Predictive Algorithms using Machine Learning for Formula 1 Races

Andrés Eduardo Nowak de Anda
Tecnologico de Monterrey
Guadalajara, México
a01638430@tec.mx

Ulises Venegas Gómez
Tecnologico de Monterrey
Guadalajara, México
a01637321@tec.mx

Sebastián Márquez Álvarez
Tecnologico de Monterrey
Guadalajara, México
a01632483@tec.mx

David Alejandro Velázquez Valdez
Tecnologico de Monterrey
Guadalajara, México
a01632648@tec.mx

Gerardo Novelo De Anda
Tecnologico de Monterrey
Guadalajara, México
a01638691@tec.mx

Esteban Sánchez García
Tecnologico de Monterrey
Guadalajara, México
a01638691@tec.mx

***Abstract*— This document discusses the use of cloud computing technology, specifically Oracle Cloud, in Formula 1 racing to develop a predictive model for race outcomes. We highlight the need for a new dataset due to inaccuracies and missing data in the current dataset, making sure to include relevant data to improve the accuracy of the predictions, also outlining the methodology, including data analysis and model development using Multiple Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) models. Presenting the expected contributions, such as an enhanced predictive model, evaluation metrics, and practical applications for teams. The results and analysis demonstrate the performance of the MLP and LSTM models using metrics like accuracy, recall, and confusion matrix. All of this with the objective of verifying if an MLP model with hystorical data has better predictions than an MLP with no historical data. The document concludes with a discussion on the potential of machine learning techniques in predicting Formula 1 race outcomes and suggests areas for further research and improvement.**

***Index terms*—MLP, LSTM, Formula 1, Historic data**

I. Antecedents

In 2021, Formula 1 racing was shaken thanks to new technology that would change racing forever, *Cloud Computing*. Initial attempts used AWS[1] and later used other tools such as Oracle Cloud. Beginning with Oracle and Team Red Bull's Racing partnership in 2021, Oracle has been able to model races in real time to help divers make split-second choices mid-race; it is also used to model strategies for future races using a prediction tool that uses previous race data and machine learning to predict a winner and strategy that can be used to ensure a victory.[2]

Currently, however, Oracle does offer a demonstration of this service for users who would like to use their cloud computing service, they also offer some insight and documentation on how it works so that new users can learn how to equip this technology. With their code and dataset being publicly available, it is possible to study and analyze their current model for predicting a winner of a race. After a careful analysis of the data used and the methods employed, a conclusion was reached: a new version or variation of the model needed to be made to more accurately predict a winner and to keep up with the incoming data demand and requests by users.

II. Understanding the problem

Oracle and Red Bull Racing have formed into a new team in Formula 1, granting Oracle access to racing info such as real-time updates on the track and historical archived data of previous races, intending to use this information predicting machine that analyzes the historical data and then finds the relation between information to make an informed prediction. It may appear, however, that some of the data used in their model are either inaccurate or missing, and some of the data that is used is irrelevant and unnecessary for the prediction being made, causing the machine always to predict the same safe answers. Not only is the data set used by the model jeopardized but the learning model itself can also be modified and switched for another learning model better suited for this type of problem. The current method used by Oracle only analyses the previous most recent race to predict

a winner, and thus would falsely claim that if someone starts in first place they are almost guaranteed to end in first place.

Here then lies an opportunity to propose a better model for predicting a winner as well as expanding and refining the data set that the better model would use. To achieve this, it is necessary to first update the dataset with new data, fix the erroneous data points, and use these new and revised points as learning material for the new learning model. It should be paramount then that in approaching this issue, we strive to prove that a historical-based learning model is better at predicting a race result rather than an individual race learning model.

III. Justification

This paper, and the learning model documented, are intended to help strengthen Oracle as a company and Formula 1 as an organization. It is also, however, made to understand the ways historical data can be used in Machine Learning to help predict historical patterns, such as races in this case, and see how it can be applied to bigger matters. By focusing on the mathematical correlation between the past and the present, the team is trying to prove how significant a different learning model can be when they not only use the most recent data but also compare it to later data that in many fields is believed to be of no interest. This paper aims to seek how important past data truly is.

IV. Expected contributions

The research project on Formula 1 Races Prediction using Machine Learning has significant expected contributions to the field of predictive analytics in the realm of high-performance racing. The primary focus of this work is on developing a novel model that can accurately predict the outcomes of Formula 1 races. The expected contributions to this project are as follows:

1) *Selecting data:*

The newly selected data was extracted and compiled from a set of tables found in Kaggle [3] that contains information on every race in Formula 1 racing spanning from the 1950s to 2023. The file downloaded is a folder filled with up-to-date information on each race from information on the racers, the constructors, the actual races themselves, and the stops made during the race. About 9 CSV files exist in this folder, and each holds different relevant information.[3] Thus a merge query needed to be made to unite the datasets. The datasets were connected using Python and were connected based on relational data that is shared between these tables, such as “driverId”, “constructorId”, “raceId”, and so on.

Afterward, what we believed to be paramount information was weather data, more specifically rain and temperature data, which are factors that drivers consider important during a race as they could impact the speed at which they travel. The dataset from Kaggle did not provide this information, necessitating the use of an external dataset or look-up tool that could find the weather in every location of a race on the date it happened and the hour it occurred. Many tools were considered but the choice ultimately landed on Open-meteo, an open-source meteorology API that has free queries and archived meteorological data. It carries weather information from around the globe, and the historical data dates back to the 1940s. A function to query the weather per every single row in the relational table was made, by using the latitude and longitude of each race, the date, and the time of day the race took place. A loop was made to go through each row, and then it was parsed to a URL and query to extract a JSON file, from which the program would compare the weather, rain, and other factors at the time of day the race took place in.

2) *Enhanced predictive model::* Through extensive data analysis, feature selection, and the utilization of advanced machine learning techniques[4], a highly refined predictive model for Formula 1 races has been developed. By using historical race data, including the constructor and driver information, race-specific variables, and weather conditions, the model aims to capture the underlying patterns and dependencies that impact race outcomes.

B. Updated and comprehensive dataset:

Correlation tests were needed to ensure which values were of worth for the predictor. With over 50 types of values, not all of them had relevance to the chances of a driver winning, and cutting some values out of the predictive model made it easier for the machine to train, as it didn’t need to focus on learning unnecessary values. Even though around 20 values were left for the sake of the user interface and program, about 12 values had an important relevance to the training model in determining a victor per race using historical data. The data used to train the model is as follows:

- For Drivers:
  - driverId
  - driverWins
  - driverPosition
  - qualifyingPosition
  - driverConfidence (**2 years prior to present race** (did not finish),
    $$\text{confidence} = \frac{\text{driverDnf}}{\text{driverAllRaces}} \quad (1)$$
)
  - driverPoints
  - Average pit stop duration (**Seconds**)
- For Constructors:

- constructorId
- constructorPoints
- constructorWins
- constructorReliability (**2 years prior to present race** (did not finish),

$$\text{reliability} = \frac{\text{constructorDnf}}{\text{constructorAllRaces}} \quad (2)$$

)

- Race and track factors:
  - Rainfall (**true or false**)

As these were the factors that had the most correlation in racing, these were carefully selected to help train the Multi-Layer Perceptron.

C. *Evaluation metrics and analysis:*

Several metrics, including f1 score, recall, accuracy, and confusion matrix, have been utilized to evaluate the performance of the predictive model. These metrics provide a comprehensive assessment of the model's effectiveness in predicting race positions, which showcases the potential of the model as a valuable tool for teams, strategists, and enthusiasts in the high-performance racing domain.

D. *Practical applications:*

If Oracle wishes to stay competitive and Formula 1 wishes to remain seen as an organization that only works with the best of the best, there is a reason then to believe that a more accurate predictor would benefit the brands of both companies. Beyond the corporate scope, the chance to create a model that is more detailed and accurate is a very worthy experiment that would push machine learning to further heights, as a new step into understanding how these tools can improve our everyday life, by teaching it history of bigger world events, future learning models will be able to predict even more serious events.

This tool will also be more useful for the racers and teams competing once it is also implemented with their real-time analysis system, helping them detect previous patterns during an ongoing race.

V. Methodology

The methodology employed in this study aimed to develop a predictive model for Formula 1 races using machine learning techniques. The key steps undertaken in the research include data analysis, model development, evaluation, and result interpretation. This provides a framework for future research in the field of high-performance racing prediction and offers practical applications for teams and stakeholders in the Formula 1 ecosystem.

VI. Data analysis

The data analysis phase played a crucial role in understanding the underlying patterns and dependencies in Formula 1 races. An extensive historical dataset was collected and processed, incorporating various sources and variables to provide a comprehensive view of race outcomes.

The dataset consisted of constructor and driver information, including performance metrics such as constructor points, driver points, wins, and reliability. Race-specific variables, such as circuit characteristics, lap times, pit duration, and qualifying position, were also included. Additionally, historical weather conditions, such as temperature, humidity, and precipitation, were incorporated into the dataset to account for their potential impact on race outcomes.

During the data analysis phase, exploratory data analysis techniques were employed to gain insights into the dataset's characteristics. Statistical measures, visualizations, and correlation analyses were utilized to identify relationships between variables and uncover potential predictors of race outcomes.

Feature selection played a crucial role in identifying the most relevant variables for the predictive models. Rigorous analysis, including feature importance ranking and correlation analysis, guided the selection process. Variables that exhibited a strong correlation with race positions and had a significant impact on the predictive power of the models were included in the final feature set as shown in Figure 2 and Figure 1. The correlation in the spearman Figure 2 method were stronger than with pearson, so this explained to us that the problem is going to be a non-linear problem instead of a linear problem.

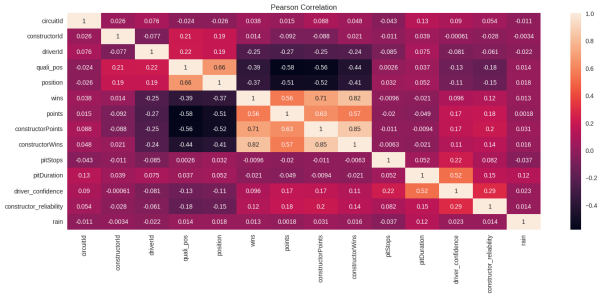

Figure 1: Pearson correlation analysis of the important variables in the dataset

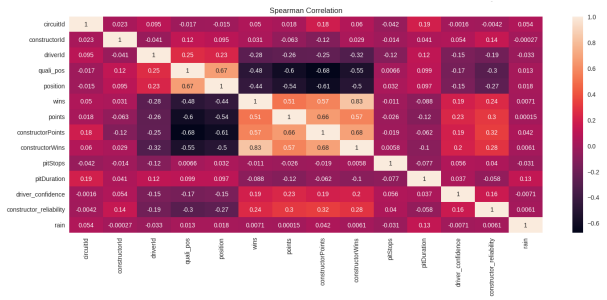

Figure 2: Spearman correlation analysis of the important variables in the dataset

Furthermore, data preprocessing techniques were applied to ensure data quality and integrity. Steps such as data cleaning, missing value imputation, and normalization were performed to create a standardized and consistent dataset for model training and evaluation.

The data analysis phase provided valuable insights into the relationships between various factors and race outcomes. It helped identify the key variables that contribute to the prediction of Formula 1 race positions, guiding the subsequent steps of model development and training. Like in this box plots Figure 3 and Figure 4, the Figure 3 box plot shows that the **mayority of changes in position** are in the last 10 to 24 position in a race. And if we reduce it to only 3 groups of poisition (Figure 4), this shows us that the **last 10 positions in a race don't really move to the first 10 positions** in a race, only some outliers escape from the last 10 position in a race to the first 10 positions.

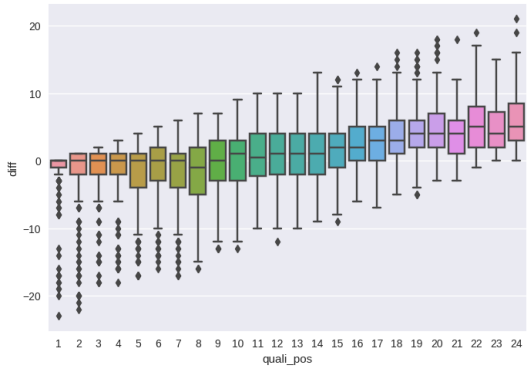

Figure 3: Box plot of the difference between qualifying position and final position using the **24 positions** that can exist in a race. Where **x** is the qualifying positions, and **y** is the difference between qualifying and final position.

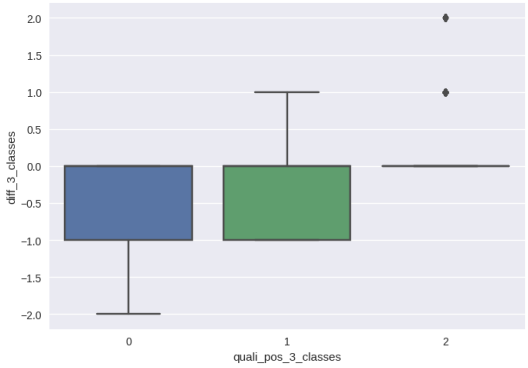

Figure 4: Box plot of the difference between qualifying position and final position reducing the **24 positions** in a race to only **3 groups of positions** (1 to 3 position is the 1st group, 4 to 9 position are the 2nd group, and from the 10 position to the last position is the 3rd group). Where **x** is the qualifying positions, and **y** is the difference between qualifying and final position.

By conducting comprehensive data analysis, our study was able to lay the foundation for the development of accurate and reliable predictive models for Formula 1 races. The insights gained from this analysis informed the feature selection process, ensuring that the models were built upon relevant and impactful variables.

VII. Models

Two primary models, Multiple Layer Perceptron (MLP), and Long Short-Term Memory (LSTM) were used for race prediction.

A. Development, Training, and Testing:

Three primary models such as one LSTM and two MLP have been employed for race prediction. The LSTM model is known for its ability to capture temporal dependencies and leverage historical race data. In the cas of MLP models, on the other hand, one of them contains variables involving past data and the other only has varibles involving present data, allowing to have a comparative analysis between the three models. All models undergo training using the processed dataset, with appropriate hyperparameter tuning and optimization techniques applied to achieve optimal performance.

To evaluate the effectiveness of the predictive models, a separate test dataset has been prepared. This test dataset contains race instances that were not used during the training phase. The models are then evaluated on this test dataset using performance metrics. The evaluation results provide insights into the models' performance on unseen data and help identify areas for further improvement.

B. MLP

Two separate MLP models were developed. The first model focused on variables involving past data, while the second model incorporated variables involving only present data. This approach allowed for a comparative analysis of the models' performance and provided valuable insights into the importance of historical data in the outcome of the final positions. Figure 5 and Table 1 shows how the MLP model is structured independent of the variables used and the hyperparameters Table 2 used for its optimal functionality.

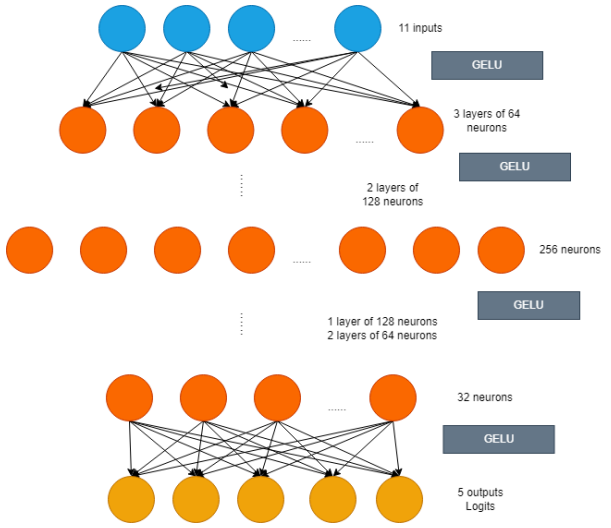

Figure 5: MLP model structure

	Layers	Neurons	Activation function
value	10	64, 64, 64, 128, 256, 128, 64, 64, 32	GELU

Table 1: MLP Design

	Optimization	Learning rate	Criterion	Batch size	l2 lambda
value	AdamW	1e-3	Cross entropy loss + l2 regularization	32	0.001

Table 2: MLP Hyperparameters

C. LSTM

The LSTM model, known for its ability to capture temporal dependencies, was employed to leverage the historical race data. By considering the sequence of past race results and other relevant variables, the LSTM model aimed to capture complex patterns and relationships in the data. The Figure 6 and Table 3 show how the LSTM model is structured and the hyperparameters (Table 4) used for its optimal functionality.

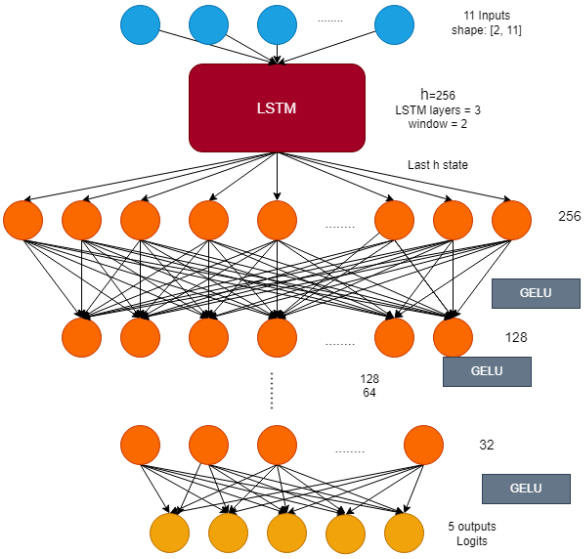

Figure 6: LSTM model structure

	Layers	Neurons or h states	Dropout	Activation function	Window
lstm	3	256	0.2	RELU	2
MLP	5	256, 128, 128, 64, 32	0	GELU	

Table 3: LSTM Design

	Optimization	Learning rate	Criterion	Batch size	l2 lambda
value	AdamW	1e-3	Cross entropy loss + l2 regularization	64	0.001

Table 4: LSTM Hyperparameters

VIII. Results and Analysis (5 classes)

The developed models were evaluated using a separate test dataset containing race instances that were not part of the training data. Performance metrics, including f1 score, recall, accuracy, and confusion matrix, were computed to assess the models' predictive capabilities.

- **Metrics used:**
  - loss: Cross Entropy Loss
  - Precision metrics: accuracy, recall, f1
  - Graph: confusion matrix

Result to compare (difference between qualification position and position). The model needs to have an accuracy of at least **46% to even be considered a reliable predictive model**.

	Percentage of difference (qualifying position - final position)
0 diff	0.465205
1 diff	0.239838
-1 diff	0.133645
2 diff	0.068525
-2 diff	0.048095

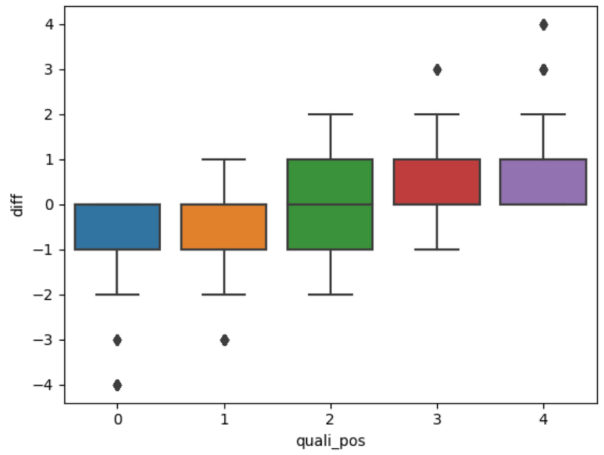

Figure 7: Box plot of the difference between qualifying position and final position, reducing the **24 positions** in a race to only **5 groups of positions** (1 to 3 position is the 1st group, 4 to 7 position are the 2nd group, 8 to 11 position are the 3rd group, 12 to 15 position is the 4th group, and 16 to the last position is the 5th group). Where **x** is the qualifying positions, and **y** is the difference between qualifying and final position.

Figure 7 shows a box plot where y is the difference between the qualifying position and the final position in the race. And x is the qualifying position. Position and qualifying position are grouped into 5 classes instead of the 20 to 24 in a race.

Table 5 demonstrate the results of the three models, showing the f1 score, recall, accuracy, and confusion matrix.

	MLP present Results	**MLP Historic Results**	**LSTM Results**
f1	0.3788	0.4651	0.4816
recall	0.3851	0.4730	0.4851
accuracy	0.3851	0.4730	0.4851

Table 5: Results of the three models

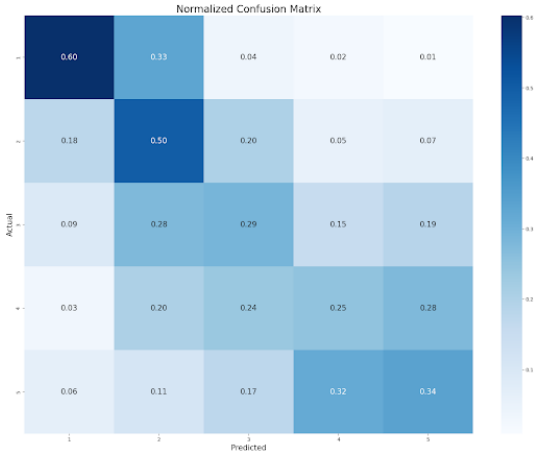

Figure 8: Result of the MLP with only present data (driverId, constructorId, qualifying position, rain).

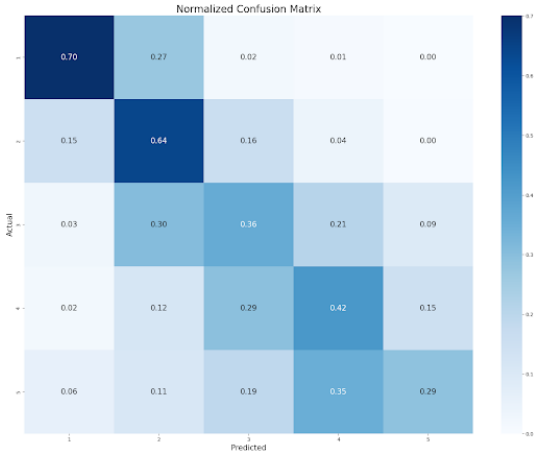

Figure 9: Result of the MLP with historic data.

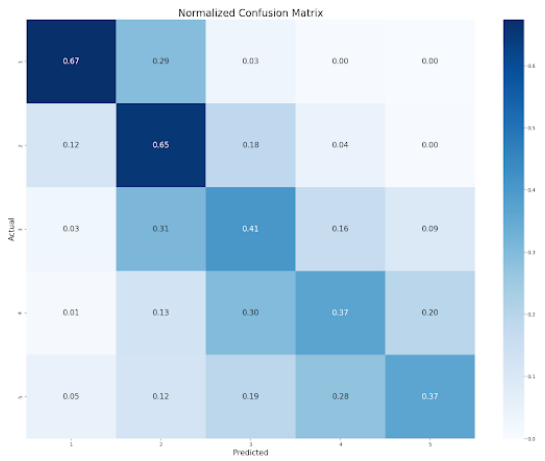

Figure 10: Result of the LSTM model with historic data and using a window of size 2.

First, we can see in the Table 5 and the Figure 8 vs. Figure 9, that a **MLP model** with historic data can better predict the final position in a race than a **MLP model** with only present data, and this confirms our hypothesis, that a **MLP model** with historic data can predict the race results better than an **MLP model** with only present data. And, for the **LSTM model** Figure 10 results Table 5, we can see that is only **1%** better than the MLP with historic data. So this tells us, that historic data helps better predict a final position in a race, but it seems, at least for now, that there isn't a pattern with past races that can help us predict better the final position in the f1 race.

IX. Discussion

The results obtained from the evaluation of the models indicate the potential of machine learning techniques in accurately predicting Formula 1 race outcomes. The inclusion of various historical variables such as constructor and driver information, race-specific data, and weather conditions contributed to the model's predictive power.

Although proven successful, further research, and refinement are required to enhance their performance. Noteworthy future additions include exploring additional variables, incorporating real-time data during races, and researching, as well as developing, more advanced machine learning algorithms.

This study demonstrates the feasibility and effectiveness of using machine learning models, specifically **MLP and LSTM**, to predict Formula 1 race outcomes through past historical data. The findings provide valuable insights for teams, strategists, and stakeholders in the high-performance racing domain, offering opportunities for improved race strategies and decision-making.

X. Conclusion

Analyzing the result that we got from this study, it has shown us that to create a more efficient predictor it would be needed a dataset that contains relevant data that is constantly being updated to include new information to improve the accuracy of this predictor, so in the future, we could look for a way to implement an improved version of the current prediction model by continuously improving the dataset and by using one of the models, like MLP or LSTM, we could pave the way of the usage of predictive analytics AI for Formula 1.

XI. Bibliography

References

[1] R. MIller, "How cloud data-crunching power accelerates the f1 racing experience," from: https://www.datacenterfrontier.com/cloud/article/11427867/how-cloud-data-crunching-power-accelerates-the-f1-racing-experience, 2021.

[2] T. Necomws, "How f1's red bull racing uses simulations to make mid-race calls," from: https://www.popularmechanics.com/cars/a40747762/f1-cloud-computing-oracle-red-bull-racing/, 2023.

[3] Vopani, *Formula 1 World Championship (1950 - 2023)*, (2023), from: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020.

[4] G. Greasley Andrew; Panchal, and A. Samvedi, "The use of simulation with machine learning and optimization for a digital twin- a case on formula 1 dss," *Wsc*, vol. 22, pp. 2198–2209, 2022.