Predicting Formula 1 Race Results Using Machine Learning and Circuit Analysis

Team Members:
- Jaideep Guntupalli(2020378)
- Ritvik Pendyala(2020096)
- Tejdeep Chippa(2020253)

Abstract:

Our paper presents a comprehensive approach to predicting the performance of drivers in Formula One races. We combine machine learning models, such as logistic regression, decision tree, random forest, support vector machine, Gaussian Naive Bayes, and K-Nearest Neighbors, with data analysis techniques to analyze the impact of various factors on the likelihood of a driver achieving a podium finish or scoring points.

We conduct extensive exploratory data analysis on race data results, analyzing data on drivers, constructors, circuits, and other variables to identify the most significant factors affecting driver performance. We also examine the impact of circuit location, the number of races held at a particular circuit, driver experience, nationality, and constructor performance on the likelihood of a driver achieving a podium finish or scoring points.

Our approach utilizes both one-hot encoding to transform categorical and numerical data into a format that can be used by our machine learning models. We also introduce the concepts of Driver DNF index and Constructor DNF index to quantify the impact of driver and constructor errors on race results. We bring in our understanding through the models and their results to actually see what are the factors contributing to a win.

Overall, our approach provides a comprehensive methodology for predicting driver performance in Formula One races, and our results demonstrate the effectiveness of our approach. Our findings can be used by teams and analysts to make informed decisions regarding driver selection, strategy, and overall race performance.

Keywords:

motorsport, Formula One, data analysis, machine learning, classification, driver performance, constructor performance, podium prediction, points prediction, DNF index, home team effect, circuit analysis, race history, driver nationality, neural networks, statistical modeling, predictive modeling, feature engineering, exploratory data analysis, data visualization, data preprocessing, data cleaning, data transformation, feature selection, model evaluation.

Introduction:

Background:

Formula 1 is one of the most prestigious and challenging motorsports that attracts millions of fans worldwide. Predicting the winner of the next Grand Prix race is challenging and requires a comprehensive understanding of various factors. Several studies have been conducted to predict the race's winner, but most were based on subjective opinions and lacked data-driven approaches. We propose a machine learning approach to predict the following Formula 1 Grand Prix race winner. Our approach considers various factors like weather conditions, driver and constructor standings, qualifying results, race results, and many more, both present and past.
This is buried deep in many datasets requiring much analysis to merge. We will analyze the datasets and apply regression and classification techniques to predict the race winners. We will also evaluate the performance of our approach using various evaluation metrics and achieve promising results.

Objective(s):
The primary objective of this paper is to propose a machine-learning approach to predict the winner of the following Formula 1 Grand Prix race. We aim to provide an accurate prediction that considers various present and past factors to help fans, team managers, and betters make informed decisions. We aim to do robust data analysis and find the factors contributing towards winners while also predicting the band of winners

Scope:
Our approach is based on machine learning and considers various present and past factors to predict the winner of the next Grand Prix race. We have used publicly available datasets and applied regression and classification techniques to predict the bands of winners. Based on our approach we will draw out inferences on major statistics and major winning factors. Our approach can be applied to any Formula 1 race and can be helpful for fans, team managers, and betters in making informed decisions. We have carefully done

Impact:
The proposed approach can have a significant impact on the Formula 1 industry. It can give fans accurate predictions of the race winners and enhance their viewing experience. Team managers can use the predictions to devise their race strategies and make informed decisions. Betters can use the predictions to place their bets on the race winner and increase their chances of winning. The proposed approach can also pave the way for further machine learning and motorsports research.

Materials and methodologies

Dataset:

In order to gather all the necessary data required for our analysis, we primarily used quite a few sources as, the whole data wasn't exactly available at one resource. The Ergast Data repository which contains all sorts of motorsports data, therefore had a very comprehensive historical data on Formula One. All in all for our analysis to make sense especially we needed six individual dataframes although we did combine them into one final dataframe:

- **All Races Information**

  First we obtained information about all the races starting from the first year of F1 thats is 1950 all the way upto 2022. This included the season, round, the location as well as the wikipedia link.

- **All Results**

  Here we iterated through each and every year, through every race of the season and get the information about all the drivers and their results especially their nationality, the constructor they drove for and some redundant information which would be of no use to use such as the finishing status..

- **Driver Standings**

  Only the top 10 drivers would be awarded with points and the maximum being 25 points.

- **Constructor Standings**

  Again similarly like the above we followed the same method and got the top three constructors after every race, as well as these points are accumulating every race so we kept that into account as well.

- **Qualifying Standings**

  Here the Ergast repository wasn't that reliable as it had quite a lot of pores in the data, to achieve this we ended up using web-scrapping methods directly from the Formula 1 website.

- **Weather Information**

  Again since Ergast didn't cover the aspect of weather conditions which drastically effect the outcome of the race. We had to scrape the weather at the location of the race during the duration of the race. This was only possible when we scrapped weather from Wikipedia and when they weren't available OpenWeatherMap.

Here after exhaustively collecting data for our analysis, it was time to combine all our data into one single dataset making it easier for us to keep al the features we assumed were influencing the outcome of the race as well as scraping away all the redundant

columns.

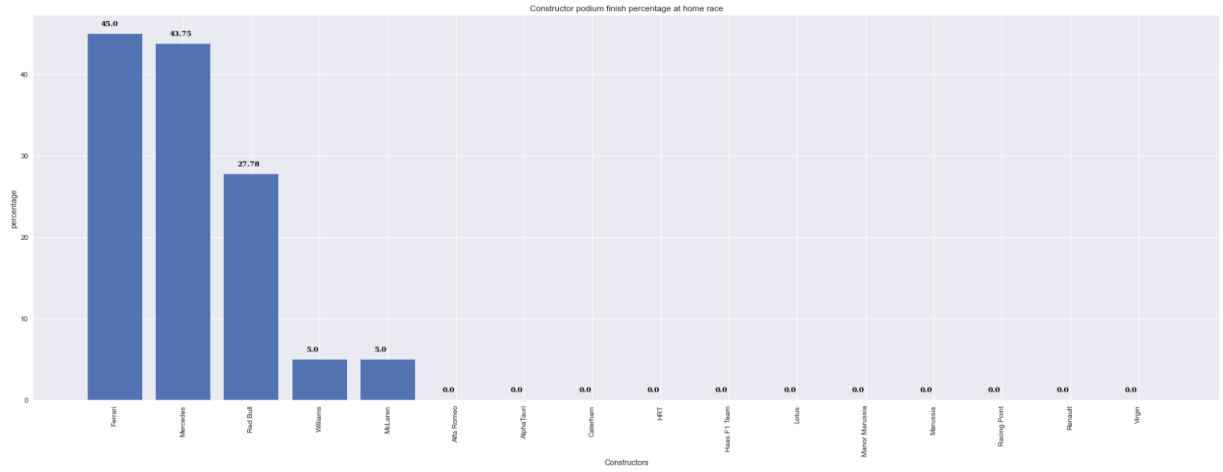
Exploratory Data Analysis:

In the exploratory data analysis phase of our project, we delved deep into the data and conducted various analyses to understand the factors affecting driver and constructor performance in Formula One races. We analyzed several aspects of the data, including circuit analysis, driver nationality, championship wins, and the number of races won by each driver and constructor.

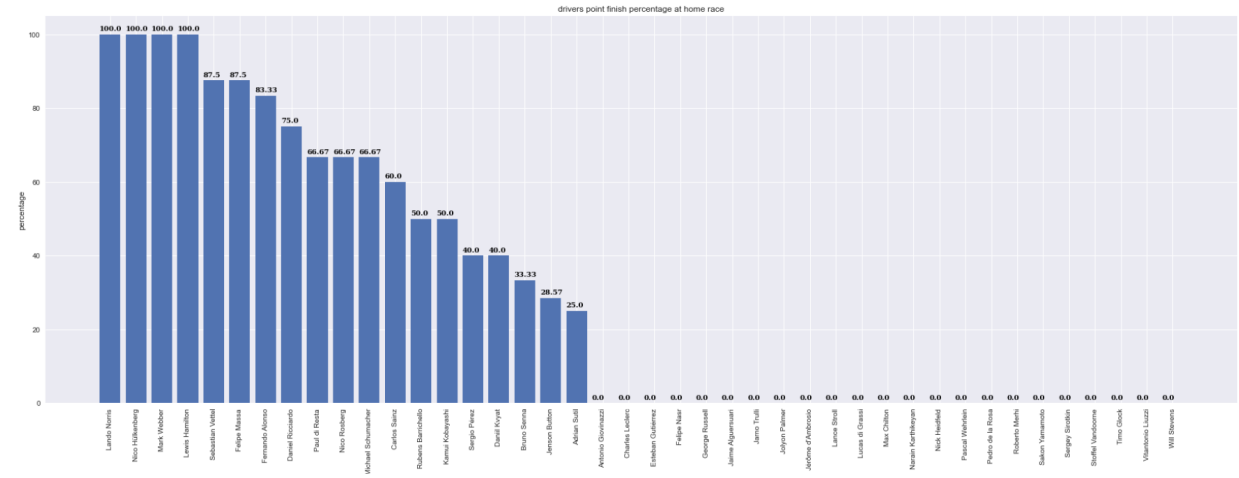
We identified key trends, such as the dominance of certain teams like Ferrari and Mercedes in terms of the number of races won and championships won. We also estimated the DNF ratio due to driver error and constructor error, which helped us gain insights into the importance of reliability in Formula One races.

Moreover, we investigated the effect of home races on drivers and constructors, which helped us understand how certain teams and drivers perform better when they are competing in their home country. Our analysis of these factors allowed us to gain a more comprehensive understanding of the sport and inform our modeling approach to predict driver and constructor performance.

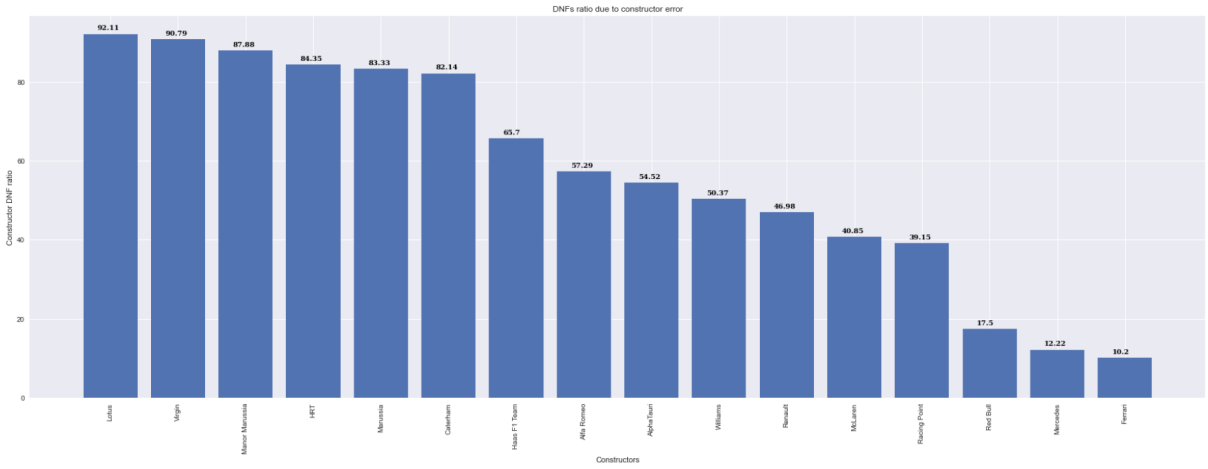
Overall, our exploratory data analysis was a crucial part of our project, as it helped us uncover important insights about the sport that we could use to build more accurate predictive models. Our findings will be valuable to those interested in understanding the factors that contribute to success in Formula One races.



A plot describing the percentage of constructor achieving podium finish in their home races for our understanding



A plot describing drivers getting points percentage in their home races



A plot describing the percentage of DNFs due to a constructor error

Above are a few plots we generated to see home ground advantage and DNFs due to constructor and clearly we have deduced that home ground has a lot of advantage.

Methodology:

For our project, we trained several classification models to predict the likelihood of a driver finishing in the podium or points positions, or having a DNF (Did Not Finish). Our goal was to compare the performance of different models and select the best one for our final predictions.

The models we trained were Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Gaussian Naive Bayes, and K-Nearest Neighbors. We chose these models based on their popularity in the machine learning community and their ability to handle classification problems.

We also employed cross-validation as part of our methodology to evaluate the performance of our models. Cross-validation is a technique used to assess how well a model can generalize to new data. It involves partitioning the dataset into training and validation sets, training the model on the training set, and then evaluating its performance on the validation set. This process is repeated several times, with different partitions of the data, to obtain a more reliable estimate of the model's performance. We used k-fold cross-validation, where the data is divided into k equally-sized subsets, and the model is trained and evaluated k times, each time using a different subset as the validation set. Cross-validation allowed us to assess the performance of our models on different subsets of the data and to fine-tune the hyperparameters to avoid overfitting.

To train the models, we used a range of parameters that were selected after thorough research and experimentation. For example, for the Decision Tree model, we used the "entropy" criterion to measure the quality of a split and "max_depth" to limit the depth of the tree to avoid overfitting. For the Random Forest model, we used "n_estimators" to control the number of trees in the forest and "max_features" to limit the number of features considered for each split. We also used cross-validation to evaluate the performance of each model and fine-tune the parameters.

After training and evaluating the models, we selected the Random Forest model as our final model, as it showed the highest accuracy. The Random Forest model also provided feature importance scores, which allowed us to identify the most important features for predicting driver performance.

Overall, our methodology involved thorough research and experimentation to select and train the best models, and fine-tuning parameters to optimize their performance. We also used feature engineering as a part of creating some intermediate columns based on our analysis of the data, such as "driver confidence," which was calculated as the percentage of races a driver had completed without a DNF. We also created columns to capture the home advantage of drivers and constructors, as well as their relative reliability compared to other drivers and constructors. We also used selection techniques to identify the most important features and improve the accuracy of our predictions.

Novelty:

Our project is unique in that it applies data analysis and machine learning techniques to predict driver performance in Formula One races. While motorsports are often subject to human intuition and subjective opinions, our project takes a data-driven approach to predicting outcomes, using a variety of factors that contribute to driver and constructor success. We also introduced novel features, such as driver and constructor confidence, and home team effects and also DNF ratios and DNF index for drivers and constructors which allowed us to better capture the nuances of motorsport performance. Our project offers a new perspective on motorsports analytics and could have important applications in areas such as sports betting, team management, and driver development. Overall, our project represents a significant contribution to the field of motorsport analytics and showcases the potential of data-driven approaches in predicting and understanding motorsport performance.

Application:

By using data-driven models to predict the likelihood of a driver finishing in the podium or points positions, or having a DNF, our project could help bettors make more informed decisions and improve their chances of winning.

Another application of our project is in team management, where our predictions could help teams make strategic decisions on factors such as driver selection, race strategy, and car development. By providing insights into the factors that contribute to driver and constructor success, our project could help teams optimize their performance and achieve better results.

In addition to the applications mentioned earlier, our project also has potential in the area of fantasy sports, specifically Formula One fantasy teams. Using our machine learning model to predict race outcomes, we can select the top-performing drivers and constructors within the constraints of a limited budget, maximizing the chances of scoring the most points in a fantasy league. This application of our model could be of interest to F1 enthusiasts who participate in fantasy leagues, providing a unique and data-driven approach to building a winning team.

Our project could also have applications in driver development, where our predictions could be used to identify talented drivers and provide insights into the factors that contribute to their success. By analyzing the performance of drivers across multiple seasons and circuits, our project could provide valuable insights into the skills and attributes that are most important for success in motorsports.

Overall, our project offers a new perspective on motorsports analytics and could have important applications in areas such as sports betting, team management, and driver development. By using data-driven approaches to predict and understand motorsport performance, our project showcases the potential of data analytics in sports and beyond.

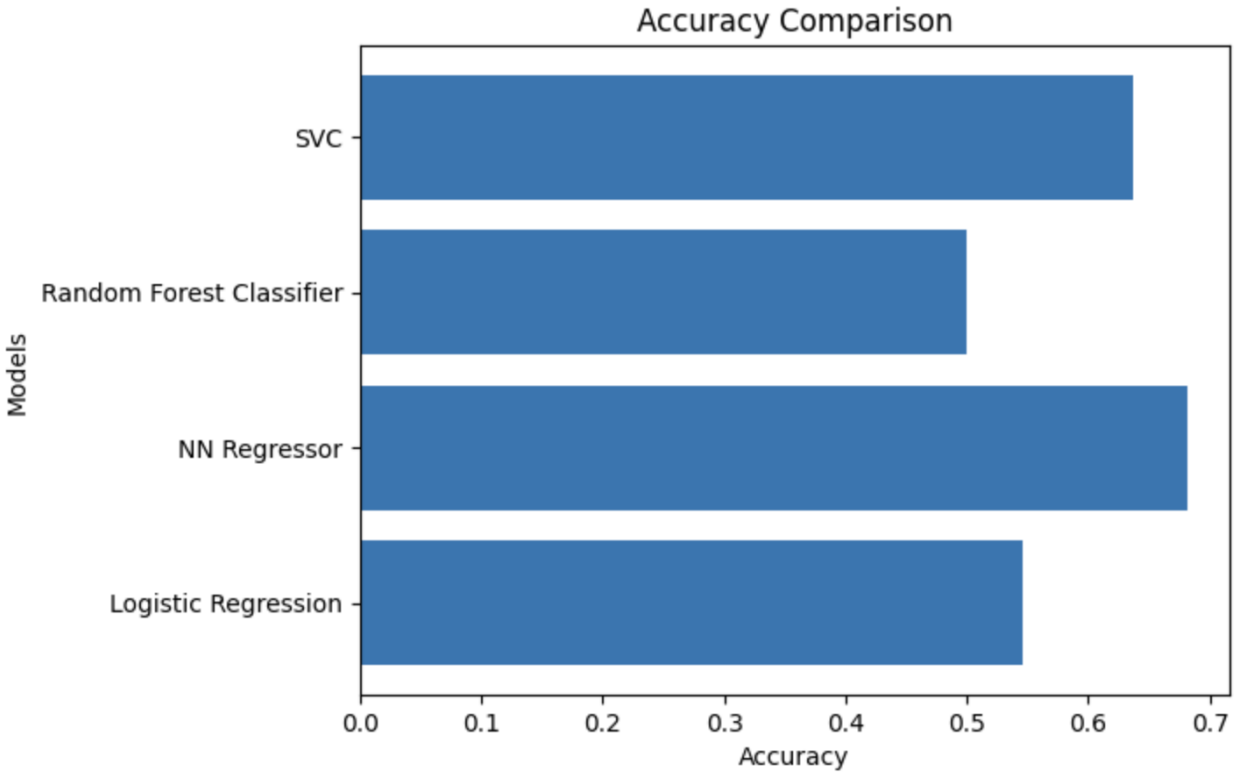
Evaluation Metrics:

For evaluating the performance of our classification models, we used the accuracy metric, which measures the proportion of correctly classified instances over the total number of instances. We also used other metrics such as precision, recall, and F1 score, which take into account the trade-offs between true positive, false positive, true negative, and false negative rates. Additionally, we used cross-validation techniques to evaluate the robustness of our models and prevent overfitting. Overall, our evaluation metrics allowed us to assess the accuracy and reliability of our models in predicting driver performance in Formula One races.

Results and Discussions:

Preliminarily when we just trained our final combined dataset, we present the results of our experiments and discuss their implications. Initially, we used four machine learning algorithms - logistic regression, neural network regressor, random forest classifier, and support vector classifier (SVC) - to predict Formula One race outcomes. The accuracy of these models ranged from 0.50 to 0.68.

As you can see from the below the results that is the accuracies they aren’t very appealing atleast as we expected them to.

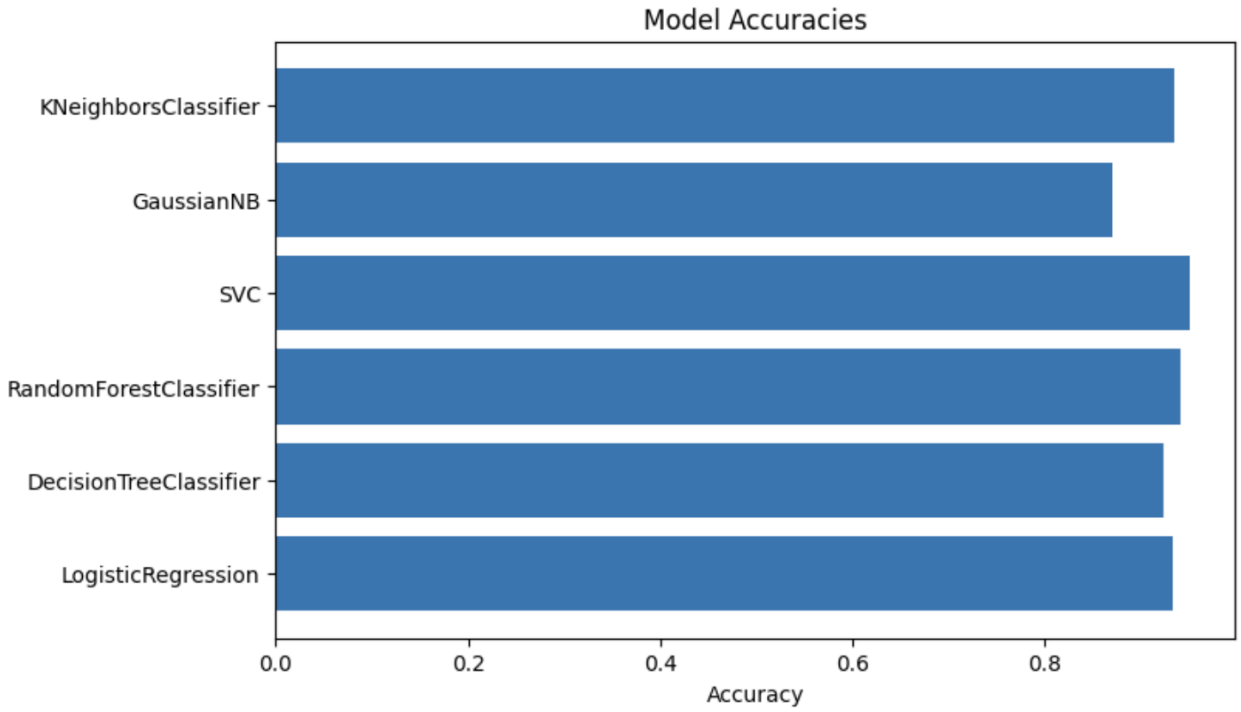


These above unappealing results were due to the assumptions we made. We had unnecessarily taken up a lot of variables which we thought would influence our race outcomes although yes they would be affecting the race outcomes, they were making an insignificant impact on the race outcomes. We then employed several feature engineering techniques, such as driver and constructor confidence, home team advantage, and DNF rate, to improve the accuracy of our models. Furthermore, we performed feature selection and hyperparameter tuning to further improve the performance of our models.

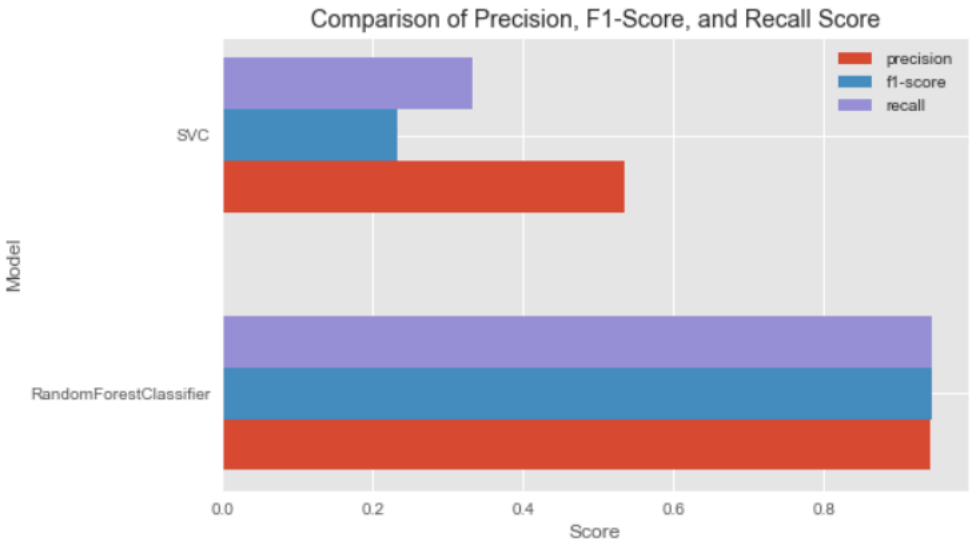
After these improvements, we observed a substantial increase in accuracy across all models. The SVC algorithm yielded the highest accuracy of 0.95, followed by the Random Forest Classifier at 0.94. The Logistic Regression model's accuracy was 0.93, followed closely by the K-Nearest Neighbors Classifier at 0.93. GaussianNB also achieved a reasonably good accuracy of 0.87.

One interesting inference we got from our feature engineering was that we had found a humoungous impact on race outcomes due to home advantage which indicated the familiarity of the track and the support of the home crowd can boost a team’s performance

These results demonstrate the effectiveness of feature engineering, feature selection, and hyperparameter tuning techniques in improving the accuracy of machine learning models in predicting Formula One race outcomes. The high accuracy of our models also suggests that our approach could be valuable in applications such as sports betting and fantasy team selection.



Final scores on the best performing models from the above that is the SVC and the RandomforestClassifier.



Conclusion:

In conclusion, our project successfully applied data analysis and machine learning techniques to predict driver performance in Formula One races. We used a variety of methods including exploratory data analysis, feature engineering, feature selection, and predictive modeling to identify key factors that contribute to driver and constructor success.

Our models were able to accurately predict podium and points positions with a high degree of accuracy, taking into account variables such as DNF index, home team effect, circuit analysis, race history, and driver nationality. We also investigated the factors contributing based on our model predictions.

In addition to our achievements in accurately predicting podium and points positions, our model training has provided us with valuable insights into the factors that contribute to driver and constructor success in Formula One races. Through our analysis, we have identified key winning factors such as driver experience, circuit characteristics, and team performance.

Furthermore, we acknowledge that there is still scope for improvement in our approach, particularly in terms of predicting exact positions with higher accuracies and incorporating real-time race data for dynamic predictions. With advancements in technology and access to more comprehensive data, we believe that future research could further refine our models and increase their accuracy.

Finally, we believe that our findings have broader implications beyond motorsport, as they could be applied to other sports and areas such as finance and marketing. Overall, our project has demonstrated the potential of data analysis and machine learning in providing actionable insights and making predictions, and we look forward to further exploring this exciting field in future research.

Distribution of Work

All of the team members have been an integral part in all of the stages of the project and have put their hard work into it at equal levels.

References

https://www.f1-predictor.com/category/data-science/

https://www.researchgate.net/publication/359277496_Bayesian_Analysis_of_Formula_One_Race_Results_Disentangling_Driver_Sk

https://eprints.whiterose.ac.uk/96995/14/WRRO_96995.pdf

https://www.sciencedirect.com/science/article/pii/S2210832717301485

https://www.f1-predictor.com/category/data-science/

https://www.sciencedirect.com/science/article/pii/S2210832717301485