F1 Pitstop Champion Prediction

This repository contains a Python pipeline to predict Formula 1 season champions using pitstop and race data from 2018 to 2024. It covers data preprocessing, feature engineering, exploratory visualization, model training, and evaluation.

Project Overview

Predict which driver will become the Formula 1 World Champion next season by leveraging pitstop performance, weather, and race metrics. We train four classifiers (Logistic Regression, Random Forest, Naive Bayes, KNN) and compare their accuracies and F1 scores.

Pipeline Steps

1. Load & Encode Drivers

- Read CSV into pandas DataFrame
- Use `LabelEncoder` (via `pd.factorize`) to convert driver names into numeric `DriverCode`

2. Compute Season Points & Champions

- Map race finishing positions to FIA points (25 for 1st, 18 for 2nd, …)
- Sum points per driver per season
- Flag the driver with max points in each season as champion (`IsChampion`)

3. Clean & Drop Columns

- Remove unneeded text columns (race name, date, location, etc.)
- This keeps only numeric and categorical features relevant to modeling

4. Handle Missing Values

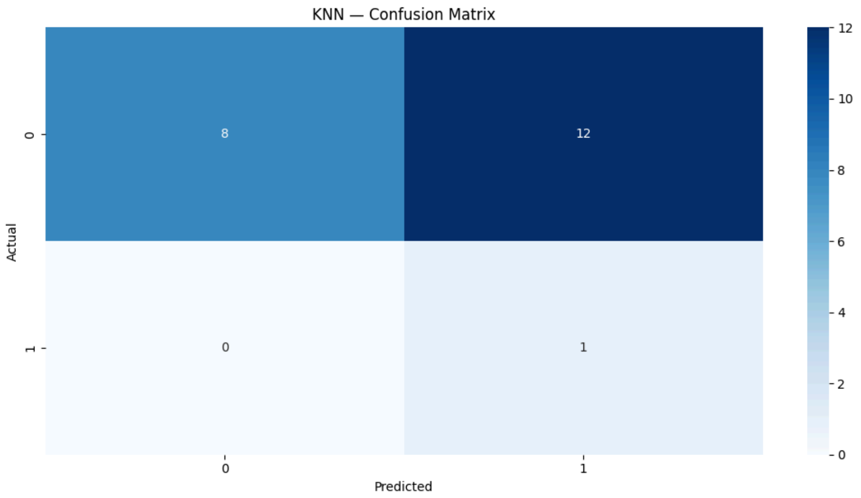
- Numeric columns: fill NAs with column medians
- Categorical columns: fill NAs with mode (most frequent value)

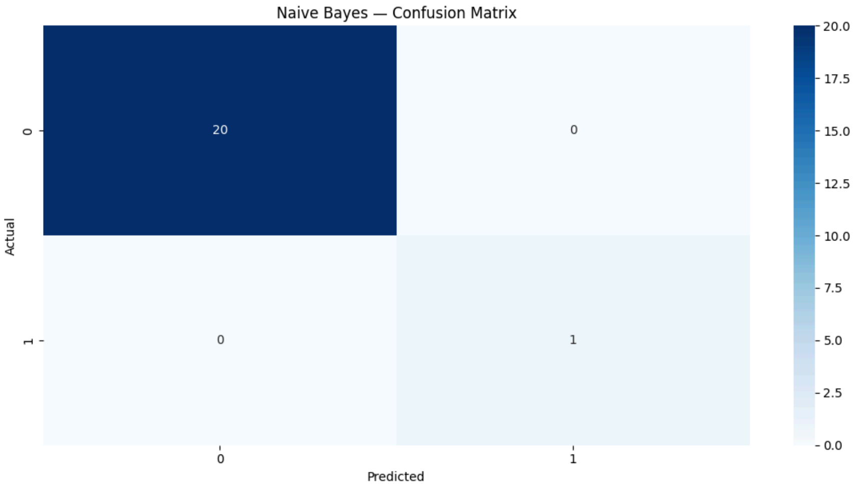
5. Aggregate Features

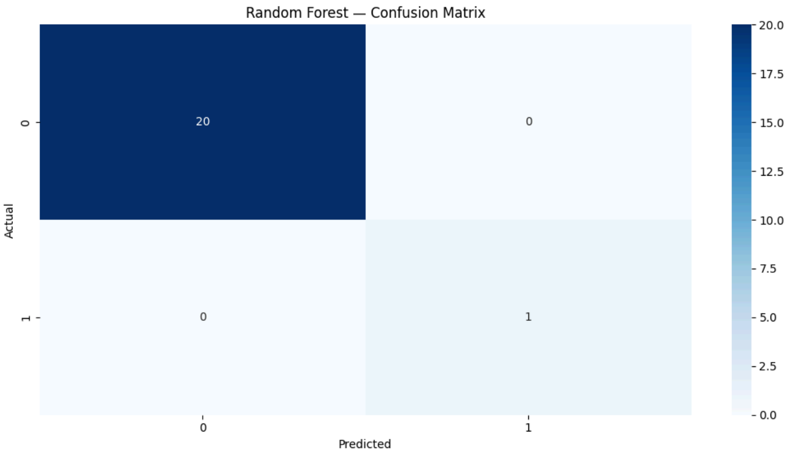
- Group by `Season` & `DriverCode` to produce one row per driver-season
- Compute:
    - **Numeric**: mean or sum for continuous metrics (avg pit stops, lap variation, total laps, etc.)
    - **Mode**: most common `Constructor` and `Tire Compound`
    - **Label**: max of `IsChampion` (1 if driver won that season)

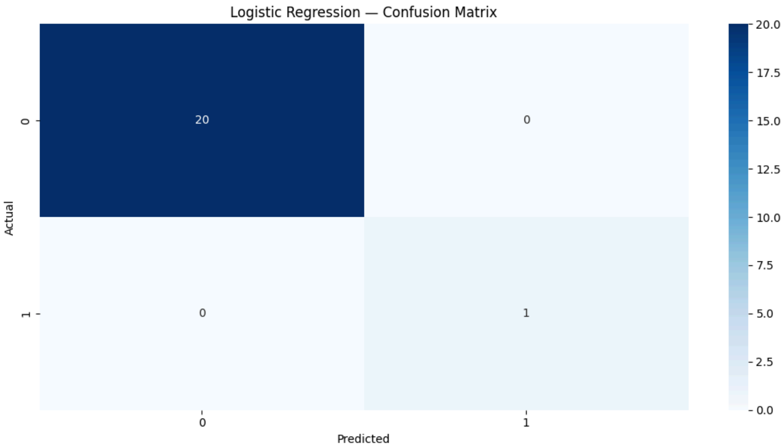
6. Exploratory Visualizations

- **Correlation Matrix**: shows inter-feature correlations

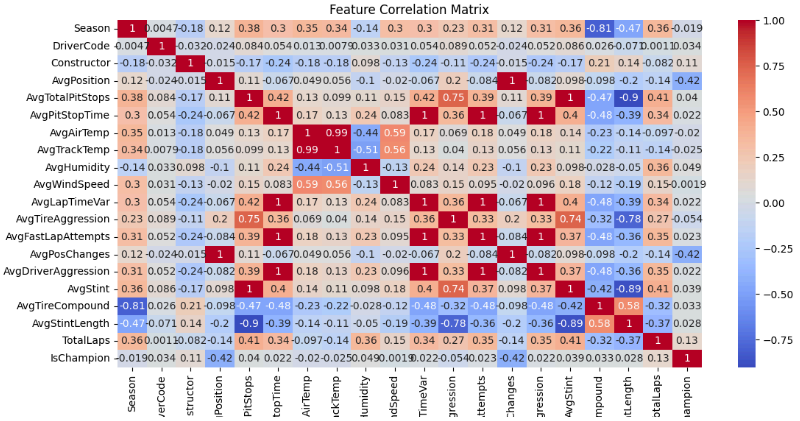








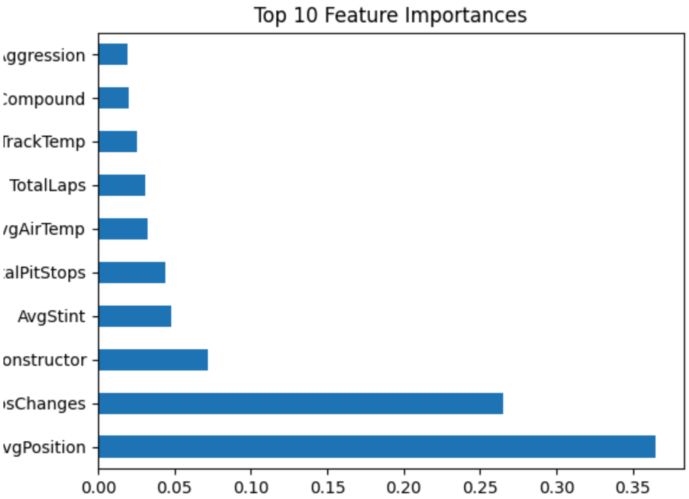
- **Feature Importances** (Random Forest): top 10 predictive features



7. Feature Selection & Importances

- Run `SelectKBest` (ANOVA F-test) to score all features

- Plot top-features from a small Random Forest



8. Prepare ML Dataset

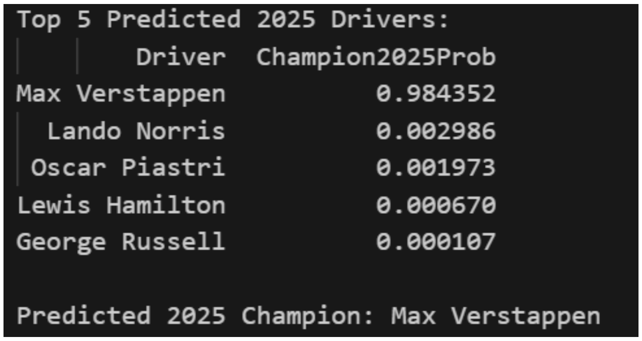
- Create `ChampionNext` label by shifting `IsChampion` to next season
- Split into training (≤2022) and test (2023)
- Drop identifiers (`Season`, `DriverCode`) from features

9. Model Training & Evaluation

- Balance classes on training set using `RandomOverSampler`
- Train four models:
  - Logistic Regression
  - Random Forest (max_depth=5)
  - Gaussian Naive Bayes
  - K Nearest Neighbors (k=5)
- Evaluate on 2023 test data:
  - Accuracy & weighted F1 score
  - Classification report
  - Confusion matrix

10. Final Model & 2025 Prediction

- Retrain best model on all data up to 2023
- Predict champion probabilities for each driver in 2024 season
- Output top-5 drivers by probability



```
Top 5 Predicted 2025 Drivers:
          Driver  Champion2025Prob
Max Verstappen          0.984352
  Lando Norris          0.002986
 Oscar Piastri          0.001973
Lewis Hamilton          0.000670
George Russell          0.000107

Predicted 2025 Champion: Max Verstappen
```