2024

CS982 – Big Data Technologies Assignment

TITLE: MACHINE LEARNING BASED PIT STOP STRATEGY PREDICTION IN FORMULA 1 RACING

Table of Contents

INTRODUCTION

Formula One is a racing series of single-seated cars, which is by many considered as a pinnacle of motorsports and the most prestigious racing competition (Formula 1, 2024). In 2023 season, the series consisted of 10 teams which were each represented by two drivers. Each of these 10 teams designs a car within a set of strict regulations issued by the Fédération Internationale de l'Automobile (FIA, 2022). The technical regulations change every year, but not every year brings significant changes, and the sport can be divided into different technical eras. Within these eras the year-on-year regulation changes are usually minor. The last major regulation change was implemented for the 2022 season when ground effect of the cars was emphasized. This contributed to the current generation of cars having more downforce, and consequently being able to follow each other more closely for longer periods of time, providing better racing (Stuart, 2021). Along with the new ruleset, the FIA has implemented the cost cap in 2021 for the first time in sports history, which limits the teams spendings (Barretto, 2020). With limited spendings, each team's focus is on being as efficient as possible. In the past, teams could have invested more in the pursuit of a better finish which usually results in a larger money prize and better sponsorship deals or product sales which are a big part of many teams' participation in the sport. It is estimated that Red Bull alone gained £1.6 billon in advertising in the period from 2000 to 2014 from Formula One (Newey, 2017, p. 122).

As teams cannot try investing more in pursuit of a better finish, they have to emphasise being efficient with the allocated resources. One important aspect of each race are pit stops. Each driver must drive on two different tyre compounds during a dry race, and hence, they are required to do at least one pit stop (FIA, 2022). However, this number usually varies anywhere between 1 and 4 and can significantly influence a drivers race and final finish position. Hence, pit stops are an important race occurrence whose understanding could benefit teams.

Therefore, in this study we have decided to build a prediction model for estimating the number of pit stops a driver will do in a race based on their initial race conditions. These conditions include weather variables, race track characteristics and drivers starting position (grid position) and tyre choice. Being able to predict the number of pit stops any driver will do, could help a team better simulate and prepare for different races which could translate into better race strategy and end result.

DATASET

DATASET DESCRIPTION

Due to major regulation changes that took place in 2022 season, our dataset only comprises of observations obtained in 2022 and 2023 seasons. As of writing, these are the only completed seasons using the current set of regulations. For these seasons, we have collected seven feature variables and the response variable, the number of pit stops. All of our data was collected from publicly available resources in a process described in the following section.

DATA COLLECTION AND VARIABLES

Prior to collecting a dataset to carry out the analysis, three groups of factors have been identified as potentially having an influence on the number of pit stops (dependent variable) a driver will do. These groups are: weather, track characteristics and driver situation variables. From publicly available dataset (Vaibhav, 2024), weather variables Track and Air temperature and Humidity were collected. The dataset contains weather variables for each minute of every race from 2018 to 2023 season. These observations were filtered to only include the two seasons of interest and only the initial (first minute) weather conditions of every race. The dependent variable, number of pit stops, as well as drivers start and end position were obtained from (Overdijk, 2022). Again, this dataset was filtered to only include the seasons of interest and we have excluded the observations of drivers who have not finished the race (DNF), did not start the race (DNS), were disqualified (DNQ) or started from the pitlane. The same dataset was also used to obtain Track type variable, which groups race tracks into one of two categories: “Race” or “Street” depending on if a track was built for racing or if it is a part of a normal street layout. The F1 page (Formula 1 - The Official F1 Website, 2023) was used to find the number of corners in each track and the starting tyre information was obtained from (Pirelli, 2024).

All collected variables were merged into one dataset with variable names shown in table 1.

Table 1. Dataset used for building the prediction model.

	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
Index	No. of pitstops	Grid position	Starting tire	AirTemp	TrackTemp	Humidity	No. of corners	Race type

The obtained dataset will be analysed by calculating its descriptive statistics and plotting the data. Once the relationship between dependent variable and independent variables is plotted and

explored, further data preprocessing such as outlier removal will be considered prior to building the prediction model.

Along with the prediction model, hierarchical clustering will be used to group drivers into different clusters depending on their start and end race positions and the number of pit stops they done.

PROPOSED METHODOLOGY

ANALYTICAL TECHNIQUES

This project comprises of both supervised and unsupervised learning parts. In the first part of the report, Decision tree algorithm is used to predict the number of pit stops a driver will do in a race. The second part of the report uses hierarchical clustering to group drivers into different clusters based on their start and finish positions and pit stop count.

DATA SPLITTING AND VALIDATION

Our supervised models are trained using training data set consisting of 70% of observations, while the remaining 30% of observations are used to evaluate the model performance.

DATA ANALYSIS AND VISUALIZATION

The dataset used in this study consists of 739 observations collected over 44 Grand Prix events raced over a period of two seasons. The summary statistics of the number of pit stops variable are shown in Table 1.

Table 1. Summary statistics for dependent variable number of pit stops.

	Number of pit stops
Count	739
Mean	2.0920
Std	1.0410
Min	1.00
Q1	1.00
Q2	2.00
Q3	3.00
Max	7.00

Per previously stated rules and shown here in Table 1., the minimum number of pit stops a driver must do in a dry session is one. Unlike the minimum, the maximum number of pit stops is not imposed by rules. However, each driver is allocated a set amount of tyres for a race weekend (Seymour, 2023). The maximum number of pit stops any driver has done in 2022 and 2023 seasons was seven. This is a very high number for modern Formula One, especially considering the mean number of pit stops for our data is 2.0920 and the standard deviation is 1.0410. To further explore the number of pit stops variable, Figure 1. was plotted.

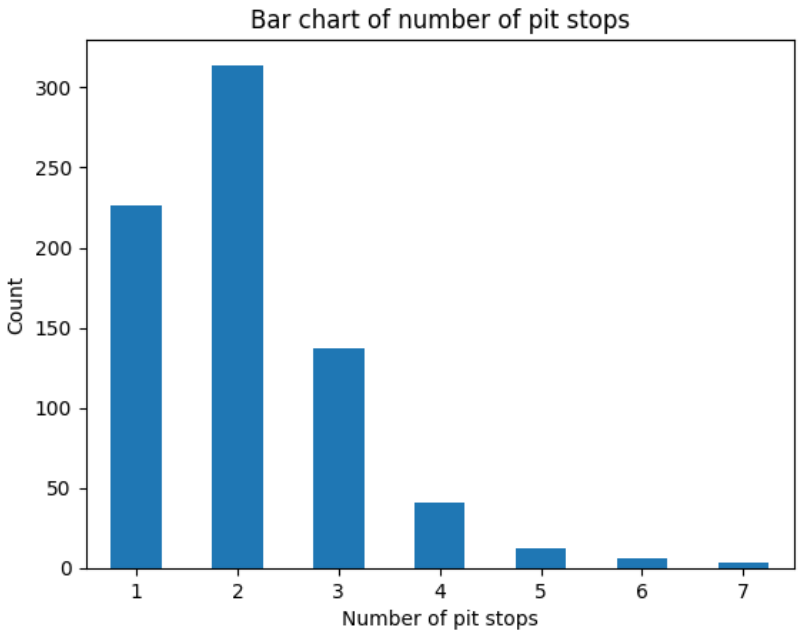


Figure 1. A bar plot showing a count for each number of pit stop occurrences in a race.

The bar plot in Figure 1. shows that four or less pit stops account for a vast majority of all observations. In fact, only 21 out of 739 observations were five or more pit stops. When looking at

which races have these observations occurred, we noted 15 of them have happened at the 2023 Netherlands Grand Prix. This was a chaotic session with constantly changing weather conditions forcing drivers to switch between wet and dry tyres multiple times (Formula 1 - The Official F1 Website, 2023). As these observation are uncommon and mostly appeared at one particular race with abnormal weather conditions, we have classified them as outliers and removed from the dataset.

STARTING TYRE ANALYSIS

To further explore the existence of any potential outliers, of interest is the distribution of starting tyres. In Formula One, every racing week teams are provided with five different tyre compounds made by Pirelli. This includes three types of dry condition tyres: softs, mediums, and hards; and two types of wet condition tyres: intermediates and wets (Pirelli, 2019). The distribution of tyres used at the start of races is shown in Table 2. The dry tyres were used far more frequently than the wet condition tyres. This is not surprising as in the recent years the sport has preferred to delay the start in wet conditions rather than use the wet conditions tyres. The most recent example of this phenomena occurred at the 2024 Brazilian Grand Prix qualifying event which got rescheduled to the following day due to a heavy rain (Benyon & Mitchell-Malm, 2024). As a consequence delaying sessions starts during wet conditions, intermediate tyres have only been used in three races at the start and wet tyres were only used once. In comparison, hard tyres were used in 25 different races at the start, while medium tyres were used in 39 of them. Therefore, due to a small variety of track and weather conditions for which wet condition tyres were used, we have decided to treat them as outliers and removed them from our dataset.

Tabel 2. Count of usage of different tyre compounds at the start of races.

Tyre type	Count
Soft	186
Medium	418
Hard	71
Intermediate	47
Wet	17

With imposing these constraints on the dataset, our problem has been reduced to classification of observations into one of four classes (representing the number of pit stops respectively) when dry condition tyres are used on race starts. After removing outliers, 654 out of 739 observations have remained in the dataset.

To explore the relationship between Starting tyre and the Number of pit stops, Figure 2. was plotted.

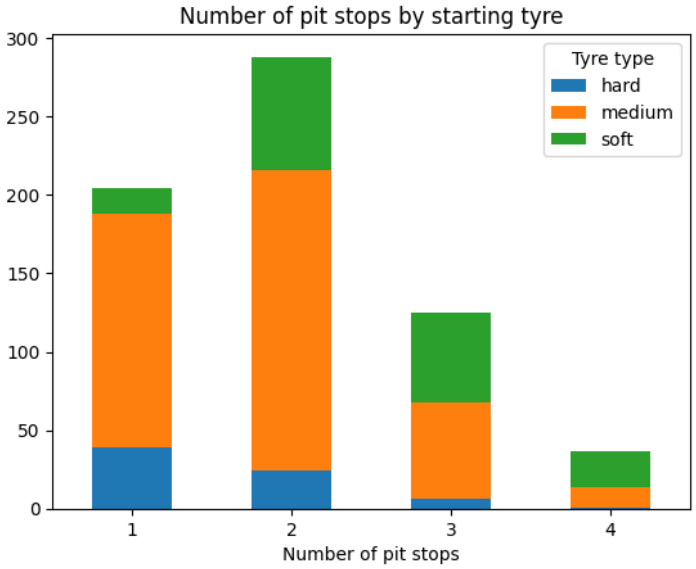


Figure 2. Two way stacked bar chart of Number of pit stops and Starting tyre type.

It is visible that the increase in the number of pit stops is followed by the increase of soft tyre usage, while the decrease is followed by the increase of hard tyre proportion.

TRACK TYPE AND THE NUMBER OF PIT STOPS

Different results were also obtained when considering Track types. On average, 2.2260 pit stops were done in race tracks and 1.5185 in street tracks. The difference across two types of tracks is visualised in Figure 3.

Table 3. Number of pit stops per two track types.

Circuit type	Count of observations	Mean number of pit stops	Std of number of pit stops
Race	438	2.2260	0.8322
Street	216	1.5185	0.6885

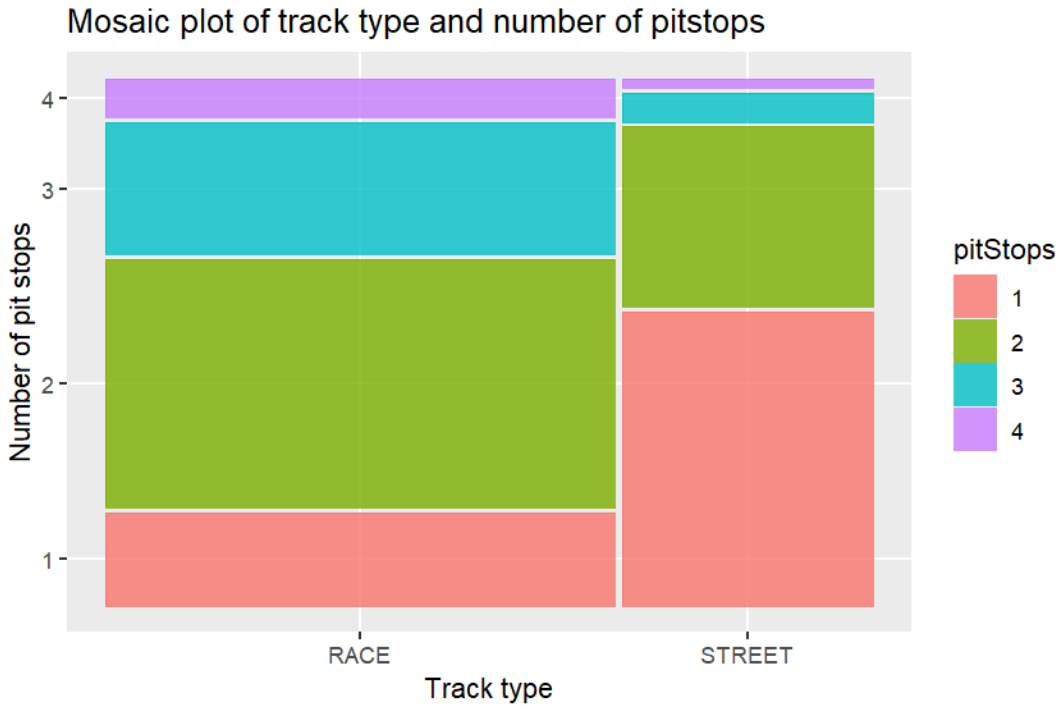


Figure 3. Mosaic plot showing the proportion of observations in race and street tracks and proportion of number of pit stops within them.

It shows that one pit stop strategies are a lot more common in street tracks than in race tracks which were dominated by two stop pit strategy. Further, three and four pit stop strategies account for noticeably smaller proportion of all observations in street tracks when compared to race tracks.

WEATHER VARIABLES AND THE NUMBER OF PIT STOPS

To explore the relationship between weather variables and the number of pit stops, Figure 4. was created. It does not show any clear relationship between our dependent and independent variables, however, Humidity and Air temperature could potentially have a small negative effect on the number of pit stops.

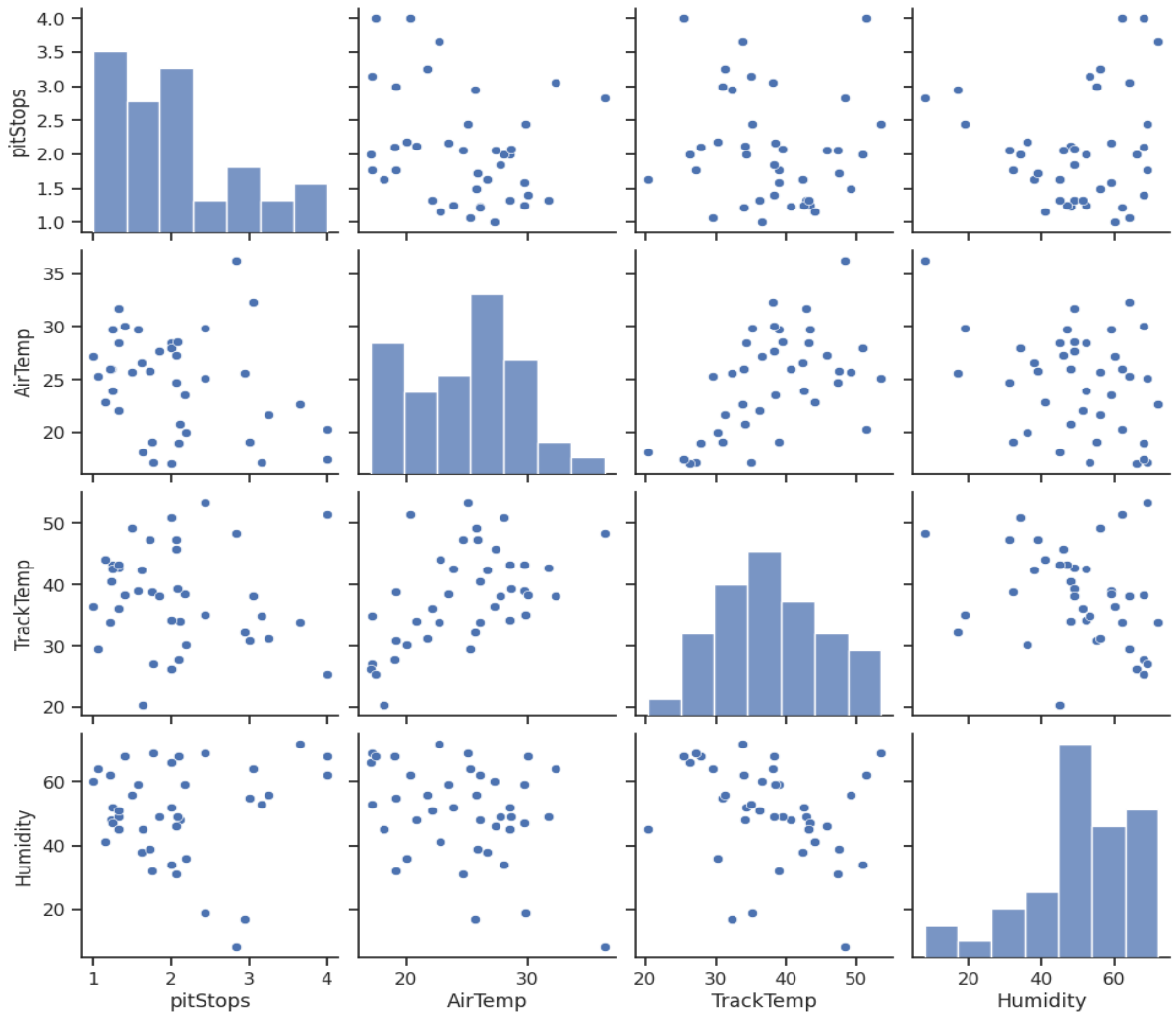


Figure 4. Scatterplots of all continuous feature variables and averaged number of pit stops per each race.

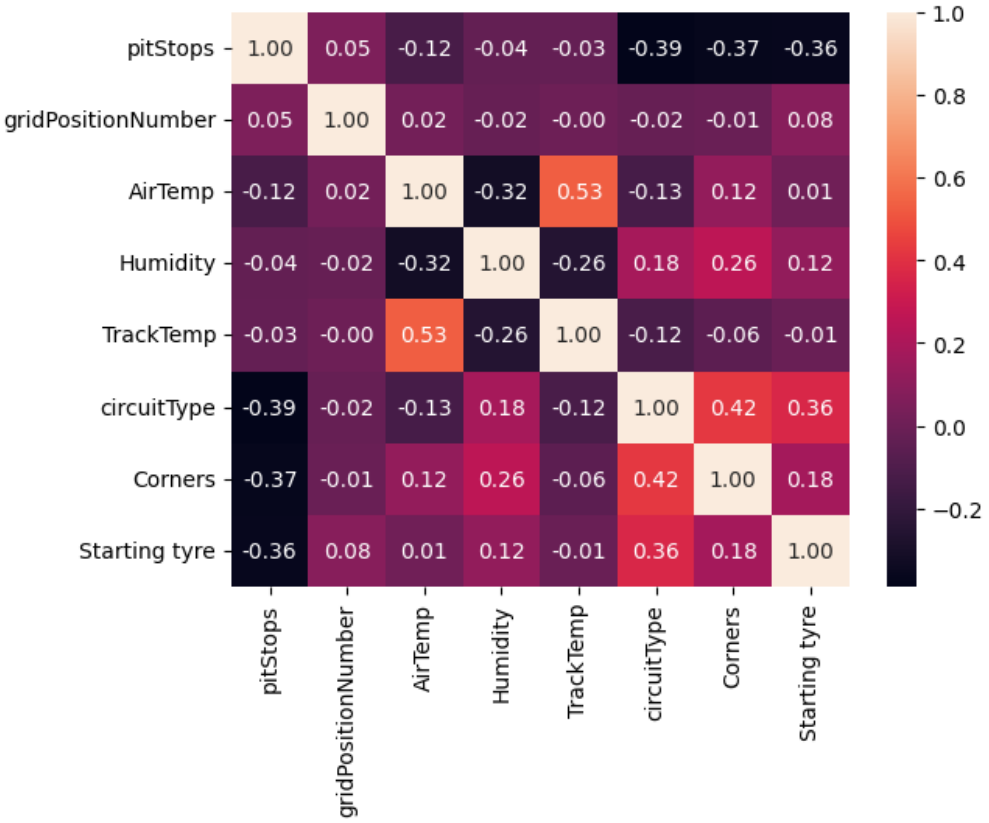
Track and Air temperature have a visible positive relationship and hence, including both of these variables in our prediction model might not add any new information. Therefore, the prediction model will be considered with both of these variables and with only one of the two. To further explore and quantify the effect our features have on the dependent variable, their correlation matrix was calculated in Table 4. The highest correlation coefficient is between Air and Track temperature as was suggested by Figure 4. The number of pit stops has a moderate negative correlation with Circuit type, Corners and Starting tyre, while the correlation with other variables is relatively small. For this reason, we have decided to train multiple models that all have the three features of moderate correlation included and contain different combinations of the remaining four variables.

ANALYSIS OF CORRELATION MATRIX

It should be noted that for the correlation matrix in Table 4., Circuit type and Starting tyre were converted into discrete variables. Circuit (track) type was transformed into a binary variable with “RACE” getting value 1 and “STREET” value 2 assigned, while Starting tyres values were labelled

1, 2 and 3 for soft, medium and hard tyres respectively. Such notation convenience should not present a problem in our analysis as Starting tyre is an ordinal categorical variable.

Table 4. Correlation matrix between the all variables of the dataset.



MODEL IMPLEMENTATION AND RESULTS

SUPERVISED LEARNING

To ensure that prediction models do not overfit to our dataset, we have decided to spilt it into two sets: the training and the testing set. The ratio that was used is 70/30 between training and testing, which is the ratio that has yielded a superior performance in some studies, such as (Nguyen, Ly, Ho, Al-Ansari, Le, Tran, Prakash & Pham, 2021). The algorithm used for building prediction models with training data was a Decision tree algorithm. The algorithm builds a tree consisting of a root, nodes and leaves. Each observation starts at a root and is passed down through nodes which have different conditions for its feature values. These conditions determining how an observation is passed down a tree until it reaches a leaf where it gets assigned a class (in classification model) (Alpaydin, 2014). As full Decision tree models are prone to overfitting (James, Witten, Hastie, Tibshirani &Taylor, 2023), our models will be pruned to obtain the best out-of-sample performing model. Pruning is a process of imposing a minimum number of observations at each node after nodes will not be further divided. This prevents decisions being made from too few observations which can result in generalisation error (Alpaydin, 2014). In table 5. the best performing out-of-sample Decision trees with different sets of feature variables are shown.

Table 5. Different prediction model with their best performing out of sample performance subject to different minimum number of nods constraint.

Excluded variable	Classification rate	Node minimum
N/A (full model)	0.7360	8
TrackTemp	0.7462	9
AirTemp	0.7157	9
AirTemp and TrackTemp	0.6954	6
Humidity	0.7157	6
Humidity and TrackTemp	0.7513	9
Humidity and AirTemp	0.7005	8
Humidity, AirTemp and TrackTemp	0.5533	16
gridPositionNumber	0.7513	8
gridPositionNumber and AirTemp	0.7563	8
gridPositionNumber and TrackTemp	0.7665	9
gridPositionNumber, AirTemp and TrackTemp	0.7310	6
gridPositionNumber, Humidity	0.7513	8
gridPositionNumber, Humidity and TrackTemp	0.7563	9
gridPositionNumber, Humidity, AirTemp	0.7360	8
gridPositionNumber, Humidity, AirTemp and TrackTemp	0.5431	2

OUT-OF-SAMPLE MODEL PERFORMANCE

The most accurate model has achieved a classification rate of **76.65%**. It did not include Grid position (gridPositionNumber in the table) nor Track temperature variables. Its peak performance

was achieved with a minimum of 9 observations at nodes constraint. The two models that excluded all three weather variables in Table 5. have performed the worst out-of-sample, managing to achieve only 55.33 and 54.31 percent accuracy. Such result suggests that weather variables in spite of their low correlation with the dependent variable still significantly improve the classification rate. However, including all three of these weather variables did not result in superior performance compared to models including only one or two weather variables. Some of these smaller models have not just matched but also outperformed models including all weather variables. This suggests that including all three weather variables is redundant, as they likely capture a lot of the same variance in the dependent variable.

UNSUPERVISED LEARNING

CLUSTER ANALYSIS FOR DRIVER PERFORMANCE

Apart from different cars, Formula One teams are differentiated by their drivers. As there are only 20 seats in the sport, drivers who do get a chance to drive in them often only have a short amount of time to prove they deserve to be there. The most recent example of a driver failing to impress a team is that of an American driver Logan Sargeant who got axed by Williams midway through his second season (Barretto ,2024). As driver performance along with cars quality plays the most important role in the sport, we have decide to build a hierarchical clustering model as a means of comparing the drivers. For this, we have calculated each drivers average start position (grid position), average end positions and the average number of pit stops which were then used to build a model shown in Figure 5. Note that drivers end position for each race was collected in our dataset, but it was not used due to our prediction model only considering initial race conditions.

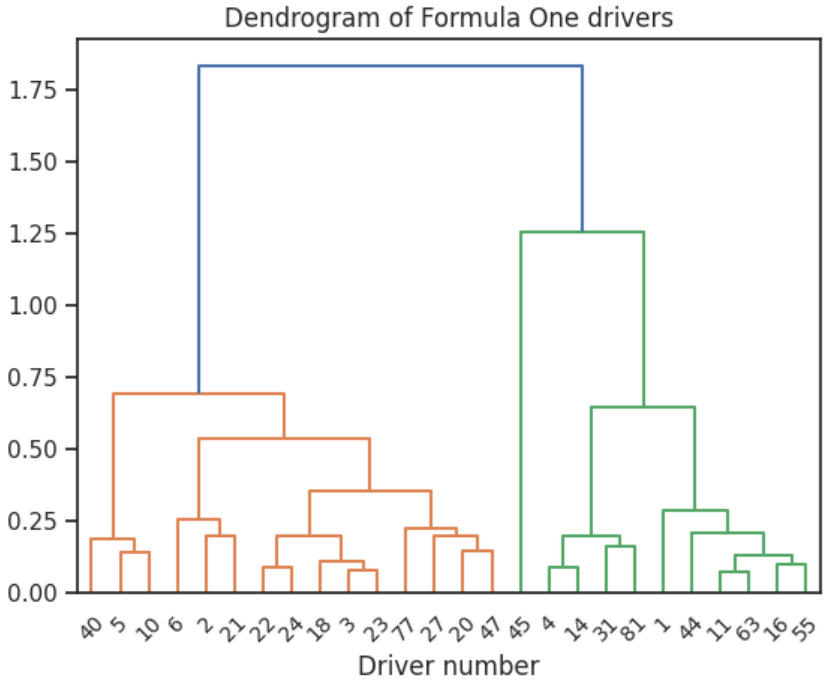


Figure 5. Dendrogram of all Formula One drivers who participated in any of 2022 or 2023 sessions. For driver names see Table A1. in the Appendix.

The dendrogram in Figure 5. clearly show two major clusters of drivers. The drivers in the right cluster coloured green mostly drive for the four best performing teams in recent years: Ferrari, Red Bull, Mercedes and McLaren. However, there are two drivers who do not belong to these teams but are joint into the same cluster at an early stage. These are Fernando Alonso (number 14) and Esteban Ocon (number 31). As the two of them are not in one of the top four finishing teams, this could suggest that their driving skills are closing the gap between their machinery and the machinery of the front cars. Another thing we can read from this graph is that Ferrari's teammates Charles Leclerc (number 16) and Carlos Sainz (number 55) are the first teammates that are clustered together. This could suggest that Ferrari has the most balanced driver lineup out of any teams.

Despite this being a relatively simple model, it shows a lot of useful information that teams can use to evaluate drivers performance relevant to other drivers.

LIMITATIONS

The dataset that was used for this study could potentially be limited by the variation of race conditions it captures, as it only consist of observations collected over two seasons of the sport. Furthermore, while the built model is effective, it uses a set of features that may not encompass all the relevant factors influencing the pit stop number, such as the tyre degradation or team decisions, which can significantly affect race outcome. The study also treats all tyre labels from different races the same, despite these not being the actual tyre compounds (which change for every race), but the relation of the three compounds used in a race. Lastly, the analysis focused solely on the number of pit stops and no context behind them was considered. These limitations suggest that further research should include a wider dataset with additional variables to enhance the model's robustness.

CONCLUSION

This study has built multiple Decision tree models using different sets of feature variables to predict a number of pit stops a driver will do in a Formula One race given the initial conditions. Three types of variables were considered for this model: weather conditions, track characteristics and driver situation variables. Track characteristics (Track type and Number of corners) and the Starting tyre variable had the highest absolute correlation with the Number of pit stops and hence were used in all the models. Weather variables and starting position (Grid position) had correlation close to 0 and different combinations of these variables were added to Decision tree models whose performance was evaluated out-of-sample. Despite, weather variables having a low corelation with the dependent variable, excluding them made model perform significantly worse. However, adding all three weather features did not improved performance when compared to models using only one or two of these variables. The best performing model achieved classification rate of 76.65%. It excluded the Grid position and Track temperature variables. Along with a Decision tree for prediction, a Dendrogram was produced using hierarchical clustering algorithm as a mean of comparing the drivers.

REFERENCES

- Alpaydin, E., IEEE Xplore , distributor and MIT Press, publisher (2014) "Decision Trees," in *Introduction to machine learning / [internet resource]*. Third edition. Cambridge, Massachusetts: MIT Press. pp.213-238.
- Barretto, L., (2024). *Why Williams decided to axe Sargeant for Colapinto*. [online] Available at: https://www.formula1.com/en/latest/article/analysis-williams-logan-sargeant-franco-colapinto-axed.5f9FpK122kaYg3Pj4fdnAK. (Accessed: 04 November 2024)
- Barretto, L. (2020). The 2021 F1 cost cap explained – what has changed, and why?. Available at: https://www.formula1.com/en/latest/article/the-2021-f1-cost-cap-explained-what-has-changed-and-why.5O1Te8udKLmkUl4PyVZtUJ. (Accessed: 04 November 2024).
- Benyon, J. & Mitchell-Malm, S. (2024). Brazilian Grand Prix F1 qualifying postponed to Sunday. Available at: https://www.the-race.com/formula-1/brazilian-f1-grand-prix-qualifying-delayed/. (Accessed: 04 November 2024).
- FIA. (2023). Formula 1 Financial Regulations.. Available at: https://www.fia.com/sites/default/files/fia_formula_1_financial_regulations_-_issue_16_-_2023-08-31.pdf.
- FIA. (2022). Formula 1: Regulations Overview. 2nd issue. Available at: fia_2023_formula_1_sporting_regulations_-_issue_2_-_2022-09-30.pdf
- Formula 1 (2024). *Everything You Need to Know about F1*. Available at: https://www.formula1.com/en/latest/article/drivers-teams-cars-circuits-and-more-everything-you-need-to-know-about.7iQfL3Rivf1comzdqV5jwc. (Accessed: 04 November 2024).
- Formula 1 - The Official F1 Website (2023). *F1 - The Official Home of Formula 1 Racing*. Available at: https://www.formula1.com/en/racing/2023. (Accessed: 04 November 2024).
- Formula 1 - The Official F1 Website. (2023). Verstappen overcomes wet-weather chaos to take Dutch GP win. Available at: https://www.formula1.com/en/latest/article/verstappen-overcomes-wet-weather-chaos-to-make-it-a-hat-trick-of-dutchgp.4VJ0ULOqjodSSN1zC6kWui. (Accessed: 04 November 2024).
- Overdijk, M. (2022). Open Source Formula 1 Database, Release v2024.21.0. Retrieved 25 October 2024. Available at: https://github.com/f1db/f1db/releases.
- Pirelli (2024). *Pirelli Global: Discover our world.* Available at: https://www.pirelli.com/global/en-ww/homepage/.
- Pirelli (2019). F1 Tires. Available at: https://www.pirelli.com/tires/en-us/motorsport/f1/tires. (Accessed: 04 November 2024).
- Seymour, M. (2023). F1 Tyres explained: the Beginner's Guide to Formula 1 Tyres. Available at: https://www.formula1.com/en/latest/article/the-beginners-guide-to-formula-1-tyres.61SvF0Kfg29UR2SPhakDqd. (Accessed: 04 November 2024).
- Stuart, G. (2021). 10 things you need to know about the all-new 2022 F1 car. Available at: https://www.formula1.com/en/latest/article/10-things-you-need-to-know-about-the-all-new-2022-f1-car.4OLg8DrXyzHzdoGrbqp6ye. (Accessed: 04 November 2024).
- Vaibhav. 03 2024. , 1st version. F1 Weather Dataset (2018-2023). Retrieved 25 October 2024. Available at: https://www.kaggle.com/datasets/quantumkaze/f1-weather-dataset-2018-2023.

APPENDIX

Our analysis was carried out exclusively using Python 3.10.12 and Google Colab apart from Figure 3. which was made using R version 4.2.1. Excel was used for manual data collection and LaTeX was used for table creation. Python code is submitted as a separate file to this report, while R code is shown in Figure A1.

```r
#packages
require(tidyverse)
require(ggmosaic)
install.packages("readxl")
library(readxl)

#set working directory and load the data
data1 <- read_excel("Final data.xlsx")

#mosaic plot
p<- ggplot(data = data1) + geom_mosaic(aes(x=product(pitStops ,circuitType),fill=pitStops))+
  labs(x="Track type", y = "Number of pit stops", title = "Mosaic plot of track type and number of pitstops")
p
```

Figure A1. R-Studio code used for this analysis.

Table A1. Driver numbers, names and teams in 2022 and 2023 seasons.

Number	Driver	Team in 2022	Team in 2023
1	Max Verstappen	Red Bull Racing	Red Bull Racing
2	Logan Sargent	Williams Racing	N/A
3	Daniel Riccardo	McLaren Racing	Scuderia AlphaTauri
4	Lando Norris	McLaren Racing	McLaren Racing
5	Sebastian Vettel	Aston Martin	N/A
6	Nicholas Latifi	Williams Racing	N/A
10	Pierre Gasly	Alpine	Alpine
11	Sergio Perez	Red Bull Racing	Red Bull Racing
14	Fernando Alonso	Alpine	Aston Martin
16	Charles Leclerc	Scuderia Ferrari	Scuderia Ferrari
18	Lance Stroll	Aston Martin	Aston Martin
20	Kevin Mangnussen	Haas F1 Team	Haas F1 Team
21	Nyck de Vries	Williams Racing	N/A
22	Yuki Tsunoda	Scuderia AlphaTauri	Scuderia AlphaTauri
23	Alexander Albon	Williams Racing	Williams Racing
24	Zhou Guanyu	Alfa Romeo	Alfa Romeo
27	Nico Hulkenberg	Aston Martin	Haas F1 Team
31	Esteban Ocon	Alpine	Alpine
40	Liam Lawson	N/A	Scuderia AlphaTauri
44	Lewis Hamilton	Mercedes AMG Petronas	Mercedes AMG Petronas
45	Nyck de Vries	N/A	Scuderia AlphaTauri
47	Mick Schumacher	Hass F1 Team	N/A
55	Carlos Sainz	Scuderia Ferrari	Scuderia Ferrari
63	George Russell	Mercedes AMG Petronas	Mercedes AMG Petronas
77	Valteri Bottas	Alfa Romeo	Alfa Romeo