GRUPPO: CORES

STATISTICA COMPUTAZIONALE – REPORT FINALE

Maccianti Federico (909656) Rapacioli Nicola (915439) Riva Pietro (908813)

1 Introduzione

Il presente report analizza un dataset ottenuto tramite il repository TracingInsights [1] Lo studio si concentra sui dati di telemetria relativi alle sessioni di qualifica della stagione di Formula 1 2025, selezionando per ciascun pilota il singolo giro migliore.

Il dataset originale include le seguenti variabili:

Variabile	Unità di misura	Supporto	Tipo
Gran premio	–	{Nomi GP}	`character`
Pilota	–	{Sigle Piloti}	`character`
Tempo dal via	s	$\mathbb{R}^+$	`numeric`
Distanza percorsa	m	$\mathbb{R}^+$	`numeric`
Distanza relativa	–	$[0, 1]$	`numeric`
Velocità	km/h	$\mathbb{N}$	`numeric`
Regime motore	RPM	$\mathbb{N}$	`numeric`
Marcia	–	$\{1, \dots, 8\}$	`integer`
Freno	–	$\{0, 1\}$	`factor`
Acceleratore	%	$[0, 100]$	`numeric`
DRS	–	$\{0, 1\}$	`factor`
Accelerazione laterale e longitudinale	g	$\mathbb{R}$	`numeric`
Coordinate spaziali x,y,z	m	$\mathbb{R}$	`numeric`
Tempo Giro	s	$\mathbb{R}^+$	`numeric`
Gomma	–	{Tipi di gomma}	`character`
Vita della gomma	–	$\mathbb{N}$	`integer`
Strategia	–	$\{1, 12, 21\}$	`integer`

Il dataset presenta le seguenti codifiche: le feature binarie assumono valore 0 in assenza dell'evento e 1 in sua presenza. La telemetria dell'acceleratore misura l'intensità dell'input del pilota, mentre per le distanze indicano la posizione progressiva rispetto allo start. La 'vita della gomma' è definita come il numero di giri già percorsi dallo pneumatico in uso. La variabile 'strategia' identifica il pattern di utilizzo delle gomme: 1 per l'uso esclusivo della Soft, 12 per la transizione Soft → Medium e 21 per la sequenza inversa Medium → Soft.

L'obiettivo del presente report è analizzare tramite il dataset descritto le qualifiche del 2025, al fine di caratterizzare il comportamento dei piloti durante il giro di qualifica attraverso le principali variabili telemetriche disponibili.

2 Analisi Esplorativa

2.1 Considerazioni sulle variabili

Poiché lo stile di guida non è riconducibile a variabili di tipo posizionale, le coordinate spaziali e le misure di distanza, sia assolute sia relative, vengono escluse dall'analisi.

In questa fase preliminare, lo stile di guida viene descritto attraverso variabili dinamiche quali l'utilizzo dell'acceleratore e del freno e le accelerazioni, longitudinale e laterale, che consentono di caratterizzare rispettivamente le modalità di decelerazione in ingresso curva, l'intensità con cui la curva viene affrontata comprendendo anche altri tipi di comportamenti riferibili allo stile di guida. A

[1] Codice di estrazione dati nell'appendice A.

sostegno delle ipotesi sopra citate, si riporta la Figura 1 che confronta nel Gran Premio degli USA le accelerazioni per i piloti: Charles Leclerc, Lando Norris e Max Verstappen. Notando infatti come nelle variazioni repentine si contraddistinguano meglio i piloti.

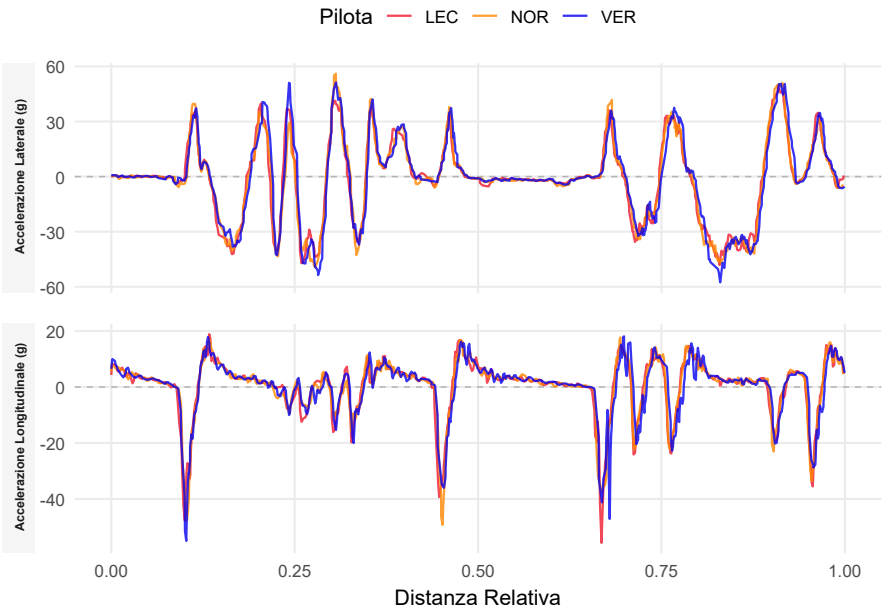

Figura 1: Accelerazioni VER, LEC, NOR.

Dalle prime analisi descrittive sul dataset originale[2], viene osservato come per il pilota Russell al Gran Premio di Miami si producano degli `NA` per quanto riguarda la variabile distanza relativa. Ricontrollando i dati grezzi ci si accorge di come probabilmente i sensori telemetrici abbiano avuto un'avaria, in quanto i record sono la maggior risulta pari a zero e le accelerazioni costanti durante l'intero giro. Per i motivi elencati sopra, si procede dunque ad eliminare il record.

Un ultimo accorgimento riguarda la variabile Tempo Giro, per alcuni piloti risulta infatti `None`, osservando i dati[2], si verifica per i piloti Tsunoda, Bearman e Hadjar rispettivamente nei Gran premi di Emilia Romagna, Australia e Stati Uniti.

Analizzando la realtà si scopre che il motivo è che non hanno terminato il giro di qualifica a causa di un incidente. I dati che si hanno a disposizione sono infatti il riscaldamento delle gomme come si nota nella figura 2. Anche se non è presente un tempo di giro si è registrata comunque la telemetria. Al fine di eliminare rumore nell'analisi si rimuovono queste osservazioni.

Figura 2: Confronto Emilia Romagna Tsunoda

[2]Estratto codice R consultabile nell'appendice B.

2.1.1 Trasformazione e creazione di nuove variabili

Al fine di rendere confrontabili i diversi tracciati dei Gran Premi, la variabile di accelerazione laterale viene riscalata[3], per ciascun Gran Premio e Pilota, nell'intervallo $[-1, 1]$. Successivamente, viene trasformata tramite valore assoluto, così da ottenere una misura dell'intensità complessiva (magnitudo), il supporto quindi risulta compreso in $[0, 1]$.

Per l'accelerazione longitudinale vengono create due nuove variabili[3].

Accelerazione: essa assimila tutti i valori positivi della variabile originaria, viene successivamente normalizzata nell'intervallo $[0, 1]$ per ciascun Gran Premio e Pilota, per riuscire a normalizzare i tracciati.

Decelerazione : essa assimila al contrario della precedente tutti i valori negativi della variabile originaria,successivamente viene applicato il valore atteso e normalizzata nell'intervallo $[0, 1]$. Per le restanti variabili telemetriche si mantiene invece la forma grezza.

Si creano tre nuove variabili di variazione percentuali[4] avendo notato dalla figura 1 che quando ci sono variazioni repentine che si contraddistingue lo stile di guida dei piloti:

- **_lag1** : variazione percentuale rispetto all'osservazione precedente
- **_lag3** : variazione percentuale rispetto a tre osservazioni precedenti
- **_lag5** : variazione percentuale rispetto a cinque osservazioni precedenti

Vengono calcolate indicando con x_t l'unità statistica al tempo t viene calcolata tramite la seguente formula:

$$\Delta = \frac{x_t - x_{t-k}}{x_{t-k}}$$

con $k = 1, 3, 5$ numero di lag per le misure di accelerazione, nel caso in cui $x_{t-k} = 0$ e $x_t = 0$ non si applica la formula e la variabile _lag assume 0 per pre-costruzione ,successivamente quando si osserva solamente $x_{t-k} = 0$ procede con la sostituzione $x_{t-k} = 0.01$ per riuscire a mantenere la validità del calcolo e continuità delle dinamiche telemetriche.

Queste variabili consentono di catturare la dinamica delle grandezze nel tempo e sono utili per caratterizzare l'evoluzione del comportamento di guida tra un istante e l'altro.

2.2 Statistiche riassuntive

Si procede al calcolo delle statistiche descrittive[5] al fine di valutare in che modo esse possano essere associate allo stile di guida, considerando separatamente ciascun Gran Premio e Pilota.

Per le misure di accelerazione laterale,longitudinale e decelerazione longitudinale vengono calcolate media e deviazione standard.

La variabile relativa alla frenata non viene invece inclusa nell'analisi descrittiva, in quanto la sua natura binaria e la forte dipendenza dalle caratteristiche specifiche del singolo Gran Premio rendono poco informative misure sintetiche come media e deviazione standard. Tali indicatori risulterebbero infatti più rappresentativi del tracciato e delle condizioni di gara che dello stile di guida del pilota.

Per le variabili di tipo _lag$k$ vengono calcolate medie e deviazioni standard distinguendo tra variazioni positive e negative, così da evidenziare eventuali asimmetrie nel comportamento dinamico del pilota.

Il nuovo dataset pertanto contiene 58 variabili, che forniscono informazione sullo stile di guida del pilota nella specifica gara.

Si procede dunque con una analisi delle componenti principali per ridurre la dimensionalità, e comprendere quali siano le variabili più significative nel fornire l'informazione.

[3]Estratto codice R consultabile nell'appendice **??**.
[4]Estratto codice R consultabile nell'appendice D
[5]Estratto codice R consultabile nell'appendice E.

2.3 Riduzione della dimensionalità

2.3.1 Analisi correlazioni

Per limitare i casi di multicollinearità ed eliminare il rumore, avendo correlazioni molto elevate ($cor(x, y) > 0.9$), attraverso la funzione `findCorrelation` implementata nella libreria : `{caret}`[6],si cerca di eliminare la ridondanza delle variabili.

In particolare la funzione analizza le coppie di variabili con correlazione eccessiva e rimuove quella che, in media, risulta più correlata con tutte le altre. Questo processo di scrematura ha permesso di ridurre la dimensionalità a 27 variabili.

2.3.2 Analisi delle componenti principali

Seppur la dimensionalità sia diminuita, 27 dimensioni sono eccessive sia ai fini di interpretabilità che per sforzo computazionale.

Si passa dunque da un analisi delle componenti principali (PCA) ai fini di ridurre la dimensionalità. Per l'applicazione della PCA si opta di inserire come input la matrice di correlazione, in quanto, le variabili hanno unità di misura diverse[7].

Dall'analisi risulta che le prime 8 componenti spiegano l'80% della varianza, i pesi associati ad ogni componente sono invece

[6]Estratto del codice R consultabile nell'appendice F

[7]Estratto del codice R consultabile nell'appendice G

Come si evince dalla Figura 3, la distribuzione ...

(a) Distribuzione Variabile X

(b) Scatterplot X vs Y

Figura 3: Analisi esplorativa iniziale delle variabili principali.

3 Modellazione

Abbiamo applicato un modello di regressione lineare:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i \tag{1}$$

A Codice R Commentato Estrazione dei dati

```r
library(jsonlite)
library(tidyverse)
rm(list=ls())
#Cartella origine
cartella <- "C:/Users/feder/Documents/datasets/Computazionale/2025"

#filtro sui file con i giri e poi qualifiche
file_giri <- list.files(path = cartella,
pattern = "laptimes\\.json$",
full.names = TRUE,
recursive = TRUE)

files_qualifiche_lap <- file_giri[grepl("Qualifying", file_giri) & !grepl("Sprint
",file_giri)]
lista_dati <- list()

#Ciclo di estrazione

for(i in 1:length(files_qualifiche_lap)) {

  cartella_giro <- files_qualifiche_lap[i]

  laptimes_data <- fromJSON(cartella_giro)

  giro <- as.numeric(laptimes_data$lap[which.min(laptimes_data$time)])

  cartella_pilota <- dirname(cartella_giro)
  if(length(giro) > 0){
    path_telemetry <- paste0(cartella_pilota, "/", giro, "_tel.json")
  } else{
    giro <- 1
    path_telemetry <- paste0(cartella_pilota, "/", giro, "_tel.json") }

  testo <- read_file(path_telemetry)
  testo <- str_replace_all(testo, "NaN", "null")
  json_data <- fromJSON(testo)[["tel"]]
  json_data <- as_tibble(json_data)

  parti_cartella <- str_split(cartella_giro, "/")[[1]]
  n <- length(parti_cartella)

  json_data$pilota <- parti_cartella[n-1]
  json_data$GP <- parti_cartella[n-3]
  json_data$gomma <- rep(laptimes_data$compound[giro],nrow(json_data))
  json_data$lap_time <- rep(laptimes_data$time[giro],nrow(json_data))
  json_data$strategy <- rep(laptimes_data$status[giro],nrow(json_data))
  json_data$life <- rep(laptimes_data$life[giro],nrow(json_data))
  json_data$life <- as.numeric(json_data$life)
  lista_dati[[i]] <- json_data
}
#merge dei dati
dataset_completo <- bind_rows(lista_dati)
#pulizia
dataset_completo <- dataset_completo %>%
select(-dataKey)
setwd("C:/Users/feder/Documents/datasets/Computazionale/F1/data")
saveRDS(dataset_completo, file = "dataset_completo_best_tel.rds")


```

B Codice R Commentato Esplorazione iniziale

```r
# A seguito del caricamento delle librerie e del dataset si inizia l'esplorazione
 con le iniziali statistiche descrittive
summary(tel)
# Sistemazione delle variabili
tel$throttle <- ifelse(tel$throttle > 100, 100, tel$throttle)

#Trattamento degli NA e None

tel[which(is.na(tel$rel_distance)),]
tel <- tel %>% filter(pilota != "RUS" & GP != "Miami Grand Prix")

tel %>%
group_by(GP,pilota) %>%
filter(lap_time == "None") %>% summarize(n(),lap_time=max(lap_time))

tel <- tel %>% filter(lap_time !="None")

```

C Codice R Commentato Trattemento variabili

```r
tel.guida <- tel %>%
group_by(GP,pilota) %>%
mutate(
  dec_x = if_else(acc_x < 0, abs(acc_x), 0),
  acc_x = if_else(acc_x > 0, acc_x, 0)) %>%
  mutate(acc_x=rescale(acc_x,to=c(0,1)),
  dec_x=rescale(dec_x,to=c(0,1),
  acc_y=abs(rescale(acc_y,to = c(-1, 1))))) %>%
          ungroup()

```

D Codice R Commentato Creazione variabili

```r
#Creazione delle variabili lag
tel.guida <- tel.guida %>%
select(GP,pilota,throttle,acc_x,acc_y,dec_x,rel_distance) %>%
arrange(GP,pilota,rel_distance) %>%
group_by(GP,pilota) %>%
mutate(
across(
c(throttle, acc_x, acc_y,dec_x),
list(
lag1 = ~round(
ifelse(
lag(.x, 1) == 0 & .x == 0,
0,
ifelse(
lag(.x, 1) == 0,
(.x-0.01)/0.01,
(.x - lag(.x, 1)) / lag(.x, 1))
),
4),
...
),
.names = "{.col}_{.fn}"
)
)  %>%
ungroup() %>%
select(-rel_distance)

```

E Codice R Commentato Statistiche Riassuntive

```r
  tel.guida_summary <- tel.guida %>%
  group_by(GP, pilota) %>%
  {
    lag_cols <- names(.) %>% .[str_detect(., "lag")]

    summarise(.,
    across(
    c(throttle, acc_x, acc_y,dec_x),
    list(
    mean = ~round(mean(.x, na.rm = TRUE), 4),
    sd = ~round(sd(.x, na.rm = TRUE), 4)
    ),
    .names = "{.col}_{.fn}"
    ),

    across(
    all_of(lag_cols),
    ~round(mean(.x[.x > 0], na.rm = TRUE), 4),
    .names = "{.col}_mean_pos"
    ),

    across(
    all_of(lag_cols),
    ~round(mean(.x[.x <= 0], na.rm = TRUE), 4),
    .names = "{.col}_mean_neg"
    ),

    across(
    all_of(lag_cols),
    ~round(sd(.x[.x > 0], na.rm = TRUE), 4),
    .names = "{.col}_sd_pos"
    ),

    across(
    all_of(lag_cols),
    ~round(sd(.x[.x <= 0], na.rm = TRUE), 4),
    .names = "{.col}_sd_neg"
    ),

    .groups = "drop"
    )
  }

```

F Codice R Commentato Analisi delle correlazioni

```r
corr <- round(cor(tel.guida_summary %>% select(where(is.numeric))),4)

#Ricerca di variabili dipendenti

variabili_dipendenti <- findCorrelation(corr, cutoff = 0.9, names = TRUE, exact = T
  ,verbose = T)

print(variabili_dipendenti)

tel.pca <- tel.guida_summary %>% select(-all_of(variabili_dipendenti))


```

G Codice R Commentato Analisi delle componenti principali

```r
PCA <- princomp(tel.pca %>%
select(where(is.numeric)),cor=T)

summary(PCA)

PCA$loadings

fviz_contrib(PCA, choice = "var", axes = 1, top = 10)
fviz_contrib(PCA, choice = "var", axes = 2, top = 10)
fviz_contrib(PCA, choice = "var", axes = 3, top = 10)
fviz_contrib(PCA, choice = "var", axes = 4, top = 10)
fviz_contrib(PCA, choice = "var", axes = 5, top = 10)
fviz_contrib(PCA, choice = "var", axes = 6, top = 10)

```