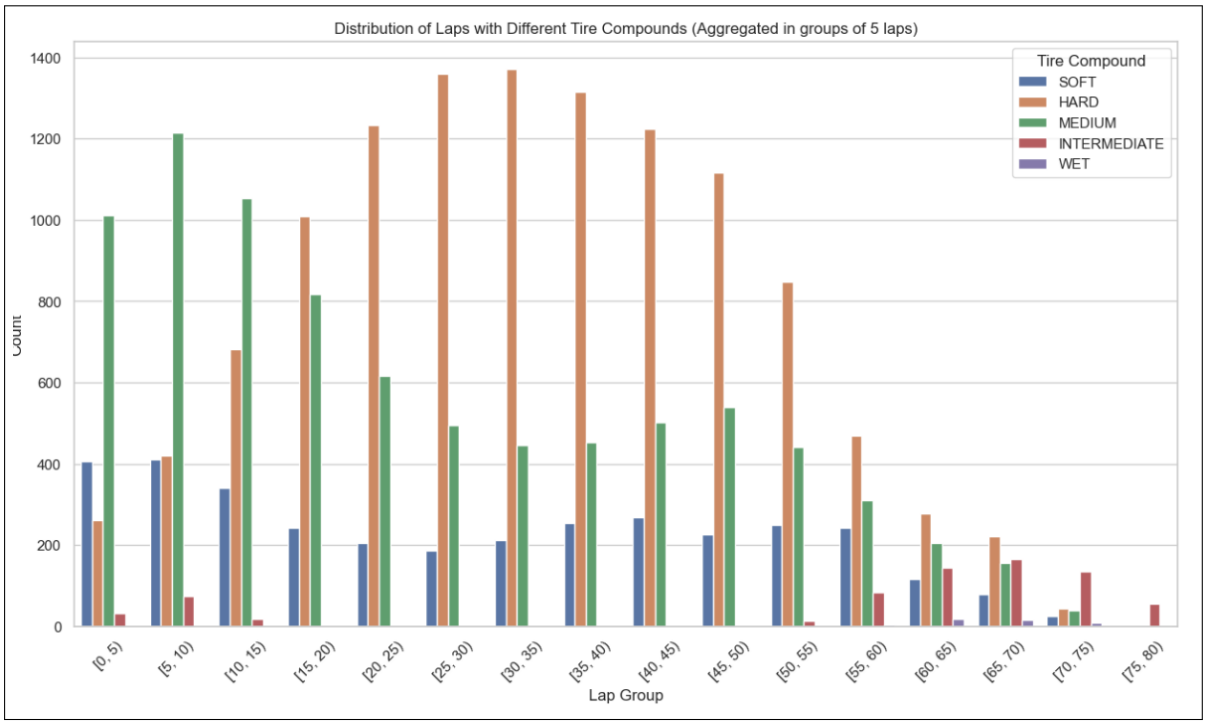

Figure 2: Lap Number Vs Air Temperature

Figure 2 is called "Distribution of Laps with Different Tire Compounds (Aggregated in groups of 5 laps)." It is a clustered bar chart. It displays the frequency of use of various tire compounds over different groups of laps in a race.

Grouped lap intervals, such as 0-5, 5-10, 10-15, and so on, are represented on the x-axis, signifying that the laps are arranged in groups of five. The number of laps completed on each tire compound at those intervals is shown on the y-axis.

The hues soft (blue), hard (brown), medium (green), intermediate (purple), and wet (red) correspond to the five distinct tire compositions. Within each lap group, each color correlates to a bar that displays the number of laps done on each tire compound during that intervals.

The tire strategies used in various racing stages may be compared using this chart. It is common in racing for softer tires to be used at the beginning of the race because they wear out more quickly than hard tires. As an example, it can be seen that the hard compound tires were used more frequently in the middle lap groups, while the soft compound tires were more frequently used in the earlier laps. Less laps are completed in changeable or wet circumstances, which is indicated by the lower usage of intermediate and wet tires. For teams to plan their pit stops and tire management for race conditions, this type of analysis is essential

3.4 Modelling

Within the dynamic landscape of Formula 1 race prediction and strategy formulation, and variety of models has been explored, with artificial neural networks (ANNs), support vector machines (SVMs), and random forests emerging as common approaches, as highlighted in the literature. However, this study takes a unique direction by embracing the power of ensemble learning. Specifically, it integrates Stacking, a sophisticated ensemble method. This technique uses diverse base learning models, including a Random Forest Classifier, a Support Vector Classifier, and a Gradient Boosting Classifier, alongside a meta-learner represented by a Logistic Regression model in this context. The base learners undergo comprehensive training on the entirety of the dataset, and subsequently, the meta-learner is strategically trained to synthesize the outputs of these individual base learners.

4 Implementation

4.1 Data Transformation and Feature Engineering

In the data preprocessing phase, several crucial steps were undertaken to enhance the quality and usability of the dataset. The initial analysis involved checking for missing values, and the results were documented in a comprehensive table showcasing the count of missing values for each variable. This step was essential to identify and address any potential gaps in the data. Additionally, a thorough examination of unique values in each column was conducted to identify inconsistencies or irrelevant features. The findings were recorded in a detailed report, outlining the uniqueness of values in various columns.

Moving on to the transformation phase, categorical columns were carefully processed. All categorical data was converted to string type to ensure uniformity and effective label encoding. The LabelEncoder from the scikit-learn library was then applied to transform string-type categorical variables into numerical representations. This step was crucial for machine learning algorithms that require numerical input. The transformed data was stored in a new dataframe, retaining the original structure.

Simultaneously, numerical columns underwent preprocessing steps as well, the StandardScaler from scikit-learn was employed to standardize the numerical features, ensuring that they were on a consistent scale. Notably, the 'PitFlag' and 'Position' columns were

excluded from the standardization process as they represented categorical and target variables, respectively.

The data preprocessing lays the foundation for an accurate model development in subsequent phases of the research. The standardized and encoded dataset is now ready for the implementation of machine learning models, facilitating effective race strategy prediction and analysis in the Formula 1.

The feature engineering process in this research involved the creation of a column 'PitFlag', which served as one of the target variables for predicting pit stop strategies. To ascertain when a driver made a pit stop, the dataset incorporated information from two columns: pit in and pit out times. By combining these times, a binary indicator 'PitFlag', was generated. Rows corresponding to the laps when a driver entered and exited the pit were marked as 1, signifying a pit stop, while all other rows were marked as 0. This binary representation effectively captured the critical periods when pit stops occurred during the race. Notably, in Formula 1 racing circuits are closed loops, and pit stops are typically located parallel to the finish line. As a result, the dataset reflected that a driver entering the pit on, for instance, the 10th lap would rejoin the track at the start of the 11th lap. Additionally, a further enhancement was made by incorporating position data for each driver in every lap, sourced from an open-source website. To handle the categorical aspect of this data, denoting positions like '1st' or '2nd,' these labels were processed to retain only the numeric information. This comprehensive feature engineering strategy provides a robust foundation for subsequent model training and race strategy predictions.

4.2 Model Building

Two distinct splits are performed for predicting different target variables: 'Position' and 'PitFlag.' Firstly, for predicting 'Position,' the dataset is divided into features (denoted as X) and the target variable 'Position' (denoted as y_position). The 'train_test_split' function from the 'sklearn.model_selection' module is then employed to create training and testing sets, with 80 percent of the data allocated for training and 20 percent for testing. Subsequently, a second split is executed for predicting 'PitFlag.' In this case, the features (denoted as X1) exclude both 'Position' and 'PitFlag' columns, while the target variable 'PitFlag' (denoted as y_pitflag) is retained. Again, an 80-20 split is applied to generate separate training and testing sets for the 'PitFlag' prediction model. This division ensures that distinct subsets of the data are available for training and evaluating the models aimed at predicting driver positions and pit stop occurrences, respectively. The 'random_state' parameter is set to 42 for reproducibility, allowing consistent results across multiple runs of the model training process.

4.3 Ensemble Model

Ensemble learning especially stacking approach involves combining multiple models to achieve better predictive performance than individual models. The stacking ensemble, in particular, incorporates diverse base learners and a meta-learner to capitalize on their collective strengths Dietterich et al. (2002). In this context, the base learners—Random Forest Classifier (rf), Support Vector Classifier (svc) with probability estimation, and Gradient Boosting Classifier (gb)—are chosen for their distinct modeling strategies. The Random Forest excels in handling complex relationships in data, the Support Vector

Classifier is effective in handling non-linear relationships, and the Gradient Boosting Classifier sequentially builds weak learners to create a robust predictive model Jozdani et al. (2019).

The stacking ensemble supports the predictions of these base learners as inputs for the meta-learner, a Logistic Regression model in this case is used. The meta-learner learns to combine the outputs of the base learners, essentially making a final prediction based on their collective insights. This meta-learning step enhances the model's overall predictive capability, allowing it to capture more intricate patterns and relationships in the data. The training of the stacking ensemble involves using the 'StackingClassifier' from the 'sklearn.ensemble' module. The 'final_estimator' parameter specifies the meta-learner, and the 'cv=5' parameter indicates 5-fold cross-validation during the training process. Cross-validation is crucial for estimating the model's performance more reliably, as it assesses the model across multiple subsets of the training data. Once trained, the model is evaluated on a separate testing set ('X_test_pf' and 'y_test_pf'). The accuracy score provides a general measure of predictive performance, while the classification report delivers a more detailed breakdown of precision, recall, and F1-score for each class (0 and 1) in the binary classification task.

In summary, the stacking ensemble combines the strengths of diverse base learners and meta-learning through a Logistic Regression model to predict pit stops in Formula 1 races. This approach enhances the model's ability to understand complex patterns in the data, resulting in a robust and accurate predictive tool for strategic decision-making in Formula 1 pit stops.

Now the focus is shifted to predicting the 'Position' of the Formula 1 drivers instead of predicting pit stops. The ensemble model is adept at handling this change in the prediction task. The evaluation metrics, namely accuracy and a detailed classification report, are calculated and printed to the console. These metrics provide insights into how effectively the model predicts the positions of the Formula 1 drivers. The accuracy score represents the proportion of correctly predicted positions, while the classification report provides a breakdown of precision, recall, and F1-score for each position category.

Accuracy: 0.8997334426901784

Classification Report:

	precision	recall	f1-score	support
1	0.98	0.96	0.97	253
2	0.93	0.94	0.94	291
3	0.94	0.92	0.93	257
4	0.92	0.88	0.90	273
5	0.90	0.91	0.91	247
6	0.91	0.95	0.93	290
7	0.92	0.86	0.89	263
8	0.86	0.91	0.88	265
9	0.87	0.85	0.86	250
10	0.91	0.87	0.89	244
11	0.86	0.87	0.87	212
12	0.85	0.89	0.87	304
13	0.86	0.91	0.88	259
14	0.86	0.89	0.88	273
15	0.92	0.83	0.87	251
16	0.86	0.94	0.90	239
17	0.94	0.85	0.89	255
18	0.86	0.91	0.88	178
19	0.92	0.94	0.93	143
20	0.98	0.97	0.97	130
accuracy			0.90	4877
macro avg	0.90	0.90	0.90	4877
weighted avg	0.90	0.90	0.90	4877

Figure 3: Classification Report

As per the classification report in Figure 4, Accuracy The reported accuracy of the model is approximately 0.8997 (or about 89.97 percent), which represents the proportion of true results (both true positives and true negatives) among the total number of cases examined.

In Classification Report, Precision indicates the ratio of correctly predicted positive observations to the total predicted positive observations. Recall (Sensitivity) measures the ratio of correctly predicted positive observations to all observations in the actual class. For class 1, the recall is 0.96, so 96 percent of actual class 1 instances were correctly identified by the model. F1-Score provides a balance between precision and recall. It's the harmonic mean of precision and recall. For class 1, the f1-score is 0.97, indicating a strong balance between precision and recall.

Overall, the model has a high level of accuracy and balanced precision, recall, and f1-scores across the classes. The model seems particularly effective in classifying certain classes where the f1-score is as high as 0.97.

5 Evaluation

5.1 Ensemble Model for PitFlag Target Variable

The training set was prepared by splitting the dataset into features (X1) and the target variable for pit stops (y_pitflag). The subsequent application of the stacking ensemble model, consisting of Random Forest, Support Vector, and Gradient Boosting classifiers, along with a Logistic Regression meta-learner, yielded impressive results. The accuracy achieved was exceptionally high at 99.79. However, this high accuracy can be attributed

to the imbalanced nature of the target variable, where about 90 of the instances represent races without pit stops (labeled as 0), and only around 10 signify races with pit stops (labeled as 1). Consequently, the model predominantly predicts instances labeled as 0 due to their majority presence in the dataset. When applying the model to a sample dataset, it correctly identified laps 13, 14, 35, and 36 as instances where the driver should make pit stops. This aligns with the understanding that pit stops typically occur during these laps in Formula 1 races, providing a promising indication of the model's predictive capabilities for the first target variable, PitFlag.

5.2 Ensemble Model for Position Target Variable

The second part of the research focuses on predicting the positions of drivers during a Formula 1 race. The dataset is split into features (X) and the target variable for position prediction (y_position). The ensemble model, comprising a Random Forest Classifier, Support Vector Classifier (SVC), and Gradient Boosting Classifier as base learners with a Logistic Regression meta-learner, is constructed and trained on this dataset.

The model exhibits a strong predictive capability, achieving an accuracy of approximately 89.97 on the test set. The classification report further highlights the model's performance across different positions. Precision, recall, and F1-score metrics indicate the model's ability to correctly identify drivers' positions across the diverse range of racing scenarios, from leading positions to mid-field and lower-ranking positions. The weighted average metrics, such as precision and recall, reinforce the model's overall effectiveness.

In a specific instance of prediction on a sample data point, the model accurately forecasts the driver's position as 1st, consistent with the earlier pit stop predictions in laps [13 14 35 36]. This alignment between the predicted position and the actual position reinforces the model's utility in strategic decision-making, providing teams with valuable insights for optimizing race strategies.

In summary, the predictive model for the second target variable, Position, demonstrates robust performance, showcasing its potential as a valuable tool for Formula 1 teams to enhance their understanding of drivers' race positions and inform strategic decisions during races.

5.3 Future Race Prediction

The final part of the research involves applying the trained models to a prediction dataset with no target variables, namely pit flags or positions. The ensemble model for pit stop prediction (stacked_model) accurately identifies the laps at which the driver should take pit stops, with predictions suggesting pit stops at laps [14 15 30 31 43 44]. This aligns with the expected behavior of Formula 1 races, where pit stops strategically occur at specific laps for optimal race performance.

The pit stop information is then incorporated into the prediction dataset by adding a 'PitFlag' column based on the pit stop predictions. Subsequently, the model for position prediction (stacked_model1) is applied to this augmented dataset. Remarkably, the model predicts the driver's position as 1st. This outcome signifies the model's generalization capability, successfully making accurate predictions even in the absence of explicit target variables in the prediction dataset.

These results underscore the robustness of the predictive model, demonstrating its adaptability to diverse datasets and scenarios. The ability to make accurate predictions

on a dataset without predefined target variables enhances the model's utility, suggesting its potential for real-world applications where comprehensive data may not always be available.

Figure 4: Actual vs Predicted Position

Figure 3, is a scatter plot that shows the difference between expected and actual placements in a dataset. The dataset probably has to do with a race or other ranking situation. The "Index (Subset)" x-axis shows a portion of the data that may have been selected at random. "Position," the label on the y-axis, denotes the order or position inside the dataset or event. The graph shows two sets of data points: red points show the expected positions, while blue points show the actual positions. The points, which are dispersed throughout the plot, demonstrate the fluctuation and correlation between the actual and expected values. While some points are isolated, illustrating differences between forecast and reality, others are overlapping, showing occasions where the predictions match the actual placements closely. The chart appears to try to allow for the visibility of overlapping points based on the use of alpha transparency in the points. The lack of ticks on the x-axis suggests that the general pattern of forecast accuracy is more important than the individual index values. The chart is useful for evaluating the effectiveness of a prediction model or procedure since the title "Actual vs. Predicted Positions" makes the plot's goal evident and the legend distinguishes between actual and projected data.

5.4 Discussion

The results collected indicate that there was a significant class imbalance problem that the model encountered when it was being trained for the 'PitFlag' target variable. The model demonstrated its ability to make correct predictions with an astounding 99 percent accuracy rate. But it's important to investigate the underlying causes of this great accuracy, especially when it comes to predicting Formula 1 racing pitstop strategy. Pit stops are tactical maneuvers that happen infrequently, usually once or twice a race in Formula 1. The dataset used to train the model has an inherent class imbalance due to this feature. The imbalance results from the difference between the percentage of the race when pit stops occur (positive class) and the percentage of the race where they do

not (negative class). The model performs well in classifying non-pitstop cases, as evidenced by its high accuracy; nevertheless, it also highlights the need to interpret accuracy measurements carefully in imbalanced datasets. Accurately capturing the minority class of pitstops is one area where additional evaluation measures like precision, recall, and F1-score become crucial for a more nuanced assessment of the model's performance. In order to improve the model's training procedure and guarantee its application to real-world situations where pit stops are tactical and rare occurrences in Formula 1 races, it is imperative that class imbalance issues in the dataset be addressed.

6 Conclusion and Future Work

The application of a pitstop race strategy produced very encouraging outcomes in this study. The predictive model accurately predicted the best times for drivers to pit stop and their final race positions, thanks to data on tire wear, weather conditions, and sensor readings. The model's strong performance is demonstrated by how well it forecasts when a driver should pit and where they will finish the race. Through the examination of several datasets that included tire performance, sensor data, and weather conditions, the model proved to have a deep comprehension of the complex dynamics affecting Formula 1 races. Its ability to produce accurate forecasts for pit stop times and the overall race rankings is indicative of its flexibility in responding to a wide range of inputs and conditions. The strategic advantage these results give Formula 1 teams makes them significant. By predicting the best pit stop tactics ahead of time, teams can make better decisions during races. This shows the possibility of obtaining a competitive edge in the highly competitive world of Formula 1 racing in addition to helping with increased race performance. The successful application of this pitstop race strategy highlights the established prediction model's real-world impact and practical usability, which represents a major advancement in the field of motorsport analytics.

The pitstop racing strategy model's implementation's effectiveness creates opportunities for further study and use in the field of motorsport analytics. Potential future applications and areas for improvement include the following:

Fine-Tuning and Optimization: By adding new features and improving the current algorithms, the predictive model's performance might be even more adjusted and maximized. Pit stop tactics and race results can be more accurately predicted by the model with continued accuracy improvements.

Real-Time Implementation: It would be beneficial to increase the model's capacity to offer forecasts in real time during ongoing races. Pitstop plans could become more dynamic and responsive by using real-time telemetry data and quickly updating projections in response to changing race conditions.

Extension to Other Motorsport Categories: The model's influence might be increased if it were applied to motorsports other than Formula 1. Customization would be necessary to adapt the model to diverse racing formats and regulations, but the benefits could extend across a wide range of racing disciplines.

References

Bopaiah, K. and Samuel, S. (2020). Strategy for optimizing an f1 car's performance based on fia regulations, *SAE International Journal of Advances and Current Practices in*

Mobility **2**(2020-01-0545): 2516–2530.

Bunker, R. P. and Thabtah, F. (2019). A machine learning framework for sport result prediction, *Applied computing and informatics* **15**(1): 27–33.

Choo, C. L. W. (2015). *Real-time decision making in motorsports: analytics for improving professional car race strategy*, PhD thesis, Massachusetts Institute of Technology.

Dietterich, T. G. et al. (2002). Ensemble learning, *The handbook of brain theory and neural networks* **2**(1): 110–125.

FRANSSEN, K. (2021). *COMPARISON OF NEURAL NETWORK ARCHITECTURES IN RACE PREDICTION*, PhD thesis, tilburg university.

Haghighat, M., Rastegari, H., Nourafza, N., Branch, N. and Esfahan, I. (2013). A review of data mining techniques for result prediction in sports, *Advances in Computer Science: an International Journal* **2**(5): 7–12.

Heilmeier, A., Thomaser, A., Graf, M. and Betz, J. (2020). Virtual strategy engineer: Using artificial neural networks for making race strategy decisions in circuit motorsport, *Applied Sciences* **10**(21): 7805.

Jozdani, S. E., Johnson, B. A. and Chen, D. (2019). Comparing deep neural networks, ensemble classifiers, and support vector machine algorithms for object-based urban land use/land cover classification, *Remote Sensing* **11**(14): 1713.

O'Hanlon, E. (2022). *Using Supervised Machine Learning to Predict the Final Rankings of the 2021 Formula One Championship*, PhD thesis, Dublin, National College of Ireland.

Piccinotti, D. (2021). Open loop planning for formula 1 race strategy identification.

Piccolomini, E. L., Evangelista, D. and Rondelli, M. (n.d.). The future of formula 1 racing: Neural networks to predict tyre strategy.

SICOIE, H. (n.d.). *MACHINE LEARNING FRAMEWORK FOR FORMULA 1 RACE WINNER AND CHAMPIONSHIP STANDINGS PREDICTOR*, PhD thesis, tilburg university.

Stoppels, E. (2017). *Predicting race results using artificial neural networks*, Master's thesis, University of Twente.

Sulsters, C. and Bekker, R. (2018). Simulating formula one race strategies, *Vrije Universiteit Amsterdam* .

Tulabandhula, T. and Rudin, C. (2014). Tire changes, fresh air, and yellow flags: challenges in predictive analytics for professional racing, *Big data* **2**(2): 97–112.