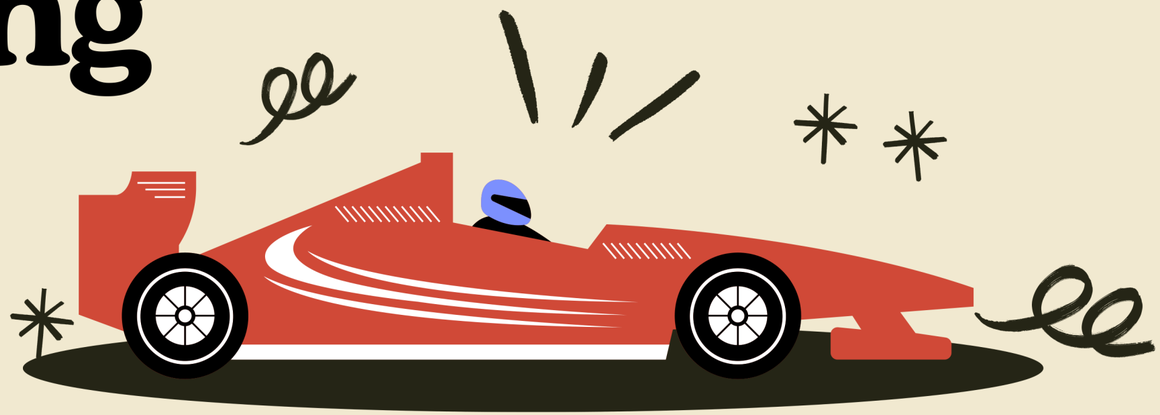
NATHAN GABRIEL C. DANAC'
MAY 21, 2024

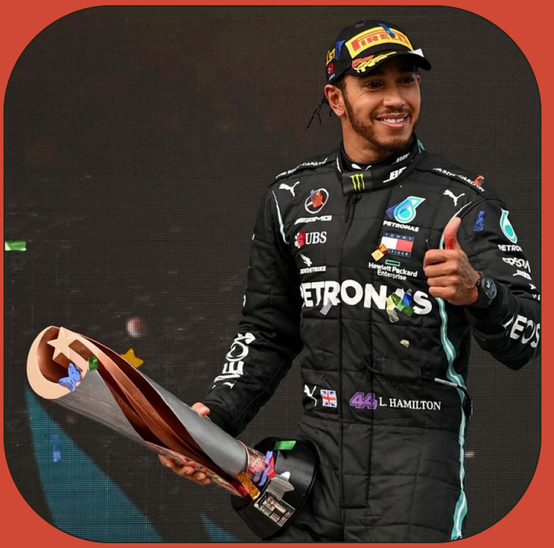
Predicting Formula 1 Race Results using Machine Learning

App Physics 157 Final Project



Formula 1

- considered the “pinnacle of motorsport”
- each team (constructor) designs their own car from scratch, adhering to a strict set of regulations
- each team consists of two drivers and 250 to 1,200 staff
    - a large portion of the team stays at the factory
    - responsible for car development and **race strategy**

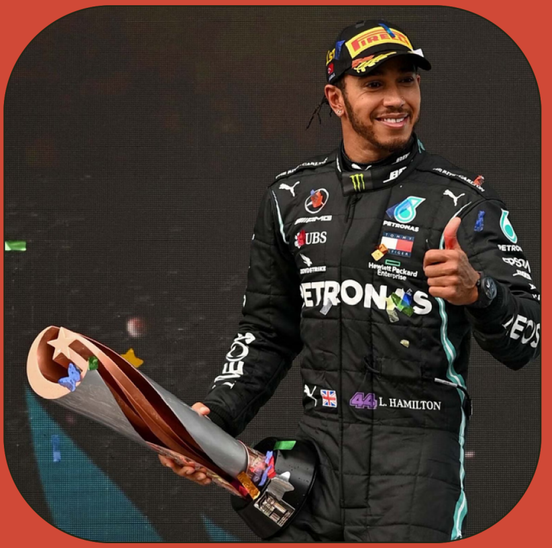


Why Machine Learning?

- there are many factors to consider in predicting a race
  - car performance
  - driver performance
  - weather
  - tyre wear
  - fuel load
  - etc.
- creating a quantitative model can be exhausting
- by representing these parameters in a dataset, we can use ML to predict race results

2022 Season

- 22 Grands Prix
- Each Grand Prix consists of:
  - 3 Free Practice sessions (FP1-FP3)
  - 3 Qualifying sessions (Q1-Q3)
  - 1 Race proper
- 20 drivers, 10 teams



2022 Formula 1 Season Dataset (Kaggle)

- contains the results for each session of each Grand Prix (e.g. 2022 Bahrain GP)
  - Driver Name (“Charles Leclerc”)
  - Car (“Ferrari”)
  - Racing # (16)
  - Laps Completed (57)
  - Time(s) (1:37:33.584)
  - Starting Position (P1)
  - Placement (P1)
  - Points Scored (25)
- imported and processed using **pandas**

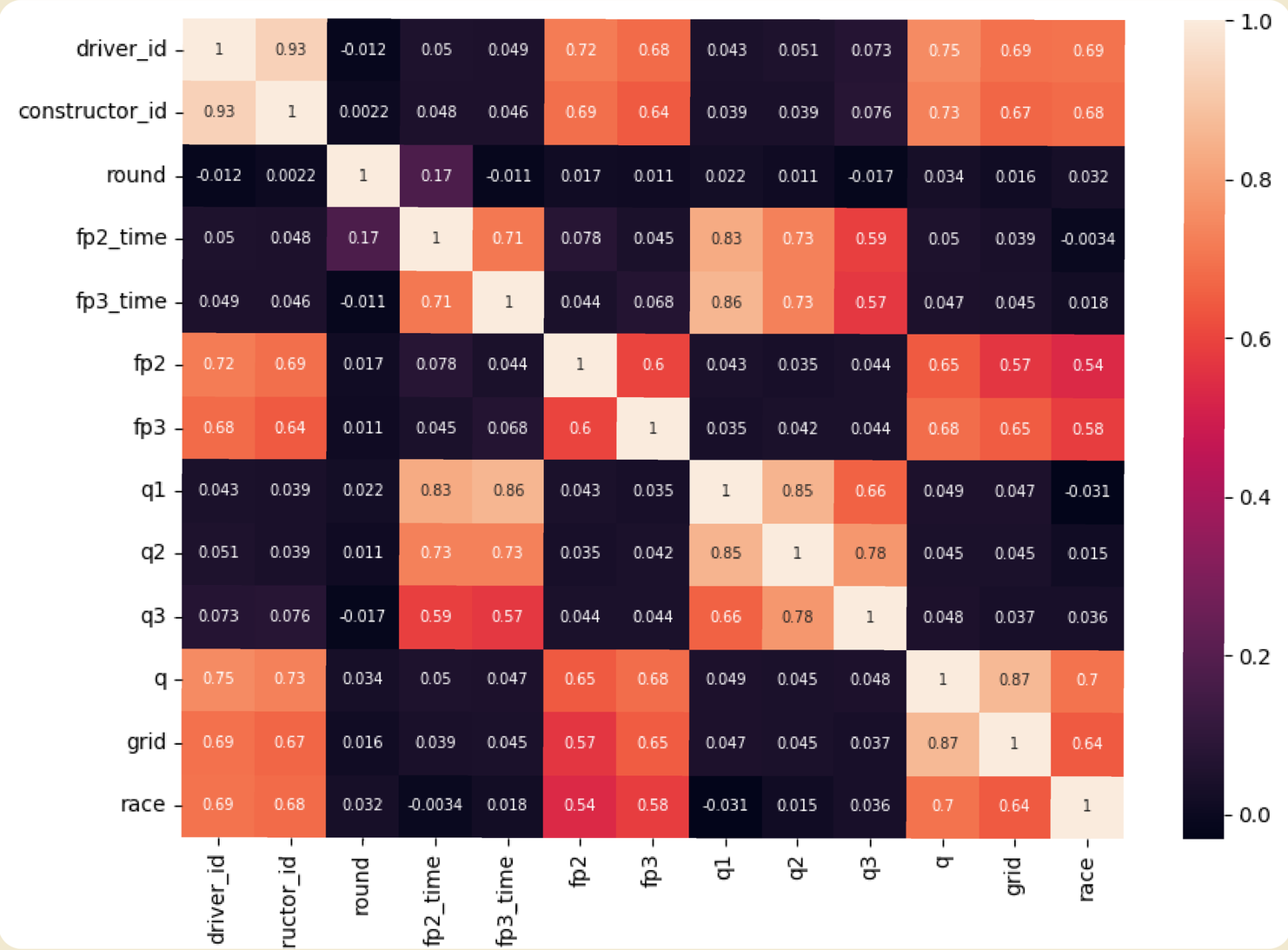
Data Pre-Processing

- re-formatted all data points:
  - Driver Name (“Charles Leclerc”) -> driver_id (based on their final drivers’ championship placement)
  - Car (“Ferrari”) -> constructor_id (based on their final constructors’ championship placement)
  - ~~Racing # (16)~~
  - ~~Laps Completed (57)~~
  - Time(s) (1:37:33.584) -> [session]_time (converted into seconds)
  - Starting Position (P1) -> grid (number only)
  - Placement (P1) -> [session] (number only)
  - ~~Points Scored (25)~~
  - round (round #)

Data Pre-Processing

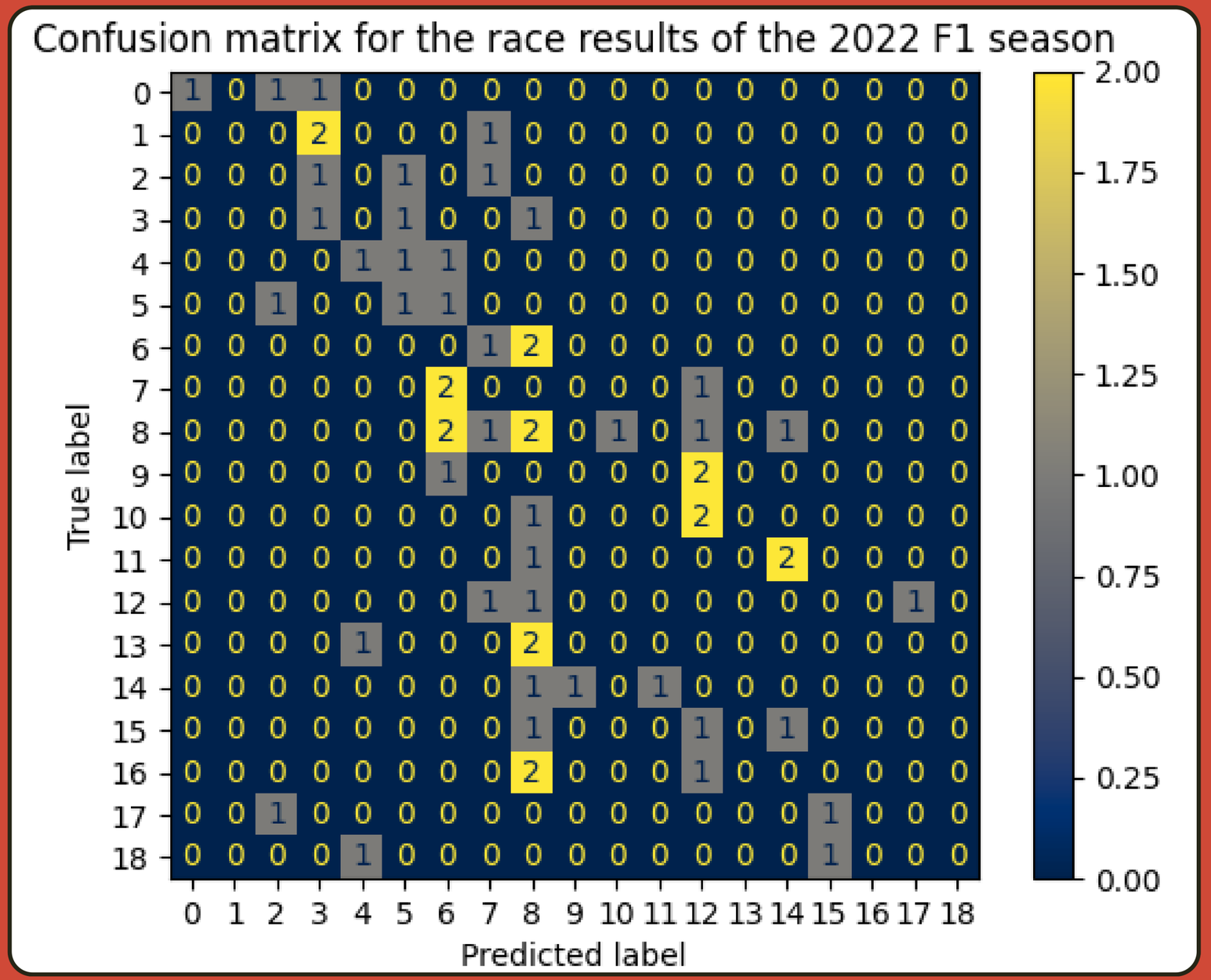
- data was missing for some sessions
  - all NaNs or incomplete data was replaced with the average of the data points for that session (e.g. if FP1 time for “Alex Albon” was NaN, it was replaced by the mean FP1 time of all drivers)
- The race result was selected as the class for our model, while all the times and placements from the other sessions were selected as the features
- *“What position will you finish in the race given your performance and placements from all the sessions from each Grand Prix throughout the season?”*
- The rounds throughout the season were partitioned as follows:
  - 1-16 (training)
  - 17-19 (cross-validation)
  - 20-22 (testing)

Correlation



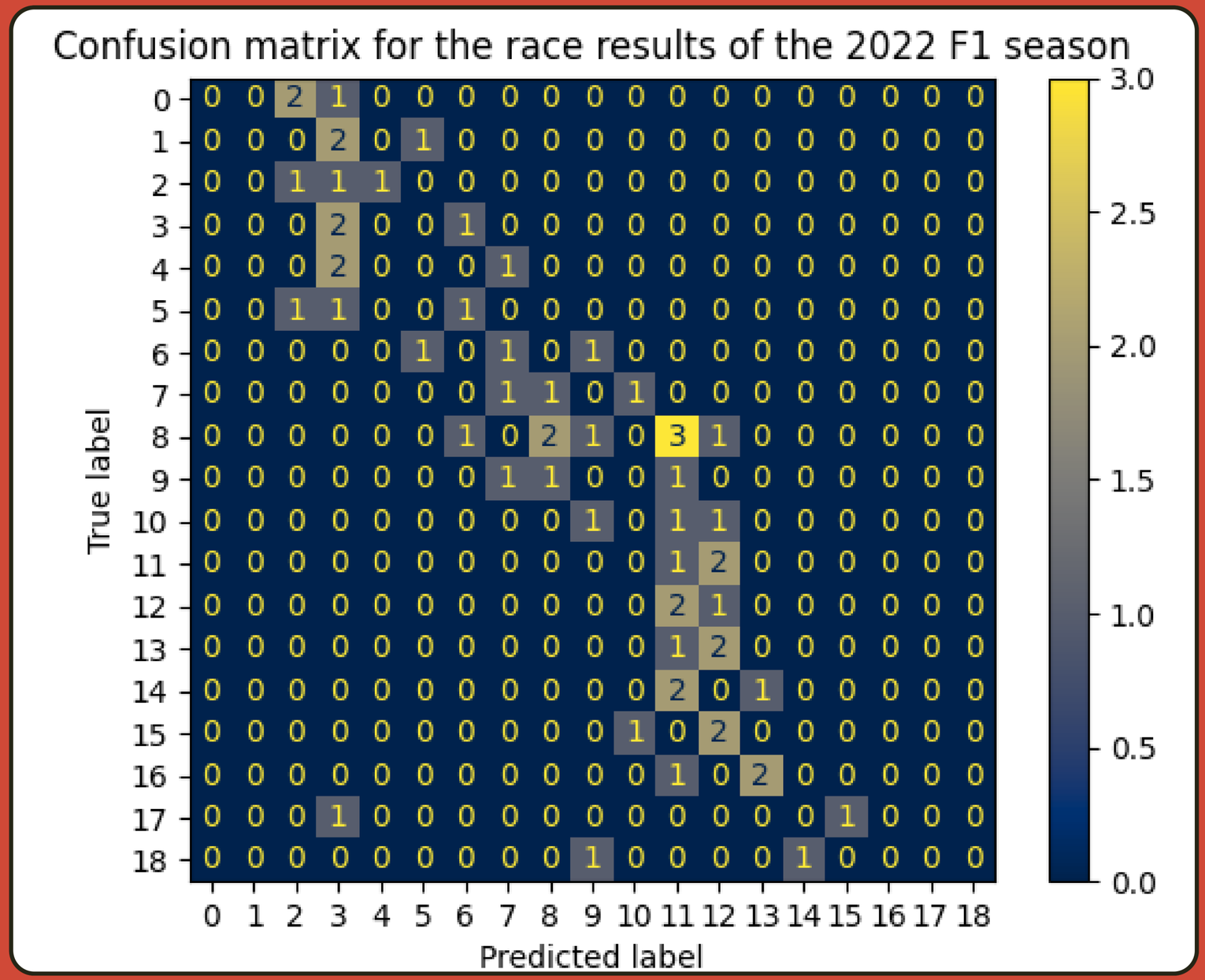
Decision Tree Classifier

- **sklearn.tree**
- accuracy = 10.00%
- was able to predict (P1, P4, P5, P6, P8)



Random Forest Regressor

- **sklearn.ensemble**
- accuracy = 13.33%
- predicted result is within a certain range of the true result



Conclusions

- F1 race results can be predicted from data from other sessions throughout the season
- The Random Forest Regressor model provided more accurate results compared to the Decision Tree Classifier model
- The model’s predicted race result usually falls within a certain range of the true race result.
- As a proof-of-concept, I’m happy with the project :-)

Future Improvements

- The scope of the data can be widened (e.g. historic data from previous seasons)
- We can potentially consider race result in intervals rather than per position
  - (e.g. classes = [”P1-3", “P4-10”, “P11-20”]
- The learning of the model can be further supervised
  - (e.g. for each race, only 1 driver must be predicted for each class, since there are no ties)

Thanks For Listening!

:-)