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. 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	Tipo
Gran premio	–	`character`
Pilota	–	`character`
Tempo dal via	s	`numeric`
Distanza percorsa	m	`numeric`
Distanza relativa	0–1	`numeric`
Velocità	km/h	`numeric`
Regime motore	RPM	`numeric`
Marcia	1–8	`numeric`
Freno	0/1	`factor`
Acceleratore	0–100 %	`numeric`
DRS	0/1	`factor`
Accelerazioni laterale e longitudinale	g	`numeric`
Coordinate spaziali x,y,z	m	`numeric`

Le variabili binarie sono codificate come 0 (non attivo) e 1 (attivo).

L'obiettivo del presente report è analizzare il dataset descritto, 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 e l'intensità con cui la curva viene affrontata. 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

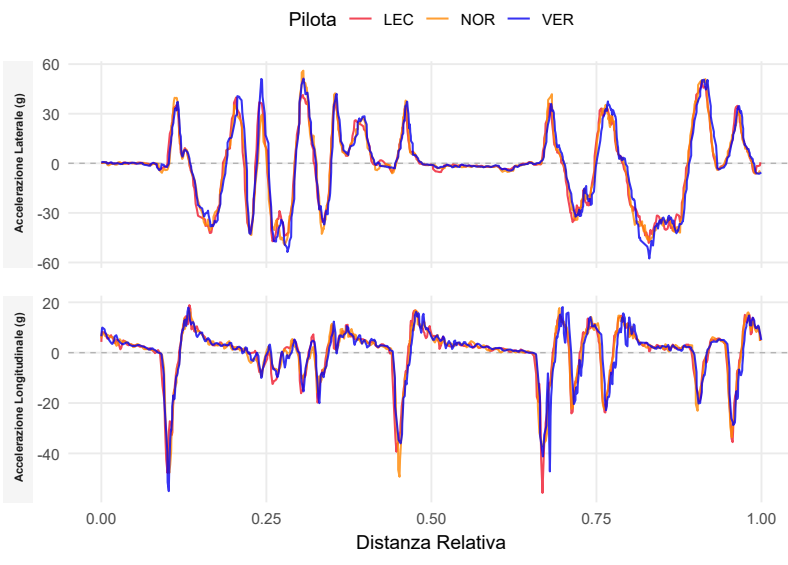


Figura 1: Accelerazioni VER, LEC, NOR.

2.1.1 Trasformazione e creazione di nuove variabili

Al fine di rendere confrontabili i diversi tracciati dei Gran Premi, le variabili di accelerazione laterale e longitudinale vengono riscalate, per ciascun Gran Premio e Pilota, nell'intervallo $[-1, 1]$. Successivamente, tali variabili vengono trasformate in valore assoluto, così da ottenere una misura della loro intensità complessiva (magnitudo), il cui dominio risulta compreso in $[0, 1]$. Per le restanti variabili telemetriche si mantiene invece la forma grezza.

Si creano tre nuove variabili di variazione percentuali 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
- **__lag2** : variazione percentuale rispetto a due osservazioni precedenti
- **__lag3** : variazione percentuale rispetto a tre 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 $k = 1, 2, 3$ numero di lag per le misure di accelerazione, nel caso in cui $x_{t-k} = 0$ si 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 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 longitudinale e laterale vengono calcolate media, deviazione standard, valore massimo e valore minimo.

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.

Anche per la variabile accelerazione non vengono considerati gli estremi minimo e massimo, poiché la misura è limitata nell'intervallo [0,100] e tali valori risulterebbero costanti per tutti i piloti e per tutti i Gran Premi, non apportando informazione discriminante.

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.

Nel calcolo delle statistiche descrittive, viene osservato come per il pilota Russell al Gran Premio di Miami si producano degli `NA`. 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. Per i motivi elencati sopra, si procede dunque ad eliminare il record.

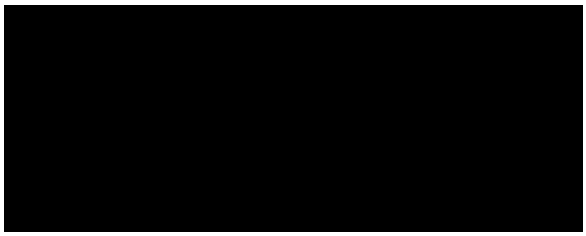
Il nuovo dataset pertanto contiene 48 variabili, che forniscono informazione sullo stile di guida del pilota nella specifica gara.

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.

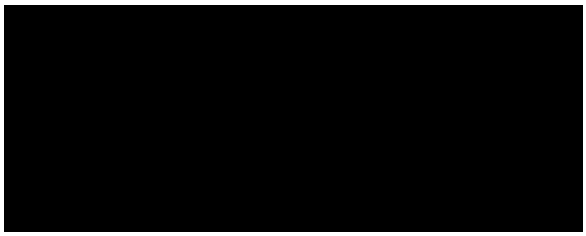
2.3 Analisi delle componenti principali

Per evitare i casi di multicollinearità, avendo correlazioni molto elevate, attraverso la funzione `findCorrelation` implementata nella libreria : `{caret}`.

Come si evince dalla Figura 2, la distribuzione ...



(a) Distribuzione Variabile X



(b) Scatterplot X vs Y

Figura 2: 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

Di seguito riportiamo lo script utilizzato per l'analisi.

```r
# Caricamento librerie
library(ggplot2)
data <- read.csv("dataset.csv")

# Analisi preliminare
summary(data)

```