GRUPPO: CORES

Statistica computazionale – Report Finale

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

1 Introduzione

Il seguente studio, si pone come obiettivo quello di individuare i diversi stili di guida presenti nella Formula 1, tramite l'analisi dei dati di telemetria (TracingInsights[1]) di ciascun pilota. Lo studio si concentra sui dati 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	–	$\{1, \ldots, 8\}$	`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`

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.

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 che relative, vengono escluse dall'analisi.

In questa fase preliminare, lo stile di guida viene descritto attraverso variabili dinamiche, quali l'utilizzo dell'acceleratore, del freno, la velocità e le seguenti accelerazioni longitudinali e laterali, che consentono, per esempio, di valutare rispettivamente le modalità di decelerazione in ingresso curva, l'intensità con cui la curva viene affrontata tutte caratteristiche che fanno riferimento allo stile di guida. A sostegno delle ipotesi sopra citate, si riporta la Figura 1 che confronta nel Gran Premio degli USA le accelerazioni e velocità per i piloti: Charles Leclerc, Lando Norris, Max Verstappen e Franco Colapinto.

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

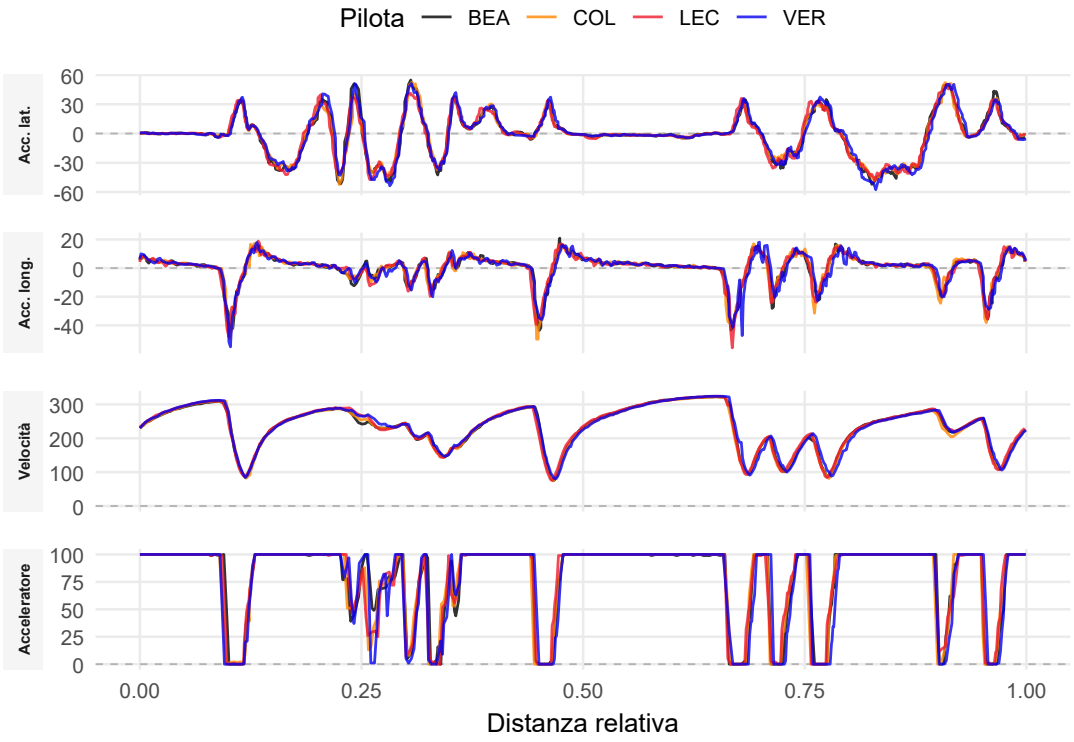

Figura 1: Accelerazioni VER, LEC, NOR, COL.

Si nota infatti come nelle variazioni repentine si possano già individuare differenze significative tra piloti.

Dalle prime analisi descrittive sul dataset[2], viene osservato come per il pilota Russell al Gran Premio di Miami siano presenti degli `NA` nella variabile `distanza relativa`. Osservando i dati grezzi si può notare come ciò sia riconducibile al malfunzionamento dei sensori telemetrici - presenza di molti zeri in molte variabili e accelerazioni costanti durante l'intero giro. Per i motivi elencati sopra, si procede dunque ad eliminare i record.

Infine, osservando la variabile `Tempo Giro`, alcuni piloti riportano alcuni valori `NA`. Osservando i dati[2], ciò si verifica per i piloti Tsunoda, Bearman e Hadjar rispettivamente nei Gran premi di Emilia Romagna, Australia e Stati Uniti.

Effettuando un'analisi di tipo qualitativo, si scopre che ciò è riconducibile ad un incidente durante il giro di qualifica. Le uniche telemetrie disponibili sono infatti riconducibili al giro di riscaldamento delle gomme, come si nota nella figura 2. Al fine di eliminare rumore nell'analisi si rimuovono queste osservazioni.

Figura 2: Confronto Australia Bearman

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

Sempre per quanto riguarda la variabile `tempo giro`, considerando il regolamento della **FIA** (39.4.b.i ; PAG:47), secondo il quale se un pilota supera il 107% del tempo giro minore viene escluso dalle qualifiche, si uniforma il dataset. Incriminati risultano essere i piloti Hamilton, Tsunoda, Antonelli, Albon e Bortoleto nel Las Vegas Grand Prix e Stroll nel Dutch Grand Prix.

Analizzando il contesto, si nota che a Las Vegas i tempi più elevati sono imputabili alle condizioni meteo avverse (pioggia intensa), seguite da un progressivo asciugamento della pista che ha avvantaggiato gli altri piloti. Nel caso di Stroll, invece, il dato si riferisce a al primo giro di lancio completo, poiché nel momento del giro di qualifica non è stato completato il giro per un incidente.

Di conseguenza per eliminare il rumore, si elimina soltanto l'occorrenza di Stroll, poiché il giro non rappresenta il "migliore" ma solamente il quello di lancio.

2.1.1 Trasformazione e creazione di nuove variabili

Per quanto riguarda l'analisi della variabile di `accelerazione laterale`, viene trasformata tramite valore assoluto. Così da ottenere una misura dell'intensità complessiva (magnitudo)[3].

Inoltre, l'`accelerazione longitudinale` viene suddivisa[3] in:

- `Accelerazione`, che comprende tutti i valori positivi;
- `Decelerazione` che al contrario della precedente comprende, tutti i valori negativi;

Successivamente viene applicato il valore assoluto e per le restanti variabili telemetriche si mantiene invece la forma originaria.

Si creano per le variabili `velocità, acceleratore, decelerazione e accelerazioni`, nuove variabili di variazione percentuali[4], per catturare l'intensità di variazione delle variabili che aiutano a distinguere lo stile di guida dei piloti, nei vari istanti di tempo:

- `_lag1` : variazione percentuale rispetto all'osservazione precedente;
- `_lag5` : variazione percentuale rispetto a cinque osservazioni precedenti;

che vengono calcolate come:

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

indicando con x_t l'unità statistica al tempo t e $k = \{1, 5\}$ numero di lag, per le misure di accelerazione e decelerazione. Nel caso in cui $x_{t-k} = 0$ e $x_t = 0$ non si applica la formula e la variabile _lag assume 0 di default, quando invece si osserva solamente $x_{t-k} = 0$ si sostituisce con $x_{t-k} = 0.01$ per riuscire a mantenere la validità del calcolo e continuità delle dinamiche telemetriche.

2.2 Statistiche riassuntive

Vengono calcolate delle statistiche descrittive[5] al fine di valutare in che modo le variabili possano essere associate allo stile di guida, considerando separatamente ciascun Gran Premio e Pilota.

Per le misure di accelerazione laterale, longitudinale, decelerazione e per gli input di frenata, accelerazione e velocità vengono calcolate media e deviazione standard.

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.

Una volta calcolate media e deviazione standard, è possibile combinare questi valori per ottenere il coefficiente di variazione (CV):

$$CV = \frac{sd(x)}{mean(x)}$$

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

che misura la variabilità relativa di una variabile rispetto alla sua media. L'impiego di tale coefficiente permette di ridurre l'influenza del setup della vettura, rendendo confrontabili tra loro variabili che, in valore assoluto, dipendono dalle regolazioni meccaniche e aerodinamiche.

Si arriva cosi ad ottenere un dataset contenente 28 variabili, che forniscono informazione sullo stile di guida del pilota nella specifica gara.

Infine, per rimuovere l'influenza del tracciato, si riscalano tutte le variabili nell'intervallo $[0, 1]$.

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

2.3 Analisi delle componenti principali

Per limitare i casi di multicollinearità ed eliminare il rumore, avendo correlazioni molto elevate $(cor(x, y) > 0.9)$[6], si cerca di eliminare le variabili rindondanti. Questo processo di scrematura permette di ridurre la dimensionalità a 26 variabili.

Seppur la dimensionalità sia diminuita, 26 dimensioni sono eccessive sia ai fini di interpretabilità che per peso computazionale per un Model Base Clustering.

Si effettua dunque un analisi delle componenti principali (PCA) ai fini di ridurre la dimensionalità. I risultati ottenuti tramite PCA[7] mostrano come le prime 5 componenti spiegano più del 70% della varianza, come si nota nella figura 3.

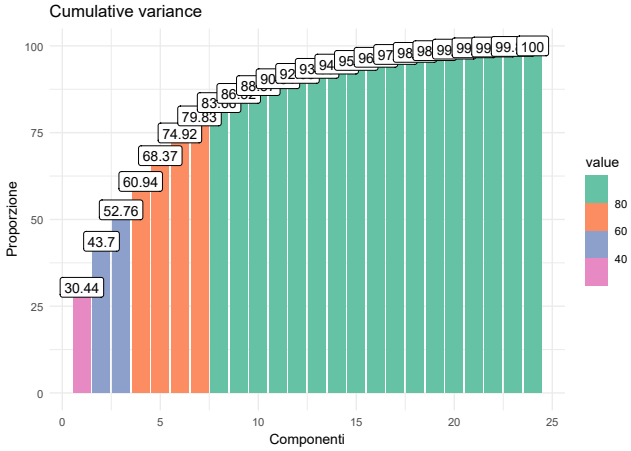

Figura 3: Varianza Spiegata Cumulata

2.3.1 Interpretazione delle componenti

Attraverso l'analisi dei pesi associati alle componenti le si interpretano:

1. La prima componente ha pesi maggiori per le variabili che descrivono la variabilità della dinamica longitudinale, sono coinvolti i coefficienti di decelerazione ed accelerazione come anche le variazioni degli input come acceleratore, freno e velocità. A valori alti viene quindi associata una guida più instabile, per valori bassi invece è associata una guida più progressiva e stabile.

2. Per la seconda componente hanno pesi maggiori le variabili per lo più che descrivono la dinamica laterale insieme all'accelerazione longitudinale. A valori elevati è associata una maggiore variabilità, indicativa di intensità laterale più irregolare e impegnativa,spesso riguarda curve più accentuate. Valori bassi sono invece associati a una dinamica laterale più regolare e controllata spesso compatibile con la percorrenza di curve con andamento costante.

3. Nella terza componente prevalgono le dinamiche sia di accelerazione longitudinale che laterale, per quanto riguarda le laterali, i coefficienti che riguardano i lag negativi hanno maggiore peso, la componente descrive quindi la dinamica transitoria,per esempio un passaggio da curva a rettilineo.

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

A valori alti viene associata una guida più instabile mentre per bassi valori viene associata una guida più progressiva e stabile.

4. Per la quarta componente i contributi più significativi sia della dinamica longitudinale, in particolare delle variabili di accelerazione e decelerazione con lag positivi, sia della dinamica laterale e della velocità. La componente descrive quindi fasi di guida dinamica in cui sono presenti variazioni coordinate di accelerazione, decelerazione e velocità, per esempio ingressi o uscite da curve. A valori elevati corrisponde una maggiore variabilità complessiva, associabile a uno stile di guida più dinamico e meno uniforme, mentre valori bassi indicano una guida più regolare e costante.

5. La quinta componente è dominata dalla dinamica laterale, con contributi rilevanti delle accelerazioni laterali con lag positivi,sono incluse anche variabili legate all'acceleratore con lag positivi. La componente può quindi essere interpretata come descrittiva di fasi di guida caratterizzate da una dinamica laterale sostenuta e prolungata. Valori elevati sono associati a una maggiore variabilità laterale persistente, mentre valori bassi indicano una dinamica laterale più contenuta e regolare.

Come si evince dalla Figura 4, la distribuzione ...

(a) Distribuzione Variabile X

(b) Scatterplot X vs Y

Figura 4: 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$lap_time <- rep(laptimes_data$time[giro],nrow(json_data))
  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),
acc_y = abs(acc_y)) %>%
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"
  )
}

tel2 <- tel %>%
group_by(GP,pilota) %>%
summarize(laptime=max(lap_time),strategy=(max(strategy)),gomma=(max(gomma))
,.groups = "drop")

tel.guida_summary$laptime <- as.numeric(tel2$laptime)
tel.guida_summary$strategy <- tel2$strategy
tel.guida_summary$gomma <- tel2$gomma

```

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)))

fviz_screeplot(PCA,choice= "variance")

summary(PCA)
print(PCA$loadings,cutoff = 0)

for (i in 1:6){
  p <- fviz_contrib(PCA, choice = "var", axes = i, top = 15)
  print(p)
}

```