Formula One: Analysis and Prediction

Statistical Learning – July 19th, 2022

G13 – Alberico Emanuele, Ludovico Lentini, Camilla Lombardi, Matteo Saba

01 **Data Collection & Transformation**

02 **Data Manipulation**

03 **Data Analysis**

04 **Modeling and comparison**

INTRODUCTION

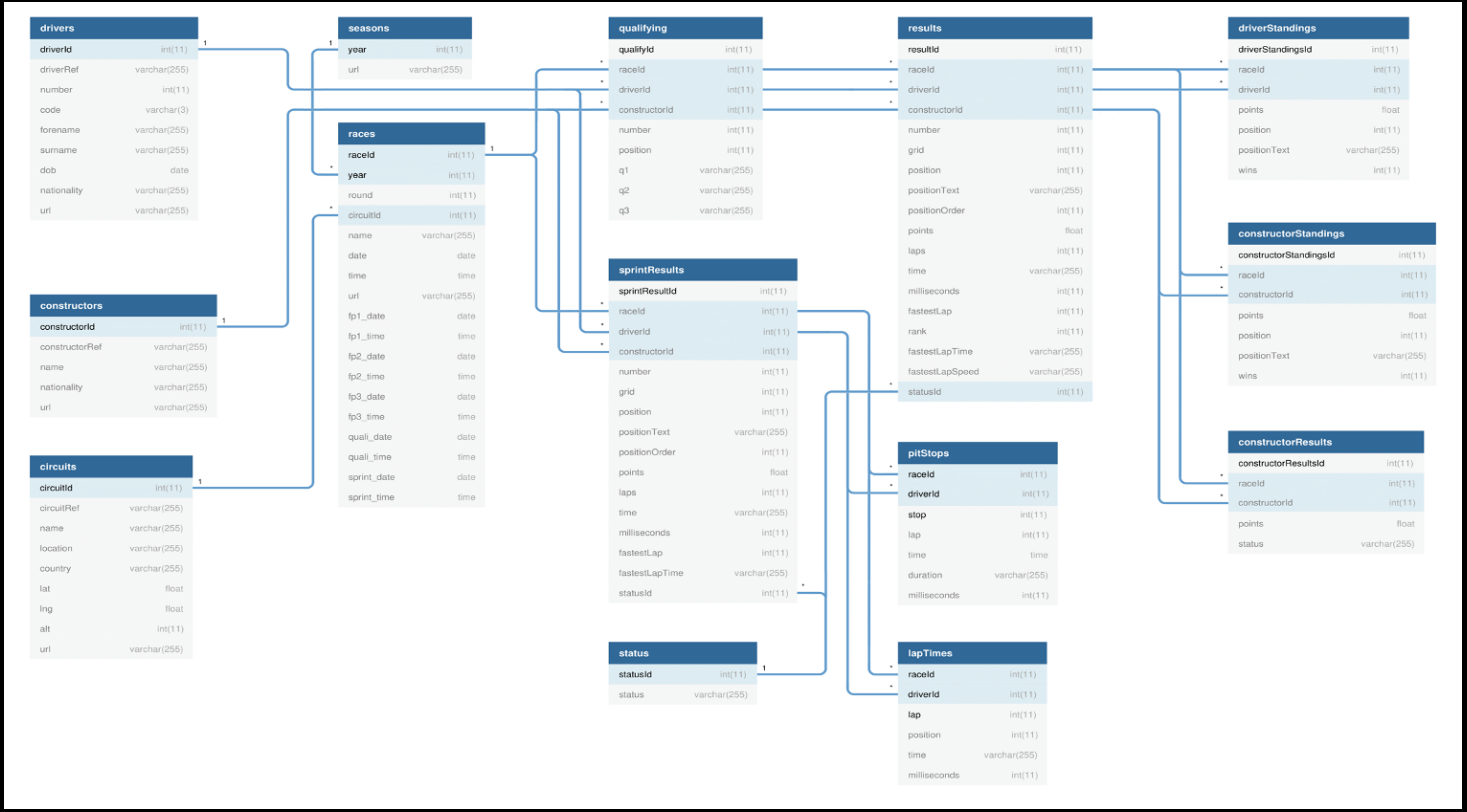
The aim of our project is to predict the finishing grid of the races divided in three classes:

- Class 1: Podiums (1^{st} - 3^{rd})
- Class 2: Positions 4^{th} – 10^{th}
- Class 3: Positions 10^{th} -20^{th}

01

DATA COLLECTION & TRANSFORMATION

EER DIAGRAM



- The dataset was stored in 14 tables containing all the information in between 1950 and 2022.
- We used this tables to create a complete dataset.
- Filter on the dataset: 2014-2021 period analyzed.

02

DATA MANIPULATION

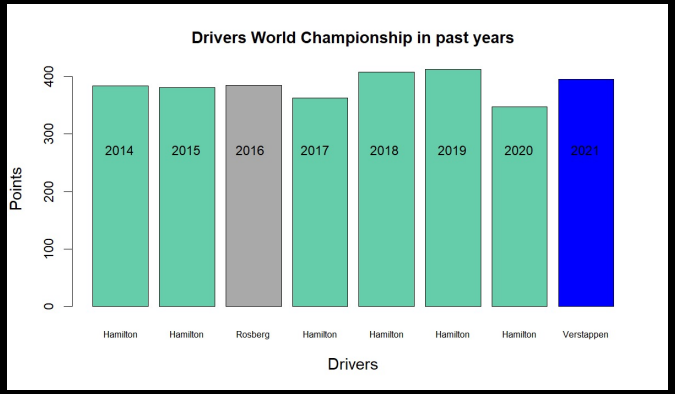
DATA MANIPULATION

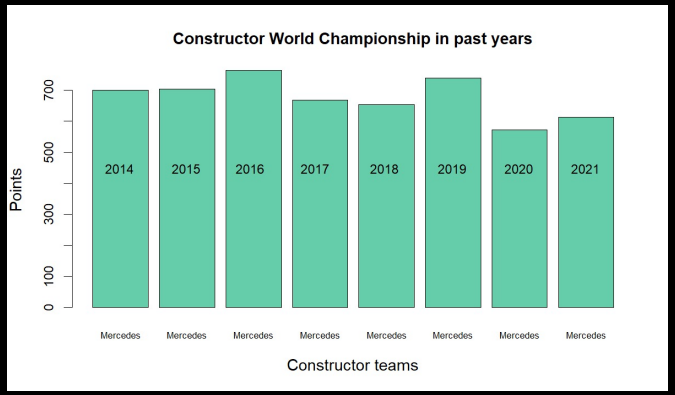
- Changing the names of some teams
- Filtering active drivers, active teams and active circuits
- Introduction of new variables:
    - DNF ratio per Team
    - DNF ratio per driver
    - Wins ratio per driver

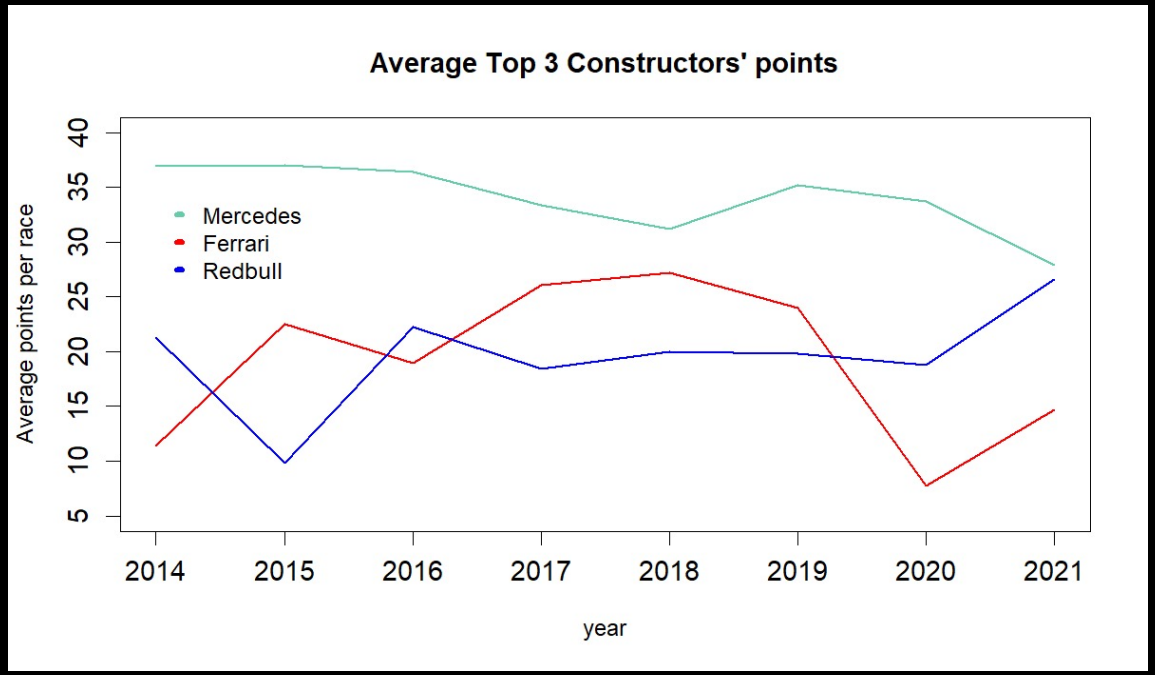
03

DATA ANALYSIS

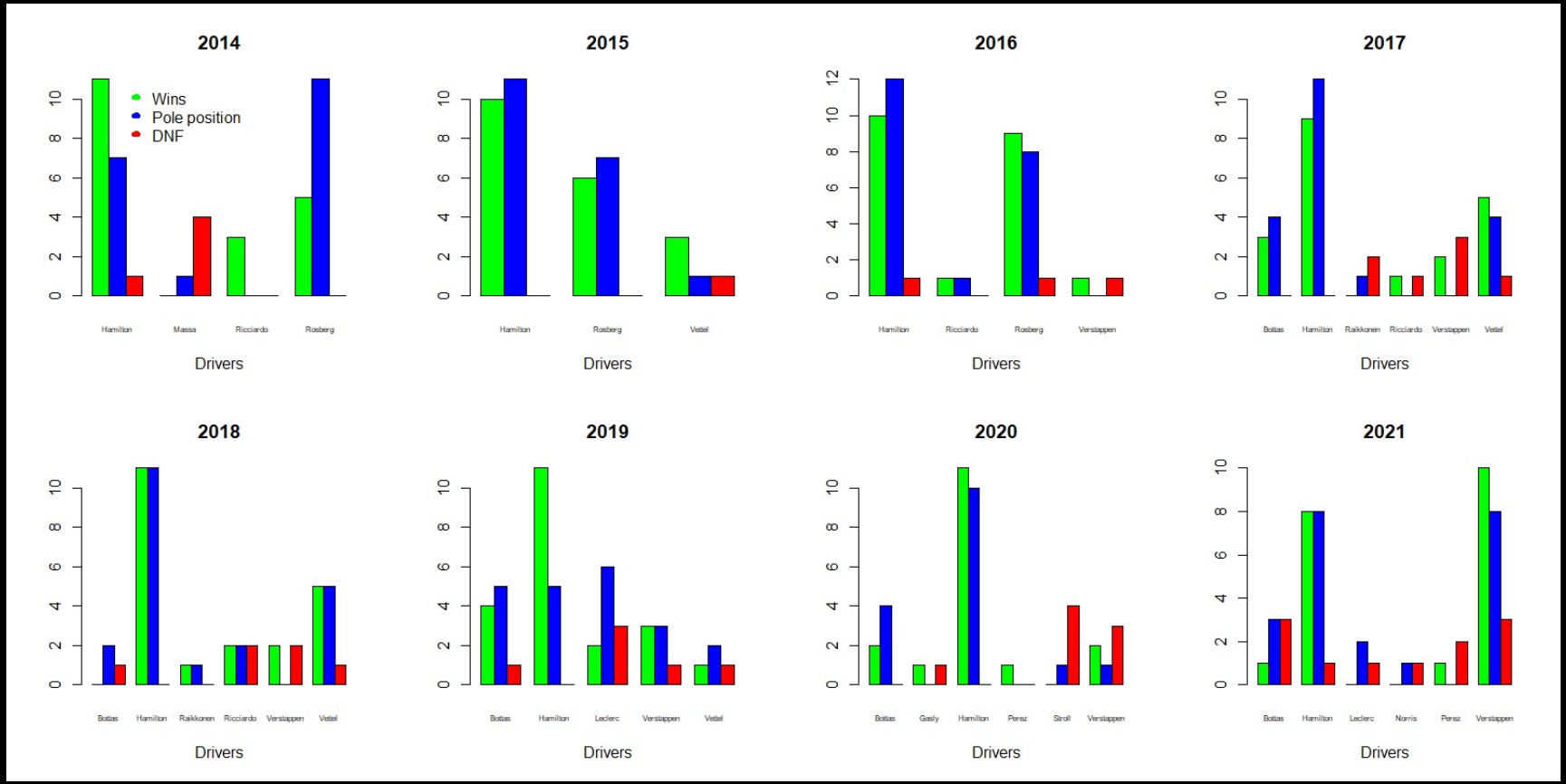
DATA ANALYSIS



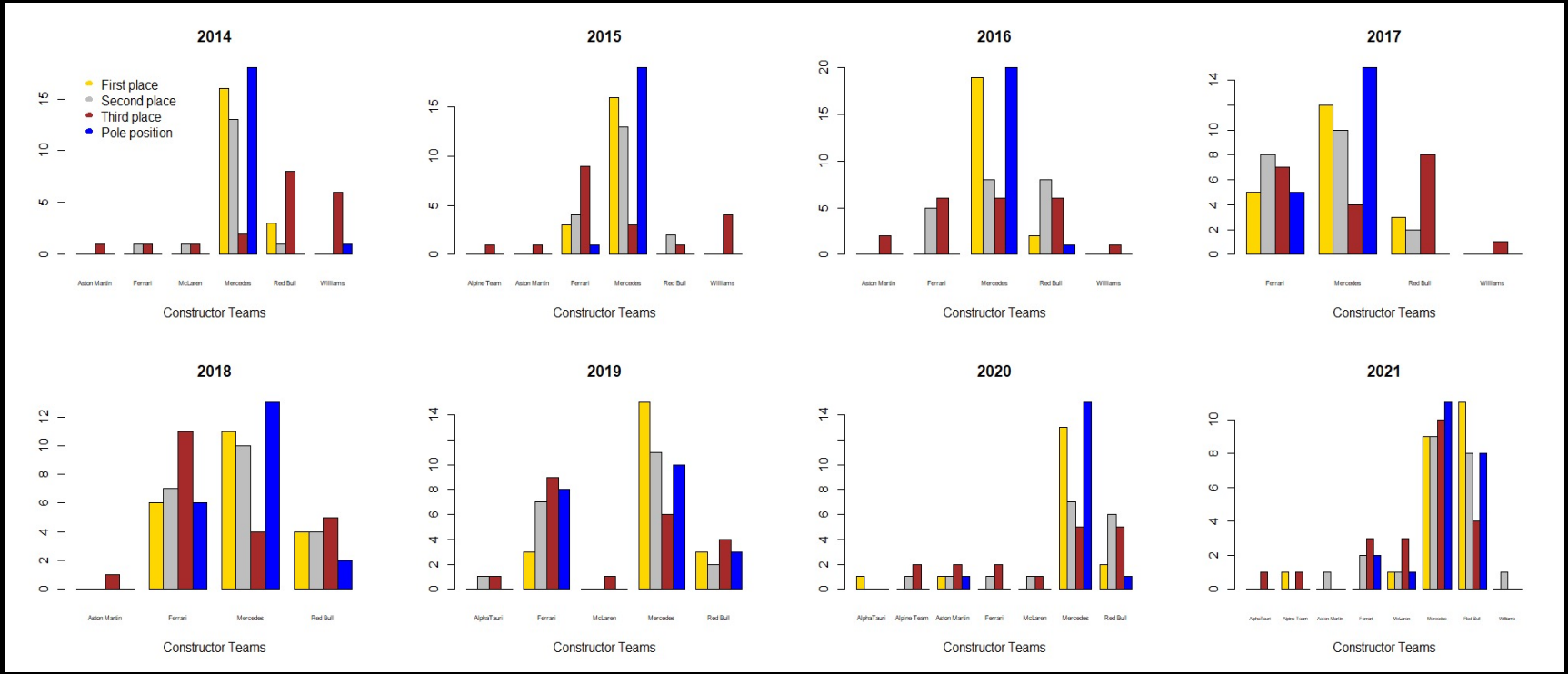




DATA ANALYSIS



DATA ANALYSIS



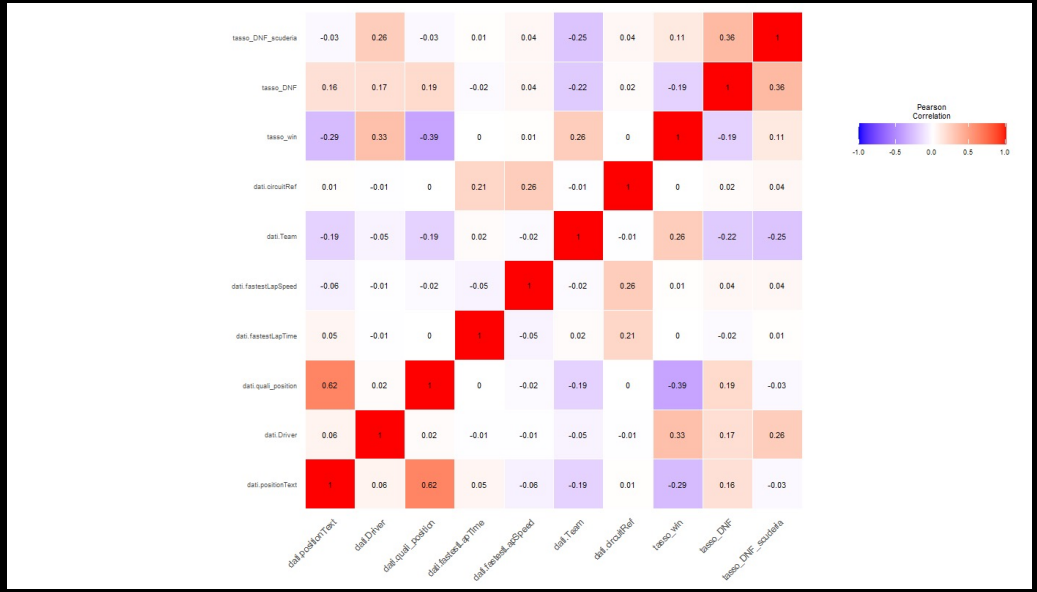
04

DATA MODELING

DATA MODELING

Independent variables:

- Driver
- Qualifying position
- Seconds of the fastest lap
- Speed of the fastest lap
- Team
- The circuit where the race took place
- Win percentage per driver
- DNF percentage per driver
- DNF percentage per team



THE MODELS

ORDERED LOGISTIC REGRESSION

The assumptions behind this model are:
· The dependent target variable has to be ordered
· One or more of the independent variables are either continuous, categorical or ordinal
· No multi-collinearity.

RANDOM FOREST

Algorithm that with an ensemble technique builds different decision trees on bootstrapped data observation and a subset of features. Each model is trained independently and generates a result: then the final output is based on majority voting after combining the results of all models.

NAÏVE BAYES CLASSIFIER

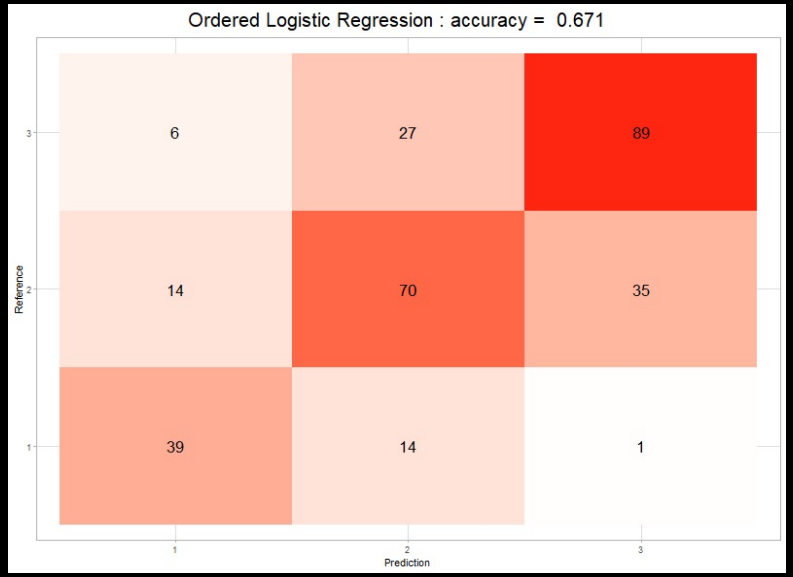
Algorithm based on Bayes' Theorem with strong hypothesis of in-dependence between the independent variables.

SVM

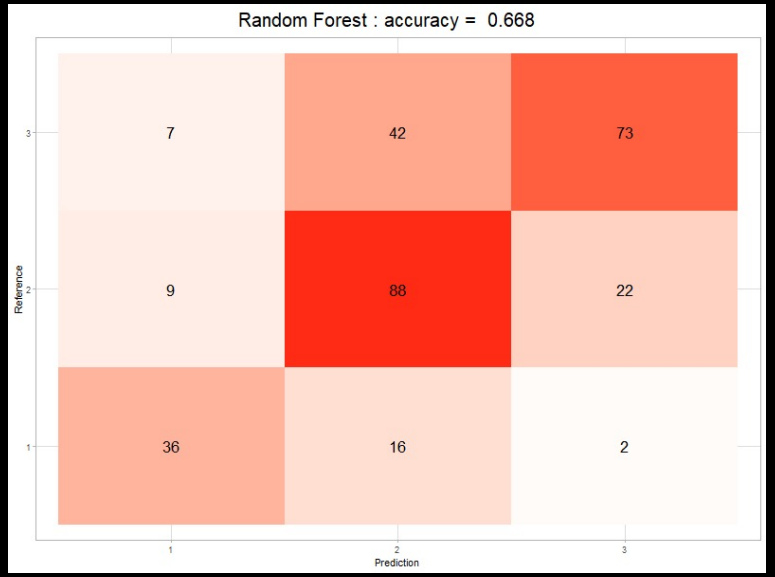
The algorithm creates a line or a hyperplane which separates the data into classes.The best hyperplane is the hyperplane whose distance to the nearest element of each tag is the largest.

MODEL RESULTS

ORDERED LOGISTIC REGRESSION

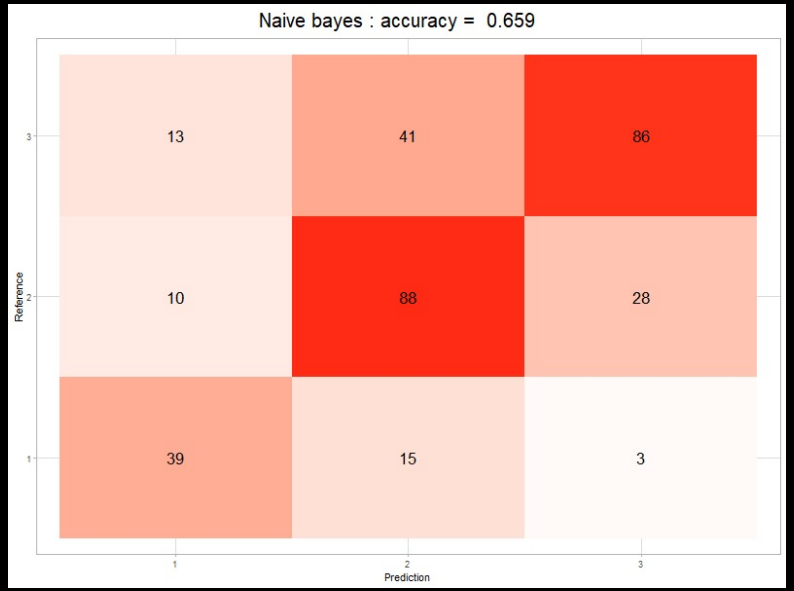


RANDOM FOREST

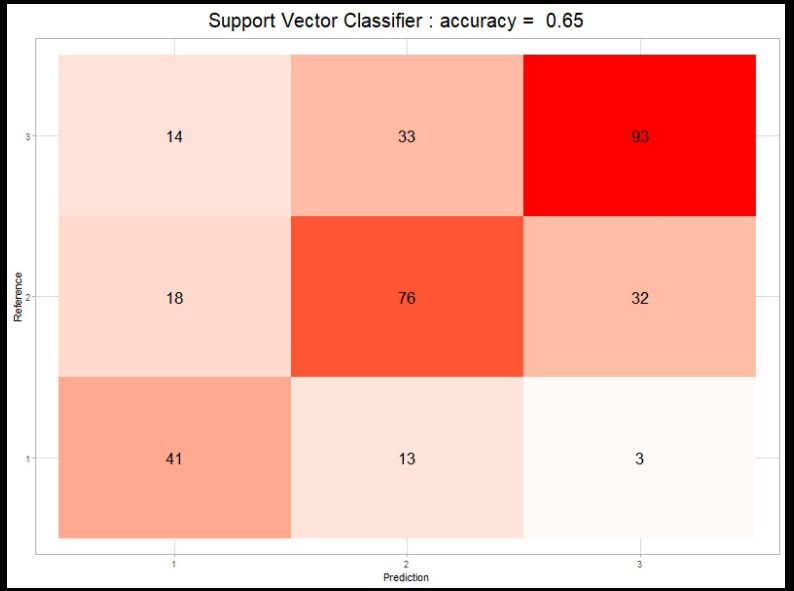


MODEL RESULTS

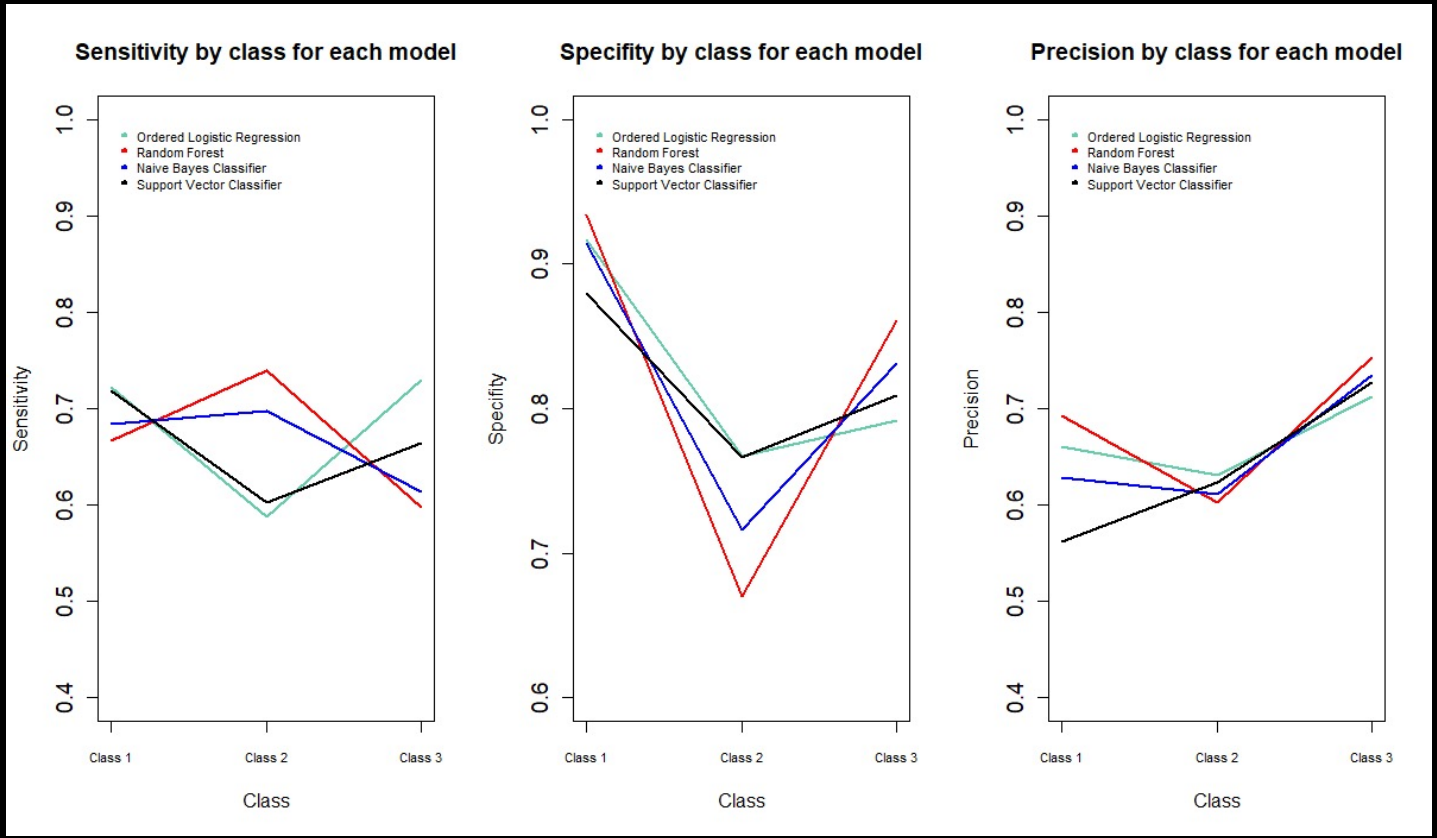
NAÏVE BAYES CLASSIFIER



SVM



MODEL COMPARISON



CONCLUSIONS

- All the four models implemented perform similarly

- We believe that the best model is the Ordered Logistic Regression thanks to the values of specificity and sensitivity per class

- It would be interesting to introduce variables which are more related to the technical parts of the car