Analyzing the effect of the 2022 FIA regulation change on race dynamics with Machine Learning

Seminar in Data Analytics
Fall 2025

Myunggong Seo

December 15, 2025

Abstract

This report quantitatively evaluates the impact of the 2022 FIA aerodynamic rule overhaul on Formula 1 race dynamics using machine learning and statistical modeling. By leveraging race data from 2019 to 2025 with FastF1 API, the project uses Random Forest Classifiers and Linear Mixed Effects Model to detect distinguishable features and cause behind performance change between pre- and post-regulation races. The mixed effects model revealed that race level factors were significant drivers of laptime variability (adjusted ICC = 0.944), with key predictors being tyre life, mean track speed, air temperature, and the regulation era itself. Post-2022 regulations were associated with substantially faster lap times. The Random Forest model achieved good performance (accuracy ~ 0.87) and confirmed that existing features can distinguish between pre- and post-regulation races. However, environmental figures contributed to overfitting and were excluded. The analysis highlights that race-level conditions and aerodynamic-related variables, rather than driver effects alone are the biggest factor in differentiating race characteristics. These findings provide a quantitative framework for evaluating regulatory changes in motorsport and establish a basis for future work incorporating more detailed telemetry, and longer season data to better understand the effects of rule changes on race dynamics.

Contents

1 Introduction

Formula 1 is recognized as one of the largest and most popular motorsports in the world in terms of revenue and scale. Governed by the Fédération Internationale de l'Automobile (FIA), F1 began in 1950 and has been around for nearly 80 years, drawing more than 827 million fans globally in 2025 [Formula 1, 2025]. FIA routinely overhauls regulations to ensure the competitiveness of the races. The latest change took place in 2022, when the FIA introduced aerodynamic regulations and 18-inch wheels to make races more exciting by reducing the amount of wake from turbulence with less reliance on clean air, which would increase the competitiveness throughout the grid [Stuart, 2021]. Existing scholarly articles extensively cover the technical aspects of F1 car design from an engineering perspective [Méndez, 2023], statistical analyses of the impact of FIA regulations, and forecasting of results using machine learning [Garcia Tejada, 2023] .

However, little work has been done applying ML to identify and quantify the factors regulatory changes have impacted. To address the gap, this project will use historical F1 race datasets from pre- and post-2022 periods and develop a machine learning model to identify the underlying variables that most significantly distinguish between the two periods. The merit of this project lies in applying common data analytics and machine learning techniques to a question that can be translated into sports analytics and regulatory science. Its broader impact is the potential to provide a more data-driven approach for governing bodies and constructors to understand how the characteristics of competition on track will change after the implementation of a major rule overhaul.

In short, this project will move beyond simply predicting the likely outcome of a race and instead ask, **"What are the distinguishing characteristics of the races that occurred after a major regulatory change compared to those that occurred before?"**

2 Domain Review

2.1 Overview of Machine Learning in Sports Analytics

The use of machine learning (ML) has emerged as a central framework in sports analytics, offering a range of tools to model and comprehend the complex dynamics of sports competitions with tremendous variations. As highlighted by Bunker and Thabtah (2019), the application of ML in sports begins with a thorough understanding of the domain and data, followed by processes such as feature extraction, data processing, and model training. Given the nature of sports events, it is crucial to differentiate between match-related and external factors. Match features refer to the data points generated directly from the competition itself, whereas external features are independent of the match outcomes. Rory and Fadi also emphasize that the conventional 7:3 train-test split may not be appropriate for sports analytics. This is particularly relevant in contexts where past season performance may not strongly correlate with future outcomes, due to significant yearly changes in team composition and lineups [Bunker and Thabtah, 2019]. For instance, a study by Chenjie on baseball prediction employed order-preserving train-test splits, such as rolling predictions by rounds or seasons, as these methods offer a more accurate representation of forecasting [Cao, 2025].

2.2 Machine Learning in Formula 1

The use of Machine Learning for Formula 1 does not deviate from conventional methods widely utilized in sports analytics. Data Science has evolved to be an inseparable part of Formula 1 operation with the increasing amount of data generated by cars during races. In fact, each car, loaded with over 300 sensors, can transmit more than 1.1 million data points per second [AWS, 2024]. Real-time analytics, from strategic decision making to predictive modeling for car development, turning data science into an essential foundation for team success in this highly competitive sport [Ambler, 2024]. Sicoie's (2022) and Tejada's (2023) literature gives a comprehensive overview of machine learning methods and some of the most common variables used for race result and behavior prediction, providing a practical framework for applying machine learning to the current race data [Sicoie, 2022, Garcia Tejada, 2023]. The study approaches the analytical challenge as a regression and classification problem, using models such as Random Forest, Gradient Boosting Regressor, and Support Vector Machine to predict a driver's

finishing position in each race. A key contribution highlighted is the emphasis on feature engineering. Sicoie enhanced the standard race data with newly created variables, such as the driver's age at the time of the race and the conversion of finish times into a time difference from the race winner, which proved to be more significant predictors. K-fold cross-validation and randomized parameter search were used for model tuning, and model performance is primarily assessed by ranking correlation metrics such as the Spearman correlation coefficient, ROC curves, and AUC.

2.3 Mixed Effects Model in sports analytics

A notable characteristic of the Formula 1 data is that it is hierarchical, with multiple observations (laps) nested within groups (drivers or teams), which does not fit well with standard linear regressions. Linear Mixed Effects Models are the appropriate statistical tool to properly account for this nested structure [Casals et al., 2025]. As highlighted by Casals et al. (2025), the use of Mixed Models in sports analytics is often incomplete, making it difficult to assess the validity of results or to replicate the method. This lack of standardized reporting is a significant methodological gap in the field. Therefore, the present study not only uses a Linear Mixed Model to answer its research question but also tackles the practice of reporting guidelines identified by Casals et al.

2.4 A breakdown of the Driver vs. Constructor Advantage

Some literature employ a more advanced statistical method to break down the races beyond predictive modeling, which yielded more information about the underlying drivers. A significant challenge in Formula 1 (F1) analytics is to deconstruct the effects of driver skill and constructor advantage.

Van Kesteren and Bergkamp (2023) address this directly by developing a Bayesian model to analyze race results from the hybrid era (2014–2021) [van Kesteren and Bergkamp, 2023]. Rather than predicting a single outcome, they used a rank-ordered logit model to estimate skill parameters for both drivers and constructors simultaneously and created independent performance ratings. This approach directly quantifies the extent of each component's contribution to race outcomes. The study concludes that the constructor advantage accounts for the vast majority of the variance in race results ($\sim 88\%$) while driver skill accounts for the remaining 12% [van Kesteren and Bergkamp, 2023]. This work is methodologically significant as it provides a statistical framework for separating individual talent from equipment advantage, which is a common issue in many sports with heavy reliance on equipment.

In fact, some studies, such as from Nagle (2022), take a step further by clustering different groups of drivers based on driving styles. The study used K-means clustering to uncover that the most successful teams tend to have drivers with consistent performance rather than sporadic wins, and paired experienced drivers with younger, rising talent [Nagle, 2022].

2.5 Technical Aspect of the 2022 FIA Regulation

The 2022 regulation introduced a significant technical overhaul centered around creating a so-called 'ground effect', which moves much of the car's downforce generation to the underbody with longer underfloor tunnels to reduce turbulent wake and improve cars' ability to maintain proximity during close contacts [Stuart, 2021]. Redesigned cars would therefore generate less turbulent wake, allowing cars to retain more downforce compared to previous ones, enabling them to stay closer to each other.

Alongside this, a new 18-inch tires with lower blanket temperature and larger rims were introduced to reduce heat sensitivity and. Together, these changes not only reshaped car performance characteristics but also formed the technical foundation for examining how race dynamics such as lap times, overtaking frequency, and overall performance patterns differ before and after the 2022 season. From a technical perspective, Méndez's (2023) work highlights the complexities of isolating and quantifying the impact of regulation on the race dynamic. Using Computational Fluid Dynamics (CFD), the study simulated and compared the wake turbulence generated by the 2021 and 2022 rear wings. Interestingly, the simulation found that the 2022 rear wing design generated a higher level of turbulent kinetic energy, failing its intended purpose.

However, Méndez notes that this finding contradicts empirical evidence, citing real-world data that shows a significant 30% increase in on-track overtakes during the 2022 season [Méndez, 2023]. The

discrepancy may be rooted in the study's assumption, where it modeled the rear wing in isolation, failing to capture its interaction with the car's complete aerodynamic profile. This conclusion is highly relevant as it underscores the limitations of simulation-based analysis and underscores the necessity of an empirical approach. This literature also creates room for statistical and machine learning models that can be used to validate the real-world impact of a regulatory overhaul.

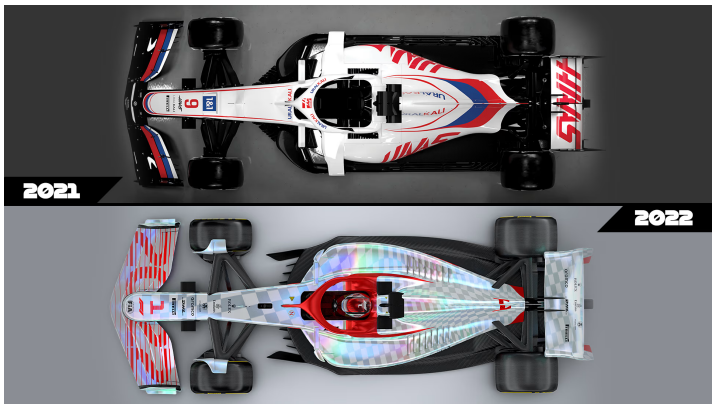
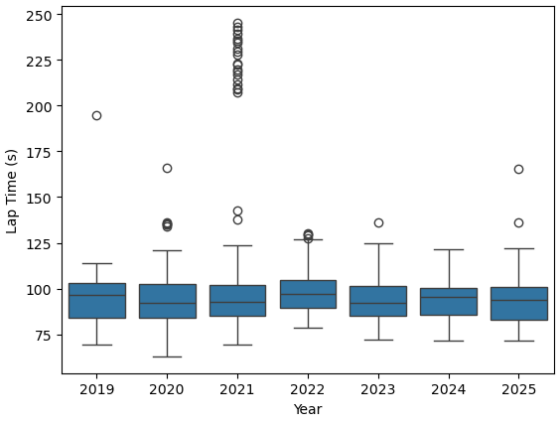


Figure 1: Top down view of the 2021 and 2022 vehicle. Redesigned vehicle is noticeable from enlarged front spoiler and rear wing with additional curvature. Retrieved from Formula 1.com

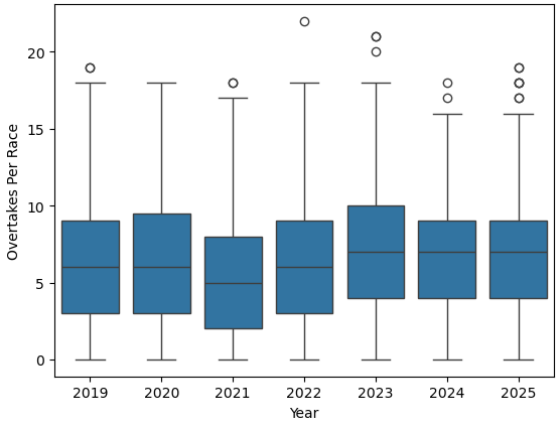
3 Methodology and Approach

3.1 Exploratory Data Analysis

Before the main analysis, an exploratory data analysis (EDA) was performed to prevent inclusion of outliers and presence of multicollinearity. EDA focused on the numeric features and identified that the general trend of LapTime stayed mostly within the interval between 75 to 125 seconds across 7 seasons. Most of the outliers were concentrated in 2021 and 2025. Interestingly, it can be seen that the general trend of Lap Time marginally increases before and after 2022.



(a) Boxplot of the lap time from all final races from 2019 to 2025.



(b) Boxplot of the number of overtakes all final races from 2019 to 2025.

Figure 2: Comparison of boxplots of the distribution of lap time and number of overtakes per race on a yearly basis.

The trend of the number of overtakes per race between 2022 to 2025 somewhat did increase, although the differences were were not significant. Checking the distribution of the frequency across the board also once again revealed that, except for a few outliers, Lap Time follows a normal distribution,

as can be observed below. This ensured that our data would be fit for a Linear Mixed Effects model. Additionally, atmospheric factors such as humidity and air temperature also followed a roughly normal distribution.

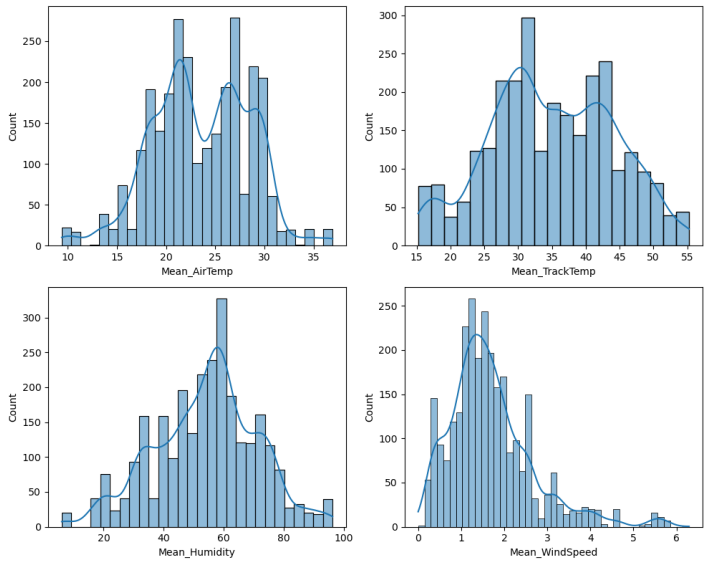


Figure 3: Visualization of various environmental factors. It can be noticed that most follow a rough normal distribution.

Lastly, the correlation between all numerical columns revealed that there was no serious multicollinearity between covariates that were thought to be highly correlated, such as the speed between different sectors and the Lap Time. Most of the correlation coefficients remained below the threshold of 0.7, except for two variables, mean air and track temperatures. Therefore, track temperature will be removed from the linear mixed effects model.

3.2 Data Collection & Aggregation

The data for this study will be sourced from the FastF1 Python library [Schaefer, 2025]. FastF1 is a publicly available API that provides access to F1 data, which is permissible for use in academic and non-commercial research. The data collection process will involve writing a Python code to iterate through each race from the beginning of 2019 through 2025. For each race event, we will load the session data using fastf1.getsession() and use the library's caching feature to store the data in a local folder to export into CSV format to reduce load times on subsequent runs. Data from the results and laps Data Frames was extracted and combined to form the primary dataset. Results were aggregated at a race level, where each row is dedicated to a lap per driver in a race. Numeric values such as speed and environmental factors were averaged, and There will be approximately 20 - 24 races per season over 5 seasons. The 2019 season is selected as the starting point as it represents a mature phase of the 2017 aerodynamic regulations, providing a stable and well-established baseline for comparison. Races that were red-flagged and not completed to at least 90% of their original distance were removed as they would fail to yield proper results for a fair comparison. Sprint races were similarly omitted as their shorter format could alter race dynamics. Additionally, we performed a feature engineering to calculate how many times a driver overtakes opponents per race by calculating the difference between lap_i and lap_{i+1} then aggregating the count per driver for every race.

3.3 Random Forest Classification

A Random Forest Classification model using Python's Scikit-Learn library was employed to predict whether a race lap occurred pre- or post-regulation [Pedregosa et al., 2011]. The model was trained on the aggregated lap dataset using most features inclduing driver, laptime, speed across sectors, number of overtakes, environemental factors, and tyre compound. To prepare the data for modeling,

categorical features (Driver, Compound) were transformed into a numerical format through one-hot encoding. The dataset was then split for training and test sets, with 80% of the data allocated for training the model and the remaining 20% for validation. Stratification was used during this split to ensure that the proportion of pre- and post-regulation lapses was maintained in both the training and testing samples to prevent class imbalance. Additionally, all lap times greater than 150 seconds were considered as outliers and removed before the analysis. Inclusion of environmental factors (AirTemp, TrackTemp, WindSpeed, Humidity) on trial run of the model resulted in a model with AUC of ~ 1 and indicated overfit. Therefore, these were removed during the real analysis to focus more on car orientated factors such as speed and tyre. Model was evaluated with a confusion matrix with scores for precision, accuracy f1, and as well as ROC curve.

3.4 Linear Mixed Effects Model

Linear Mixed Effects Model was chosen over a standard linear regression, given the nature of the Formula 1 data, which contains multiple, non-independent observations (laps) from the same drivers and different drivers under one data [Bolker et al., 2009]. The fixed effects covariates included tyre degradation (TyreLife), compound choice (FinalStintCompound), environmental conditions (Mean_AirTemp), race outcome variables (FinalPosition, Overtakes per race) and race level pace (mean speed). We also included an interaction between regulation era and mean speed at straight line to capture differential effects of regulation changes on them. Random intercepts were set for both Driver and Track (RaceID) to account for unseen correlation within drivers and races.

$$\begin{aligned}\text{LapTime} \sim\ & \text{TyreLife} + \text{FinalStintCompound} + \text{Mean\_AirTemp} + \text{FinalPosition} + \text{Overtakes\_Per\_Race} \\ & + \text{RegulationEra} \times \text{Mean\_SpeedST} + (1 \mid \text{Driver}) + (1 \mid \text{Race\_ID})\end{aligned}$$

Driver and Track (RaceID) are given random intercepts because they are categorical variables rather than numerical values we're directly interested in. As noted by Bolker et al. (2009), treating these factors as random allows the model to quantify between drivers and support the inference that extends beyond specific drivers and races in the dataset. Random effects were evaluated using ICC, studentized residuals and multicollinearity. R programming language was used run the mixed effects model [Pedregosa et al., 2011].

Table 1: Variables Used in Linear Mixed Effects Model

Variable Name	**Data Type**	**Interpretation**
LapTime	Continuous (numeric)	Time to complete a lap (response variable).
TyreLife	Continuous (numeric)	Number of laps completed on the current tyre set.
FinalStintCompound	Categorical (factor)	Tyre compound used in the final stint.
Mean_AirTemp	Continuous (numeric)	Average air temperature during the race.
FinalPosition	Continuous (numeric)	Driver's finishing position in the race.
Overtakes_Per_Race	Continuous (numeric)	Total overtakes normalized per race.
RegulationEra	Categorical (factor)	Regulatory era under which the race occurred.
Mean_SpeedST	Continuous (numeric)	Average speed on the main straight.
Racename	Categorical (factor)	Unique ID per race (random effect).

4 Results

4.1 Results from the Linear Mixed Effects Models

Table 2: Linear Mixed-Effects Model Fixed Effects

Variable	Estimate	Std. Error	t value	p-value
Intercept	183.144	3.882	47.18	$< 0.001^{***}$
TyreLife	-1.123	0.090	-12.45	$< 0.001^{***}$
FinalStintCompound (HARD)	1.402	1.359	1.03	0.302
FinalStintCompound (INTERMEDIATE)	6.581	1.519	4.33	$< 0.001^{***}$
FinalStintCompound (MEDIUM)	1.383	1.354	1.02	0.307
FinalStintCompound (SOFT)	1.163	1.363	0.85	0.394
FinalStintCompound (UNKNOWN)	83.990	12.954	6.48	$< 0.001^{***}$
FinalStintCompound (WET)	6.120	2.000	3.06	0.002^{**}
Mean_AirTemp	-3.141	1.011	-3.11	0.002^{**}
FinalPosition	0.223	0.013	17.09	$< 0.001^{***}$
Overtakes_Per_Race	0.013	0.019	0.71	0.479
RegulationEra	-26.754	4.952	-5.40	$< 0.001^{***}$
Mean_SpeedST	-0.322	0.011	-29.38	$< 0.001^{***}$
RegulationEra × Mean_SpeedST	0.100	0.015	6.66	$< 0.001^{***}$

The linear mixed-effects model revealed significant variability in lap times, with the ICC of 0.944 indicating that most variance can be rooted to race-level differences after accounting for fixed effects. Random intercept variance was larger for races than for drivers.

Some fixed effects had strong correlation with lap time. For instance, tire life was negatively associated with lap time ($coef = -1.12$, $p < 0.001$), indicating faster laps with increasing tire age within stints. Mean straight-line speed was also negatively associated with lap time ($coef = -0.32$, $p < 0.001$). The post-2022 regulation era was associated with lower lap times ($coef = -26.75$, $p < 0.001$).

Mean air temperature was associated with slightly slower lap times ($coef = -3.14$, $p = 0.002$). Final race position showed a positive association with lap time. Model had no issues regarding multicollinearity as diagnostic indicated low VIF most predictors, with moderate collinearity appearing only on the regulation and speed interaction term.

Table 3: VIF

Term	VIF	Adj. VIF
TyreLife	1.21	1.10
FinalStintCompound	1.14	1.07
Mean_AirTemp	1.06	1.03
FinalPosition	1.09	1.05
Overtakes_Per_Race	1.16	1.08
Mean_SpeedST	2.13	1.46
RegulationEra	5.33	2.31
RegulationEra × Mean_SpeedST	6.36	2.52

4.2 Results from the Random Forest Classification Model

Some of the most significant features were all speed related, with lap time being the biggest differentiator, following by the speed captured at the finish line. Most driver factors were significantly less important than other features, with the exception of intermediate tire compound.

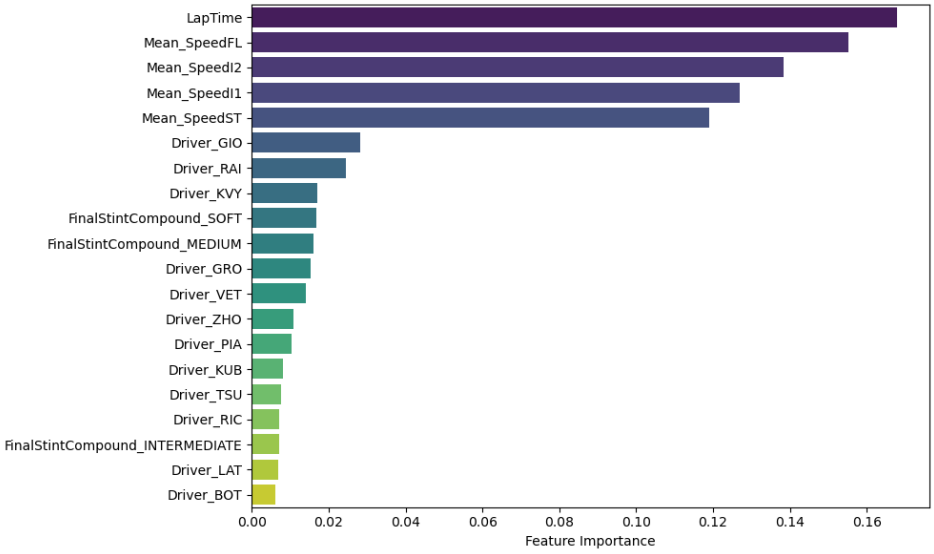


Figure 4: Visualization of outstanding features from the random forest model. It can be noticed that speed related variables were most significant.

The random forest model performed relatively well. It achieved an accuracy of 0.87, with balanced precision and recall across both classes: F1-scores 0.83 and 0.90 for classes 0 and 1, respectively. The confusion matrix indicated a low rate of false negatives for the positive class. The AUC score was also roughly 0.94.

Table 4: Confusion Matrix

	Predicted Class	
Actual Class	0	1
0	172	52
1	19	316

Results from the cross validation confirmed that model did not over or underfit, marked by consistent performance across folds with accuracy of 0.871 and F1 score of 0.869. Low variability implies that the model generalizes well and is likely not sensitive to folded tests.

Table 5: Cross-Validation Result

Metric	**Mean ± Std. Dev.**
Accuracy	0.8710 ± 0.0044
Precision	0.8726 ± 0.0030
Recall	0.8710 ± 0.0044
F1-score	0.8693 ± 0.0052

5 Conclusion

5.1 Discussion

This research investigated how race characteristics are different following 2022 FIA rule change, moving beyond simple race prediction to identify distinguishing features of post-regulation races. The linear mixed-effects model revealed that race-level were largest in explaining variability in lap time. Adjusted ICC of 0.944 indicates that the majority of variance is can be traced to conditions unique to each race, rather than drivers skill or style. Some of the covariates most strongly associated with lap-time were tyre life, track speed, and air temperature. Longer tyre life and higher straight line speeds were consistently associated with faster laps. It is worth noting that, the post-2022 regulatory era was associated with substantially lower lap times ($\beta = -26.75$), and the positive interaction with mean start/finish speed suggests that drivers capitalized differently on track speed under the new regulations. However, limitations does exist for this model. While we assume that tyre degradation is linear, the residual plots shows non linear residual patterns with occasional spikes [Hartig, 2025], suggesting that a simple linear slope may not fully capture the rate of the loss of tyre performance over a lap with no pit stops.

The classification also resulted in a relatively good predictive ability (test-set accuracy of ~ 0.87), indicating that races from two periods with the features above can be quantitatively distinguished [Bunker and Thabtah, 2019, Jafri, 2024, Garcia Tejada, 2023]. There were moderate collinearity for the interaction between regulation era and speed, which is expected when examining post-regulatory changes that influence multiple aspects of performance [Kim, 2019]. In short, the mixed model quantified how lap times and performance metrics are affected by the regulatory shift, while the random forest model confirmed that these differences are clearly distinguishable.

Our results suggest that post-regulation races are to some extent faster and imply a change in tyre management, speed, and driver performance which justifies the potential impact of rule changes on race dynamics [Méndez, 2023, Stuart, 2021]. These findings will have implications for race strategy, vehicle setup, and regulatory evaluation from the perspective of FIA.

5.2 Future Work

Future work can explore driver specific "response" to regulations using random slopes, and as well as team "response" including pit-stop timing. At the same time, we could include additional factors into account and change the equation as seen on work by [Jafri, 2024]. This includes building a new relevant feature, such as "Number of Overtakes" or "Changes in Max/Min Velocity" could also help us obtain meaningful results, as witnessed by previous work by Sicoie and Tejada. Environmental factors can also be a subject of examination by only isolating factors such as temperature or humidity against results or speed using telemetry data.

One further change to validate the significance of the results is to conduct the analysis on a single selected track (e.g., the Spanish Grand Prix) across seven seasons and expand the analysis to a lap-level instead of a race-level, without all outliers omitted. This approach would help control environmental factors and allow a deeper investigation, as the model could be run on a more granular level that takes lap by lap variations into account. Finally, expanding the dataset across multiple season by choosing a different past regulation change would allow for a more comprehensive assessment of regulatory impacts.

References

[Ambler, 2024] Ambler, W. (2024). How data analysis transforms f1 race performance — catapult. Catapult.

[AWS, 2024] AWS (2024). F1 insights powered by aws — formula 1 uses amazon web services.

[Bolker et al., 2009] Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology Evolution*, 24:310–334.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

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

[Cao, 2025] Cao, C. (2025). Sports data mining technology used in basketball outcome prediction. ARROW@TU Dublin. Accessed: 2025-12-15.

[Casals et al., 2025] Casals, M., Fernández, D., Zumeta-Olaskoaga, L., Sánchez, A., and Zuccolotto, P. (2025). Reporting of generalized linear mixed models (glmm) in sports sciences: A scoping review. *Journal of Sports Analytics*, 11:22150218251384557.

[Formula 1, 2025] Formula 1 (2025). Everything you need to know about f1. Accessed: 2025-12-15.

[Foundation, 2025] Foundation, P. S. (2025). Python language reference, version 3.13.7. [Computer software]. Accessed: 2025-12-15.

[Garcia Tejada, 2023] Garcia Tejada, L. (2023). Applying machine learning to forecast formula 1 race outcomes.

[Harris et al., 2020] Harris, C., Millman, K., van der Walt, S., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M., Brett, M., Haldane, A., del R'io, J., Wiebe, M., Peterson, P., G'erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. (2020). Array programming with numpy. *Nature*, 585:357–362.

[Hartig, 2025] Hartig, F. (2025). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. CRAN. R package vignette, version accessed from CRAN: https://rdrr.io/cran/DHARMa/f/vignettes/DHARMa.Rmd, Accessed: 2025-12-15.

[Jafri, 2024] Jafri, A. (2024). Predicting formula 1 race outcomes: A machine learning approach.

[Jowett, 2022] Jowett, R. (2022). Overtaking fundamentals: Action packed 2022 f1 season?

[Keertish Kumar and Preethi, 2023] Keertish Kumar, M. and Preethi, N. (2023). Formula one race analysis using machine learning. In Gunjan, V. K. and Zurada, J. M., editors, *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, pages 533–540, Singapore. Springer Nature Singapore.

[Kim, 2019] Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, 72:558–569.

[McKinney, 2010] McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445. Accessed: 2025-12-15.

[Méndez, 2023] Méndez, L. A. (2023). Quantifying the impact of the 2022 formula one technical regulations on wake turbulence: A numerical analysis.

[Nagle, 2022] Nagle, D. (2022). Racing your rival: Cluster analysis of formula 1 drivers.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

[R Core Team, 2021] R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Accessed: 2025-12-15.

[RMetS, 2011] RMetS (2011). F1 and the weather. Accessed: 2025-12-15.

[Schaefer, 2025] Schaefer, P. (2025). Fastf1, version 3.6.1. MacOS GitHub. Accessed: 2025-12-15.

[Sicoie, 2022] Sicoie, H. (2022). Machine learning framework for formula 1 race winner and championship standings predictor. *Bachelor's thesis, Tilburg University, School of Humanities and Digital Sciences*.

[Stuart, 2021] Stuart, G. (2021). 10 things you need to know about the all-new 2022 f1 car — formula 1®.

[van Kesteren and Bergkamp, 2023] van Kesteren, E.-J. and Bergkamp, T. (2023). Bayesian analysis of formula one race results: disentangling driver skill and constructor advantage. *Journal of quantitative analysis in sports*, 19(4):273–293.

6 Appendix

6.1 Appendix A: List of variables from the dataset from FastF1 API documentation:

Table 6: Columns in the Aggregated Race-Level Dataset

Column Name	Description
Year	Year of the race event.
RaceName	Name of the Grand Prix.
Driver	Driver abbreviation.
DriverNumber	Official driver number.
TotalLapsCompleted	Total laps completed by the driver in the race (Max LapNumber).
IsPersonalBest	Whether the driver set a personal best lap (Any True).
TyreLife	Maximum tire life achieved in a single stint.
RegulationEra	**0**: Pre-2022 Rules; **1**: Post-2022 Rules.
LapTime	Fastest Lap Time achieved in the race.
FinalStintCompound	Tire compound used in the driver's final race stint.
Team	Driver's team (constructor) name.
Mean_SpeedI1	Average speed at Intermediate Trap 1 (km/h).
Mean_SpeedI2	Average speed at Intermediate Trap 2 (km/h).
Mean_SpeedFL	Average speed at the Flying Lap line (km/h).
Mean_SpeedST	Average speed at the Start/Finish line (km/h).
Mean_AirTemp	Average ambient air temperature (C).
Mean_TrackTemp	Average track temperature (C).
Mean_WindSpeed	Average wind speed.
Mean_Humidity	Average humidity.
Mean_TrackStatus	Average track status code during the race.
Overtakes_Per_Race	Total number of successful on-track overtakes.
FinalPosition	Driver's official final finishing position.
FinalRaceTime	Time when the driver crossed the finish line.

6.2 Appendix B: Additional Figures and Tables from results

Table 7: Random Effects and Model Diagnostics

Component	**Variance**	**Std. Dev.**
Race_ID (Intercept)	157.07	12.53
Driver (Intercept)	0.13	0.35
Residual	9.26	3.04
Observations	2791	
Race_ID Groups	143	
Driver Groups	39	
REML Criterion	14984.2	

Table 8: Intraclass Correlation Coefficients (ICC)

Metric	**Value**
Adjusted ICC	0.944
Unadjusted ICC	0.468

Table 9: VIF table with all variables

Term	**VIF**	**Adj. VIF**	**Tolerance**
TyreLife	1.21	1.10	0.83
FinalStintCompound	1.14	1.07	0.88
Mean_AirTemp	1.06	1.03	0.95
FinalPosition	1.09	1.05	0.91
Overtakes_Per_Race	1.16	1.08	0.86
Mean_SpeedST	2.13	1.46	0.47
RegulationEra	5.33	2.31	0.19
RegulationEra × Mean_SpeedST	6.36	2.52	0.16

Table 10: Random Forest Performance

Class	**Precision**	**Recall**	**F1**	**Support**
0	0.90	0.77	0.83	224
1	0.86	0.94	0.90	335
Accuracy	-	-	0.87	559
Macro Avg	0.88	0.86	0.86	559
Weighted Avg	0.88	0.87	0.87	559

6.3 Appendix C: Additional Figures from EDA and Model Results

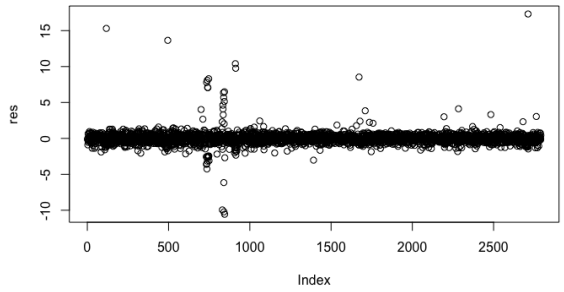


Figure 5: Visualization of the residual from the linear mixed model

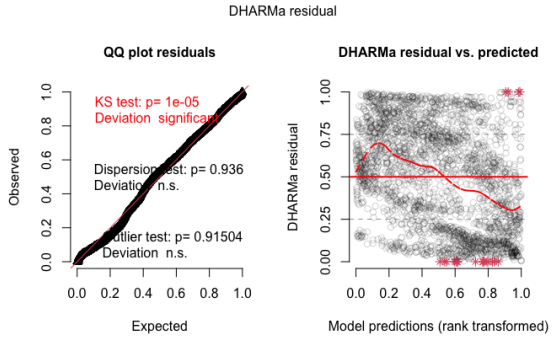


Figure 6: DHARMa residuals from linear mixed model

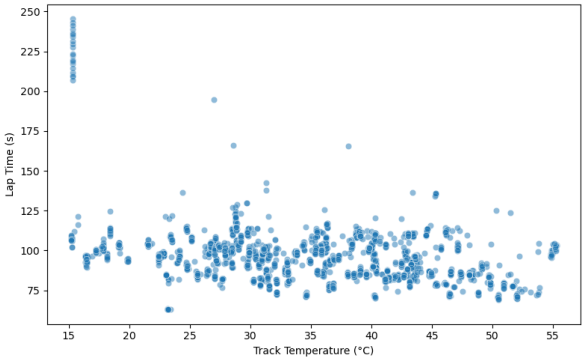


Figure 7: Scatterplot of the relation between track temperature and lap times

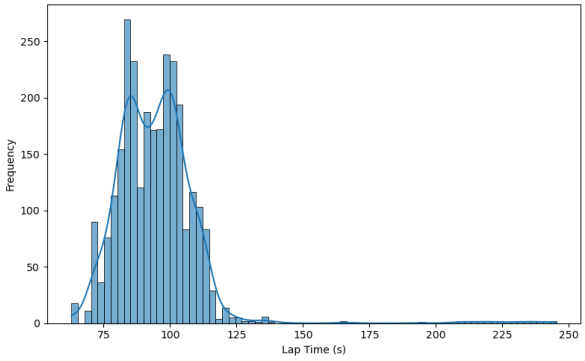


Figure 8: Barplot of time times for all races across 7 seasons, roughly following a normal distribution.

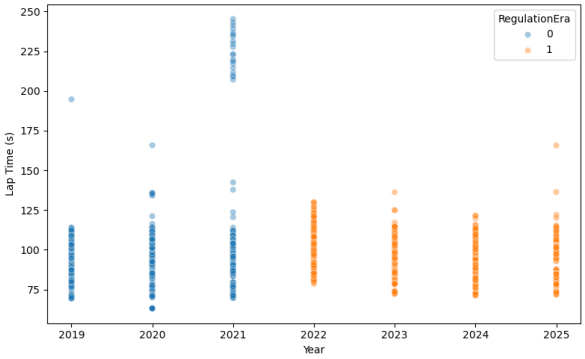


Figure 9: Scatterplot of the distribution of lap time by on a yearly basis.