

# Analyzing the Determinants of Success in Formula 1

Rody Bertolini

CSPB

CU Boulder

Denver, CO

[pabe9903@colorado.edu](mailto:pabe9903@colorado.edu)

## Motivation

Formula 1 has long been the “engineering racing series,” where technological innovation and clever design often overshadows the individual skill of the driver. The performance gap between constructors (the teams in F1, each fielding 2 cars) can make the car itself appear more important than the driver behind the wheel. Yet in the sport’s 70 year history there have been several drivers who had championship success across multiple teams. For instance, this season Max Verstappen has consistently placed on the podium while his teammate rarely even makes the top 10, indicating the car itself is flawed but Verstappen is able to overperform regardless. There is clearly a complex interaction between car performance, driver ability, and contextual race factors like the circuit type and weather that determine who will be successful in an F1 race.

This project seeks to apply data mining techniques such as preprocessing, association, classification and clustering to determine if there is one factor that most strongly influences a race. By analyzing decades worth of data this project uncovers patterns and relationships between driver statistics, car performance, and circuit characteristics.

The findings could provide insight into whether driver development or car development contributes more to constructors success in F1. These insights may benefit teams in their resource allocation, analysts and commentators when giving their predictions for a given race weekend, fantasy sports players who draft based on performance indicators, and gamblers who

seek to understand probability for profit. This project will use data mining techniques to extract meaningful knowledge from a complex domain, such as that presented by F1 data.

## Literature Survey

There has been some examination of Formula 1 using data-mining techniques. Sicoie (2022) developed a machine-learning framework employing supervised regression models to predict race winners and championship standings. Using the same Ergast API dataset, Sicoie performed preprocessing and feature extraction, achieving a strong correlation ( $\rho = 0.90$ ) between predicted rankings and actual results.

O’Hanlon (2022) also explored Formula 1 ranking prediction using linear regression and neural networks, focusing on the 2010–2021 seasons. O’Hanlon found that data preparation and feature engineering had a substantial effect on model performance.

These studies demonstrate that Formula 1 outcome prediction benefits from the application of data-mining and machine-learning methods. However, existing research primarily emphasizes predictive modeling rather than exploring which underlying attributes most strongly lead to success. This project expands on prior work by using association-rule mining and clustering to reveal relationships among drivers, constructors, and circuit characteristics, aiming to uncover the structural patterns behind consistent performance rather than simply forecasting results.

## References

- [1] Horatiu Sicoie. 2022. *Machine Learning Framework for Formula 1 Race Winner and Championship Standings Predictor*. Bachelor's Thesis, Tilburg University, Netherlands. Available at: <https://arno.uvt.nl/show.cgi?fid=157635>
- [2] Emma O'Hanlon. 2022. *Using Supervised Machine Learning to Predict the Final Rankings of the 2021 Formula One Championship*. National College of Ireland. Available at: <https://norma.ncirl.ie/6628/1/emmaohanlon.pdf>

## Proposed Work

This project will use basic data-mining methods to analyze the full historical Formula 1 dataset in order to identify the variables that have the greatest influence on race outcomes. The work will proceed in four primary phases: data collection (completed), preprocessing, pattern discovery, and evaluation.

### *Data Collection*

Data come from the Kaggle Formula 1 dataset, which covers all Formula 1 seasons from 1950 through 2024. These tables include races, drivers, constructors, qualifying results, race results, and pit stops. The data will be combined into a single relational schema to facilitate integration across all tables.

### *Preprocessing and Feature Engineering*

The dataset will undergo standard cleaning procedures, including the standardization of variable names and the conversion of categorical IDs into human-readable labels. Missing values for “Did Not Finish” (DNF) entries and similar cases will be resolved using the status table. Additional derived features will include attributes such as each driver's age at the time of a race and recent-form indicators, such as average points or grid position over the previous few events.

### *Pattern Discovery*

Association-rule mining will be applied to uncover high-lift relationships among key variables, including grid position, constructor performance, and final results.

Classification models will then be developed to predict whether a driver will finish in the points or achieve a podium result. This will be implemented using logistic regression and decision-tree classifiers for interpretability.

Clustering methods will group drivers and constructors based on performance profiles, revealing consistent patterns such as “qualifying specialists” or “race-pace performers.”

### *Evaluation and Comparison*

Each model will be tested through cross-validation and compared against baseline predictors. Evaluation metrics will include accuracy, F1 score, support, and confidence. Visualizations such as confusion matrices, association graphs, and cluster maps will accompany the analysis to highlight key findings.

This integrated approach combines both predictive and exploratory data-mining methods to provide interpretable insight into the interplay among driver skill, constructor performance, and circuit characteristics in Formula 1 racing.

## Dataset

This project uses the public Formula 1 World Championship dataset available on Kaggle (<https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020>) and updated through 2024. The dataset originates from the Ergast API and includes detailed records of every Formula 1 race from 1950 onward. It contains over 1,000 races, 800 drivers, and 200 constructors.

The data are divided into multiple tables:

**races.csv** – Information about each race, including circuit, date, and weather conditions when available.

**drivers.csv** – Biographical details for each driver, such as name, nationality, and date of

birth.

**constructors.csv** – Information about teams (constructors), including nationality and constructor name.

**results.csv** – Finishing positions, points, grid positions, and fastest laps.

**qualifying.csv** – Qualifying times and ranks.

**pit\_stops.csv** – Lap-by-lap pit stop counts and durations.

**circuits.csv** – Circuit details including location, length, and altitude.

**status.csv** – Encoded descriptions of race completion status (e.g., finished, DNF, gearbox failure).

## Evaluation Methods

Model performance will be assessed using standard data-mining metrics to ensure both accuracy and interpretability. Cross-validation will be used to reduce overfitting and verify model stability. Association-rule mining will be evaluated using support, confidence, and lift to measure the reliability and strength of discovered relationships. Clustering quality will be assessed using the silhouette coefficient and visual inspection to identify meaningful groupings based on performance profiles. Results will be compared to baseline models to gauge improvement. This approach should allow for the discovery of patterns that are both statistically valid and meaningful in explaining Formula 1 success.

## Tools

Analysis and modelling will be done via Python. Key libraries that will be utilized are Pandas, Numpy, scikit-learn, mixtend, matplotlib and seaborn.

## Milestones

From *October 13–26*, data integration and preprocessing will be completed, including cleaning, feature engineering, and preparing the dataset for analysis.

Between *October 27 and November 9*, classification models will be developed and refined.

From *November 10–23*, clustering and association-rule mining will be performed, along with preliminary visualizations and analysis.

Between *November 24 and December 5*, results will be finalized, figures and evaluation metrics prepared, and the final report and presentation completed for submission.