Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per le tencologie e le scienze

METODI STATISTICI PER L'ANALISI DI DATI SULLA FORMULA 1

Relatore: prof. Erlis Ruli
Dipartimento di Scienze Statistiche

Laureando: Riccardo Schiavon
Matricola n. 2003666

Anno Accademico 2022/2023

Indice

Capitolo 1

Introduzione alla Formula Uno

La Formula Uno è la più alta classe di competizioni automobilistiche che coinvolge vetture monoposto a ruote scoperte. Questa categoria è ufficialmente regolamentata e governata dalla Federazione Internazionale dell'Automobile (FIA). L'organizzazione e la gestione della competizione sono, ad oggi, sotto la responsabilità del "Formula One Group", che è controllato da Liberty Media. La categoria della Formula Uno ha avuto origine nel 1948 ed è diventata una competizione mondiale a partire dal 1950. La parola "Formula"[1] si riferisce a un insieme di normative stringenti a cui tutti i partecipanti, comprese le vetture e i piloti, devono aderire rigorosamente. Queste regole sono state introdotte per garantire uniformità e prevenire eccessive differenze tecniche tra le vetture, limitare il loro sviluppo e, soprattutto, massimizzarne la sicurezza. Le scuderie partecipanti competono per cercare di vincere il Campionato, che ha luogo ininterrottamente, con cadenza annuale, dal 1950. Questo Campionato si svolge solitamente da fine marzo a dicembre in circuiti situati nei cinque continenti del mondo. Ogni scuderia è formata da due piloti titolari e in palio ci sono due titoli: il Titolo Costruttori e il Titolo Piloti assegnati alla scuderia e pilota con il punteggio più alto alla fine della stagione.
Nel corso della sua storia, la Formula 1 ha attraversato numerose evoluzioni, le principali riguardanti due aspetti fondamentali in questo sport: l'aerodinamica e il motore. L'aerodinamica è la scienza che si occupa dello studio del comportamento dell'aria e delle forze che questa esercita sui corpi statici o in movimento rispetto ad essa.[2] Nel contesto della Formula 1, l'aerodinamica ha un ruolo cruciale nella progettazione e nell'ottimizzazione della forma delle vetture con l'obiettivo di massimizzare la performance. Il motore rappresenta un altro pilastro fondamentale nell'evoluzione di questo sport. L'evoluzione dei motori è stata caratterizzata principalmente da variazioni sostanziali nella cilindrata e nella tecnologia utilizzata, per migliorarne le prestazioni, l'efficienza e l'affidabilità. E' di nostro interesse specificare i grandi cambiamenti al regolamento a partire dalla stagione 2014 con l'inizio dell'era ibrida. Il cambiamento più radicale ed influente è stato l'introduzione delle cosiddette 'Power Unit' (PU) ibride, compiendo un passo verso la sostenibilità e l'efficienza energetica. Le nuove power unit ibride hanno sostituito i tradizionali motori V8 aspirati con motori V6 turboalimentati, accoppiati a sistemi di recupero dell'energia (ERS) che sfruttano l'energia termica e cinetica proveniente dalla vettura. Questo approccio ibrido ha portato a un notevole aumento dell'efficienza dei motori, riducendo contemporaneamente l'impatto ambientale delle gare. Altri cambiamenti "secondari"

[1] Wikipedia, Formula 1; `https://it.wikipedia.org/wiki/Formula_1`
[2] Ralph-DTE; Aerodinamica (approfondimenti); `https://www.ralph-dte.net/aerodinamica`

hanno influenzato aspetti come la distribuzione del peso delle vetture, gli pneumatici e il sistema frenante.[3] L'era ibrida si è conclusa al termine della stagione 2021. Nel 2022 c'è stato un altro significativo cambiamento nell'approccio tecnico. Il focus si è spostato verso l'implementazione dell'effetto suolo, una tecnologia che mira a migliorare la stabilità delle vetture al suolo attraverso un miglioramento della pressione aerodinamica e una maggiore aderenza alla pista.
Negli ultimi decenni, l'uso dei dati è diventato l'elemento centrale nel mondo della Formula 1. Con l'evoluzione della tecnologia e dei sistemi di rilevamento dati avanzati, tutte le categorie di motorsport hanno intrapreso un percorso di trasformazione digitale senza precedenti. Una delle collaborazioni cruciali in questo ambito è stata la partnership tra la Formula Uno e AWS[4] (Amazon Web Services). AWS ha fornito una piattaforma avanzata per l'analisi dei dati in tempo reale. Combinando questi dati con i "trackside data" (dati provenienti da sensori piazzati in tutto il circuito) si consente agli ingegneri e ai data scientist delle scuderie di raccogliere, elaborare e interpretare una quantità straordinaria di dati utili su molti aspetti tra i quali usura gomme e consumo carburante, parametri cruciali del motore, telemetria della vettura e variazioni delle condizioni meteorologiche lungo il tracciato.
Questa sempre più crescente dipendenza dai dati nella Formula 1 apre un'evidente connessione con una disciplina scientifica, la statistica. La statistica può svolgere un ruolo cruciale nella gestione e nell'analisi di questa enorme quantità di dati. In particolare, può contribuire in modo significativo in molti ambiti, per esempio con lo sviluppo di modelli predittivi. L'obiettivo della presente Tesi è fornire un'analisi statistica e sviluppare alcuni metodi statistici avanzati per l'analisi di dati sulla Formula 1, tra cui l'applicazione di algoritmi di machine learning e di modellistica predittiva.

La seguente Tesi si articola in cinque capitoli. Il Capitolo 1 è dedicato all'introduzione alla Formula Uno. Nel Capitolo 2 verrà affrontata la fase di pre-processing che comprende la gestione dei dati e creazione dei dataset utilizzati per le analisi. Seguono, nel Capitolo 3, alcune analisi esplorative dei dati a disposizione. Il Capitolo 4 è dedicato all'applicazione di metodi di classificazione per i piloti che saliranno o meno sul podio. Infine, il Capitolo 5 affronta alcuni modelli statistici per analizzare, modellare e prevedere i tempi sul giro durante un Gran Premio.

Le analisi dei dati sono state eseguite utilizzando il software R (versione 4.2.0) e, talvolta, Python (versione 3.9).

[3] F1ingenerale; Categoria Formula 1; `https://f1ingenerale.com/`
[4] Amazon Web Services (AWS); `https://aws.amazon.com/it/sports/f1/`

Capitolo 2

Pre-processing dei dati

Nel seguente Capitolo verranno inizialmente descritte le principali fonti di variabilità nel mondo della Formula 1 e successivamente le strutture dei datasets creati per le analisi svolte nei capitoli seguenti.

2.1 Analisi della variabilità in un Gran Premio

La Formula Uno rappresenta uno degli sport più sofisticati al mondo. È caratterizzato da una complessità sorprendente dovuta a un'immensa quantità di variabili che influenzano ogni aspetto della competizione. Le principali fonti di variabilità in un Gran Premio riguardano i piloti, le vetture e le condizioni della pista.
Un Gran Premio con il formato tradizionale si svolge interamente in un weekend e prevede tre sessioni di Prove Libere, tra il venerdi e il sabato, che permettono ai team di acquisire dati sulle prestazioni delle vetture per ricercare l'assetto migliore. Al sabato segue la sessione di qualifica dove si determina l'ordine di partenza dei piloti nella gara della domenica.
Uno tra gli aspetti più fondamentali nell'evoluzione di una gara è la strategia. La strategia di gara include molteplici componenti, tra cui la gestione dei pit stop, la scelta e gestione delle gomme, la pianificazione delle soste e la gestione del carburante da parte del pilota. La strategia di gara viene pianificata prima della gara per ogni pilota sulla base di tutti i dati raccolti dalla squadra. Tuttavia molte decisioni devono essere prese in tempo reale durante l'evolversi della gara e ogni scelta sbagliata può avere un impatto decisivo sul risultato finale.
Durante i Campionati presi in considerazione, Pirelli, l'unica fornitrice di pneumatici per la Formula Uno, ha offerto diverse opzioni di mescole di gomma. Vengono codificate dalla A1, la più resistente, alla A7, la più morbida (in ordine Superhard, Hard, Medium, Soft, Supersoft, Ultrasoft, Hypersoft) [1]. Per le condizioni bagnate inoltre sono disponibili pneumatici intermedi per piste umide o leggermente bagnate e pneumatici da bagnato estremo noti come "Full Wet". Tranne che per il 2018 dove sono state offerte tutte le opzioni di mescole elencate, negli altri anni sono state messe a disposizione dalle tre alle cinque mescole, in aggiunta a quelle da bagnato. Le mescole più dure consentono ai piloti di percorrere più giri e di conseguenza la possibilità di fare meno soste ai box. Tuttavia, queste mescole non consentono tempi sul giro veloci come consentono le mescole più morbide.
Un'altra fonte di variabilità che può decidere l'esito di un Gran Premio per un pilota

[1] F1 Timing Database `https://github.com/TUMFTM/f1-timing-database`

è l'affidabilità della vettura. Come già evidenziato, l'affidabilità ha un ruolo di notevole importanza nella determinazione dei risultati in una gara, infatti un guasto in un qualsiasi componente cruciale come il motore, il cambio o il sistema elettronico, può portare al ritiro del pilota dalla gara.
Un altro fenomento di variabilità all'interno delle gare di Formula 1 è rappresentato dall'ingresso in pista della Safety Car o del segnale di Virtual Safety Car, o in casi estremi di bandiera rossa e interruzione della gara. La Safety Car è la vettura di sicurezza che si posiziona davanti al pilota di testa e attende che tutto il gruppo si ricompatti in attesa che la pista venga dichiarata libera. La Virtual Safety Car, invece, è un sistema virtuale che non prevede l'ingresso in pista di alcuna autovettura di sicurezza ma richiede ai piloti di comportarsi come se ci fosse, e di conseguenza rallentare.[2] La presenza della Safety Car o del segnale di Virtual Safety Car comporta quindi un rallentamento della velocità delle vetture e questo apre alla possibilità di differenziare le strategie, in quanto una sosta ai box per il cambio gomma farebbe perdere meno tempo. Sono tra i fenomeni meno prevedibili di questo sport ma possono stravolgere l'esito di una gara.
Infine, le condizioni meteorologiche rappresentano un'altra importante fonte di variabilità. La temperatura dell'aria e dell'asfalto, la presenza di pioggia e talvolta il vento sono tutti fattori che possono avere un impatto significativo sulle prestazioni delle vetture e sulla strategia di gara. Nonostante gli ingegneri siano in grado di monitorare in tempo reale gli aggiornamenti meteorologici, la decisione di effettuare un pit stop per sostituire le gomme da asciutto con gomme da bagnato o viceversa, non è del tutto scontata. Spesso infatti, la decisione finale spetta al pilota, che basandosi sulla sua esperienza e sensibilità, può optare per una strategia diversa rispetto a quanto suggerito dall'ingegnere.

2.2 Creazione dei datasets

Per svolgere tutte le seguenti analisi, sono stati creati vari datasets. Il primo di questi, a cui si farà riferimento con il nome di "Podium", considera i dati relativi ai Gran Premi dal 2014 al 2021 dove ogni osservazione rappresenta i dettagli delle prestazioni di una vettura in una gara di Formula 1. Questo dataset è una rielaborazione dei dati disponibili nell'Ergast Developer API[3] che è un servizio web sperimentale che fornisce un record storico dei dati sulle corse automobilistiche.

Il secondo dataset, a cui si farà riferimento con il nome di "Race", contiene tutti i tempi sul giro dei Gran Premi svolti dal 2014 al 2021. Questo dataset è stato ottenuto attraverso un'accurata rielaborazione. In particolare, si è ottenuto con l'unione di due sotto-dataset, il primo contenente i dati dal 2014 al 2019 ottenuti con una revisione del database "F1 Timing"[4], sviluppato dal Prof. Alexander Heilmeier presso l'Institute of Automotive Technology. F1 Timing utilizza dati storici nel campo del motorsport provenienti dal servizio web sperimentale Ergast, che fornisce accesso a tali informazioni. Il secondo sotto-dataset comprende i dati delle stagioni 2020 e 2021, disponibili direttamente dal sito web dell'Ergast Developer API.

[2] Wikipedia; `https://it.wikipedia.org/wiki/Safety_car`
[3] Ergast Developer API; `http://ergast.com/mrd`
[4] Alexander Heilmeier; `https://github.com/TUMFTM/f1-timing-database`

Il terzo e ultimo dataset, "Weather", verrà approfondito nel paragrafo 1.3 e comprende i dati del dataset "Race" nei Gran Premi dal 2018 al 2021 ma con l'aggiunta di alcune variabili meteorologiche.

2.2.1 Descrizione del dataset "Podium"

Il dataset Podium è stato creato per condurre alcune analisi esplorative dettagliate relative alle statistiche dei piloti, delle scuderie e dei circuiti che verranno approfondite nel Capitolo 2. Inoltre è il dataset utilizzato per le analisi svolte nel Capitolo 3. Nella Tabella 2 vengono descritte tutte le variabili del dataset Podium con la specificazione del loro tipo (numerica, dummy o categoriale).

2.2.2 Descrizione del dataset "Race"

Il dataset Race è stato creato per condurre alcune analisi esplorative e verrà utilizzato per le analisi svolte nel Capitolo 4. Nella Tabella 3 vengono descritte tutte le variabili del dataset Race con la specificazione del loro tipo (numerica, dummy o categoriale).

2.2.3 Descrizione del dataset "Weather"

Il dataset Weather è stato creato per svolgere alcune analisi esplorative dettagliate su variabili atmosferiche e successivamente per le analisi nei paragrafi 2.3 e 4.2.3. Si sottolinea che i dati meteorologici disponibili tramite l'API FastF1, sono relativi alle stagioni 2018-2021 e non comprendono tutte le stagioni tenute in considerazione negli altri datasets. Nella tabella 1 vengono riportate le variabili d'interesse atmosferiche in aggiunta alle variabili già presenti nel dataset Race.

	Nome	**Tipo**	**Descrizione**
1	*AirTemp*	numerica	Temperatura dell'aria (in gradi Celsius)
2	*Humidity*	numerica	Umidità relativa (in percentuale)
3	*Pressure*	numerica	Pressione atmosferica (in hPa)
4	*TrackTemp*	numerica	Temperatura dell'asfalto (in gradi Celsius)
5	*Rainfall*	dicotomica	Assume valore 1 se piove durante il giro in corso
6	*WindSpeed*	numerica	Velocità del vento (in m/s)

Tabella 1: Descrizione delle variabili del dataset Weather

Tabella 2: Descrizione delle variabili del dataset Podium

	Nome	**Tipo**	**Descrizione**
1	*Poleman*	dicotomica	Assume valore 1 se il pilota ha effettuato la Pole Position
2	*Circuit*	categoriale	Nome del circuito dove si è svolta la gara
3	*Last_race_FoP*	dicotomica	Assume valore 1 se il pilota è andato a podio nella gara precedente
4	*DriverRef*	categoriale	Nome del pilota
5	*grid*	numerica	Posizione nella griglia di partenza
6	*Team*	categoriale	Nome della scuderia
7	*Podio*	dicotomica	Variabile riposta assume 1 se la vettura è terminata sul podio in quella gara
8	*Podiums*	numerica	Numero di podi stagionali del pilota fino alla gara precedente
9	*Last_pos*	numerica	Posizione del pilota nella classifica finale al termine della stagione precedente
10	*Last_team_pos*	numerica	Posizione della scuderia nella classifica finale al termine della stagione precedente
11	*Sc_vsc*	dicotomica	Assume valore 1 se nella gara c'è stato regime di Virtual o Safety Car
12	*Year*	numerica	Anno stagionale del Gran Premio

	Nome	**Tipo**	**Descrizione**
1	*Driver*	categoriale	Cognome del pilota (sigla di tre lettere)
2	*Circuit*	categoriale	Nome del circuito
3	*LapTime*	numerica	Tempo sul giro (in secondi)
4	*LapNumber*	numerica	Numero del giro della gara
5	*Pit_out*	dicotomica	Assume valore 1 se il pilota effettua il giro dopo la sosta
6	*Pit_in*	dicotomica	Assume valore 1 se il pilota al termine del giro effettua la sosta
7	*Compound*	categoriale	Mescola di gomma
8	*TyreLife*	numerica	Vita della gomma (in giri)
9	*Sc_vsc*	dicotomica	Assume valore 1 se il giro è effettuato sotto regime di Virtual/Safety Car
10	*Team*	categoriale	Scuderia del pilota
11	*Year*	numerica	Anno stagionale a cui fa riferimento il giro

Tabella 3: Descrizione delle variabili del dataset Race

Capitolo 3

Analisi esplorative

3.1 Piloti e scuderie dal 2014 al 2021

Per rendere più chiara l'analisi, viene inizialmente fornita, in Tabella 4, una panoramica dei piloti titolari e delle scuderie che hanno partecipato ai Campionati nel periodo di interesse. I dati utilizzati si riferiscono agli otto Campionati svolti tra il 2014 e il 2021, periodo noto nella Formula Uno come "era ibrida". Una scuderia in particolare è riuscita a conquistare tutti i Titoli disponibili dal 2014 al 2020 e si tratta della Mercedes. Tuttavia, nel 2021, sebbene la Mercedes abbia ottenuto il Titolo Costruttori, il Titolo Piloti è stato vinto, in modo controverso all'ultimo giro dell'ultima gara, dal pilota di punta della Red Bull, Max Verstappen.
Nella Formula Uno il numero di scuderie partecipanti può variare da stagione a stagione, recentemente il numero è rimasto attorno alla decina. Tra le scuderie principali che hanno una presenza costante in Formula 1 negli ultimi decenni possiamo citare la Scuderia Ferrari, la Mercedes-AMG Petronas Formula One Team, la Red Bull Racing, la McLaren e molti altri team storici. In aggiunta ai due piloti titolari esiste anche la figura del "terzo pilota" all'interno delle scuderie, che può svolgere un ruolo importante durante la stagione. Il terzo pilota può essere impiegato per effettuare test privati o in pista, nelle sessioni di prove libere del venerdì al posto di un pilota titolare o come riserva nel caso in cui uno dei piloti titolari non possa partecipare a una gara. Questo garantisce una flessibilità e copertura alla scuderia in caso di imprevisti.
Il numero di gare disputate in ciascuna stagione è variabile, per esempio la stagione 2014 ha previsto un numero complessivo di 19 gare mentre il 2021 è stato il Campionato più lungo di sempre, ad oggi, con 22 gare complessive.

Tabella 4: Piloti e Scuderie dal 2014 al 2021

Scuderia	**Stagione**	**Piloti**
Scuderia Ferrari	2014-2021	Raikkonen (2014-2018), Alonso (2014), Vettel (2015-2020), Leclerc (2019-), Sainz (2021-)
Mercedes-AMG	2014-2021	Hamilton (2013-), Bottas (2017-2021), Rosberg (2010-2016), Russel (2020-)
Red Bull Racing	2014-2021	Verstappen (2016-), Ricciardo (2014-2018), Vettel (2014), Kvjat (2015-2016), Albon (2019-2020), Perez (2021-)
McLaren	2014-2021	Ricciardo (2021-), Button (2010-2017), Alonso(2015-2018), Magnussen (2014-2015), Vandoorne (2016-2018), Sainz(2019-2020), Norris(2019-)
Williams Racing	2014-2021	Massa (2014-2017), Bottas(2014-2016), Di Resta (2017), Stroll (2017-2018), Sirotkin (2018), Kubica (2019), Russell (2019-2021), Latifi (2020-)
Toro Rosso (AlphaTauri dal 2020)	2014-2021	Vergne (2014), Kvyat (2014, 2016- 2017, 2019-2020), Sainz (2015-2017), Verstappen (2015-2016), Hartley (2017- 2018), Gasly (2017-2021), Albon (2019), Tsunoda (2021)
Force India (Racing Point 2019-2020 e Aston Martin 2021)	2014-2021	Perez (2014-2020), Hulkenberg (2014-2016, 2020), Ocon (2017-2018), Stroll (2019-), Vettel (2021-)
Haas F1 Team	2016-2021	Grosjean (2016-2020), Gutierrez (2016), Magnussen (2017-2020), Mazepin (2021), Schumacher (2021-)
Sauber (Alfa-Romeo Sauber dal 2018)	2014-2021	Sutil (2014), Gutierrez (2014), Nasr (2015-2016), Ericsson (2015 2018), Wehrlein (2017), Leclerc (2018), Raikkonen (2019-2021), Giovinazzi (2017,2019-2021)
Lotus F1 Team (Renault 2016-2020 e Alpine 2021)	2014-2021	Grosjean (2014-2015), Maldonado (2014-2015) Magnussen (2016), Palmer (2016-2017), Hulkenberg (2017-2019), Sainz (2017-2018), Ricciardo (2019-2020), Ocon (2020-), Alonso (2020-)
Caterham	2014	Ericsson, Lotterer, Stevens
Marussia-Manor	2014-2016	Chilton (2014), Bianchi (2014), Stevens (2015), Rossi (2015), Mehri (2015), Wehrlein (2016), Haryanto (2016), Ocon (2016)

La dicitura "anno-" indica che il pilota fa attualmente parte della rispettiva Scuderia nella stagione 2022.

3.2 Analisi e statistiche per piloti, scuderie e circuiti

Svolgeremo ora delle analisi esplorative per esaminare alcune statistiche principali riguardanti piloti, scuderie e circuiti nel periodo di interesse. L'obiettivo principale è mettere in luce piloti e scuderie che hanno ottenuto le performance più significative in quest'era ibrida.
Poichè l'analisi svolta nel capitolo successivo riguarda il concetto del podio nella Formula Uno, ci concentreremo ora a esaminare in dettaglio questa dinamica specifica tramite delle analisi esplorative. Per determinare i piloti e le scuderie più performanti sono stati effettuati tre diagrammi a barre. Il primo, a sinistra in figura 1, mette in evidenza le cinque scuderie con più podi conquistati. Il secondo, a destra in figura 1, mette in evidenza i 10 piloti con più podi ottenuti in questo periodo.
Come già accennato, i due grafici evidenziano il netto dominio di una scuderia in particolare, la Mercedes. In particolare si nota il contributo di uno dei piloti più forti della storia della Formula Uno, Lewis Hamilton, Campione del Mondo per ben sei anni in questo periodo (2014, 2015, 2017, 2018, 2019 e 2020) battuto solamente nel 2016 dal suo compagno di squadra Nico Rosberg e nel 2021 dal pilota rivale della Red Bull, Max Verstappen. Si nota inoltre un rendimento discreto per la Ferrari che, nonostante non conti su nessun Titolo vinto, ha un numero di podi conquistati molto simile alla Red Bull. L'unico pilota fuori dai top3 Team capace di rientrare nella classifica è stato Felipe Massa, pilota della scuderia Williams nel periodo 2014-2017 capace di conquistare in totale sei podi. Inoltre, Sergio Perez, pilota dal 2014 al 2020 della scuderia Force-India poi diventata Racing-Point ha conquistato ben sette podi in questo periodo e altri cinque nella stagione 2021 guidando per la scuderia Red Bull Racing, al fianco del campione del mondo Max Verstappen. Infine, mettendo in relazione i piloti e le scuderie, in figura 2, si evidenzia il contributo per pilota nella conquista dei podi per le tre migliori scuderie (Mercedes, Ferrari e Red Bull).

Il dataset Race ha permesso, inoltre, di determinare un'altra statistica interessante sulla prestazione dei piloti in qualifica e in gara. Nel grafico a sinistra, in figura 3, viene raffigurata la posizione di partenza mediana dei piloti presi in esame precedentemente. Nel grafico a destra, invece, si determina la posizione mediana occupata dai piloti durante le gare.
Si evidenzia ancora una volta il netto dominio della Mercedes in particolare con i due piloti di punta, Lewis Hamilton e Nico Rosberg. Da questi due grafici si può notare inoltre come, apparte Sebastian Vettel che ha lottato con Lewis Hamilton per il Titolo nel 2018, la Red Bull abbia delle statistiche vantaggiose rispetto alla Ferrari.

Figura 1: Numero di podi conquistati dalle scuderie e dai piloti, 2014-2021

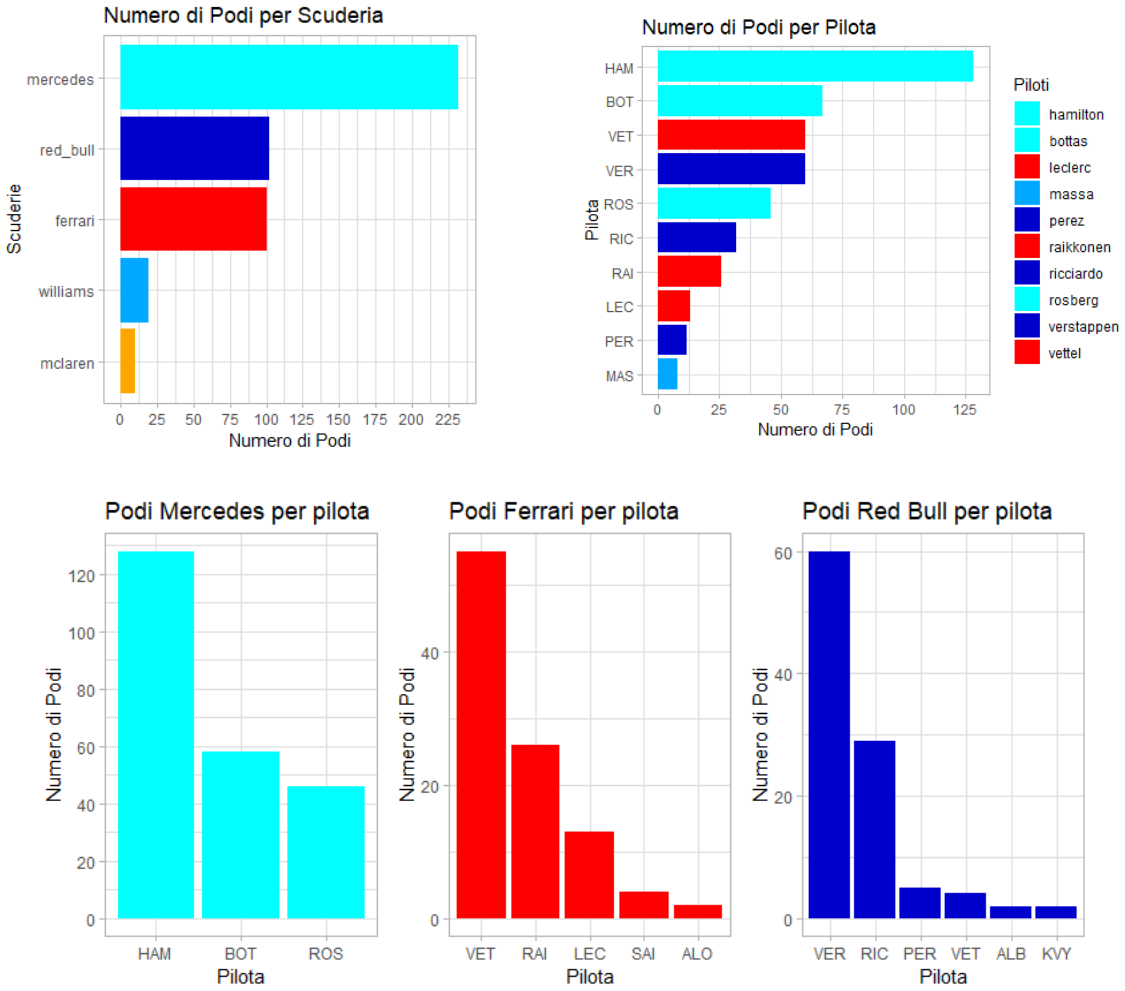

Figura 2: Numero di podi conquistati dai piloti per scuderia, 2014-2021

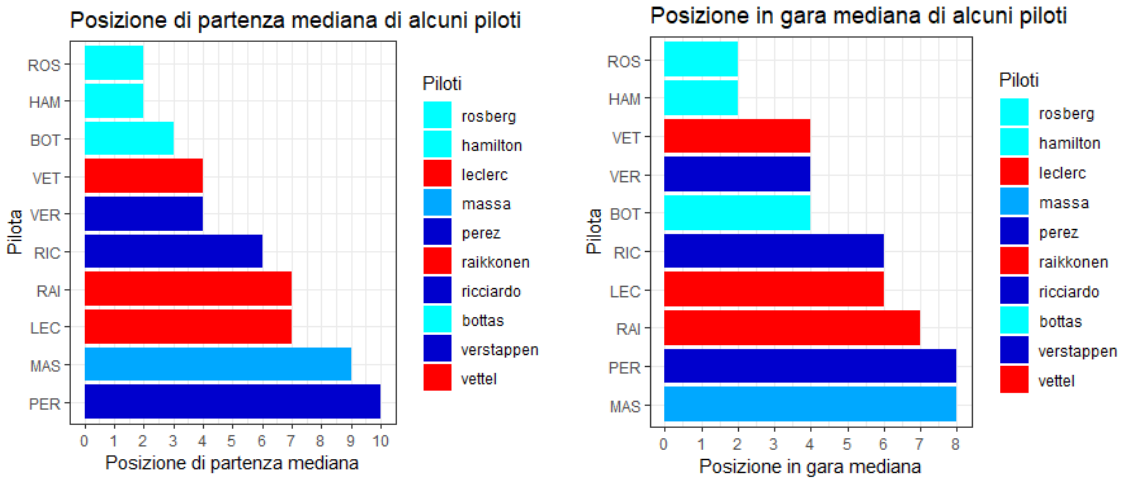

Figura 3: Posizione mediana di partenza e in gara di alcuni piloti, 2014-2021

La successiva analisi riguarda in modo spcifico la posizione di partenza che, come già anticipato, potrebbe risultare molto importante nell'ottica del piazzamento finale in gara. Viene quindi svolta una particolare analisi statistica per determinare in quali circuiti la qualifica è risultata essere più discriminante. In Figura 4, viene visualizzato il tasso di conquista dei podi in gara in relazione ai piloti che partivano dalla Pole Position, cioè dalla prima posizione. Sono stati considerati tutti i circuiti dal 2014 al 2021 in cui si sono corsi almeno quattro Gran Premi in questo periodo. I circuiti con tasso più basso sono rispettivamente l'Hockenheimring in Germania, l'Hermanos Rodriguez in Messico e Baku in Azerbaijan. Ciò significa che in questi tre circuiti conquistare la Pole Position non è garanzia della conquista del podio in gara. Al contrario, si può notare chiaramente che in alcuni circuiti la Pole Position è di grande aiuto nel conquistare il podio in gara. I circuiti con il tasso al 100% sono sono rispettivamente Villeneuve in Canada, Suzuka in Giappone, Spa-Francorchamps storico circuito in Belgio, Interlagos in Brasile, Austin in America e il cricuito del Bahrain. Da queste analisi è evidente come la posizione di partenza sia una tra le variabili più discriminanti per la conquista del podio.
Per approfondire i tre circuiti con tasso minore elencati in precedenza viene rappresentato, in figura 5, il confronto tra il tasso già analizzato (Pole Position-Podio) e un nuovo tasso, Top3-Podio, per determinare se nei circuiti in cui la Pole Position non è garanzia di podio in gara sia tuttavia fondamentale partire tra le prime tre posizioni. Si evidenzia come il tasso aumenti considerevolmente nei circuiti in Germania e in Messico, mentre nel circuito di Baku l'aumento è significativo tuttavia la Top3 in qualifica non certifica l'arrivo a podio in gara.

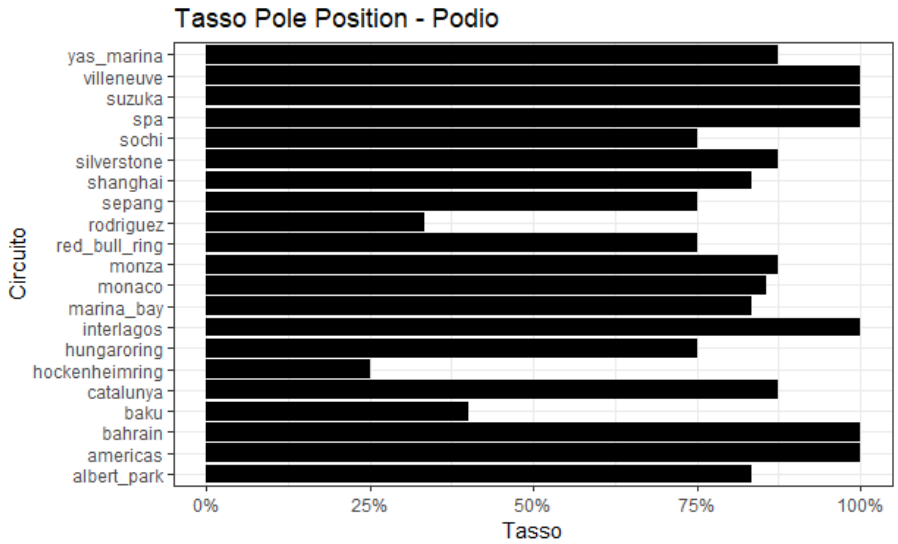

Figura 4: Proporzione di podi partendo dalla Pole Position, per circuito, 2014-2021

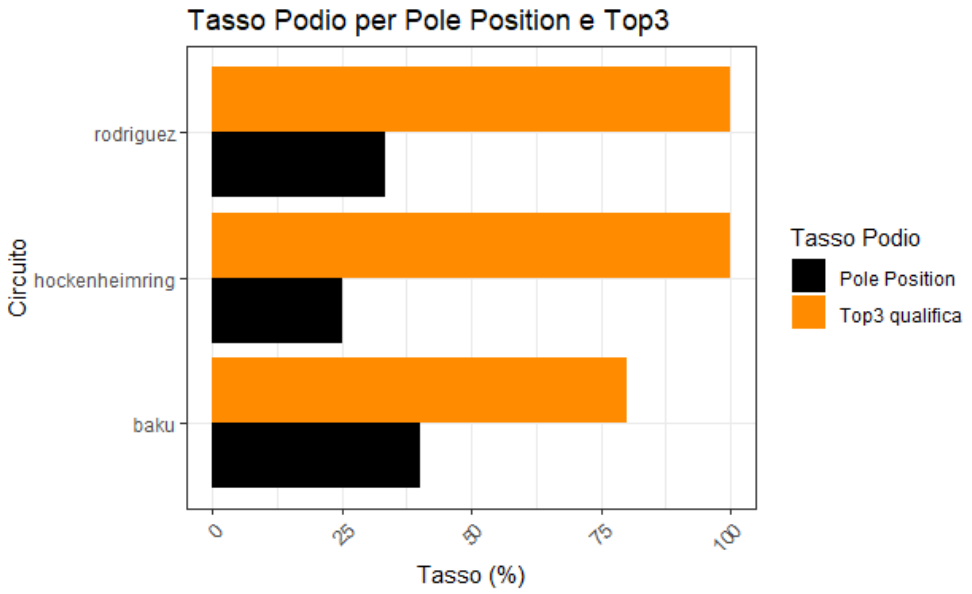

Figura 5: Confronto proporzioni di podi partendo dalla Pole Position e dalle prime tre posizioni, per ogni circuito, 2014-2021

Per approfondire maggiormente l'analisi per il circuito di Baku e per valutare che incidenza ha la posizione di qualifica sulla conquista del podio, al netto di altre variabili, è stato stimato un modello di regressione logistica. Esso assume la variabile risposta dicotomica "podio" che assume valore 1 se il pilota ha terminato la gara a podio e 0 altrimenti. L'unica variabile esplicativa inserita nel modello è la posizione di partenza del pilota.

La struttura del modello è la seguente:[1]

$$g(\pi_i) = \beta_0 + \beta_1 x_i$$

dove $g(\cdot)$ rappresenta la funzione di legame (logit), π_i rappresenta la probabilità che podio = 1 per l'i-esima unità statistica, x_i rappresenta la posizione del pilota alla partenza e i coefficienti di regressione sono rappresentati da $\beta = (\beta_0, \beta_1)$.
Il valore del coefficiente stimato β_1 = -0.31 indica che la probabilità di raggiungere il podio è moltiplicata per un fattore approssimativamente pari a 0.73 ($e^{-0.83}$) per ogni incremento unitario della posizione di partenza . Vengono riportate in Tabella 5 le stime puntuali delle probabilità di arrivare a podio per le prime 10 posizioni di partenza. Si nota come la probabilità non sia elevata già dalla prima posizione di partenza e scenda all'aumenare della posizione in griglia di partenza.

[1]Salvan, A., Sartori, N., and Pace, L. (2020). Modelli Lineari Generalizzati. Springer Milan.

Posizione di partenza	Baku Probabilità stimata
1	0.57
2	0.49
3	0.41
4	0.34
5	0.27
6	0.21
7	0.17
8	0.13
9	0.09
10	0.07

Tabella 5: Probabilità di podio nel Gran Premio di Baku per le prime 10 posizioni in griglia di partenza.

3.3 Analisi di dati meteorologici

Nel mondo della Formula Uno, l'analisi delle condizioni meteorologiche durante il weekend di gara è estremamente importante. Tra tutti i fattori che possono influenzare un weekend di gara, la pioggia è senza dubbio il più significativo. Altri fattori atmosferici da non trascurare sono le temperature, l'umidità, la pressione atmoferica e l'intensità del vento. Procederemo quindi con un'analisi dei dati meteorologici dei Gran Premi dal 2018 al 2021, in quanto si hanno a disposizione tutte le informazioni elencate solamente per queste stagioni. La prima analisi svolta si concentra per determinare una possibile relazione tra la temperatura dell'aria e la temperatura dell'asfalto. Si è quindi realizzato un grafico di dispersione tra la temperatura media dell'asfalto in relazione alla temperatura media dell'aria per stabilire l'esistenza di un'eventuale relazione come sarebbe logico aspettarsi. Dalla figura 6, si nota infatti un discreto trend lineare: al crescere della temperatura dell'aria cresce anche quella dell'asfalto. Inoltre, a sostegno di questa prima osservazione, è stato calcolato il coefficiente di correlazione di Pearson per valutare la relazione lineare ed è risultato positivo pari a 0.73 indicando una discreta relazione tra le due variabili.
La successiva analisi riguarda invece un altro possibile fattore, l'umidità. Sono stati realizzati due grafici di disperione tra l'umidità e le due temperature già menzionate, per determinare eventuali relazioni. Sia dai grafici in figura 7 che dai valori dei coefficienti di correlazione calcolati, si evidenzia che al contrario dell'analisi precedente c'è una relazione negativa. La relazione maggiore risulta tra l'umidità e la temperatura dell'asfalto con un coefficiente di correlazione pari a -0.68.

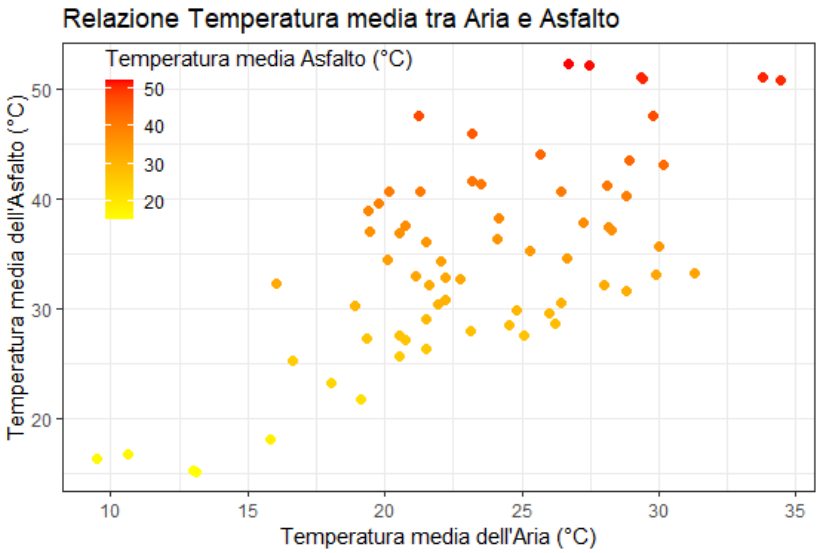

Figura 6: Grafico di dispersione tra temperatura media dell'aria e dell'asfalto, 2018-2021

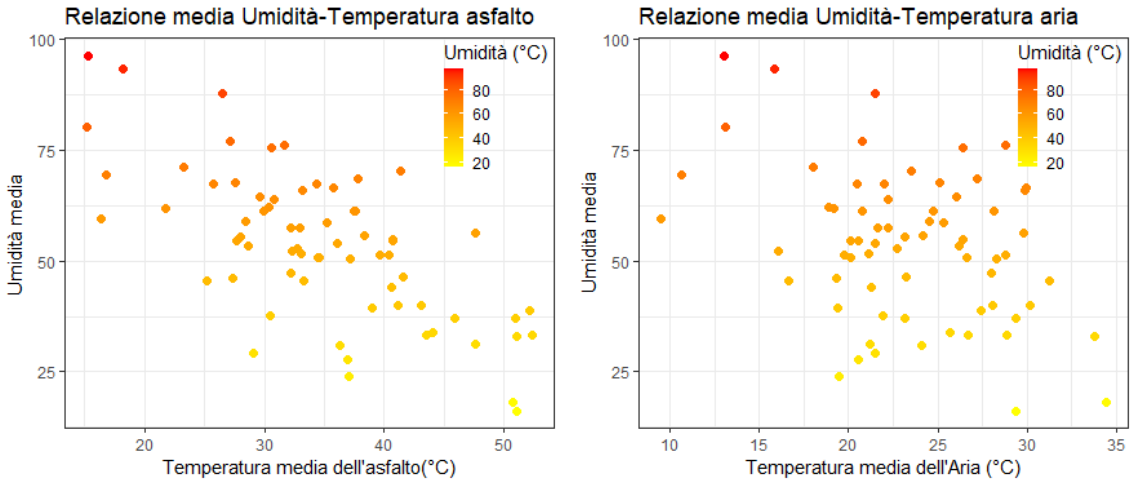

Figura 7: Confronto grafici di dispersione tra umidità e temperatura, 2018-2021

La temperatura dell'asfalto è riconosciuta come uno dei principali fattori in grado di condizionare l'usura degli pneumatici in una vettura di Formula Uno. La seguente analisi mette a confronto la temperatura media dell'asfalto nel corso dei giri percorsi nel 2019 in relazione alle varie mescole di gomma utilizzate.

I boxplot in figura 8 evidenziano che le gomme più dure ("Hard") sono utilizzate con temperature mediamente maggiori rispetto alle altre mescole più morbide. Com'era intuibile, infatti, la mescola più dura è in grado di sopportare maggiormente le temperature elevate. Si nota che l'utilizzo della mescola più dura avviene con un range

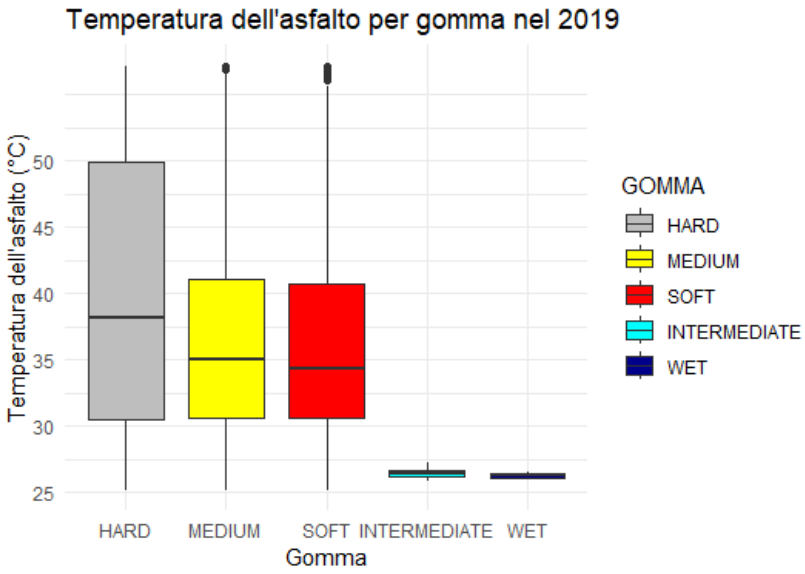

Figura 8: Boxplot della temperatura dell'asfalto per ogni mescola di gomma nel 2019

di temperatura dell'asfalto maggiore rispetto alle altre mescole. Essa offre presumibilmente un'ottima prestazione, durata e consistenza rispetto alle altre mescole in condizioni di temperatura elevata e al tempo stesso in condizioni di temperature basse una buona tenuta con basso degrado a discapito di un po' di performance rispetto alle altre mescole più morbide

3.4 Analisi esplorative del dataset "Race"

In quest'ultimo paragrafo, vengono fornite alcune analisi esplorative del dataset Race che contiene tutti i tempi sul giro effettuati da tutti i piloti nelle gare dal 2014 al 2021. Il Gran Premio di riferimento per le seguenti analisi grafiche è il Gran Premio d'Italia che si è svolto a Monza e ha visto protagonista il pilota della Ferrari, Charles Leclerc, capace di portare la Scuderia Ferrari nel gradino più alto del podio.
Il grafico di dispersione in figura 9 rappresenta i tempi sul giro di Leclerc durante la gara. Ogni punto sul grafico rappresenta un singolo giro e il colore dei punti è associato alle diverse mescole di gomma utilizzate. Questo grafico permette di osservare il passo gara e la strategia adottata dal pilota della Ferrari. Si sottolinea che si sono trascurati i giri effettuati sotto regime di Safety Car o Virtual Safety Car e i giri influenzati dai pit-stop per una miglior rappresentazione.

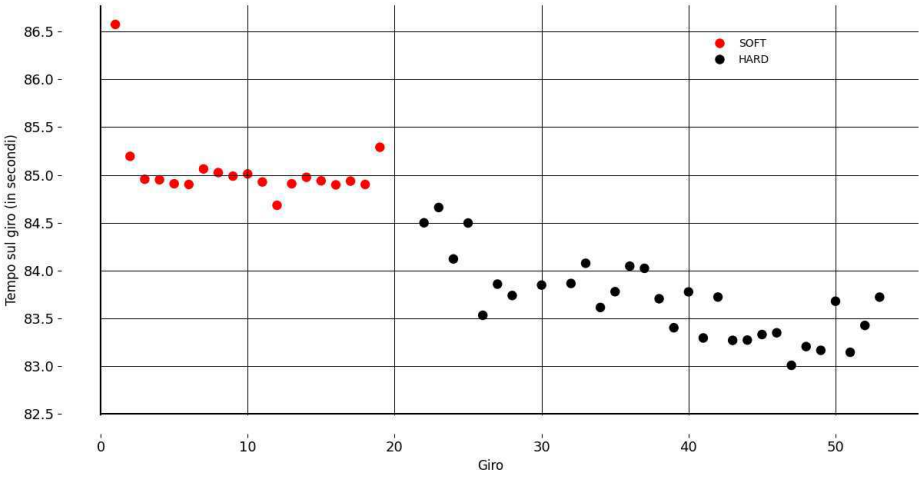

Figura 9: Distribuzione dei tempi sul giro di Leclerc, Monza 2019

Il secondo grafico, in figura 10, rappresenta la distribuzione dei tempi sul giro dei sei piloti più veloci durante il Gran Premio d'Italia (da sinistra a destra: Leclerc, Bottas, Hamilton, Ricciardo, Hulkenberg, Albon). Questo grafico consente di confrontare le prestazioni dei piloti evidenziando anche la distribuzione dei tempi sul giro effettuati da ciascun pilota per ogni mescola di gomma utilizzata.
Il grafico in figura 11 mostra invece il cambiamento delle posizioni, giro per giro, per ogni pilota durante la gara. Si ottiene un'ottima rappresentazione dinamica delle variazioni nelle posizioni dei piloti durante l'intera gara con un confronto della classifica finale presente a destra del grafico.

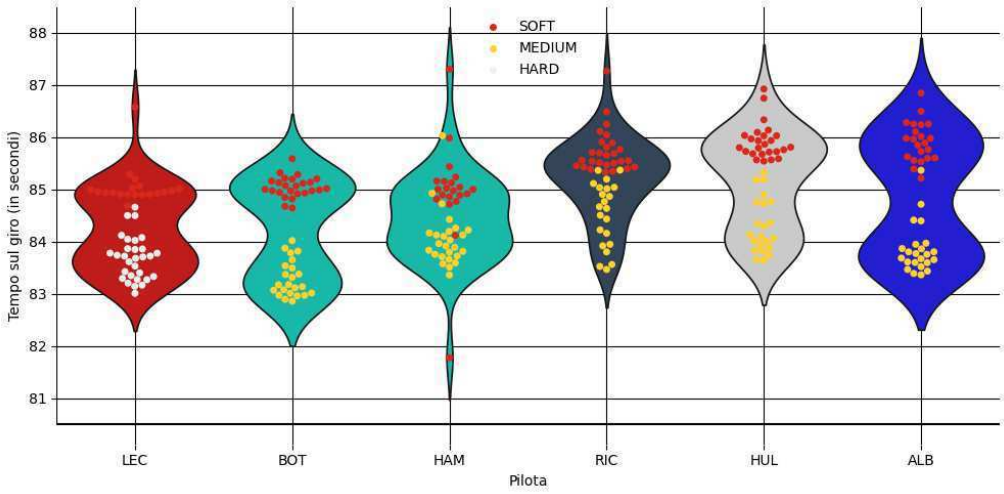

Figura 10: Violin-plot della distribuzione dei tempi sul giro dei piloti più veloci, Monza 2019

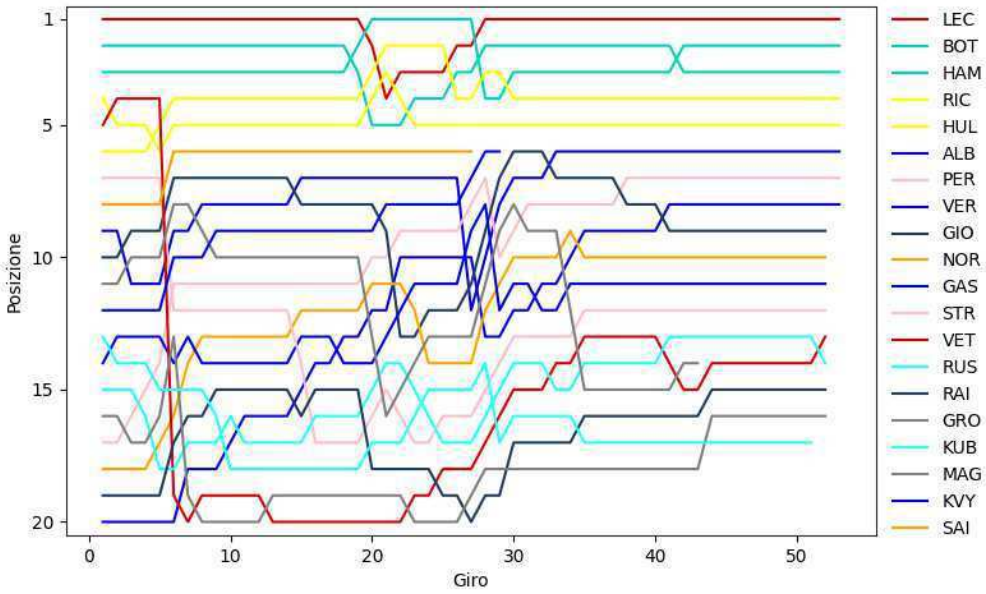

Figura 11: Cambiamento delle posizioni per ogni pilota, Monza 2019

Capitolo 4

Metodi di classificazione

Nel seguente capitolo, affronteremo un aspetto cruciale dell'analisi dei dati nel contesto delle competizioni di Formula 1: la classificazione del podio in una gara. L'obiettivo è identificare correttamente i podi dei Gran Premi del 2021, dei quali si conoscono i risultati, utilizzando i dati disponibili dal 2014 al 2020. Per raggiungere questo obiettivo verranno utilizzati due metodi di classificazione. Verrà prima introdotto l'albero di classificazione per poi approfondire la nostra analisi con un metodo più avanzato, la Random Forest.

4.1 Albero di classificazione

Il primo metodo di classificazione che verrà affrontato è l'albero di classificazione. La scelta di utilizzare una tecnica non parametrica è dovuta a una serie di vantaggi come la gestione di variabili categoriali, la flessibilità rispetto a tecniche parametriche, la robustezza e alla riduzione del rischio di overfitting.
Una strategia semplice per approssimare qualsiasi funzione consiste nell'utilizzare una funzione a gradini. L'idea è analoga per gli alberi di regressione, tuttavia negli alberi di classificazione la funzione incognità è rappresentata da $p(x)$, dove l'indicatore di gruppo vale $Y = 0$ o $Y = 1$. La funzione $p(x) = \mathbb{P}\{Y = 1 \mid X = x\}$ viene approssimata mediante [1]

$$\hat{p}(x) = \sum_{j=1}^{J} P_j I\left(x \in R_j\right)$$

con $P_j \in (0, 1)$ e appresenta la probabilità che Y=1 per la regione R_j. Questi P_j sono ancora stimati da

$$\hat{P}j = M\left(y_i : x_i \in R_j\right) = \frac{1}{n_j} \sum i \in R_j I\left(y_i = 1\right)$$

cioè la frequenza relativa di elementi 1 nella regione R_j. Poichè Y_i è dicotomica (assume valore 1 o 0), il criterio di accostamento sarà la minimazzione della devianza. Tuttavia, nei problemi di classificazione è spesso usato anche il numero di casi erratamente classificati come misura di discrepanza al posto della minimizzazione della devianza. Avendo a disposizione un numero considerevole di osservazioni si procederà con la crescita dell'albero su un insieme di stima (*"train set"*) e verrà minimizzata la devianza calcolata sull'insieme di verifica (*"test set"*) che comprende

[1] Azzalini, A. and Scarpa, B. (2012). Data analysis and data mining: An introduction. OUP USA.

le osservazioni non incluse nel *train set*. L'algoritmo inizia con un procedimento di ottimizzazione "passo a passo", procedendo per suddivisioni successive, in cui l'insieme originario viene diviso in due parti, le quali a loro volta saranno divise in due e così iterativamente. In questo modo l'albero sta "crescendo" e si formerebbe un albero binario con J nodi terminali (detti foglie). Procedendo così indefinitamente si otterrebbe un albero con n foglie e quindi per evitare l'overfitting bisogna ricorrere a una potatura dell'albero per ridurre a $J < n$ foglie. Il criterio per la selezione di J è quindi la minimizzazione della devianza. La devianza negli alberi di classificazione è, a meno di una costante, una media del grado Q di impurità delle foglie pesata con le numerosità delle foglie, infatti [2]

$$D = 2n \sum_j \frac{n_j}{n} Q\left(\hat{P}_j\right)$$

dove l'impurità delle foglie è misurata dall'entropia o come nel nostro caso con l'indice di Gini

$$Q\left(P_j\right) = \sum_{k=0,1} P_{jk}\left(1 - P_{jk}\right).$$

La rappresentazione grafica finale dell'albero è una struttura di tipo gerarchica con un nodo alla radice dal quale si ramificano i collegamenti agli altri nodi e le foglie sono i nodi terminali. Ogni nodo rappresenta uno "split" determinato da una variabile e divide l'albero in due. Di conseguenza, le variabili utilizzate per le suddivisioni sono considerate le più importanti nel determinare la variabile di risposta. Sarà quindi altrettanto interessante capire quali sono state le variabili più discriminanti nella nostra analisi.

4.1.1 Costruzione del dataset "Tree"

Il dataset "Tree", utilizzato in questo capitolo per affrontare la classificazione, deriva dal dataset Podium. Le variabili prese in considerazione e inserite nel modello sono indicate e descritte in Tabella 6. Si assume la variabile risposta "podio" (Y), dove

$$Y = \begin{cases} 1 & \text{il pilota ha terminato la gara a podio} \\ 0 & \text{altrimenti} \end{cases}$$

[2] Canale, A. Metodi statistici per i big data, "Problemi di classificazione", 2022/2023

Tabella 6: Descrizione delle variabili del dataset Tree

	Nome	**Tipo**	**Descrizione**
1	*poleman*	dicotomica	Assume valore 1 se il pilota ha effettuato la Pole Position
2	*circuit*	categoriale	Nome del circuito dove si è svolta la gara
3	*last_race_FoP*	dicotomica	Assume valore 1 se il pilota è andato a podio nella gara precedente
4	*driverRef*	categoriale	Nome del pilota
5	*grid*	numerica	Posizione nella griglia di partenza
6	*team*	dicotomica	Nome della scuderia del pilota
7	*podio*	dicotomica	Variabile riposta assume 1 se il pilota ha terminato la gara a podio
8	*podiums*	numerica	Numero di podi stagionali del pilota fino alla gara precedente
9	*last_pos*	numerica	Posizione del pilota nella classifica finale al termine della stagione precedente
10	*last_team_pos*	numerica	Posizione della scuderia nella classifica finale al termine della stagione precedente

4.1.2 Addestramento dell'albero di classificazione e risultati

Per addestrare il modello sono stati utilizzati i dati delle stagioni dal 2014 al 2020 e l'obiettivo è prevedere i podi dei Gran Premi del 2021 trascurando l'ordine dei piloti delle prime tre posizioni. I parametri dell'albero sono stati configurati con una cross-validation in modo da bilanciare l'accuratezza e precisione della classificazione ed evitare l'overfitting (un adattamento eccessivo ai dati) che renderebbe il modello inaffidabile per la previsione di nuovi dati futuri.
Di seguito, verranno esaminati i principali risultati emersi dall'albero di classificazione. Come primo risultato viene fornita la matrice di confusione generata dalle previsioni del modello rispetto ai risultati effettivi dei Gran Premi del 2021 ed offre una panoramica generale delle prestazioni del modello.

	Previsione: 0	Previsione: 1
Effettivo: 0	353	20
Effettivo: 1	20	46

Matrice di confusione dell'albero di classificazione

Si riportano di seguito alcune osservazioni sui risultati ottenuti dalla matrice di confusione e il grafico della curva ROC ottenuta:

- il modello classifica correttamente 399 osservazioni (353+46) su un totale di 439, il tasso di corretta classificazione è pari al 90.88%
- il modello classifica erratamente 40 osservazioni (20+20), il tasso di errata classificazione è di circa 9.12%
- il modello prevede in totale 64 podi, di cui 46 effettivi e 18 errati. La precisione, l'indice che misura la frazione di previsioni corrette del modello, è di circa il 72%
- la sensibilità, l'indice che misura la capacità del modello di identificare correttamente tutti i casi "positivi" (podio), è circa pari al 70%
- la specificità, l'indice che misura la capacità del modello di identificare correttamente tutti i casi "negativi" (non podio), è pari al 94.66%
- l'indice F1-Score che è una media armonica tra precisione e sensibilità e fornisce una misura complessiva delle prestazioni del modello con valori compresi tra 0 e 1, vale 0.707

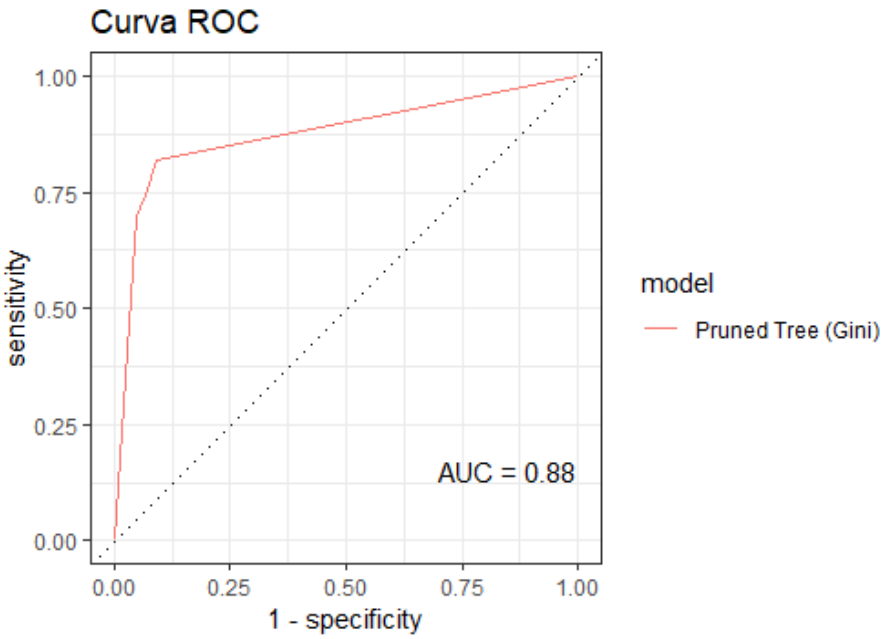

Figura 12: Curva ROC

In generale, il modello sembra avere delle buone prestazioni, con un tasso di corretta classificazione elevato e una bassa errata classificazione. Il modello sembra essere valido e con buone capacità predittive, inoltre la curva ROC, con un valore di AUC pari a 0.88, evidenzia una medio-alta capacità di discrimazioni tra le due classi.
Viene riportato, in Tabella 7, l'elenco di tutti i Gran Premi della stagione 2021 e successivamente, in Tabella 8, tutti i podi delle gare del 2021 previsti dal modello affiancati ai risultati reali. Vengono riportate le sigle dei nomi dei piloti contrassegnati dal colore verde se la previsione del modello è corretta e dal colore rosso se la previsione del modello è errata. Si sottolinea che il Gran Premio di Stiria ("STI") si è svolto sullo stesso circuito del Gran Premio d'Austria, nel Red Bull Ring, ma nel 2021 si sono corsi effettivamente due Gran Premi distinti.

Round	**Nome del Gran Premio**	**Località**	**Sigla**
1	*Gran Premio del Bahrain*	Bahrain	BHR
2	*Gran Premio dell'Emilia-Romagna*	Imola (Italia)	IMO
3	*Gran Premio del Portogallo*	Portimao	POR
4	*Gran Premio di Spagna*	Catalunya	SPA
5	*Gran Premio di Monaco*	Monte Carlo	MON
6	*Gran Premio d'Azerbaigian*	Baku	AZE
7	*Gran Premio di Francia*	Le Castellet	FRA
8	*Gran Premio di Stiria*	Spielberg	STI
9	*Gran Premio d'Austria*	Spielberg	AUS
10	*Gran Premio di Gran Bretagna*	Silverstone	GBR
11	*Gran Premio d'Ungheria*	Budapest	UNG
12	*Gran Premio del Belgio*	Spa-Francorchamps	BEL
13	*Gran Premio d'Olanda*	Zandvoort	OLA
14	*Gran Premio d'Italia*	Monza	ITA
15	*Gran Premio di Russia*	Soči	RUS
16	*Gran Premio di Turchia*	Istanbul	TUR
17	*Gran Premio degli Stati Uniti*	Austin	USA
18	*Gran Premio di Città del Messico*	Città del Messico	MEX
19	*Gran Premio di San Paolo*	San Paolo (Brasile)	BRA
20	*Gran Premio del Qatar*	Doha	QAT
21	*Gran Premio d'Arabia Saudita*	Gedda	ARB
22	*Gran Premio di Abu Dhabi*	Yas Marina	ABU

Tabella 7: Elenco dei Gran Premi del Campionato 2021

BHR-T	BHR	IMO-T	IMO	POR-T	POR	SPA-T	SPA
BOT	BOT	HAM	HAM	HAM	HAM	BOT	BOT
VER	HAM	VER	VER	BOT	BOT	HAM	HAM
HAM	VER	LEC	NOR	VER	VER	VER	VER

MON-T	MON	AZE-T	AZE	FRA-T	FRA	STI-T	STI
LEC	VER	LEC	PER	HAM	HAM	BOT	BOT
VER	NOR	VER	VET	BOT	VER	HAM	HAM
BOT	SAI	HAM	GAS	VER	PER	VER	VER

AUS-T	AUS	GBR-T	GBR	UNG-T	UNG	BEL-T	BEL
HAM	NOR	HAM	HAM	HAM	HAM	BOT	BOT
VER	VER	BOT	BOT	BOT	SAI	HAM	HAM
BOT	BOT	VER	LEC	VER	OCO	RIC	RUS

OLA-T	OLA	ITA-T	ITA	RUS-T	RUS	TUR-T	TUR
BOT	BOT	HAM	NOR	HAM	HAM	BOT	BOT
HAM	HAM	BOT	BOT	NOR	VER	VER	VER
VER	VER	VER	RIC	SAI	SAI	HAM	PER

USA-T	USA	MEX-T	MEX	BRA-T	BRA	QAT-T	QAT
BOT	PER	BOT	PER	HAM	HAM	BOT	ALO
HAM	HAM	HAM	HAM	BOT	BOT	HAM	HAM
VER	VER	VER	VER	VER	VER	VER	VER

ARB-T	ARB	ABU-T	ABU
BOT	BOT	HAM	HAM
HAM	HAM	VER	VER
VER	VER	NOR	SAI

Tabella 8: Podi previsti dal modello e reali per i Gran Premi del Campionato 2021. La sigla "-T" indica i risulati previsti dall'albero ("Tree")

Ricordando l'imprevidibilità di questo sport, i risultati sembrano complessivamente buoni e soddisfacenti. Il modello prevede correttamente l'intero podio finale per ben sette Gran Premi su 22. In altri 11 Gran Premi prevede correttamente due piloti su tre, in tre Gran Premi un solo pilota su tre a podio e infine solamente in un Gran Premio sbaglia interamente il podio. Analizzando in modo approfondito, le difficoltà sono state riscontrate principalmente nei Gran Premi di Monaco, Azerbaijan, Ungheria e a Monza, nel Gran Premio d'Italia. Nel Gran Premio di Monaco il modello sbaglia la classificazione di due piloti a podio per due situazioni imprevedibili. Infatti, prevede a podio il pilota della Ferrari Charles Leclerc che però non prenderà parte al Gran Premio poichè prima di schierarsi sulla griglia di partenza ha dovuto rinunciare alla gara a causa di un guasto al semiasse sinistro della sua vettura. Inoltre, la seconda previsione sbagliata dal modello riguarda il pilota della Mercedes Valtteri Bottas, il quale è stato costretto a ritirarsi per un problema alla gomma anteriore destra al momento della sosta.

Il Gran Premio di Azerbaijan del 2021 è stato uno dei Gran Premi più imprevedibili negli ultimi anni. Il modello pronostica a podio Lewis Hamilton, Max Verstappen e Charles Leclerc. La gara subisce un colpo di scena a quattro giri dalla fine quando il pilota della Red Bull, Max Verstappen, subisce l'esplosione della gomma posteriore destra, perde il controllo della vettura e termina la gara contro il muro ritirandosi. La direzione di gara decide di stoppare la gara con bandiera rossa e alla ripartenza Lewis Hamilton tenta il sorpasso per la prima posizione ma a causa di un severissimo bloccaggio delle ruote anteriori perde il comando della gara concludendo la gara penultimo. Successivamente lo stesso pilota ha affermato di aver commesso un errore alla ripartenza e di aver attivato un pulsante che sposta completamente il bilanciamento dei freni verso la parte anteriore, inducendolo all'inevitabile errore.
Nel Gran Premio d'Ungheria, la pioggia al via ha causato un incidente al primo giro che ha coinvolto ben nove piloti costringendo al ritiro quattro di questi (tra cui Bottas che veniva pronosticato a podio dal modello). Max Verstappen, coinvolto nell'incidente, ha subito un danno notevole nel fondo della vettura e ha concluso la gara solo in nona posizione.
Infine, il Gran Premio d'Italia è stato caratterizzato dallo spettacolare incidente tra i due contendenti al Titolo Mondiale, Hamilton e Verstappen, causando il ritiro di entrambi i piloti.
La variabile nettamente più rilevante e importante per l'albero di classificazione è risultata la posizione di partenza e conferma le precedenti osservazioni fatte sulla base delle analisi esplorative.

4.2 Random Forest

Nel paragrafo precedente, si è applicato un albero di classificazione per prevedere il podio dei piloti nelle gare di Formula 1. Tuttavia, nonostante i medio-buoni risultati forniti, è importante sottolineare che, come qualsiasi modello, ha i suoi limiti. In particolare, è molto elevato il rischio di overfitting e non è un modello particolarmente robusto. L'idea è quindi di esplorare una tecnica più avanzata di classificazione che estende i concetti dell'albero, la Random Forest.
La Random Forest ("foresta casuale") è algoritmo supervisionato di machine learning che estende il concetto di albero di classificazione, consentendo una maggiore precisione e robustezza. Infatti, combina i risultati di più alberi, per migliorare la capacità predittiva. Le foreste casuali hanno l'obiettivo di ridurre la varianza diminuendo la correlazione tra i vari alberi allenati.[3] Invece di considerare ad ogni passo tutte le variabili d'ingresso, durante la costruzione dell'albero le foreste casuali selezionano casualmente solo un sottoinsieme di queste variabili per la suddivisione di ciascun nodo e questo aiuta a ridurre la correlazione tra gli alberi migliorando le prestazioni del modello.[4]

4.2.1 Addestramento della Random Forest e confronto dei risultati

Il dataset utilizzato per implementare la Random Forest è lo stesso utilizzato per l'albero di classificazione singolo e le variabili sono già state specificate preceden-

[3] Hastie, T., Tibshirani, R., & Friedman, J. 2016. The Elements of Statistical Learning. Springer
[4] IBM; Random Forest; `https://www.ibm.com/it-it/topics/random-forest`

temente in Tabella 6. I parametri di controllo delle Random Forest configurati attraverso una cross-validation sono:

- il numero n_trees di alberi, che specifica quanti alberi decisionali vengono inclusi nell'algoritmo e aumentando il numero di alberi può aumentare la precisione ma comporta un maggiore costo computazionale
- il numero q di variabili che vengono considerate in ogni suddivisione (split)

I risultati ottenuti dall'addestramento della Random Forest sono molto simili a quelli ottenuti dall'albero di classificazione. Lo si evince dalla matrice di confusione che risulta molto simile alla matrice di confusione dell'albero singolo. Gli unici due miglioramenti ottenuti riguardano le previsioni del Gran Premio degli Stati Uniti, a Austin, e del Gran Premio del Messico.
Nel primo il podio viene previsto completamente corretto in quanto la Random Forest prevede correttamente a podio il pilota della Red Bull Sergio Perez al posto della previsione dell'albero che pronosticava a podio Valtteri Bottas.
Analogamente, nel Gran Premio del Messico viene previsto correttamente il podio di Perez al posto del pilota della Mercedes, Bottas (Vedi Tabella 9).
Per tutti gli altri Gran Premi non si sono evidenziati miglioramenti significativi. Questo suggerisce che, nonostante l'utilizzo di un modello più sofisticato, rimane comunque difficile la previsione a causa di molta imprevedibilità che caratterizza questo sport.

USA-RF	USA-T	USA	MEX-RF	MEX-T	MEX
HAM	HAM	HAM	PER	BOT	PER
VER	VER	VER	HAM	HAM	HAM
PER	BOT	PER	VER	VER	VER

Tabella 9: Podi previsti dal modello Random Forest, albero di classificazione e reali per i Gran Premi degli Stati Uniti e Messico, 2021. La sigla "-RF" indica i risultati previsti dalla Random Forest mentre "-T" i risulati previsti dall'albero ("Tree")

	Previsione: 0	Previsione: 1
Effettivo: 0	353	20
Effettivo: 1	18	48

Matrice di confusione della Random Forest

4.2.2 Importanza delle variabili

Per determinare i fattori che influenzano maggiormente le previsioni del modello viene attribuita un'importanza alle variabili presenti. Viene calcolata tramite l'accumulo dell'importanza attribuita a ciascuna variabile in ogni suddivisione effettuata per ciascun albero. Grazie a questo indice è stato possibile costruire un grafico a barre per visualizzare chiaramente questo concetto. In questo modo, la variabile con valore di importanza più elevato avrà un indice normalizzato pari a 1, mentre le altre variabili avranno valori compresi tra 0 e 1 in base alla loro importanza relativa rispetto alla variabile più importante (Figura 13).
Come si era già sottolineato per l'albero di classificazione, anche nella Random Forest la variabile *grid* risulta nettamente la più importante per prevedere un podio. Tra le altre variabili influenti spiccano in ordine di importanza *podiums*, *last_team_pos*, *last_pos* e infine, relativamente, anche *circuit*.

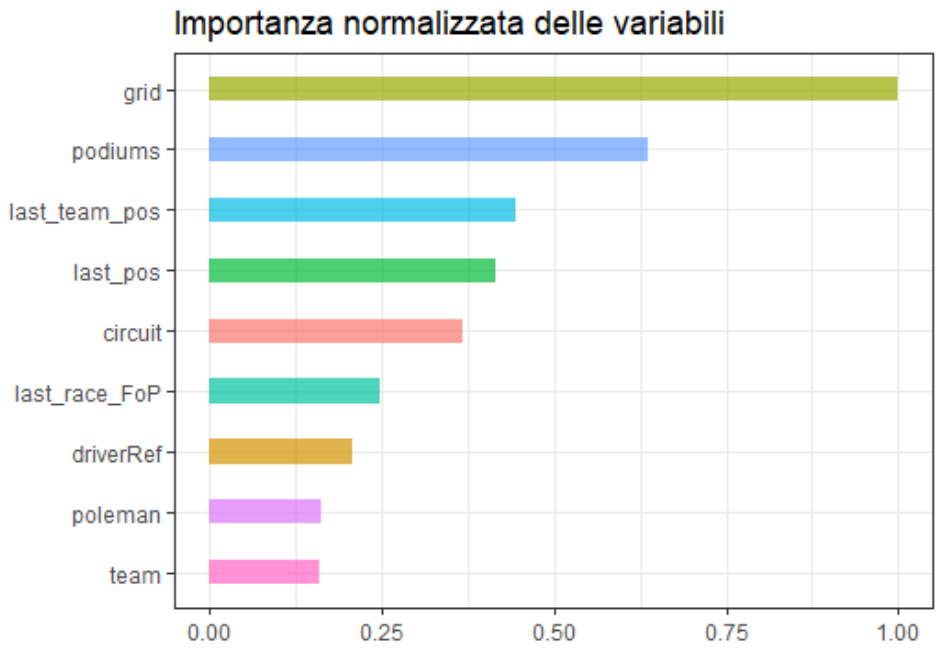

Figura 12: Importanza normalizzata delle variabili presenti nella Random Forest

Capitolo 5

Modellazione dei tempi sul giro

In quest'ultimo Capitolo verrà affrontata un'altra dinamica interessante e ben più analitica di questo sport. L'attenzione viene posta sulla variabile più rappresentativa degli sport a motori, il tempo sul giro. Verranno utilizzati i dataset "Race" e "Weather", ricordando che il primo dataset comprende i tempi sul giro relativi al periodo 2014-2021 mentre il secondo comprende i dati relativi solo al periodo 2018-2021. L'obiettivo è modellare i tempi sul giro e comprendere quali siano le variabili più discriminanti per questa dinamica. Un ulteriore obiettivo sarà quello di fornire una previsione, più accurata possibile, per i tempi sul giro per un determinato Gran Premio della stagione 2021, utilizzando i dati relativi agli anni precedenti.

Il tempo sul giro è la variabile che maggiormente riflette l'unione delle performance di piloti e vetture. Il suo valore è influenzato da una lunga serie di fattori, tra cui appunto le capacità dei piloti e la performance delle vetture, ma anche dalle strategie di gara e di conseguenza dalla gestione del carburante e delle gomme e, talvolta, dalle condizioni meteorologiche. L'insieme di tutti i tempi sul giro percorsi in una gara da un determinato pilota viene chiamato passo gara. Per comprendere al meglio questo concetto, viene proposto in figura 13 il confronto tra i passi gara di tre piloti (Lewis Hamilton, Max Verstappen e Lando Norris) nel Gran Premio di Russia svolto a Sochi nel 2021. La decisione di rappresentare questa gara viene da una serie di motivi. In primo luogo, questa gara comprende una piccola ma decisiva frazione di giri percorsi sotto la pioggia (dal giro 48) che ha stravolto e deciso il risultato finale, portando alla vittoria n°100 Lewis Hamilton, primo e unico pilota nella storia a raggiungere 100 vittorie. Questo grafico è molto rappresentativo del termine "passo gara" poichè mette in evidenzia l'insieme dei giri percorsi dai tre piloti in questione. Il pilota della Mclaren, Lando Norris, conduce da Leader la corsa fino al giro 48 quando improvvisamente inizia a piovere. L'intensità della pioggia inizialmente è debole e nessun pilota pensa all'eventuale cambio gomme a "soli" 6 giri al traguardo. Tuttavia, dopo soli due giri, l'intensità aumenta notevolmente portando a uno stravolgimento della gara. Lewis Hamilton decide di disobbedire il team ed effettuare la sosta per montare gomme intermedie, perfette in queste condizioni miste, al contrario di Lando Norris che decide di proseguire con le gomme d'asciutto fino a commettere un inevitabile errore che gli compromette la sua prima vittoria in carriera e consegna ad Hamilton la sua centesima vittoria storica.

Com'è logico aspettarsi, dal giro 48 con l'arrivo della pioggia e l'aumento dell'intensità, i tempi sul giro vengono nettamente influenzati alzandosi notevolmente e

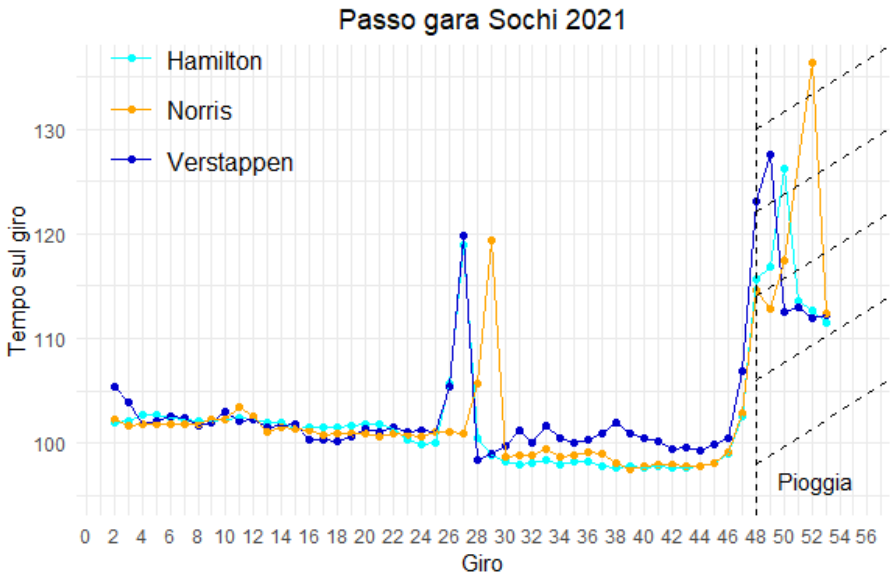

Figura 13: Distribuzione dei tempi sul giro, Russia 2021

dimostra come il possesso dei dati meteorologici risulti fondamentale in queste situazioni. Analizzando i giri precedenti all'arrivo della pioggia si evidenzia che nel primo stint di gara (parte di gara effettuata con una singola mescola di gomma) i tre piloti abbiano avuto un passo gara molto simile tra di loro, mentre dopo la sosta fino all'arrivo della pioggia Verstappen non è riuscito a tenere il passo dei primi due. Un'altra osservazione importante riguarda l'andamento dei tempi in quanto è evidente un andamento di tipo decrescente nell'arco della gara (escludendo l'ultima frazione influenzata dalla pioggia). Generalmente ciò è dovuto principalmente da due fattori: il peso della macchina che diminuisce con l'avanzare dei giri poichè il consumo del carburante permette alle vetture di alleggerirsi e dalla gommatura dell'asfalto, in quanto giro dopo giro l'aderenza della pista aumenta anche grazie al consumo delle gomme che rilasciano piccoli pezzi di gomma sull'asfalto aumentando il grip della pista. Di conseguenza, in generale si ottiene un abbassamento progressivo del tempo sul giro all'aumentare dei giri durante la gara. Per avvicinarci inizialmente al vasto mondo della modellistica si partirà come base di partenza da due modelli, la regressione lineare e la regressione robusta. Successivamente verranno proposti due modelli con differente approccio, parametrico e non parametrico.

5.1 Regressione lineare e robusta

Per le seguenti analisi si farà riferimento ad un altro Gran Premio del 2021, il Gran Premio degli Stati Uniti svolto a Austin. Questa decisione è dovuta al fatto che è stato uno dei Gran Premi più combattutti nella stagione dai rivali per il titolo, Hamilton e Verstappen, arrivando al traguardo distaccati da un solo secondo e con un passo gara pressochè identico. Il loro netto dominio in questa gara è dimostrato dal distacco subito da Sergio Perez che ha concluso in terza posizione a 42 secondi dal vincitore Max Verstappen. Ci concentreremo ora sul confronto del rispettivo passo gara.

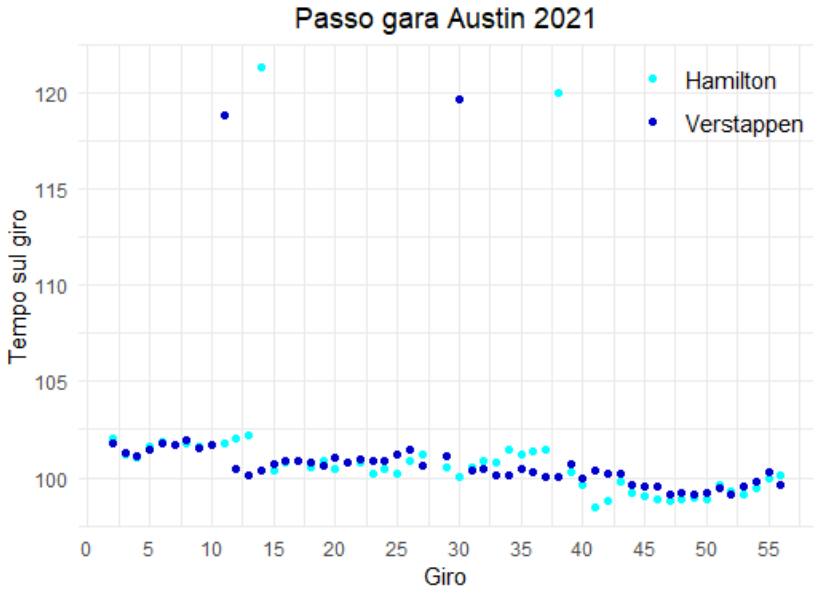

Figura 14: Distribuzione dei tempi sul giro, Austin 2021

Il grafico in figura 14 mostra il passo gara simile dei due piloti e mostra un andamento generale decrescente all'aumentare del numero dei giri di gara. I rispettivi due giri per pilota con tempo elevato corrispondono al giro influenzato dalla sosta effettuata e si può notare come in entrambe le soste Verstappen si sia fermato ai box anticipatamene rispetto a Hamilton effettuando l'undercut. Questa strategia consiste nell'anticipare la sosta, rispetto al rivale davanti, per il cambio gomme e sfruttare gli pneumatici nuovi per avere più prestazione nei giri successivi e superare il pilota rivale al momento della sua sosta. Questa strategia si è poi rivelata vincente.
Si considera inizialmente un modello di regressione lineare semplice per analizzare la relazione tra il tempo sul giro e il numero del giro di gara. La regressione lineare consente di stimare quanto l'incremento dei giri influenzi direttamente i tempi sul giro e se esiste una relazione lineare tra queste variabili. Si è consapevoli che tale modello non è rappresentativo ma è volto a confermare l'ipotesi che l'andamento di un passo gara all'aumentare dei giri sia decrescente. Vengono riportati in tabella 10 i coefficienti del modello di regressione lineare stimato dove ci si aspetterebbe, per quanto detto in precedenza, un valore del coefficiente relativo alla variabile sul numero del giro di gara negativo.

	Hamilton	Verstappen
Intercetta	102.967	102.988
LapNumber	-0.060	-0.065

Tabella 10: Modello di regressione lineare semplice per i tempi sul giro di Hamilton e Verstappen, Austin 2021

Il valore dell'intercetta rappresenta il tempo di base stimato e si osserva che per Hamilton è minore di due centesimi di secondo rispetto a Verstappen. Tuttavia, il coefficiente più importante è quello relativo al numero del giro di gara (LapNumber) ed è negativo per entrambi i piloti, quindi si stima che il tempo sul giro diminuisca all'aumentare dei giri di gara. In particolare si nota che per Verstappen il coefficiente risulta leggermente inferiore e ciò significa che in media, per ogni aumento unitario del numero del giro, ci si aspetta che i tempi sul giro di Verstappen si abbassino di una quantità leggermente superiore rispetto a quelli di Hamilton. Per visualizzare graficamente il modello di regressione adattato rappresentiamo lo stesso grafico precedente con sovrapposta la retta stimata della regressione lineare (Figura 15). A conferma dei coefficienti ottenuti e dal passo gara molto simile si ottiene una sovrapposizione delle due rette stimate.

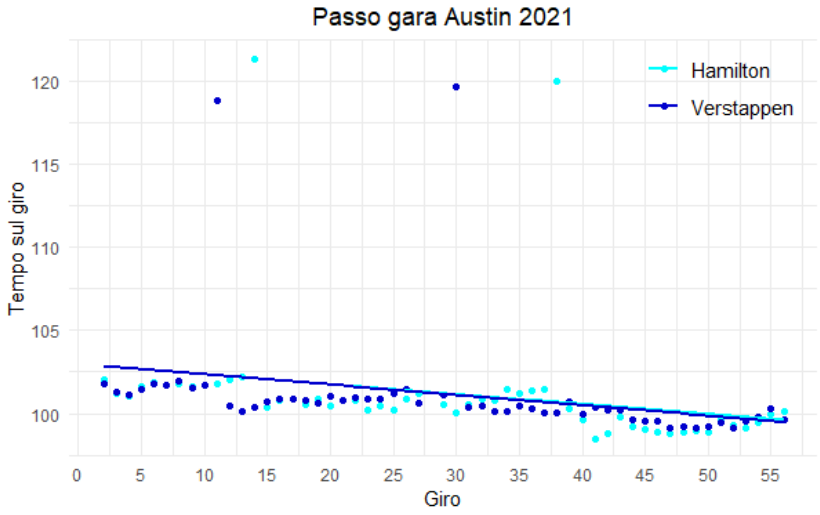

Figura 15: Distribuzione dei tempi sul giro e retta di regressione lineare stimata per Hamilton e verstappen, Austin 2021

Tuttavia, è importante sottolineare che un modello di regressione lineare semplice non rappresenti al meglio la dinamica dei tempi sul giro in una gara di Formula 1. Ci sono molte variabili importanti trascurate che influenzano le prestazioni dei piloti e che non sono state considerate. Inoltre, la regressione lineare è fortemente influenzata dai valori anomali ed è evidente che la retta non interpola bene le osservazioni poichè è "attratta" dalle due osservazioni anomale riferite ai giri delle soste ai box. Pertanto, la forma del modello sarà invariata ma procederemo ora con l'adattamento di un modello di regressione robusta volta a minimizzare l'effetto dei valori anomali sulle stime dei coefficienti di regressione. Essa, infatti, attribuisce meno peso a queste osservazioni anomale e cerca di fornire una stima più accurata dei parametri del modello, ottenendo delle stime più robuste e dei risultati più affidabili. Vengono riportate in Tabella 11 le stime dei coefficienti dei modelli adattati per entrambi i piloti e come in precedenza vengono rappresentate graficamente le rette di regressione robusta stimate (Figura 16).

Coefficienti	**Hamitlon**	**Verstappen**
Intercetta	102.01	101.78
LapNumber	-0.051	-0.043

Tabella 11: Modello di regressione robusta per i tempi sul giro di Hamilton e Verstappen, Austin 2021

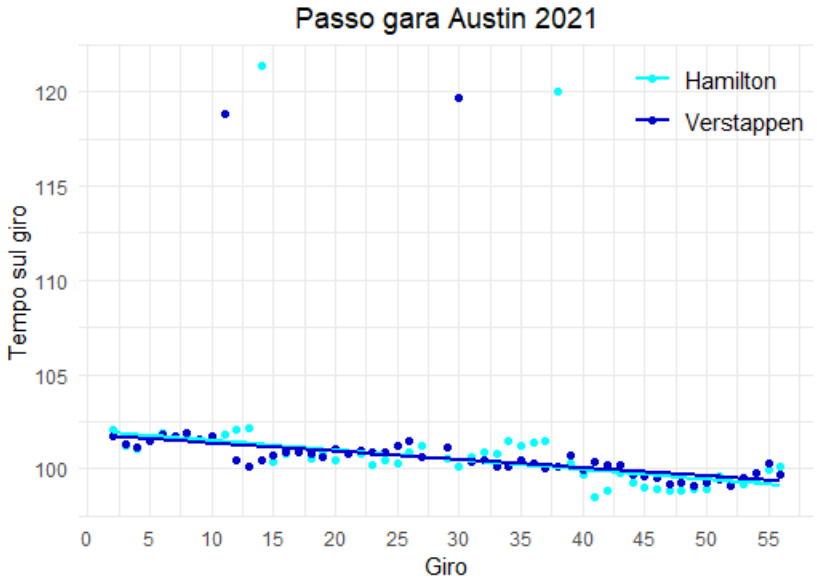

Figura 15: Distribuzione dei tempi sul giro e retta di regressione robusta stimata per Hamilton e verstappen, Austin 2021

I coefficienti stimati dalla regressione robusta e dalla regressione lineare sono differenti e ciò suggerisce che i due modelli forniscono stime diverse per la relazione tra il tempo sul giro e il numero del giro di gara. La prima osservazione riguarda l'intercetta stimata che ora risulta inferiore per Verstappen, al contrario di quanto

stimato in precedenza dalla regressione lineare. Inoltre, si evidenzia allo stesso modo che, a differenza dei risultati ottenuti nella regressione lineare, con la regressione robusta il coefficiente stimato relativo al numero dei giri risulta, seppur di poco, inferiore per Hamilton. Il grafico evidenzia un adattamento piuttosto simile per i due piloti, tuttavia nell'ultima frazione di gara Hamilton, costretto a rimontare alla ricerca della vittoria, sembra essere stato leggermente più veloce del rivale. L'interpolazione dei dati mediante la retta di regressione robusta è migliorata rispetto alla retta di regressione lineare indicando una maggiore robustezza nei confronti dei valori anomali. Tuttavia, è importante ricordare che anche questo modello di regressione non tiene conto di molte altre variabili fondamentali nello svolgimento di una gara. Proviamo ora a stimare per gli stessi dati un modello di regressione robusta tenendo conto delle altre variabili a disposizione riguardanti le mescole di gomma utilizzate, l'età della gomma e variabili relative alla sosta ai box. In seguito vengono riportati, in tabella 12, le stime dei coefficienti dei modelli ottenuti.

Coefficienti	**Hamilton**	**Verstappen**
Intercetta	101.21	101.28
LapNumber	-0.060	-0.048
Pit_ out	21.01	18.65
Pit_ in	0.35	0.46
Compound_ medium	0.038	0.024
Compound_ hard	-0.038	-0.025
TyreLife	0.08	0.03

Tabella 12: Modello di regressione robusta per i tempi sul giro di Hamilton e Verstappen, Austin 2021

Confrontando le stime ottenute dai due modelli, è possibile metterli a confronto con lo scopo di valutare gli effetti delle variabili considerate e identificare similitudini o differenze significative nelle loro prestazioni. L'intercetta stimata risulta molto simile e indica il tempo "base" stimato. Una prima differenza la si evince dai coefficienti ottenuti relativi alla variabile *LapNumber* in quanto, al netto di altre variabili, il tempo sul giro tende a diminuire più per Hamilton rispetto a Verstappen. Dalle due variabili relative alle soste ai box si evince come la variabile che più influenza il tempo sul giro è la variabile *Pit_ out* che rappresenta il giro all'uscita dalla sosta. Questa variabile, al netto delle altre, aumenta in media il tempo sul giro di circa 21 secondi per Hamilton e meno di 19 secondi per Verstappen. La differenza è probabilmente dovuta al minor tempo di esecuzione dei pit-stop per il pilota della Red Bull. Analizzando invece i coefficienti delle variabili relative alle gomme, i coefficienti stimati della variabile *TyreLife* indicano che al netto della mescola utilizzata, all'aumentare unitario dei giri di gara il tempo sul giro per Hamilton aumenta in media di circa otto centesimi di secondo mentre il tempo sul giro per Verstappen aumenta in media di soli tre centesimi di secondo. Verstappen, conducendo da leader per più della meta della corsa, ha probabilmente svolto più gestione gomma rispetto al pilota della Mercedes che, al contrario, ha svolto più giri in rimonta subendo presumibilmente più usura delle gomme. I coefficienti relativi alle mescole utilizzate invece indicano che Verstappen è stato in media più veloce con le gomme Medie nel

primo stint di gara mentre Hamilton è stato più performante negli altri due stint di gara con le gomme Hard nonostante non sia riuscito a completare la rimonta.

Vogliamo ora entrare nello specifico e trovare il modo per modellare i tempi sul giro singolarmente affrontando due diversi approcci. Il primo approccio, di natura più 'statistica', si concentrerà sull'adattamento di un modello parametrico. Questo offre il vantaggio di fornire stime interpretabili dei parametri, consentendo una migliore comprensione dei fattori in relazione ai tempi sul giro. Il secondo, invece, sarà un approccio di tipo non parametrico e sfrutterà un algoritmo di machine learning avanzato. Spesso questo metodo è più efficace nella fase predittiva rispetto ai modelli parametrici tradizionali grazie alla capacità di catturare relazioni complesse e non lineari nei dati.

5.2 Modelli Misti Lineari

I modelli lineari a effetti misti (Linear Mixed Models, LMM) rappresentano una classe di modelli statistici parametrici utilizzati per affrontare le situazioni in cui i dati presentano una struttura di tipo gerarchica. La caratteristica principale che contraddistingue questi modelli è la loro capacità di gestire contemporaneamente sia degli effetti fissi che degli effetti casuali. Gli effetti fissi sono in grado di catturare le relazioni sistematiche tra le variabili indipendenti e la variabile risposta, mentre gli effetti casuali catturano le variazioni casuali dovute a diversi raggruppamenti. Gli effetti casuali vengono trattati come variabili casuali e sono modellati tramite una distribuzione normale multivariata. Ciò permette di tener conto delle correlazioni possibili entro i gruppi e tra i gruppi in modo efficace.
In generale, un modello lineare normale con effetti misti può essere scritto nella forma:[1]

$$Y_{ij} = \mathbf{x}_{ij}\beta + z_{ij}\mathbf{u}_i + \varepsilon_{ij},$$

con β vettore p-dimensionale di effetti fissi, $\mathbf{u}_i \sim N_q(0, \Sigma_u)$ vettore q-dimensionale di effetti casuali, mentre marginalmente $\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$, indipendente da $\mathbf{u}_i$. Il modello prevede dunque che $E(Y_{ij}) = \mu_{ij} = \mathbf{x}_{ij}\beta$. Il termine $z_{ij}\mathbf{u}_i$ descrive la variabilità tra le osservazioni (o gruppi), mentre ε_{ij} descrive la variabilità interna tra le osservazioni (o gruppi). Nel nostro modello, l'effetto casuale sarà rappresentato dalle caratteristiche individuali relative al singolo pilota. Sarà fondamentale quindi catturare la possibile variabilità tra i piloti e soprattutto eventuali correlazioni nei tempi sul giro dello stesso pilota.

5.2.1 Adattamento del modello ai dati

Il dataset utilizzato per queste analisi è il dataset Race con alcune modifiche. É stato eliminato il primo giro di ogni gara per ciascun pilota poichè è evidentemente influenzato dalla procedura di partenza e inoltre sono stati rimossi tutti i giri sotto regime di Safety car o Virtual Safety Car. Le variabili inserite nel modello sono elencate in tabella 13.

[1] Salvan, A., Sartori, N., and Pace, L. (2020). Modelli Lineari Generalizzati. Springer Milan.

	Nome	Tipo	Descrizione
1	*Driver*	categoriale	Cognome del pilota (sigla di tre lettere)
2	*LapTime*	numerica	Tempo sul giro (in secondi)
3	*LapNumber*	numerica	Numero del giro della gara
4	*Pit_out*	dicotomica	Assume valore 1 se il pilota effettua il giro dopo la sosta
5	*Pit_in*	dicotomica	Assume valore 1 se il pilota al termine del giro effettua la sosta
6	*Compound*	categoriale	Mescola di gomma
7	*TyreLife*	numerica	Vita della gomma (in giri)
8	*Team*	categoriale	Scuderia del pilota

Tabella 13: Variabili incluse nel modello LMM

Inizialmente sono stati modellati i tempi realizzati durante il Gran Premio degli Stati Uniti del 2021. Alcuni dei risultati più significativi ottenuti dopo l'analisi vengono riportati in Tabella 14.

Coefficiente	Stima	Coefficiente	Stima
Intercetta	100.83	Pit_out1	20.46
Compound = MEDIUM	0.16	Pit_in1	1.83
Compound = SOFT	0.73	TyreLife	0.07
Team = Mercedes	0.99	Team = Ferrari	0.85
Team = McLaren	1.50	Team = Haas	4.45

Tabella 14: Alcune stime dei parametri del modello LMM

Si sottolinea che per la variabile Compound il livello di riferimento è la mescola HARD mentre per la variabile relativa alle scuderie (*Team*) il riferimento è la Red Bull. Di seguito si riportano le principali osservazioni dei parametri del modello stimati :

- analizzando le due variabili relative ai pit-stop, le stime risultano positive impattando il tempo sul giro in maniera evidente soprattutto con la variabile *Pit_out*
- al netto di tutte le altre variabili, la variabile *TyreLife* indica che all'aumentare unitario dell'età della gomma il tempo sul giro subisce un incremento poichè il coefficiente è risultato positivo
- focalizzandosi sui coefficienti relativi alla variabile *Compound* si ottegono dei risultati apparentemente inaspettati in quanto ci si potrebbe aspettare che al netto delle altre variabili le gomme più morbide (Medium e Soft) siano in grado di offrire prestazioni superiori in termini di tempo sul giro. Ciò però non accade probabilmente perchè le mescole Soft e Medium sono state utilizzate maggiormente nello primo stint di gara, con vettura più carica di carburante, e quindi questo dettaglio potrebbe aver influito maggiormente rispetto alla

morbidezza della gomma portando a tempi sul giro più alti rispetto ai giri effettuati con gomma Hard nella seconda parte di gara a vettura più leggera.

- analizzando invece i coefficienti della variabile *Team* si nota che tutti i coefficienti risultano positivi. Questo è dovuto dal fatto che il team di riferimento è la Red Bull e quindi che nessun team è risultato più veloce.

5.2.2 Analisi predittiva del modello

Per analizzare la capacità predittiva del modello, l'analisi dei dati relativi a tutti i Gran Premi svolti nel periodo 2014-2020 avrebbe richiesto risorse computazionali considerevoli. Di conseguenza, ci siamo focalizzati nuovamente su un'approfondita analisi del Gran Premio di Austin, utilizzando i dati di tutti i Gran Premi svolti in questo circuito. L'obiettivo sarà quindi quello di adattare un modello su questi dati e prevedere il passo gara di Lewis Hamilton a Austin nel 2021. Questa scelta ha permesso di semplificare il modello considerando solo i dati relativi ad una specifica pista ed eliminando la variabile che tiene conto dei vari circuiti. In figura 16 viene riportato il confronto tra i tempi sul giri previsti dal modello LMM e osservati.

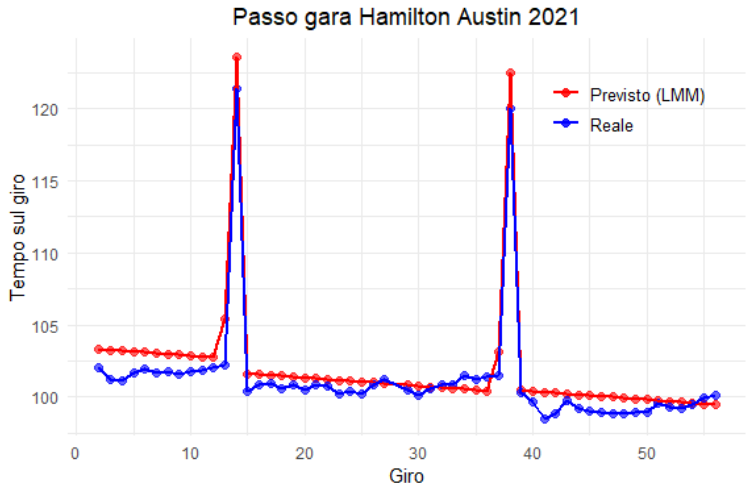

Figura 16: Confronto dei tempi sul giro di Hamilton previsti (in rosso) dal modello LMM e osservati (in blu), Austin 2021

Rispetto alle osservaznio reali le previsioni ottenute non sono del tutto soddisfacenti. Si nota un evidente distacco nel primo stint di gara ma anche nell'ultima frazione di gara. Si può notare anche come il trend sia continuamente discendente, non catturando il possibile fenomeno del degrado gomme il quale in un singolo stint di gara induce i tempi sul giro ad alzarsi lievemente. I modelli lineari a effetti misti, nonostante siano adatti per l'analisi di dati con struttura di tipo gerarchica, presentano delle limitazioni legate all'assunzione di linearità nelle variabili. Questo rappresenta un forte limite del modello soprattutto in ottica previsiva non catturando in modo ottimale le complesse relazioni nei dati. Per affrontare questa limitazione e provare

a ottenere una maggiore capacità predittiva implementeremo un modello di machine learning avanzato, noto come XGBoost (Extreme Gradient Boosting), capace di catturare relazioni non lineari nei dati mediante l'uso di algoritmi di boosting e alberi decisionali.

5.2.3 Estensione del modello utilizzando i dati meteorologici

Siamo interessati ora a integrare nel modello i possibili effetti delle variabili meteorologiche disponibili nel dataset Weather. Analizzeremo tutti Gran Premi della stagione 2021, adattando la stesso tipo di struttura del modello LMM con l'aggiunta di alcune variabili meteo per valutarne gli effetti e darne un'interpretazione. Oltre alle variabili riportate precedentemente vengono inserite le variabili meterologiche specificate in precendenza in Tabella 1.

L'adattamento del modello ai dati della stagione 2021 suggerisce la non significatività delle variabili *Humidity* e *Pressure*. Il coefficiente stimato relativo alla variabile *WindSpeed* risulta positivo e significativo indicando in media che il tempo sul giro aumenti al crescere dell'intensità del vento. Il risultato ottenuto non è sorprendente poichè nell'era ibrida le vetture di Formula Uno sono diventate estremamente suscettibili all'impatto del vento. Com'era logico aspettarsi, anche il coefficiente della variabile *Rainfall*, è risultato positivo indicando che il tempo sul giro viene nettamente influenzato dall'eventuale presenza della pioggia. Infine, risultano signficativi anche i coefficienti delle temperature. Il coefficiente relativo a *AirTemp* è risultato positivo suggerendo che il tempo sul giro sia influenzato in modo negativo all'aumentare della temperatura dell'aria. Il fattore affidabilità gioca un ruolo fondamentale in un Gran Premio e la temperatura dell'aria è forse la variabile che più mette in pericolo questo aspetto. Al contrario, il coefficiente relativo a *TrackTemp* è risultato negativo indicando che, in media, all'aumentare della temperatura dell'asfalto il tempo sul giro diminuisca. Questo risultato è conforme con le aspettative generali poichè una bassa temperatura implica che gli pneumatici diventino più rigidi e meno aderenti. Al contrario, al netto delle altre variabili, con temperature mediamente più alte dell'asfalto si ottiene una condizione favorevole poichè gli pneumatici tendono a diventare più morbidi e flessibili aumentando l'aderenza e generando più grip in pista.

5.3 Extreme Gradient Boosting

Per migliorare le prestazioni predittive ottenute con il modello lineare a effetti misti (LMM) approfondiremo un algoritmo potente di machine learning che offre una serie di vantaggi rispetto ai precedenti approcci, l'Extreme Gradient Boosting, abbreviato come XGBoost. Le Random Forest, essendo un modello di bagging, costruiscono alberi decisionali e calcolano le loro previsioni in modo parallelo. Il boosting, invece, è una tecnica che agisce con un approccio diverso. Si costruisce un modello complesso finale combinando una serie di modelli deboli costruiti in sequenza in modo da migliorare progressivamente le prestazioni del modello correggendo gli errori commessi dai modelli precedenti. Il Gradient Boosting è un algoritmo che si basa sul "potenziamento del gradiente", cioè utilizza un algoritmo di discesa del gradiente per ridurre al minimo l'errore di previsione quando si costruiscono i nuovi modelli. Viene definito ora l'algoritmo generico di Gradient Boosting per la regressione.

Algoritmo Gradient Boosting [2]

1. Inizializza $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.
2. Per $m = 1$ a M:
   (a) Per $i = 1, 2, \ldots, N$ calcola
   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right] f = f_{m-1}.$$
   (b) Adatta un albero di regressione ai residui r_{im}, ottenendo regioni terminali R_{jm}, $j = 1, 2, \ldots, J_m$.
   (c) Per $j = 1, 2, \ldots, J_m$ calcola
   $$\gamma_{jm} = \arg\min_\gamma \sum x_i \in R_{jm} L(y_i, f_{m-1}(x_i) + \gamma).$$
   (d) Aggiorna $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
3. Output $\hat{f}(x) - f_M(x)$.

Il primo passo dell'algoritmo consiste nell'inizializzare il modello costante ottimale, un albero con un solo nodo terminale. Le componenti del gradiente sono calcolate nel secondo passo e sono chiamate residui generalizzati o pseudo-residui. Successivamente nel secondo passo, dopo aver calcolato questi residui, iterativamente viene costruito un nuovo albero di regressione basato sui residui calcolati in precedenza. Questo nuovo albero cerca di catturare le discrepanze tra i dati di addestramento e le previsioni correnti del modello. Per ogni regione dell'albero viene calcolato il γ ottimale tale che venga minimizzata la funzione di perdita per i dati all'interno di ciascuna regione terminale. Infine si aggiorna il modello corrente aggiungendo i contributi dei singoli alberi creati ad ogni iterazione. Questi contributi sono pesati

[2] Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning, Second Edition. Springer

dai valori γ ottenuti ad ogni iterazione. Si ottiene alla fine un modello complessivo che è una combinazione dei modelli stimati ad ogni iterazione. Tutti le varianti specifiche del Gradient Boosting, come XGBoost, sono ottenute mediante l'inserimento opportuno di diverse funzioni di perdita $L(y, f(x))$
In particolare, per l'XGBoost la funzione è la seguente: [3]

$$\mathcal{L}(\gamma, \lambda, \tau) = \sum_i l\left(\hat{y}_i, y_i\right) + \sum_k \Omega\left(f_k\right)$$
$$\text{con } \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

con:

- λ e τ sono i parametri che regolano il compromesso varianza-distorsione
- T rappresenta il numero di nodi terminali (foglie) presenti nell'albero
- w rappresenta una funzione per le foglie dell'albero il cui valore cresce quanto più accurate saranno le previsioni

XGBoost può affrontare efficacemente sia problemi di classificazione sia di regressione ed è significativamente più veloce di altri algoritmi di potenziamento del gradiente. Un altro notevole vantaggio rispetto alle Random Forest è la sua maggiore resistenza all'overfitting. Infatti, XGBoost non costruisce gli alberi nella loro massima profondità ma vengono potati secondo un criterio specifico ed è inoltre dotato di una potente tecnica di regolarizzazione che aiuta a prevenire l'overfitting. Sebbene spesso raggiunga una precisione maggiore rispetto ad altri modelli, XGBoost è un algoritmo significativamente complesso e ottimizzare i numerosi parametri diventa fondamentale per massimizzare le prestazioni. Inoltre, sebbene XGBoost spesso raggiunga una precisione maggiore rispetto ad altri modelli, un suo evidente limite è la scarsa capacità di interpretazione che altri modelli possiedono.

5.3.1 Analisi predittiva del modello

L'implementazione di questo algoritmo rappresenta un passo avanti significativo rispetto all'approccio precedente in cui avevamo limitato l'analisi di un circuito specifico. Questo limite era principalmente di natura computazionale, in quanto l'adattamento del modello ai dati completi avrebbe richiesto notevoli risorse. Sfruttando appieno l'efficienza e la potenza computazionale di XGBoost possiamo ora utilizzare l'intero dataset a disposizione senza limitazioni per addestrare il modello e ottenere previsioni più accurate del passo gara obiettivo della nostra analisi. Poiché XGBoost tratta variabili solo di tipo numeriche, per gestire le variabili categoriali è stata utilizzata una tecnica chiamata codifica one-hot, che trasforma le variabili categoriali in una serie di variabili binarie (0 o 1) corrispondenti alle diverse categorie. Per ottenere la massima prestazione del modello, è stata eseguita un'ottimizzazione accurata dei parametri attraverso la tecnica della cross-validation. Le previsioni del modello XGBoost sono notevolmente migliorate rispetto al modello precedente. Nel modello lineare visto in precedenza, si era osservata una tendenza costantemente decrescente dei tempi lungo la gara. Tuttavia, con l'implementazione del modello XGBoost si

[3] Chen, T., Guestrin C. e et al. (2016). Xgboost: A Scalable Tree Boosting System. In arXiv:1603.02754

nota che la distribuzione dei tempi sul giro è più aderente ai tempi realmente osservati. Si sottolinea la capacià del modello XGBoost di catturare le variazioni nel tempo del giro che possono includere sia diminuzioni che aumenti creando una sorta di "sali-scendi". Questa capacità è probabilmente dovuta dall'efficace cattura di effetti non lineari da parte del modello in questione consentendo di ottenere previsioni più realistiche. L'adattamento è molto buono per il primo e ultimo stint di gara, mentre nello stint centrale di gara si nota che prevede abbastanza efficacemente i primi dieci giri (dal giro 15 al 25 circa) salvo poi prevedere una diminuzione dei tempi sul giro fino alla seconda sosta. Ciò non corrisponde alla realtà in quanto ha subito un forte ed inaspettato degrado della gomma dovendo inoltre compiere due doppiaggi che gli hanno costato alcuni decimi di secondo al giro.

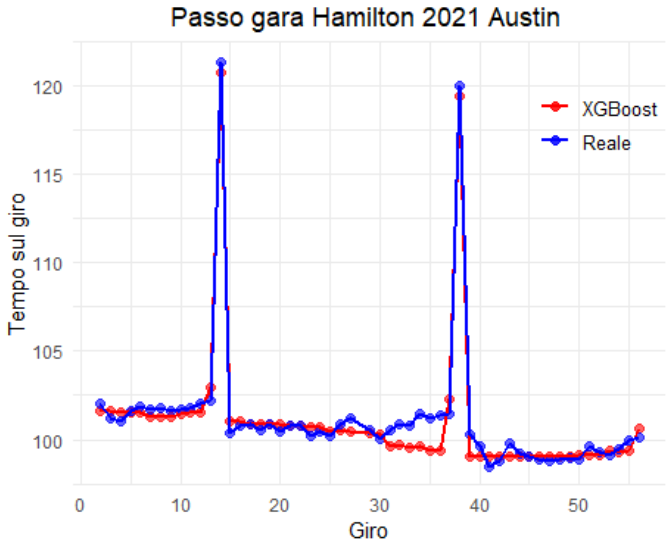

Figura 17: Confronto dei tempi sul giro di Hamilton previsti (in rosso) dal modello XGBoost e osservati (in blu), Austin 2021

5.3.2 Importanza delle variabili

Come per le Random Forest, l'implementazione dell'algoritmo XGBoost permette di calcolare degli indici per valutare l'importanza delle variabili in ottica previsiva. Gli indici principali sono tre:

- **Gain**: misura l'importanza relativa di una variabile basandosi sul contributo di ciascuna variabile per ciascun albero nel modello. Un valore più alto indica una maggiore importanza della variabile nella capacità predittiva del modello.
- **Cover**: valuta l'importanza delle variabili in base alla "copertura" dell'albero decisionale che viene influenzato da ciascuna variabile. Le variabili con una vasta copertura nell'albero sono considerate le più importanti.
- **Frequency**: indica la frequenza con cui una variabile si verifica negli alberi decisionali del modello. Le variabili con una frequenza più alta sono state utilizzate più spesso per prendere le decisioni.

Il *Gain* è l'indice più rilevante per interpretare l'importanza relativa di ciascuna variabile. Viene quindi rappresentato un grafico per determinare quali sono le variabili che hanno contribuito maggiormente alla capacità predittiva del modello (figura 18). Prima di commentare i risultati ottenuti si ricorda la codifica effettuata per le variabili categoriali e lo si nota anche dal grafico in quanto la variabile relativa al circuito è ora "suddivisa" in una serie di variabili binarie ciascuna relativa ad un circuito specifico. Si evidenzia quindi che le variabili relative ai circuiti contribuiscono maggiormente alle capacità predittive. Tra le più importanti inoltre si notano altre quattro variabili, in ordine di importanza *TyreLife*, *LapNumber*, la variabile binaria relativa alla mescola di gomma intermedia e infine *PitOutTime*

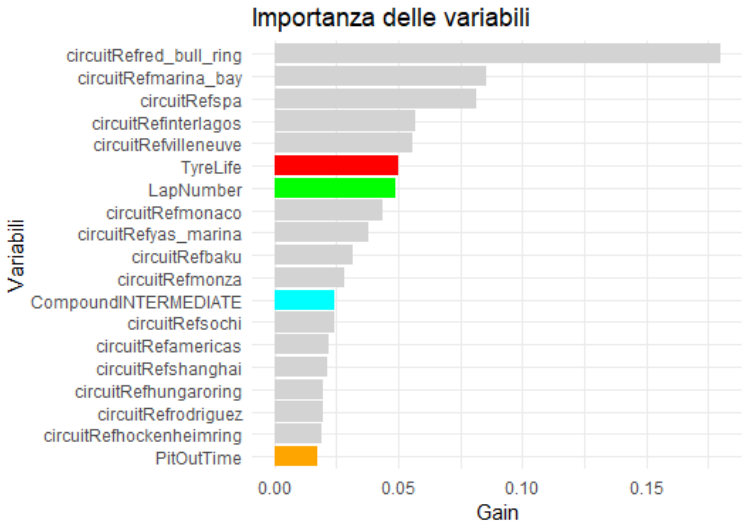

Figura 18: Importanza delle variabili presenti nell'XGBoost

Capitolo 6

Conclusioni

Nel corso di questa Tesi, abbiamo affrontato un'analisi approfondita dei dati relativi ai Campionati di Formula 1 dal 2014 al 2021. Le analisi preliminari hanno permesso di evidenziare il netto dominio da parte di una scuderia in particolare, la scuderia Mercedes-AMG.
In seguito, tramite l'uso di due tecniche non parametriche, albero di classificazione e Random Forest, si è cercato di prevedere il podio dei Gran Premi della stagione 2021 utilizzando tutti i dati degli anni precedenti. I risultati ottenuti sono stati discretamente precisi probabilmente anche grazie al fatto che nelle forze in gioco non c'è stato un cambiamento significativo rispetto agli anni precedenti. La scuderia Mercedes, infatti, ha di gran lunga dominato questo periodo, salvo poi essere stata battuta nel titolo Piloti del 2021 dal pilota della Red Bull Max Verstappen. L'albero, di conseguenza, prevede a podio molto spesso i piloti di punta relativi a queste due scuderie. Un altro limite di questo modello è stato sicuramente quello di aver dovuto inserire il terzo pilota a podio per due Gran Premi (Imola e Belgio). Infatti, a causa della soglia di probabilità impostata, il modello prevedeva solamente due piloti a podio. L'implementazione della Random Forest non ha evidenziato notevoli miglioramenti se non in due casi specifici. Questo ci suggerisce che, nonostante l'utilizzo di un modello più sofisticato come la Random Forest, la previsione del podio in un Gran Premio di Formula 1 rappresenta una sfida notevole a causa dell' imprevedibilità che caratterizza questo sport. Inoltre, entrambi i metodi utilizzati suggeriscono che la variabile più discriminante per questa dinamica è la posizione di partenza. Tuttavia, in alcuni Gran Premi, si sono mostrate discriminanti altre variabili riguardo la scuderia del pilota, il numero di podi conquistati dal pilota e il circuito.

Infine, l'attenzione si è focalizzata sulla modellazione dei tempi sul giro in un Gran Premio specifico mediante l'uso di modelli parametrici e successivamente di un algoritmo di machine learning. Il modello parametrico utilizzato per offrire una completa interpretazione delle variabili è stato il modello lineare a effetti misti (LMM). L'adattamento di questo modello ha permesso di stabilire delle relazioni tra le variabili esplicative e il tempo sul giro evidenziando anche le differenze di perfromance tra mescole di gomma utilizzate e tra le varie scuderie. Si è potuto confermare inoltre il fenomeno del degrado gomma con l'inserimento della variabile relativa all'età dei giri degli pneumatici. L'estensione di questo modello ai dati ridotti ma comprensivi

delle variabili meteorologiche ha permesso di stabilire che le variabili significative per questa dinamica riguardano principalmente l'eventuale presenza di pioggia e le temperature. Non sono risultate significative le variabili relative alla pressione atmosferica e all'umidità al contrario della variabile relativa all'intensità del vento. Tuttavia, questo approccio ha evidenziato un limite nella capacità predittiva e tutte le intepretazioni fornite si basano su relazioni lineari. Pertanto, per migliorare questi due aspetti, è stato implementato un algoritmo di machine learning, noto come XGBoost, che si è rivelato nettamente più efficace in ottica previsiva.

In alternativa al modello LMM, sarebbe stato possibile esaminare i modelli GAM o GAMM (Generalized Additive Models o Generalized Additive Mixed Models). Questi due modelli sono particolarmente efficaci nel catturare relazioni non lineari e complesse tra le variabili. In particolare, consentono l'uso di funzioni di base lisce e flessibili che avrebbe probabilmente permesso di modellare più efficacemente le relazioni, specialmente nel caso di interazioni complesse tra le variabili.
In conclusione, l'adattamento di questi modelli si è rivelato discretamente buono trascurando però una componente fondamentale durante i Gran Premi di Formula 1, la presenza della Safety Car. Riconosciamo che, data la complessità e l'imprevedibilità di particolari eventi come la presenza della Safety Car e altre situazioni rare, costruire un modello "perfetto" risulta essere un obiettivo praticamente inarrivabile.

Bibliografia

- Salvan A.; Sartori N.; Pace L.; (2020). *Modelli Lineari Generalizzati*. Springer Milan
- Azzalini A.; Scarpa B.; (2012). *Data analysis and data mining: An introduction*. OUP USA.
- Hastie T.; Tibshirani R.; Friedman J.; (2016). *The Elements of Statistical Learning*. Springer
- James G.; Witten D.; Hastie T.; Tibshirani R.; (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer
- Chen T.; Guestrin C. e et al. (2016). *Xgboost: A scalable tree boosting system*. In: arXiv preprint arXiv:1603.02754