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

Scienze Statistiche

Statistica e golf: modelli predittivi per il calcolo degli *strokes gained*

Relatore: Prof. Francesco Lisi

Correlatore: Prof. Federico Cacciafesta

Laureando: Mirko Gabriel Briglia

Matricola n. 2082052

Anno Accademico 2023/2024

Ai miei genitori
Daniela e Franco

In memoria di Pier Paolo Briglia

Indice

1 Introduzione **9**
1.1 Storia delle analisi statistiche nel golf 9
1.2 Broadie, *benchmark* e *strokes gained* 14
1.3 Domanda di ricerca 18

2 Il golf **21**
2.1 Panoramica del gioco 21
2.2 Regole fondamentali 23
2.2.1 Sequenza di gioco 23
2.2.2 Principali tipologie di colpi 24
2.2.3 Equipaggiamento e terminologia di base 26

3 I *dataset* **29**
3.1 Dati 2023 29
3.1.1 Descrizione delle variabili 30
3.1.2 Analisi esplorative su `data_2023` 31
3.2 *Dataset* 2024 36
3.2.1 QGis e Golf Club Ca'Amata 37
3.2.2 Analisi univariate 40
3.2.3 Analisi di correlazione 44

4 Modellazione: dati 2023 **47**
4.1 *Tee shot benchmark* 47
4.1.1 Modello lineare 48
4.2 *Approach shot benchmark* 53
4.2.1 Generalized Additive Models (GAM) 53
4.2.2 Applicazione ai dati 55
4.3 *Putting shot benchmark* 57
4.4 Esempio di applicazione completa 59

5 Modellazione: dati 2024 **63**
5.1 Classificazione logistica ordinale 65
5.1.1 Risultati del modello HCP 5/15 67
5.1.2 Risultati del modello HCP 15/30 70
5.2 Classificazione logistica multinomiale 72
5.2.1 Risultati del modello HCP 5/15 74
5.2.2 Risultati del modello HCP 15/30 76
5.3 MARS (Multivariate Adaptive Regression Splines) . . . 78
5.3.1 Risultati del modello HCP 5/15 79
5.3.2 Risultati del modello HCP 15/30 82
5.4 Alberi di classificazione con splines naturali 84
5.4.1 Risultati per il modello HCP 5/15 88
5.4.2 Risultati per il modello HCP 15/30 90
5.5 *Random Forest* 91
5.5.1 Risultati per il modello HCP 5/15 93
5.5.2 Risultati per il modello HCP 15/30 94
5.6 Strokes gained 96
5.6.1 *Danger zone* 98

6 Conclusioni **103**

A Descrizione delle Variabili **107**

B Metodo di raccolta **113**
B.1 Tecniche di raccolta 113
B.2 *Shiny Dashboard* 114

Bibliografia **117**

Capitolo 1

Introduzione

In questo capitolo, dapprima viene presentata la storia dell'analisi dei dati nel golf, partendo dagli inizi del secolo scorso fino ad arrivare alle ultime scoperte scientifiche in questo campo. Successivamente viene presentato il lavoro principale dal quale è scaturita l'idea per questa ricerca, spiegando in seguito il contenuto della stessa. Il capitolo 2 presenta, invece, la spiegazione dettagliata di tutta la terminologia tecnico-sportiva, necessaria per comprendere al meglio ogni aspetto di questa tesi.

1.1 Storia delle analisi statistiche nel golf

Durante il corso del XX secolo si è assistito a un continuo miglioramento nelle *performance* degli atleti, in tutte le discipline sportive, grazie a un incremento nell'utilizzo delle tecnologie nelle gare, a un aumento degli investimenti sia nella sfera medico-nutrizionale degli sportivi, sia nella sfera delle tecniche di *coaching*.

Una delle discipline interessate a questo miglioramento è stato il **golf**, uno sport che ad occhio esterno può sembrare semplice ed intuitivo, ma al suo interno nasconde molte complessità.

Qualsiasi aspetto del golf può essere ridotto al "dato", la selezione della mazza, il tipo di pallina usata, la temperatura sul campo e in aria, la direzione del vento sono solo esempi delle decisioni che un golfista deve prendere prima di colpire.

Nel corso degli anni '60 e '70, l'analisi statistica nel golf era ancora in una fase embrionale. Le statistiche disponibili erano limitate e si concentravano principalmente su metriche di base, facilmente calcolabili e comprensibili sia dai giocatori che dagli appassionati. Durante questo periodo, le principali statistiche utilizzate per valutare le *performance*

dei giocatori includevano il punteggio medio, il numero di *fairway* colpiti, il numero di *green* in regolazione e il numero di *putt per round*.

Il **punteggio medio** rappresentava la media dei punteggi totali dei giocatori su un determinato numero di *round* o tornei, offrendo una misura diretta della capacità di un giocatore di completare il percorso con il minor numero di colpi possibile.

Il **numero di *fairway* colpiti** indicava la percentuale di volte in cui un giocatore riusciva a posizionare la pallina sul *fairway* con il colpo di partenza, riflettendo così la precisione e la consistenza dal *tee*.

Allo stesso modo, il **numero di "*green* in regolazione"** misurava la percentuale di volte in cui un giocatore raggiungeva il *green* in un numero di colpi pari al *par* della buca -2, evidenziando la capacità di avvicinarsi alla buca nei tempi previsti. Infine, il **numero di *putt* per *round*** calcolava il totale dei *putt* eseguiti in un *round*, indicando l'efficacia nel gioco corto sul *green*.

Queste statistiche erano raccolte principalmente attraverso l'osservazione diretta e la registrazione manuale durante i tornei. Giocatori e allenatori utilizzavano queste informazioni per ottenere una comprensione generale delle aree di forza e debolezza, oltre che per confrontare le *performance* nel tempo o rispetto ad altri giocatori.

Tuttavia, nonostante l'utilità di queste metriche di base, esse presentavano diverse limitazioni che ne riducevano l'efficacia nell'analisi approfondita delle *performance*.

In primo luogo, le statistiche tradizionali fornivano una visione superficiale del gioco senza tenere conto della complessità intrinseca del golf. Non consideravano, ad esempio, la difficoltà specifica delle buche o le condizioni ambientali che potevano influenzare le *performance*. Parametri come il numero di *fairway* colpiti non tenevano conto della lunghezza dei colpi o della posizione sul *fairway*; un colpo corto ma preciso e uno lungo ma leggermente fuori dal *fairway*, potevano avere impatti diversi sul punteggio, ma le statistiche non riflettevano queste sfumature.

Inoltre, le metriche di base erano incapaci di misurare l'"abilità reale" del giocatore in situazioni specifiche. Un giocatore poteva mancare il *green* in regolazione, ma eseguire un colpo di recupero eccezionale, salvando il *par*; questa abilità non era adeguatamente catturata dalle statistiche tradizionali. L'aggregazione dei dati non considerava la variabilità delle *performance* su diverse tipologie di campi o in condizioni di gioco differenti, limitando la capacità di trarre conclusioni significative.

Queste limitazioni evidenziavano la necessità di sviluppare metodi di analisi più sofisticati, in grado di fornire una comprensione più dettagliata

e accurata delle prestazioni dei giocatori.

Con l'avvento della tecnologia informatica negli anni '90, il golf iniziò a beneficiare di strumenti più avanzati per la raccolta e l'analisi dei dati. Questo periodo segnò una transizione significativa dall'uso di semplici statistiche aggregate a metodologie più dettagliate e analitiche. Il progresso esponenziale nelle capacità di elaborazione dei computer permise di gestire quantità di dati molto più grandi e complessi.

Nel golf ciò si tradusse nell'introduzione di sistemi di raccolta dati elettronici, che consentirono di ottenere informazioni più dettagliate su ogni colpo eseguito dai giocatori durante i tornei. La possibilità di archiviare grandi quantità di dati storici permise l'analisi di tendenze e pattern nelle *performance* dei giocatori su lunghi periodi. Inoltre, emerse una nuova generazione di software specializzati in grado di elaborare i dati raccolti, generando statistiche avanzate e visualizzazioni grafiche delle *performance*.

Un punto di svolta fondamentale fu l'introduzione del sistema **ShotLink** da parte del PGA Tour nel 2001. ShotLink è un sistema avanzato di raccolta dati che utilizza una combinazione di tecnologie, tra cui laser, GPS e segnali radio, per tracciare ogni colpo di ogni giocatore in tempo reale durante i tornei. Ogni colpo è tracciato con una precisione di pochi centimetri, registrando la posizione esatta della pallina dalla partenza all'arrivo. Le informazioni sono disponibili quasi istantaneamente, permettendo analisi e aggiornamenti durante lo svolgimento del torneo. Oltre alla posizione della pallina, vengono registrati dati come la distanza del colpo, il tipo di bastone utilizzato, le condizioni del vento e altri parametri ambientali.

L'introduzione di ShotLink ebbe un impatto significativo sul golf professionistico: con dati così dettagliati, fu possibile sviluppare nuove metriche statistiche che tenevano conto di molte più variabili rispetto al passato. Le trasmissioni televisive e i media poterono fornire informazioni più ricche e dettagliate agli spettatori, rendendo il gioco più coinvolgente e comprensibile. L'accesso a dati granulari permise ai giocatori, e ai loro team, di identificare con maggiore precisione le aree di miglioramento, personalizzando l'allenamento e le strategie di gioco.

Un esempio emblematico dell'utilizzo di ShotLink è il lavoro di Mark Broadie, professore alla *Columbia Business School (New York)*, che sviluppò il concetto di *strokes gained*, portando a un grosso passo avanti nell'analisi statistica del gioco e che verrà analizzato nel dettaglio nel capitolo 1.2.

Questa metrica confronta il rendimento di un giocatore su ogni colpo

con quella media degli altri giocatori, permettendo di quantificare esattamente quanto ogni colpo contribuisca al punteggio totale. L'introduzione di ShotLink rappresentò una rivoluzione nell'approccio alle statistiche nel golf, permettendo di superare le limitazioni delle metriche tradizionali e fornendo una comprensione molto più profonda delle performance dei giocatori e delle dinamiche del gioco. Questo aprì la strada a un'era di *data-driven golf*, in cui le decisioni sono sempre più basate su analisi quantitative avanzate.

In linea con questa evoluzione, Heiny e Heiny (2014) hanno impiegato una catena di Markov assorbente per modellare la stagione del PGA Tour 2012. Nel golf, i giocatori si muovono da uno *state*, descrivibile come distanza e posizione rispetto alla buca, al successivo, con una certa probabilità (*probabilità di transizione*) fino all'assorbimento rappresentato dalla pallina che entra nella buca. Le probabilità di transizione sono stimate empiricamente utilizzando i dati ShotLink™, che forniscono dettagli su ogni colpo, come la posizione di partenza, la distanza dalla buca e la posizione finale. Per ogni buca viene creato un modello separato per i *par* 3, *par* 4 e *par* 5, poiché il comportamento del gioco cambia in base alla lunghezza e alla difficoltà. In questo studio la funzione di *Benchmark* è stata sviluppata attraverso questi diversi modelli di Markov assorbenti per calcolare il punteggio atteso da qualsiasi posizione. Essendo la catena di Markov assorbente un modello stocastico, il punteggio medio osservato da qualsiasi posizione è semplicemente una realizzazione del processo stocatico, nel quale il punteggio atteso può essere interpretato come la rappresentazione del comportamento asintotico del processo stesso [16].

Anche il *putting*, una componente fondamentale del gioco, è stato oggetto di approfondite analisi grazie ai dati ShotLink. Yousefi e Swartz (2013) hanno introdotto una metrica avanzata che tiene conto non solo della distanza dalla buca, ma anche della direzione e della difficoltà del *putt*. Utilizzando tecniche di statistica spaziale, hanno costruito mappe che stimano il numero atteso di *putt* necessari da varie posizioni sul green. Questa metrica, che si basa su un approccio bayesiano, permette di confrontare la *performance* effettiva di un giocatore rispetto a quella attesa in condizioni simili. Il loro lavoro rappresenta un'estensione del concetto di *strokes gained*, applicandolo in modo specifico al *putting*, uno degli aspetti più tecnici e delicati del gioco, e migliorando significativamente la capacità di valutare la performance dei giocatori sul *green* [30].

Parallelamente, Stöckl, Lamb e Lames (2011) hanno sviluppato il metodo ISOPAR, un approccio innovativo per analizzare le prestazioni nel golf. Questo metodo si basa su mappe che rappresentano il numero medio

di colpi necessari per completare una buca da una determinata posizione. Utilizzando i dati di ShotLink raccolti durante tornei professionali, il metodo ISOPAR fornisce una valutazione visiva della difficoltà di diverse aree del campo, simile alle mappe topografiche utilizzate in meteorologia. La qualità del colpo viene determinata confrontando la posizione iniziale e finale della pallina, permettendo così di quantificare con precisione l'impatto di ogni singolo colpo sul punteggio totale [25].

Un altro contributo fondamentale in questo ambito è stato fornito da Connolly e Rendleman (2011), i quali hanno esaminato le probabilità di successo nel qualificarsi per il PGA Tour utilizzando simulazioni avanzate basate sui dati ShotLink. Il loro studio ha evidenziato come le prestazioni dei giocatori siano spesso influenzate da variazioni casuali e imprevedibili, sottolineando l'importanza di metriche avanzate come gli *strokes gained* per distinguere tra la vera abilità di un giocatore e gli effetti dovuti al caso. Le loro simulazioni hanno dimostrato che, nonostante l'abilità dei giocatori rimanga il fattore chiave per il successo, fattori come la variazione della *performance* in situazioni di alta pressione possono influenzare significativamente i risultati. Questo studio rafforza ulteriormente l'idea che l'utilizzo di dati dettagliati come quelli forniti da ShotLink sia essenziale per una valutazione accurata delle abilità e delle probabilità di successo dei giocatori. [8].

Drappi e Co Ting Keh (2019) hanno sviluppato un modello predittivo basato su dati ShotLink che permette di stimare, colpo per colpo, la probabilità di raggiungere un determinato punteggio su ogni buca. Utilizzando una rete neurale addestrata con un vasto set di dati, il loro modello è in grado di aggiornare in tempo reale le probabilità di successo di un giocatore, tenendo conto delle caratteristiche del campo e delle condizioni esterne come vento, altezza dell'erba e velocità del green. Questo modello rappresenta un'ulteriore evoluzione nell'analisi statistica del golf, in quanto consente non solo di prevedere con maggiore precisione i punteggi, ma anche di migliorare la gestione strategica del gioco, suggerendo le scelte migliori da fare su ogni colpo [11].

Infine, è importante sottolineare gli enormi passi avanti fatti proprio dalle aziende del settore golf come *Garmin, Bushnell, Trackman e FlightScope* che possiedono già nel proprio catalogo, *devices* che producono analisi sofisticate e funzionano come *caddy* automatici, suggerendoti il bastone in base a diverse variabili come condizioni atmosferiche, tipologia di terreno e difficoltà del campo.

1.2 Broadie, *benchmark* e *strokes gained*

Broadie (2008) ha introdotto l'analisi degli *strokes gained* principalmente per determinare l'abilità che separa il gioco dei professionisti da quello degli amatori [2]. Fearing et al. (2010) hanno esteso quest'analisi per modellare la performance del *putt* sul circuito professionistico PGA TOUR [13].

Per definire questa metrica, Broadie (2008) ha considerato una funzione $J(d, c)$, dove:

- d rappresenta la distanza dalla buca dalla posizione corrente (non la distanza del colpo);
- c indica la condizione attuale della posizione della pallina, ossia il tipo di terreno: *green*, *tee*, *fairway*, *rough*, *bunker* o situazione di *recovery*.

La funzione $J(d, c)$ esprime il numero medio di colpi che un giocatore del PGA TOUR impiega per completare la buca a partire dalla posizione (d, c). Per semplicità, J viene definita come il *benchmark*.

Si definisce quindi lo *strokes gained* del colpo i-esimo su una buca, che inizia dalla posizione (d_i, c_i) e termina in (d_{i+1}, c_{i+1}), come:

$$S_i = J(d_i, c_i) - J(d_{i+1}, c_{i+1}) - 1 \tag{1.1}$$

Questa formula quantifica il valore del colpo confrontando il numero medio di colpi necessari per completare la buca dalla posizione iniziale e dalla posizione finale del colpo, sottraendo 1 per il colpo appena eseguito.

Interpretazione della formula

- $J(d_i, c_i)$: numero medio di colpi necessari per terminare la buca dalla posizione iniziale (d_i, c_i).
- $J(d_{i+1}, c_{i+1})$: numero medio di colpi necessari per terminare la buca dalla nuova posizione (d_{i+1}, c_{i+1}) dopo il colpo.
- -1: rappresenta il colpo appena effettuato.

Il valore di S_i indica quanto il colpo i-esimo ha avvicinato il giocatore alla conclusione della buca rispetto al giocatore medio del PGA TOUR. Un valore positivo di S_i significa che il giocatore ha guadagnato colpi rispetto alla media, mentre un valore negativo indica una perdita.

Esempio applicativo

Supponiamo che un giocatore si trovi a una distanza di 150 m dalla buca sul *fairway* ($c_i = \textit{fairway}$), e il valore $J(150, \text{fairway})$ sia 2.8 colpi. Dopo aver eseguito il colpo, la pallina si trova a 3 metri dalla buca sul *green* ($c_{i+1} = \textit{green}$), con $J(3, \text{green})$ pari a 1.1 colpi. Applicando la formula:

$$\begin{aligned} S_i &= J(150, \text{fairway}) - J(3, \text{green}) - 1 \\ &= 2.8 - 1.1 - 1 \\ &= 0.7 \end{aligned}$$

Il giocatore ha quindi guadagnato 0.7 colpi rispetto alla media, indicando un colpo migliore rispetto allo standard del PGA TOUR.

Il *benchmark* viene calcolato utilizzando i dati di riferimento del circuito professionistico, ad esempio si prenda un colpo che arrivi a 6 metri dalla buca, mediamente si completa con 1 colpo il 59% delle volte, con 2 colpi il 28% delle volte e con 3 colpi il 13% delle volte, allora il *benchmark* di riferimento da quella distanza sarà pari a 1.54 (59% (1) + 28% (2) + 13% (3))

Gli *strokes gained*, che per comodità da ora in poi verranno definiti nel testo come SG presentano un'importante **proprietà additiva**: gli SG di un gruppo di colpi è pari alla somma degli SG dei colpi individuali. Supponendo che un golfista impieghi n colpi per terminare una buca, allora gli SG totali per quegli n tiri è pari a

$$\begin{aligned} \sum_{i=1}^{n} S_i &= \sum_{i=1}^{n} \left(J(d_i, c_i) - J(d_{i+1}, c_{i+1}) - 1 \right) \\ &= J(d_1, c_1) - n \end{aligned} \tag{1.2}$$

Tabella 1.1: Calcolo degli *strokes gained* per i golfisti A e B

Golfista	**Colpo**	**Posizione**	$J(d,c)$	***Strokes Gained***
A	1	Da: 175 m, tee A: 8 m, green	$J(175, \text{tee}) = 3{,}4$ $J(8, \text{green}) = 1{,}4$	$S_1^A = 1{,}0$
	2	Da: 8 m, green A: 3 m, green	$J(8, \text{green}) = 1{,}4$ $J(3, \text{green}) = 1{,}1$	$S_2^A = -0{,}7$
	3	Da: 3 m, green A: In buca	$J(3, \text{green}) = 1{,}1$ $J(0, \text{in buca}) = 0$	$S_3^A = 0{,}1$
B	1	Da: 175 m, tee A: 40 m, rough	$J(175, \text{tee}) = 3{,}4$ $J(40, \text{rough}) = 2{,}6$	$S_1^B = -0{,}2$
	2	Da: 40 m, rough A: 1 m, green	$J(40, \text{rough}) = 2{,}6$ $J(1, \text{green}) = 1{,}0$	$S_2^B = 0{,}6$
	3	Da: 1 m, green A: In buca	$J(1, \text{green}) = 1{,}0$ $J(0, \text{in buca}) = 0$	$S_3^B = 0{,}0$

Come si evince dalla Tabella 1.1 i due golfisti A e B hanno completato la buca in due modi completamente diversi seppure con lo stesso risultato: il golfista A ha eseguito un *tee shot* ottimo, piazzandosi molto bene in *green* ($SG = 1$), un *putt* verso la buca sbagliato ($SG = -0.7$) per un totale di $SG = 0.4$; il golfista B ha eseguito un *tee shot* non perfetto, tirando in *rough* a 40 metri dalla buca ($SG = -0.2$), ma l'approccio è stato praticamente perfetto ($SG = 0.6$), lasciandosi così un *tap-in* impossibile da sbagliare.

Grazie alla proprietà 1.2, si nota che il valore di SG è lo stesso sia per A che per B. Si può quindi andare a definire le caratteristiche dei due giocatori, A ha un buon gioco lungo, mentre B eccelle negli approcci (colpi verso il green), andando così a decomporre gli SG di un intero giro di 18 buche nelle varie categorie di gioco (gioco lungo, gioco corto, approcci) e sottocategorie (ferri lunghi, ferri corti, etc.) Così facendo il golfista può andare a vedere in maniera analitica dove andare a operare in fase di allenamento per ottenere risultati migliori.

Strokes gained sul putt

Per quanto riguarda l'adattamento degli *strokes gained* sul *putt*, Fearing et al. (2010) propongono un approccio modellistico, andando a definire la probabilità di imbucare con un colpo come:

$$P[Y=1|d] = \left[1 + e^{-\left(\beta_0+\beta_1 d+\beta_2 d^2+\beta_3 d^3+\beta_4 d^4+\beta_5 \log d\right)}\right]^{-1} \tag{1.3}$$

dove Y è il numero di *putt* per concludere e i termini β sono i coefficienti di un polinomio di quarto ordine rispetto alla distanza d.

Questa regressione logistica viene mescolata con un modello che si basa su una distribuzione Gamma che, data la distanza di partenza, mostra la probabilità di arrivare entro una certa distanza se non si imbuca il *putt*.

Il modello finale, quindi, è una semplice catena di Markov, dove lo spazio degli stati è la retta reale positiva che rappresenta la distanza dalla buca, combinata con uno stato assorbente, H, che rappresenta la buca.

Da qualsiasi stato $d \in \mathbb{R}^+$, la probabilità di passare allo stato H è uguale alla probabilità di imbucare, e la densità di probabilità associata al passaggio allo stato $z \in \mathbb{R}^+$ è uguale alla probabilità di mancare il primo *putt* moltiplicata per la densità rappresentata dal modello della distanza da percorrere.

In termini del modello di Markov, la probabilità di raggiungere la buca in n *putt* dalla distanza d corrisponde alla probabilità di raggiungere per la prima volta lo stato assorbente, H, in esattamente n transizioni dallo stato d.

$$P[Y=n|d] = \begin{cases} P[Y=1|d] & n=1 \\ P[Y>1|d] \int_0^\infty P[Y=1|z] f(z|d) dz & n=2 \\ P[Y>1|d] \int_0^\infty P[Y=2|z] f(z|d) dz & n=3 \\ \vdots \end{cases} \tag{1.4}$$

dove, come sopra Y è il numero di *putt*, d è la distanza di partenza del *putt*, z è la distanza rimanente e $f(z|d)$ è la densità della funzione gamma.[13]

Funzione *benchmark*

Nella formula 1.1 viene presentata la funzione J, che permette di ottenere il numero medio di colpi dei professionisti del circuito (*benchmark* di riferimento) per finire una buca, che tipicamente cresce con la distanza dalla buca e dipende dalle condizioni del campo al punto in cui atterra la palla. Considerando che i tiri dal *rough* sono più complicati dei colpi nel

fairway, il *benchmark* presenterà valore più alto in quanto è più comune che ci siano ostacoli che blocchino il passaggio verso la buca, e il terreno sia più difficoltoso per eseguire il colpo.

Broadie (2012) per costruire la funzione J utilizza i dati osservati dal sistema ShotLink con più di 8 milioni di tiri effettuati dai professionisti dal 2003 al 2010, in un database che racchiude varie variabili fra cui la distanza dalla buca e la condizione della palla.

I metodi utilizzati nella sua ricerca sono 2: o *model-based* o mediante approccio statistico (interpolazione, regressione lineare, *splines* di lisciamento, regressione locale con *kernel* radiale, etc.). Alla fine nella ricerca vengono utilizzate delle funzioni polinomiali definite a tratti per comporre la forma del *benchmark*, ad eccezione degli SG dove viene utilizzato un approccio *model-based* come in Broadie and Bansal (2008). La forma di queste funzioni non è esplicita, in quanto Broadie, nel suo lavoro, la specifica solamente per il *Tee Shot*. Come variabili che compongono la regressione vengono utilizzate, come già citate prima, la distanza dalla buca e la condizione in cui si trova la pallina (*fairway*, *rough*, *sand*, etc.).

1.3 Domanda di ricerca

Ciò che vuole presentare questa tesi è un metodo alternativo per calcolare questo *benchmark*, mediante l'esplorazione di varie tecniche di ***Data Mining*** e l'implementazione di ulteriori variabili per adattare meglio il modello alle condizioni del campo e dei singoli giocatori. Per tale scopo, vista l'impossibilità di accedere ai dati ShotLink che sono privati (e l'acquisto di API per la raccolta risulta troppo oneroso), si procede mediante l'utilizzo di due tipologie di dati diverse: in primis viene analizzato un *dataset* di dati raccolti da un singolo soggetto durante il corso del 2023, dove ogni unità statistica si riferisce a una buca giocata. Successivamente si analizzerà un *dataset* con i dati "colpo per colpo", dove ogni unità statistica rappresenta un colpo eseguito da parte di 4 soggetti di diverso livello di abilità, che si sono prestati a registrare i propri risultati durante un periodo che va dal 9 dicembre 2023 al 30 settembre 2024. Questo *dataset* in fase di modellistica verrà diviso in 2 parti per poter modellare 2 livelli di abilità differenti. Inoltre l'utilizzo di dati "empirici" raccolti non su professionisti, ma su amatori il quale HCP varia da 5 a 30, permette di ottenere una panoramica degli *strokes gained* e un *benchmark* fuori dal PGA TOUR, dove gli atleti hanno distanze diverse e un livello di precisione più alto.

Capitolo 2

Il golf

In questo capitolo verranno esplorate le fondamenta di questo gioco, fornendo una panoramica completa che va dalla comprensione delle regole fondamentali all'analisi dell'equipaggiamento necessario e verranno approfondite le diverse tipologie di colpi e la terminologia essenziale per comprendere bene questa ricerca.

2.1 Panoramica del gioco

Il golf è uno sport di precisione, un insieme di strategia e tecnica che si basa sull'abilità del giocatore di colpire una pallina utilizzando una serie di bastoni appositamente progettati, con l'obiettivo di infilare la pallina in una sequenza di buche disposte lungo un percorso prestabilito.

L'obiettivo fondamentale del golf è completare l'intero percorso, solitamente composto da 18 buche, impiegando il minor numero possibile di colpi. Questo richiede una combinazione di competenze tecniche, controllo fisico, pianificazione strategica e concentrazione mentale non indifferenti.

Il gioco del golf si svolge all'aperto, su un'area estesa e attentamente progettata, che copre in genere una superficie tra i 50 e 150 ettari, in campi caratterizzati da una varietà di terreni e ostacoli naturali o artificiali, come prati, sabbia, acqua e vegetazione varia.

Ogni buca (solitamente sono 18) presenta sfide specifiche, richiedendo al giocatore di usare la migliore strategia in base alla quale deve scegliere il bastone più appropriato per ciascun colpo.

Figura 2.1: Buca 4 - Golf Club Ca'Amata (TV)

Le componenti principali di una buca sono:

- **Area di partenza (*teeing ground*)**: è il punto di inizio di ogni buca, dove il giocatore effettua il primo colpo noto come *tee shot*. Solitamente è una superficie piana e ben curata, dotata di diversi *tee markers* che indicano le diverse posizioni di partenza in base al livello di abilità o alla categoria dei giocatori (professionisti, dilettanti, donne, *junior*).

- ***Fairway***: è la porzione del percorso che collega l'area di partenza al *green*. Il *fairway* è caratterizzato da erba tagliata corta e uniforme, che facilita il gioco e premia i colpi precisi. La larghezza e la conformazione del *fairway* possono variare, presentando talvolta restringimenti o curvature (*dogleg*) che aumentano la difficoltà strategica.

- ***Rough***: situato ai lati del *fairway*, il *rough* è costituito da erba più alta e fitta, oltre a possibili ostacoli naturali come arbusti o alberi. Concludere il colpo nel *rough* rende più complicato il successivo, a causa delle condizioni meno favorevoli del terreno e della vegetazione, che può interferire con il movimento del bastone e inficiare il volo della palla.

- **Ostacoli (*hazards*)**: gli ostacoli sono elementi inseriti nel campo progettati per aumentare la sfida del gioco. Si dividono principalmente in:

  – ***Bunker***: aree riempite di sabbia, posizionate strategicamente lungo il percorso o intorno al *green*. Richiedono una tecnica

specifica per essere superati, poiché la sabbia altera la dinamica del colpo.

  - **Ostacoli d'acqua**: comprendono laghi, stagni, ruscelli o fossati. Se la pallina finisce in un ostacolo d'acqua, il giocatore deve, secondo le regole, applicare una penalità e "droppare" (lasciare cadere vicino a sè) una nuova pallina in una posizione definita.

- ***Green***: è l'area finale di ogni buca, dove si trova la buca stessa (*hole*). Il *green* è caratterizzato da un manto erboso estremamente curato e tagliato molto corto, permettendo alla pallina di rotolare in modo uniforme. Sul *green*, i giocatori utilizzano il *putter*, un bastone specifico progettato per colpi precisi e delicati.

- **Buca (*hole*)**: la buca è un foro cilindrico nel terreno con un diametro standard di 108 mm e una profondità di almeno 100 mm. È spesso contrassegnata da un'asta con una bandiera che aiuta i giocatori a individuarne la posizione da lontano.

Il golf non è solo una competizione contro altri giocatori, ma rappresenta principalmente anche una sfida personale a migliorare costantemente le proprie prestazioni e gestire le proprie emozioni sotto pressione.

Uno degli aspetti distintivi del golf è il suo sistema di **handicap**, che permette a giocatori di diversi livelli di abilità di competere equamente tra loro. L'handicap riflette il potenziale di punteggio di un giocatore e viene utilizzato per bilanciare le differenze di abilità, rendendo il gioco inclusivo e accessibile a un'ampia gamma di partecipanti.

2.2 Regole fondamentali

2.2.1 Sequenza di gioco

Il golf è uno sport che segue una sequenza strutturata per ogni buca, ripetuta lungo l'intero percorso. Comprendere questa sequenza è essenziale per apprezzare le decisioni strategiche che i giocatori devono prendere durante il gioco.

Il gioco di ogni buca inizia dall'area di partenza, o *teeing ground*, dove il giocatore esegue il primo colpo noto come *tee shot*. L'obiettivo di questo colpo iniziale è posizionare la pallina in un punto favorevole sul *fairway*, preparandosi per i colpi successivi.

Dopo il *tee shot*, i giocatori avanzano lungo il *fairway* verso il *green*, eseguendo colpi aggiuntivi per avvicinarsi alla buca. Durante questa fase, vengono possono essere utilizzati diversi bastoni in base alla distanza da coprire e alle condizioni del terreno, con l'obiettivo di mantenere la pallina in gioco ed evitare ostacoli.

Una volta che la pallina raggiunge il *green*, i giocatori impiegano il *putter* per eseguire gli ultimi (o ultimo) colpi, chiamati *putt*, con l'intento di mandare la pallina in buca. Il *green* richiede particolare attenzione per via delle sue superfici perfettamente curate e delle possibili pendenze che influenzano la traiettoria della pallina.

La buca si considera completata quando la pallina entra nella buca stessa, e viene registrato, per il punteggio, il numero di colpi effettuati. Questo processo si ripete per ogni buca del percorso, costituendo così l'intero *round*.

2.2.2 Principali tipologie di colpi

Nel golf esistono diverse tipologie di colpi fondamentali, ciascuna con caratteristiche tecniche specifiche e utilizzate in base alla situazione sul campo. L'attenzione sarà focalizzata sui colpi principali, escludendo quelli più particolari come il *pitch*, il *chip* e il *punch*.

Il ***Tee shot*** è il colpo di apertura su buche lunghe. Generalmente si utilizza il *driver* (legno 1), caratterizzato da una testa voluminosa e un basso angolo di *loft*, solitamente compreso tra 8° e 13°. L'obiettivo del drive è massimizzare la distanza del colpo per avvicinarsi il più possibile al *green* o posizionarsi favorevolmente per il colpo successivo. Questo colpo richiede una combinazione di potenza e precisione; l'ampiezza dello *swing* (movimento del bastone) e la velocità del bastone sono cruciali per generare la distanza desiderata.

Il ***fairway shot*** comprende i colpi eseguiti dal *fairway* per coprire distanze intermedie tra il punto di arrivo del *drive* e il *green*. A seconda della distanza da coprire, e delle condizioni del terreno, si utilizzano legni da *fairway* (come il legno 3 o 5), ibridi o ferri dal n.2 al n.7. L'obiettivo è far avanzare la pallina verso il *green* con precisione, mantenendola in gioco ed evitando ostacoli. Questi colpi richiedono un buon controllo dello *swing* e una corretta selezione del bastone per adattarsi alla situazione specifica.

L'***approach shot*** è il colpo che mira a raggiungere il *green* da una distanza medio-corta. Si utilizzano ferri corti (come l'n.8 o il n.9) o *wedge* (come il *pitching wedge*). L'obiettivo è posizionare la pallina sul *green* il

più vicino possibile alla buca, facilitando il successivo *putt*. Questo colpo richiede precisione e controllo sulla distanza e sulla traiettoria: spesso si adotta una parabola alta per far atterrare la pallina dolcemente sul *green*.

Il ***bunker shot*** è un colpo eseguito dalla sabbia di un *bunker*, un ostacolo specifico del campo. Si utilizza il *sand wedge*, un bastone progettato con una suola ampia e un *loft* elevato (circa 56°) per interagire efficacemente con la sabbia. L'obiettivo è sollevare la pallina dalla sabbia e portarla sul *green* o in una posizione favorevole. Questo colpo richiede una tecnica specifica, colpendo la sabbia dietro la pallina per farla sollevare da terra, ed è fondamentale il controllo della forza e dell'angolo di uscita.

Il ***recovery shot*** viene eseguito da situazioni difficili, come dal *rough* denso o da posizioni con ostacoli che impediscono un colpo diretto. I bastoni utilizzati variano in base alla situazione; spesso si usano bastoni con *loft* adeguato a superare ostacoli o bastoni che permettono di mantenere la pallina bassa. L'obiettivo è tornare in una posizione favorevole sul *fairway* o avvicinarsi al *green*, minimizzando il potenziale impatto negativo sul punteggio. Questo colpo richiede adattabilità e una valutazione accurata delle opzioni disponibili; spesso si devono modificare la posizione del corpo e lo *swing* per eseguirlo correttamente.

Infine, il ***putt*** è il colpo eseguito sul *green* con l'obiettivo di mandare la pallina in buca. Si utilizza il *putter*, uno specifico bastone con un'estremità piatta progettata per far rotolare la pallina sul *green*. L'obiettivo è inserire la pallina in buca con precisione, minimizzando il numero di colpi sul *green*. Questo colpo richiede un'elevata sensibilità e controllo; la lettura delle pendenze e della velocità del *green* sono essenziale per determinare la linea e la forza del ***putt***.

Importanza strategica dei colpi

La scelta del tipo di colpo e del bastone appropriato è fondamentale nel golf, e dipende da vari fattori. Ogni bastone è progettato per coprire una certa distanza, e la corretta selezione gli consente al giocatore di raggiungere l'obiettivo desiderato. Le condizioni del campo, come ostacoli, vento e caratteristiche del terreno, influenzano significativamente la scelta del colpo e del bastone. Inoltre, la posizione della pallina (*lie*) può limitare le opzioni disponibili e richiedere adattamenti nella strategia. In base alla situazione di gara e al proprio punteggio, un giocatore può optare

per scelte più conservative o aggressive, influenzando così l'andamento della partita.

2.2.3 Equipaggiamento e terminologia di base

Il golf richiede un equipaggiamento specifico e una terminologia propria che sono fondamentali per comprendere e praticare il gioco. Di seguito verranno descritti gli elementi essenziali dell'equipaggiamento, in primis un approfondimento sul *set* di bastoni e sulla terminologia specifica utilizzata nel golf, fornendo una panoramica utile anche per i non addetti ai lavori.

Set di bastoni

Ogni giocatore può portare in campo fino a un massimo di 14 bastoni durante una partita, come stabilito dalle regole del golf. Il *set* di bastoni è progettato per permettere al giocatore di affrontare una vasta gamma di situazioni che si presenteranno in campo, offrendo diverse opzioni in termini di distanza, traiettoria e controllo. I bastoni si differenziano per lunghezza, angolo della faccia (*loft*) e design, e si suddividono principalmente nelle seguenti categorie:

- **Legni**: utilizzati per colpi lunghi dal *tee* o dal *fairway*, i legni hanno teste grandi e una faccia con basso *loft*, permettendo di colpire la pallina a lunga distanza. Il *driver* (legno 1) è tipicamente usato per il *tee shot*.

- **Ferri**: impiegati per una vasta gamma di distanze, i ferri offrono maggiore controllo rispetto ai legni. Sono numerati dal 2 al 9, e, all'aumentare del numero, corrispondono a un maggiore *loft*, un minore volo di palla.

- ***Wedge***: si tratta di ferri specializzati con *loft* elevato, utilizzati per colpi corti e precisi vicino al green. Includono il *pitching wedge*, il *sand wedge* e il *lob wedge*.

- **Ibridi**: combinano caratteristiche dei legni e dei ferri, offrendo facilità d'uso e versatilità e sono spesso utilizzati in sostituzione dei ferri lunghi.

- ***Putter***: utilizzato sul *green*, il *putter* è progettato per far rotolare la pallina verso la buca con precisione. Ha una faccia piatta e un design che facilita il controllo della direzione e della velocità.

La scelta del bastone appropriato in una determinata situazione è fondamentale per eseguire il colpo desiderato. Ai giocatori spetta il compito di valutare fattori come la distanza dalla buca, le condizioni del vento, la posizione della pallina e gli ostacoli presenti per decidere quale bastone utilizzare.

Terminologia di gioco

Il termine ***par*** indica il numero standard di colpi assegnato a una buca, basato sulla sua lunghezza e difficoltà. Le buche possono essere classificate come *par* 3, *par* 4 o *par* 5. Completare una buca con un numero di colpi inferiore al *par* è considerato un risultato positivo; ad esempio, terminare una buca *par* 4 in tre colpi viene chiamato ***birdie***. Se si completa la buca con due colpi sotto il *par*, si ottiene un ***eagle***. Al contrario, impiegare più colpi del *par* rappresenta un risultato meno desiderabile: un colpo sopra il *par* è un ***bogey***, due colpi sopra il *par* un ***double bogey***.

L'**handicap** è un sistema che misura il livello di abilità di un giocatore, permettendo a golfisti di diverso livello di competere equamente. Esso rappresenta il numero di colpi che un giocatore può sottrarre dal suo punteggio totale, equilibrando le differenze tra giocatori più o meno esperti.

La terminologia del golf include anche termini relativi alla **traiettoria** della pallina. Ad esempio, per i giocatori destrorsi ,uno ***slice*** è una traiettoria indesiderata in cui la pallina curva eccessivamente verso destra per un giocatore destrorso, mentre un ***hook*** è una curva eccessiva verso sinistra. Comprendere questi termini aiuta a identificare e correggere problemi tecnici nel proprio gioco.

Infine, esistono diverse **modalità di gioco** nel golf, tra cui lo ***stroke play*** e il ***match play***. Nello *stroke play*, vince il giocatore con il minor numero totale di colpi su tutte le buche. Il *match play*, invece, è uno scontro fra due giocatori dove si conta il numero di buche vinte, perse o pareggiate, indipendentemente dal numero totale di colpi.

Capitolo 3

I *dataset*

In questo capitolo vengono presentati i due *dataset* derivanti dalla raccolta dati, ne vengono descritte le variabili e presentate delle analisi esplicative mirate a fornire una panoramica dei dati raccolti.

3.1 Dati 2023

I primi passi in questa ricerca sono stati mossi dal sottoscritto, che ha deciso di raccogliere più variabili di gioco possibile, sia durante le sessioni di allenamento sia durante le competizioni.

Figura 3.1: Esempio di *scorecard*

Tutti i dati sono stati scritti, durante il gioco, sugli *scorecards*, dei fogli prestampati dai vari circoli contenenti le informazioni sulle buche (Figura 3.1), utilizzati solitamente per segnare il punteggio.

Questo *dataset*, che per comodità verrà chiamato `data_2023`, contiene i dati raccolti per 46 sessioni diverse, con inizio il 4 febbraio 2023 e fine

il 23 settembre 2023, per un totale di 530 osservazioni e 15 variabili alle quali vanno aggiunte altre 9 riguardanti le caratteristiche immutabili del campo.

In questo caso ogni osservazione è una buca diversa, il che non permette il calcolo completo degli SG.

3.1.1 Descrizione delle variabili

Di seguito viene fornita una descrizione dettagliata delle variabili contenute nel *dataset* `data_2023`.

- `Data` (*Data*): data della sessione di gioco (formato AAAA-MM-GG).
- `Tee Partenza` (*Categoriale*): *tee* di partenza utilizzato per la buca (es. Giallo, Rosso).
- `BUCA` (*Numerica*): numero della buca giocata.
- `Shots` (*Numerica*): numero di colpi effettuati sulla buca.
- `Putt` (*Numerica*): numero di *putt* effettuati sulla buca.
- `Distance Putt in m` (*Numerica*): distanza del *putt* in metri.
- `Slope` (*Numerica*): indice di difficoltà del campo (*slope rating*).
- `Fairway in Regulation` (*Categoriale*): indica se il *fairway* è stato colpito in regolazione (1 = sì, 0 = no).

  Colpire il *fairway* in regolazione significa che il *tee shot* è atterrato direttamente sul *fairway*
- `Green in Regulation` (*Categoriale*): indica se il *green* è stato raggiunto in regolazione (1 = sì, 0 = no).

  Raggiungere il *green* in regolazione significa che il golfista è arrivato sul *green* con un numero di colpi pari al *par* della buca -2, ovvero ha la possibilità mediante un *one-putt* di chiudere la buca con un *birdie*.
- `Distance to green in m` (*Numerica*): distanza al *green* dall'ultimo colpo prima di entrare in *green*, espressa in metri.
- `FWR Err` (*Categoriale*): direzione dell'errore dal *tee* (es. sinistra, destra, lungo, corto).

- `GIR Err` (*Categoriale*): direzione dell'errore nell'approccio al *green*.
- `Penality` (*Numerica*): numero di penalità ricevute sulla buca.
- `CIRCOLO` (*Stringa*): nome del circolo in cui è stata giocata la buca.
- `Gara` (*Categoriale*): indica se la sessione è una gara ufficiale (1 = sì, 0 = no).

Le 9 variabili aggiuntive sono, per ogni buca, la distanza dei vari *tee* di partenza, l'indice di difficoltà della buca (*HCP index*) e il *par*.

3.1.2 Analisi esplorative su `data_2023`

Come mostrato in Figura 3.2, la distribuzione dei colpi presenta una media μ generale di colpi pari a 4.534 e una varianza σ^2 pari a 1.111.

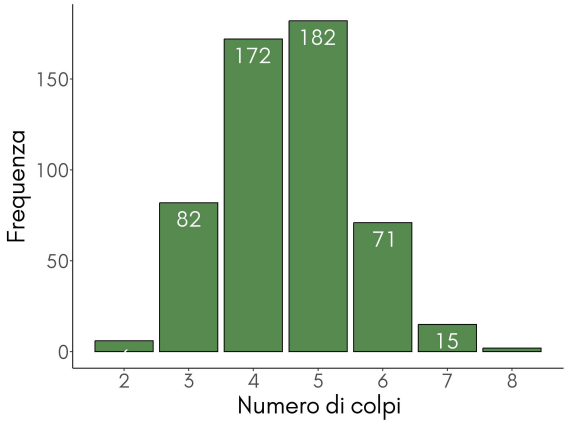


Figura 3.2: Distribuzione dei colpi presenti in `data_2023`

Due variabili importanti per determinare il numero di colpi che un giocatore impiega per terminare la buca, sono la *Fairway in Regulation* e la *Green in Regulation* come mostra la Figura 3.3.

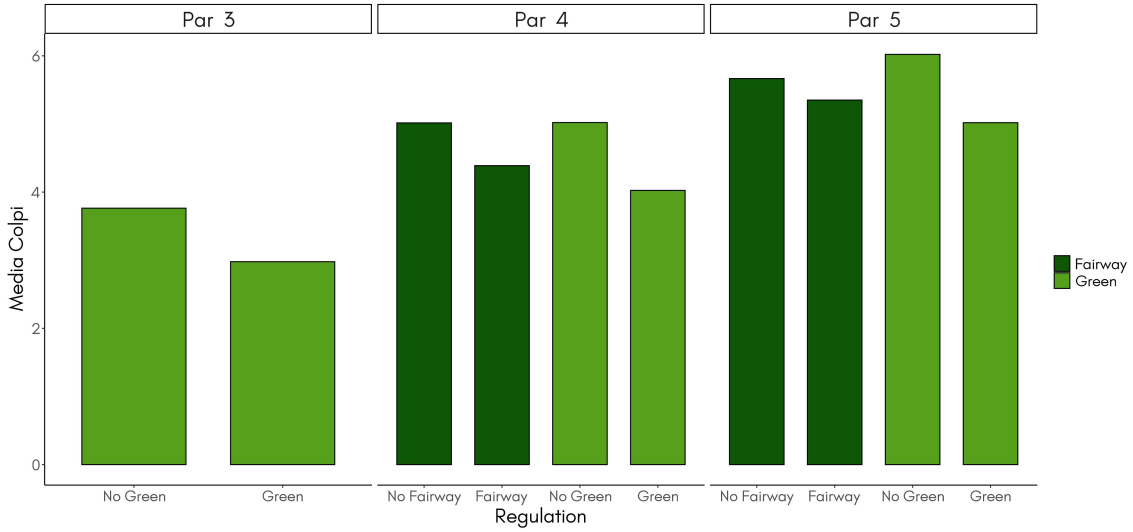


Figura 3.3: Numero di colpi per terminare la buca in base al *Fairway in Regulation* e il *Green in Regulation*

Per quanto riguarda la variabile riferita al *Fairway in Regulation*, aver preso il *fairway* con il colpo dal *tee shot* ha assicurato un guadagno medio di colpi pari a 0.6277 nei *par* 4 e di 0.3158 nei *par* 5.[1]

Riuscire ad arrivare in *green* con un numero di colpi pari al *par* meno 2 (*Green in Regulation*), ha permesso un guadagno medio di 0.7856 colpi nei *par* 3, 0.9957 colpi nei *par* 4 e 1.003 colpi nei *par* 5.

Fra le due variabili prese in considerazione quella che sembra maggiormente discriminante per il numero di colpi è *Green in Regulation* in quanto, da un punto di vista del gioco, se con il primo tiro la palla non finisce in *fairway* ma in *rough*, è comunque possibile riuscire a raggiungere il *green* per poter ottenere il *Green in Regulation*.

[1] I dati per i par 3 relativi a questa variabile non sono stati raccolti in quanto corrisponderebbero alla variabile *Green in Regulation*

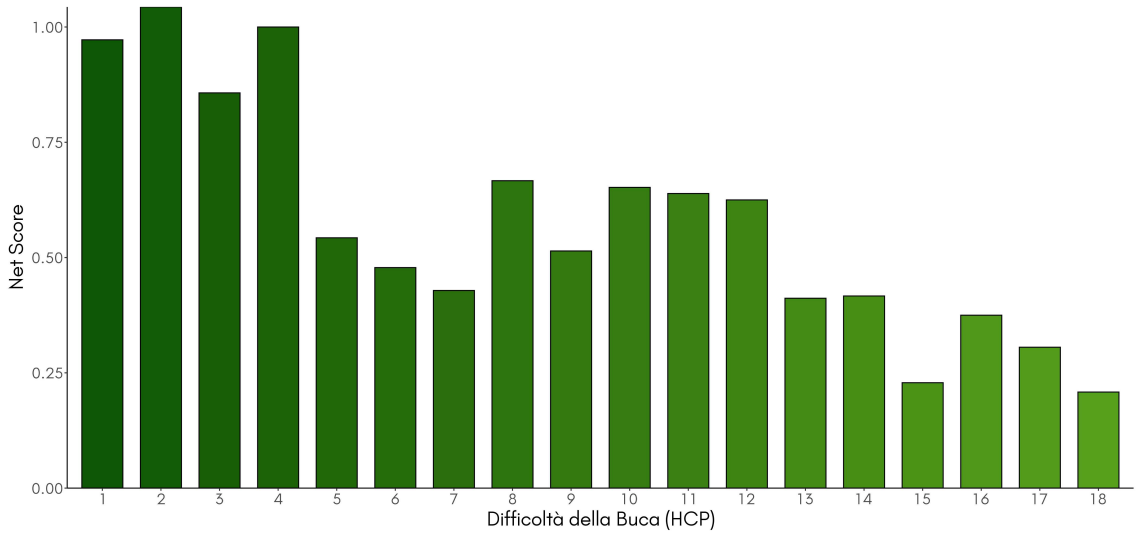


Figura 3.4: Punteggio netto *Net Score* medio in base alla difficoltà della buca (HCP)

Un'altra variabile significativa nel gioco del golf è la difficoltà della buca (`HCP Index`): buche che vengono definite a priori dal *Course Rater* più complicate, è logico che portino a un numero di tiri superiore. A tale scopo è stata creata la variabile `net_score` (punteggio netto) che rappresenta il numero di colpi impiegato dal giocatore − *par*, per indicare come ha performato rispetto alla singola buca, quindi se ha ottenuto un *birdie*, *bogey* o altro.

La Figura 3.4 rappresenta come varia il *net score* in base alla difficoltà buche, e risulta emblematico come buche più difficili abbiamo portato a un numero di colpi medio maggiore, e buche più semplici abbiano portato a un numero di colpi medio minore, come era lecito attendersi.

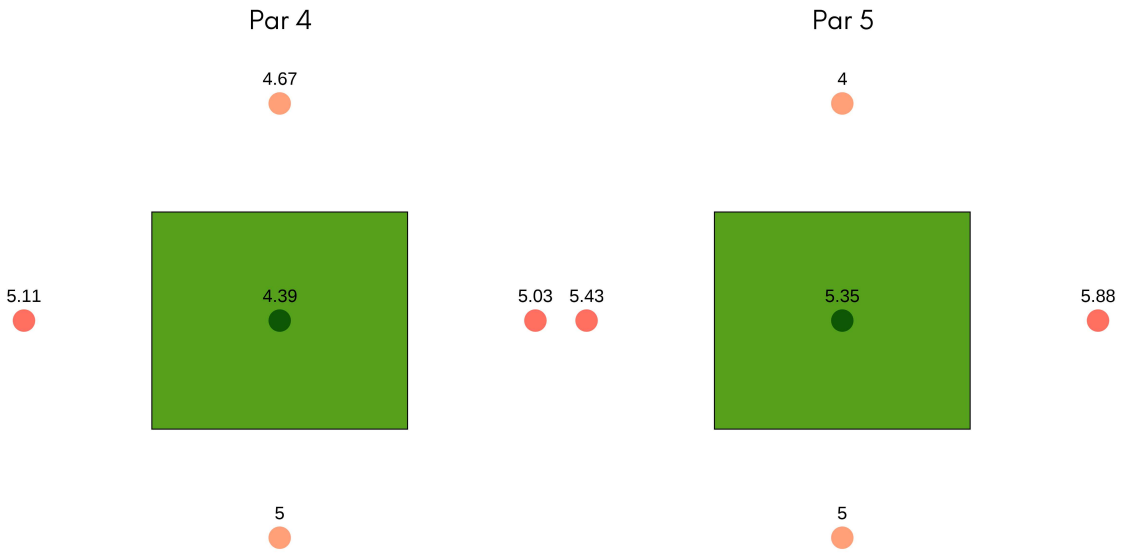


Figura 3.5: Effetto dell'errore sul primo colpo in *fairway* (`FWR Err`) sulla distribuzione dei colpi medi

La Figura 3.5 rappresenta una schematizzazione di un *fairway* dove il rettangolo verde è il *fairway*, quindi il valore della variabile *FWR Err* pari a 0, mentre fuori dal rettangolo ci sono i 4 possibili *outcomes* di errore (*Long, Short, Right, Left*). Nei *par* 4 la differenza è più marcata in quanto tirare il *driver* all'interno del *fairway* ha permesso di ottenere mediamente 0.72 colpi in meno che "sbagliare a sinistra". C'è una differenza anche nella direzione in quanto, per i dati in possesso, tirare il primo colpo a destra ha fatto guadagnare mediamente 0.08 colpi rispetto che tirare a sinistra.

Nei *par* 5, complice anche la scarsa numerosità sui valori *long* e *short* della variabile `FWR Err`, la differenza è più sottile; tirare il primo colpo nel *fairway* comunque permette mediamente di finire la buca in meno colpi.

Infine, la Figura 3.6 mostra la *heatmap* di correlazione fra le variabili quantitative, escluse le distanze dei vari *tee* che sono inutili all'analisi

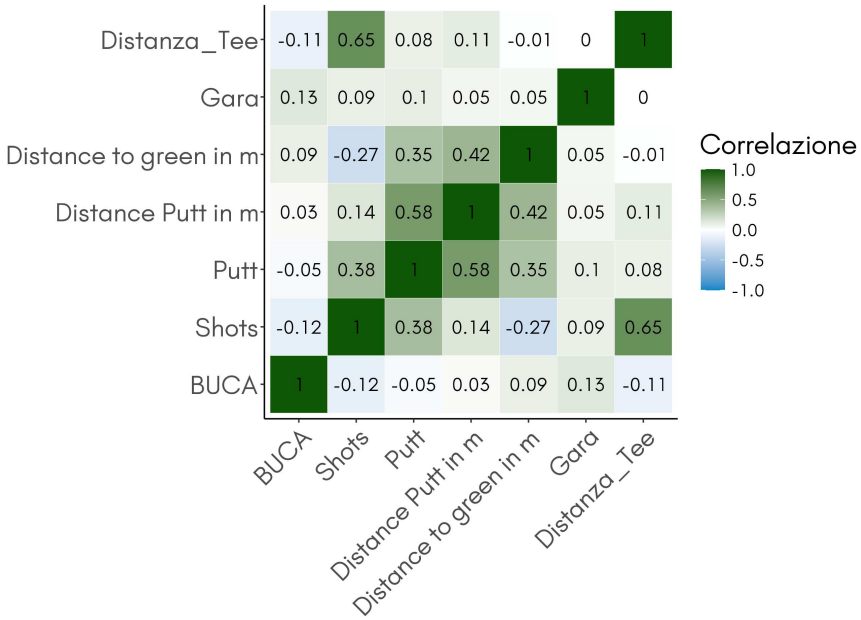


Figura 3.6: *Heatmap* di correlazione tra le variabili quantitative di `data_2023`

Come era lecito aspettarsi, le variabili maggiormente correlate sono `Distance Putt in m` e `Putt`, quindi più aumenta la distanza dalla buca, più il numero di *putt* per completarla aumenta [13]. La variabile maggiormente correlata con la variabile risposta è la lunghezza della buca, definita con `Distanza_Tee`. Da un punto di vista golfistico è anche logico che maggiore è la distanza dalla buca, maggiore è il numero di colpi che serve per impiegare a terminarla.

Le sezioni 4.1, 4.2 e 4.3 presenteranno 3 modelli diversi per implementare gli *strokes gained* utilizzando i dati presenti in questo *dataset*.

Nella prossima sezione verrà descritto il *dataset* proveniente dalla raccolta dati dei 4 soggetti che hanno partecipato allo studio, con conseguente analisi esplorativa e suddivisione in base al livello di abilità.

3.2 *Dataset* 2024

Considerando che per una trattazione approfondita del calcolo degli *SG* servono i dati "colpo per colpo", dove ogni unità statistica è un colpo effettuato da un determinato giocatore in una certa posizione, era necessario cambiare strumento di raccolta dati. A tale scopo si è deciso di avvalersi di un *software* di terze parti per *mobile*, ***GolfPad***, preferito ai suoi *competitors* grazie alla comoda *Graphic User Interface* e alla possibilità di esportare i dati facilmente.

Questa applicazione (Figura 3.7) offre una schermata dove, una volta scelto il campo in cui si vuole giocare, sono mostrate le distanze, rispettivamente a fine, centro e inizio *green*. Oltre a mostrare lo *scorecard* in alto, il *software* ti dà la possibilità di tracciare i propri colpi [2].

Figura 3.7: *Screenshot* dell'applicazione

Per maggiori informazioni sulle tecniche di raccolta dati si rimanda all'appendice B

Il numero di *rounds* totali registrati è 61, rendendo così il *dataset* `shots` con 2337 osservazioni.

È possibile eseguire una suddivisione del *dataset* in base all'abilità dei giocatori che vi hanno contribuito:

[2] La spiegazione sul funzionamento dell'applicazione viene fornita ai soggetti dello studio mediante un approfondito documento esterno

- `5 - 15 HCP` includendo i giocatori il cui handicap varia da 5 a 15
- `15,1 - 30 HCP` includendo i giocatori il cui handicap varia da 15,1 a 30.

Questo porta ad avere 2 *dataset shots* diversi, rispettivamente di 1075 e 1262 osservazioni.

3.2.1 QGis e Golf Club Ca'Amata

Si è deciso, per lo scopo di questa ricerca, di focalizzarsi sui *rounds* giocati solamente al circolo **Golf Club Ca'Amata**, ubicato a Castelfranco Veneto (TV), in quanto rappresentano il 78% dei dati registrati.

A tal proposito il campo è stato mappato nella sua interezza attraverso il *software* `QGIS`, che permette di analizzare ed editare dati spaziali e di generare cartografia.

La mappatura spaziale del campo è servita per implementare nel *dataset* `shots` le variabili relative alle coordinate X e Y dei colpi, utili successivamente nei modelli.

Di ogni singola buca sono stati selezionati e rappresentati sette elementi fondamentali:

- *Tee*
- *Fairway*
- *Bunker*
- *Water Hazard*
- *Green*
- *Flagpole*
- *Skeleton* che rappresenta la linea mediale del *fairway*, calcolata prendendo l'insieme di punti equidistanti dal perimetro dello stesso.

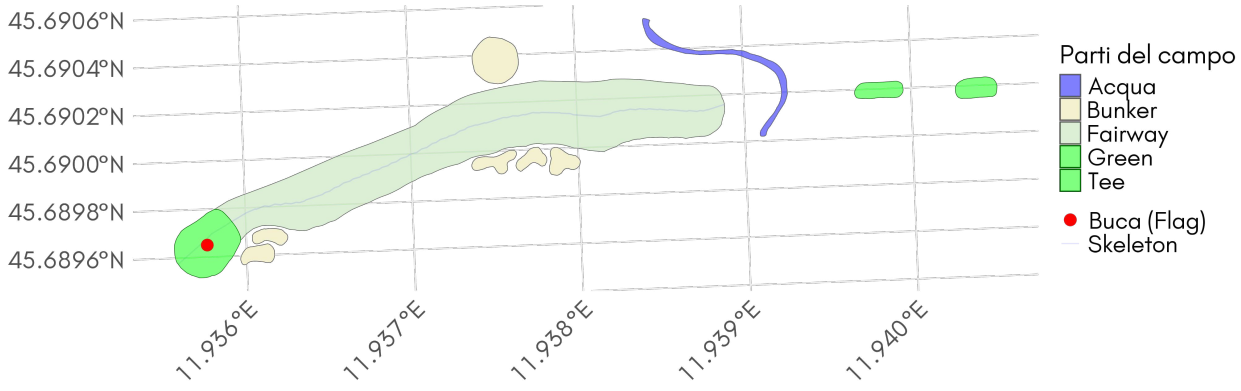


Figura 3.8: Rappresentazione della buca 2

Ogni elemento è stato salvato in un file singolo con estensione `.shp` (*shapefile*) contenente il/i poligono/i che lo compongono, con annesse le informazioni di CRS (*Coordinate Reference Systems*). Per la zona del circolo si è scelto il sistema di riferimento di coordinate *Universal Transverse Mercator* (UTM zone 32N) che ha la sua origine sull'Equatore a una longitudine specifica dove i valori Y aumentano verso sud e i valori X aumentano verso ovest.

Nei dati che *GolfPad* esporta non sono presenti le esatte coordinate geografiche dei singoli colpi, perciò queste informazioni sono state ricavate utilizzando i dati forniti, ovvero l' `outcome` del colpo, la lunghezza del colpo (`shot length meters`) e la distanza laterale in metri dal centro del *fairway* (`dist_cen`).

Per capire come vengono calcolate le coordinate di ogni colpo, viene riportato di seguito un esempio.

Supponiamo che nella prima buca venga effettuato il primo tiro dal *tee shot*, del quale conosciamo le coordinate (essendo un poligono si prende il centroide come riferimento), che atterra bene nel *fairway* (del quale conosciamo le coordinate dello scheletro) e percorre 224 metri.

Tabella 3.1: Esempio dei dati del colpo

hole number	**shot number**	**lie**	**outcome**	**dist_cen**	**shot length meters**
2	1	Tee	Fairway	-5	224

Dal punto di vista teorico nel golf, quando si esegue un colpo, esso copre una certa distanza in linea d'aria, ma può farlo in tutte le direzioni. Quindi è possibile schematizzare il punto di atterraggio della pallina mediante una circonferenza con raggio pari alla distanza che percorre, escludendo l'effetto di rimbalzi dovuti ad ostacoli.

Quindi per trovare le coordinate del primo tiro viene disegnata la circonferenza, viene individuato il punto in cui interseca lo *skeleton* del *fairway* e poi l'informazione `dist_cen` dice di quanti metri ci si deve muovere sulla circonferenza (segno positivo = senso orario, segno negativo = senso antiorario) trovando così il punto di atterraggio del colpo.

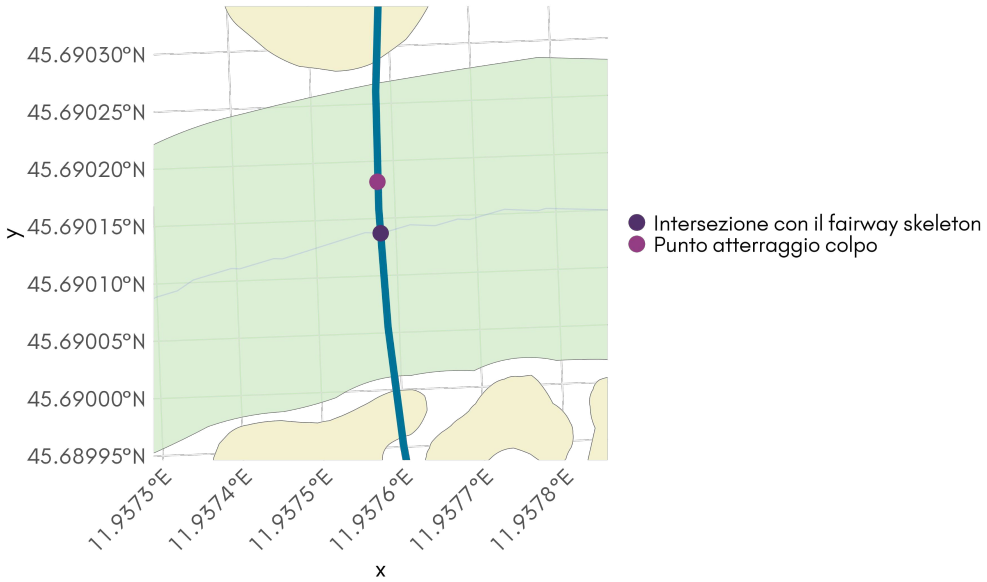


Figura 3.9: Rappresentazione di un colpo di esempio alla buca 2

In tal modo si è riusciti a mappare tutti i colpi presenti all'interno del dataframe `shots`, per il circolo Golf Club Ca'Amata (Figura 3.12).

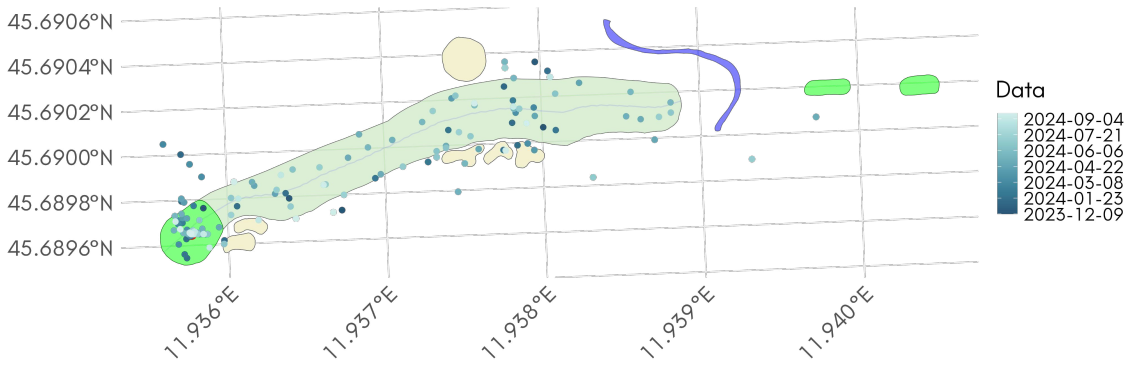


Figura 3.10: Rappresentazione di tutti i colpi raccolti alla buca 2

3.2.2 Analisi univariate

In questo paragrafo si procede analizzando i dati raccolti per i due *dataset* diversi. La prima cosa che salta all'occhio è la naturale differenza di distanza percorsa dai vari colpi in base al proprio HCP.

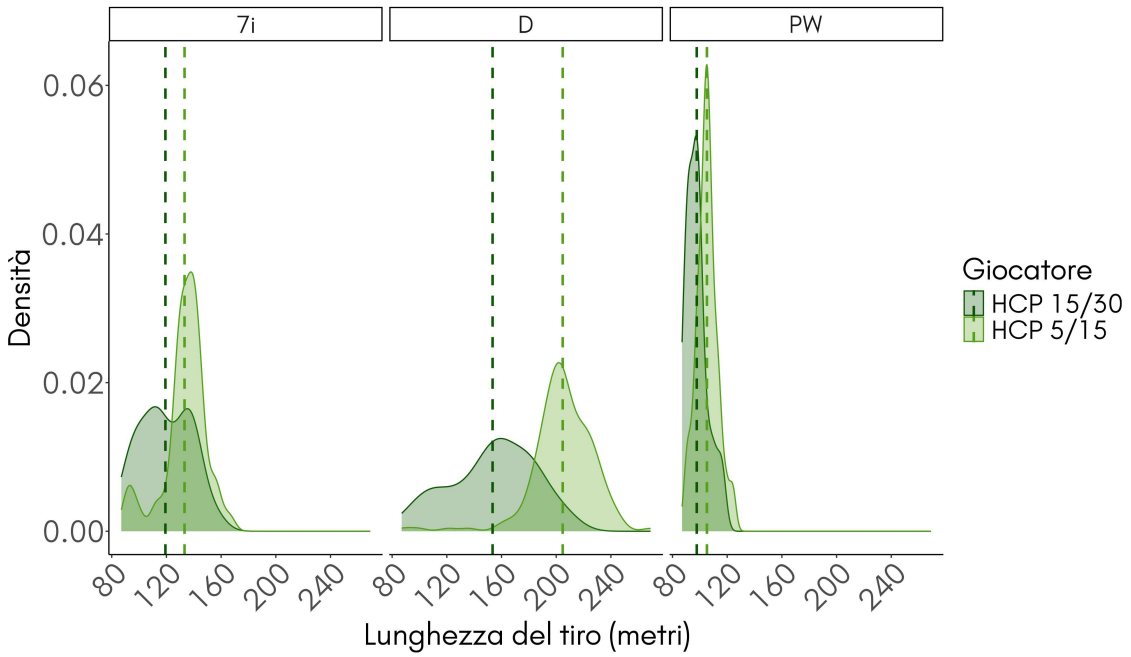


Figura 3.11: Differenze nella media della distanza percorsa per 3 mazze diverse: *driver*, *ferro 7*, *pitching wedge*

I giocatori con HCP bassi, ai quali corrisponde un livello di gioco più alto, tendono a tirare più lungo rispetto ai giocatori con HCP alti, come si può evincere dal grafico in Figura 3.13. Maggiore è il ferro giocato, maggiore risulta essere la differenza in distanza percorsa dalle due tipologie di giocatore. Da questi grafici di densità è possibile notare anche la differenza in variabilità dei colpi: un giocatore con HCP alto tenderà ad avere varianza più bassa rispetto a un giocatore con HCP basso, ovvero sarà più preciso e tenderà a tirare i colpi tutti alla stessa distanza. In questa rappresentazione sono stati eliminati i colpi di *recovery* per queste mazze, quindi quei colpi, volutamente fatti con esse, che percorrono poca distanza e che servono per recuperare da situazioni di svantaggio.

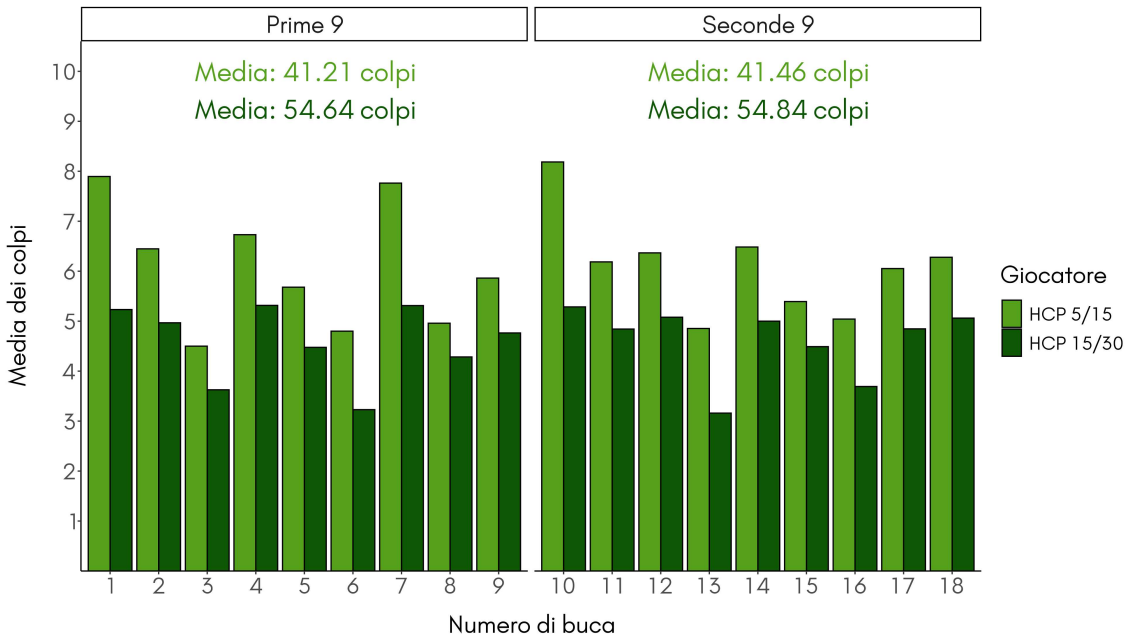


Figura 3.12: Distribuzione dei colpi medi in base al numero di buca

La Figura 3.12 illustra come le due tipologie di giocatori hanno performato sulle 18 buche, con una suddivisione fra le prime 9 (*Front 9*) e le seconde 9 (*Back 9*).

Essendo l'HCP una misura di "quanti colpi aggiuntivi ti fornisco in base al tuo livello", è possibile notare che il giocatore fra 5 e 15 di HCP percorre il campo mediamente in +11 rispetto al *par* del campo (71), mentre il giocatore con HCP fra 15 e 30 in +37.

Questo rispecchia il livello dei giocatori, in quanto in fase di "gara" il HCP viene leggermente modificato per sovvertire alle difficoltà del campo. Per valori di HCP alto questa modifica è più visibile.

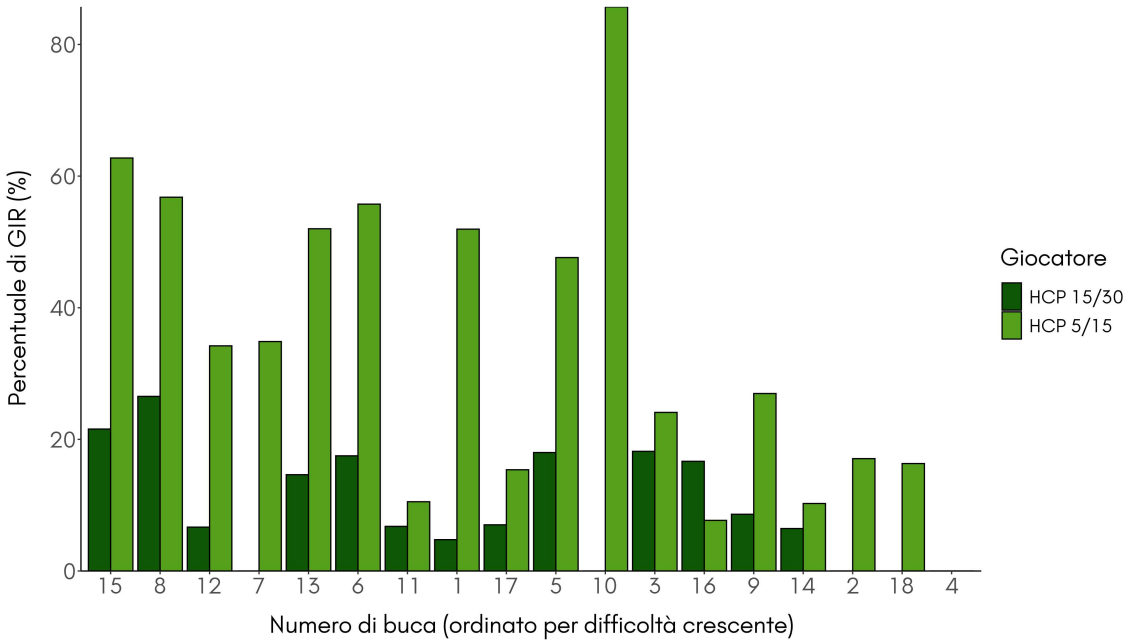


Figura 3.13: Distribuzione della variabile `Green in Regulation` per ogni giocatore a ogni buca

La Figura 3.13 mostra, per entrambi i giocatori, la distribuzione della *Green in Regulation*, ovvero se la buca è stata raggiunta con un numero di colpi pari al *par* della buca − 2.

Il giocatore con il livello più basso ha un valore medio di questa metrica più basso in praticamente tutte le buche, con un valore massimo alla buca 8, pari al 26,53%, mentre per alcune buche (le più difficili del campo come la 2, la 4 e la 18) non presenta alcun *Green in Regulation*. Una peculiarità in questo dato è che il giocatore con HCP fra 15 e 30 ha avuto maggiore difficoltà a registrare *Green in Regulation* nelle buche con *par* 5 (la buca numero 7, 10 ad esempio) rispetto alle buche con *par* 4 o *par* 3.

Il giocatore con HCP più basso invece mostra una buona percentuale di *Green in Regulation*, in quanto la percentuale media totale è pari al 33%, dove spicca la buca 10, definita dal campo come la ottava buca più complicata, con una percentuale altissima del 85.71%.

La buca più ostica è la numero 4, un *par* 4 di oltre 400 metri con un *dog-leg* a sinistra, e vari altri ostacoli come acqua, alberi e fuori limite che ne caratterizzano la complessità. Difatti il punteggio medio in questa buca è il più alto fra tutti i *par* 4 e più alto di alcuni *par* 5 per entrambi i giocatori, con 6.731 colpi per il giocatore con HCP alto e 5.318 per il giocatore con HCP basso.

3.2.3 Analisi di correlazione

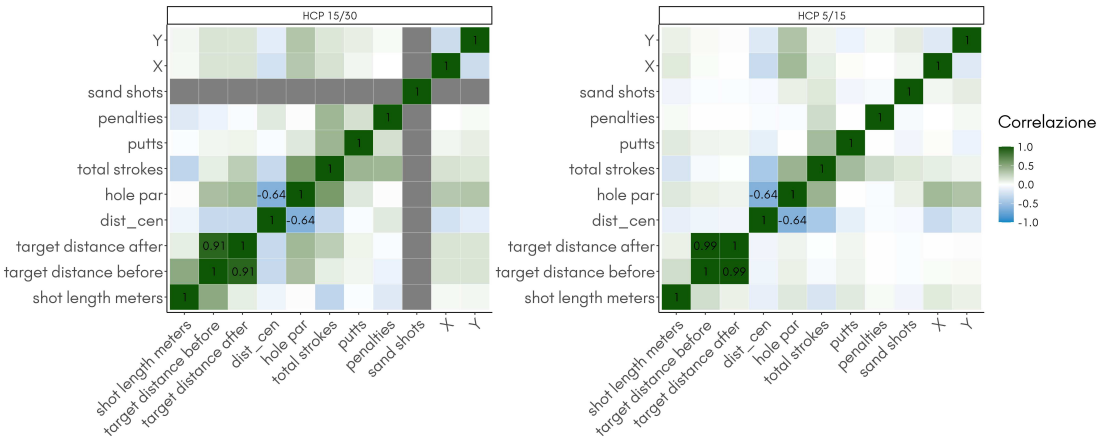


Figura 3.14: *Heatmap* di correlazione fra le variabili del *dataset*

La Figura 3.14 mostra la correlazione fra le variabili del *dataset* ed è possibile evidenziare che le variabili prese in considerazione non presentano forti correlazioni, se non `target distance after` `target distance before`, collineari in quanto la loro differenza produce la variabile `shot length meters`.

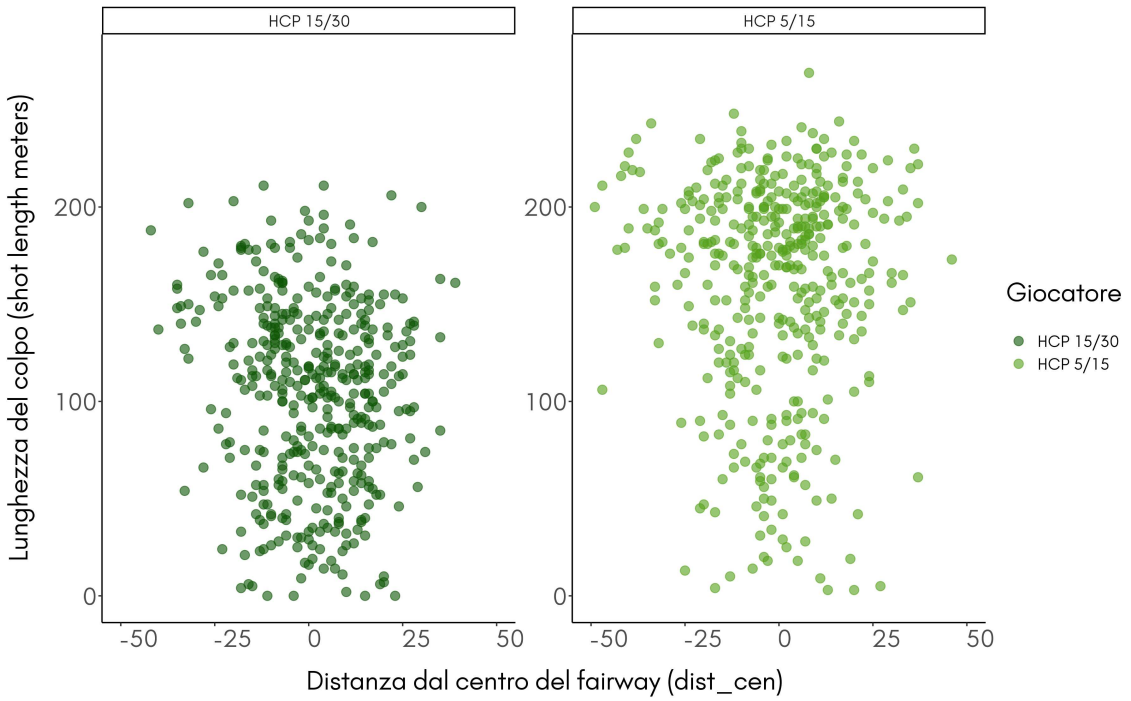


Figura 3.15: Grafico di dispersione fra la lunghezza del colpo e la distanza dal *fairway*

Con la Figura 3.15 si nota la differenza fra gli stili di gioco dovuti alla differenza di livello; il giocatore con HCP più basso presenta una dispersione laterale dei colpi dal *fairway* inferiore rispetto al giocatore con HCP più alto.

In generale si vede come più aumenta la distanza percorsa dal colpo e più aumenta la distanza dal centro del *fairway*, quindi diminuisce la precisione. Da un punto di vista del fisico e di questo sport è un qualcosa di naturale, più strada c'è da percorrere e minore è la probabilità di raggiungere un bersaglio che resta sempre delle stesse dimensioni.

Capitolo 4

Modellazione: dati 2023

In questo capitolo verranno proposti 3 modelli diversi per il calcolo degli *strokes gained*, in base alla diversa situazione diversa di gioco. Come già presentato nella sezione 1.2, essi sono una somma telescopica, e in quanto tale, scindibile nelle singole parti. Essendo questi dati non comprensivi di tutti i colpi, ma delle buche, e di informazioni generali su alcuni colpi, verranno definiti modelli diversi per *tee shot*, *approach shot* e *putt shot*. Ogni modello verrà introdotto da una breve descrizione teorica dello strumento statistico utilizzato.

Considerata l'esigua dimensione del *dataset*, per ogni modello è stata eseguita prima una verifica tramite convalida incrociata a 10 *fold*, per poter scegliere il *fold* migliore.

4.1 *Tee shot benchmark*

Per iniziare l'analisi con i dati raccolti nel 2023 si parte andando a costruire un modello per adattare il *benchmark* a partire dal primo colpo, ovvero il *tee shot*. Quindi si va a definire per ogni buca con quanti colpi mediamente ci si aspetta di finirla. Da un punto di vista logico, il numero di colpi per terminare una buca dovrebbe tendenzialmente crescere in base alla distanza dalla bandiera.

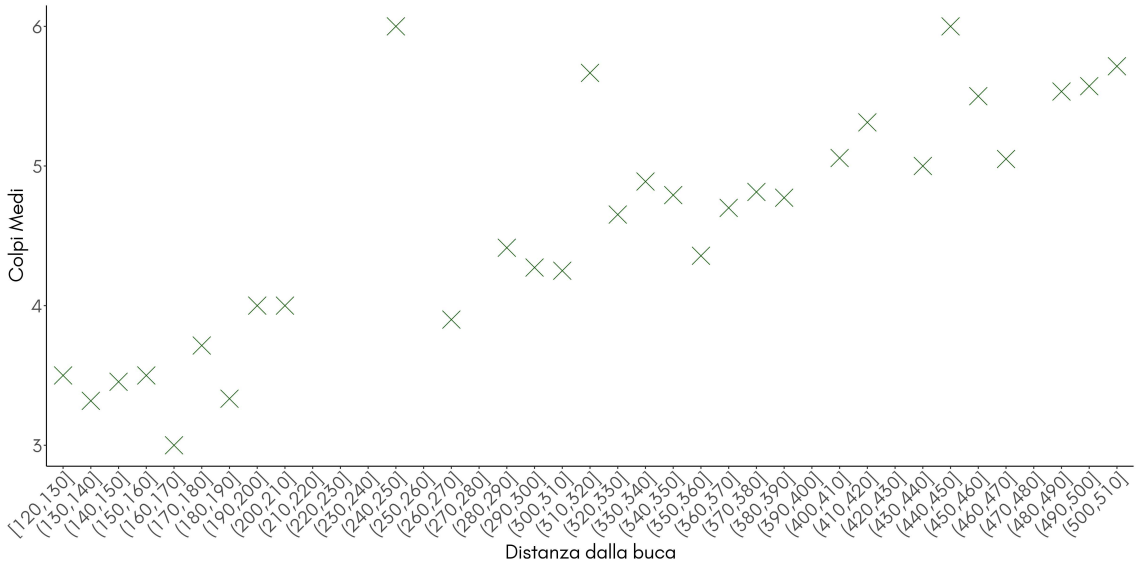


Figura 4.1: Colpi medi per terminare la buca in base alla distanza dalla bandiera

4.1.1 Modello lineare

Per analizzare la relazione tra il numero di colpi necessari per completare una buca e le varie caratteristiche del gioco, si utilizza un modello di regressione lineare multipla. Questo modello permette di descrivere la relazione tra una variabile dipendente continua e una o più variabili indipendenti.

Il modello di regressione lineare multipla è definito come:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i \tag{4.1}$$

dove:

- y_i è la variabile dipendente per l'osservazione i (ad esempio, il numero medio di colpi per terminare la buca)
- x_{ji} è la variabile indipendente j-esima per l'osservazione i (ad esempio, distanza dalla bandiera, condizioni meteorologiche, tipo di terreno, etc.)
- β_0 è l'intercetta del modello, rappresenta il valore atteso di y quando tutte le $x_j = 0$
- β_j è il coefficiente di regressione associato alla variabile x_j, indica la variazione attesa in y per una variazione unitaria in x_j, mantenendo costanti le altre variabili

- ε_i è il termine di errore, una variabile aleatoria che rappresenta la differenza tra il valore osservato e quello previsto dal modello. Si assume che ε_i sia distribuito normalmente con media zero e varianza costante σ^2.

L'obiettivo del modello di regressione lineare multipla è stimare i coefficienti $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ che minimizzano la somma dei quadrati degli errori (ε_i). Questo metodo è noto come *Metodo dei Minimi Quadrati Ordinari* (*Ordinary Least Squares*, OLS).

La funzione di perdita da minimizzare è:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ji} \right)^2 \tag{4.2}$$

dove n è il numero totale delle osservazioni e p è il numero delle variabili indipendenti.

Una volta stimati i coefficienti $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, il modello può essere utilizzato per prevedere il valore atteso di y per nuovi valori delle variabili indipendenti:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{4.3}$$

Per selezionare le variabili è stata attuata una procedura *stepwise* ibrida con doppia direzione: si è partiti dal modello nullo senza predittori e si aggiungeva la variabile maggiormente rilevante eliminando di conseguenza quella che portava meno informazione. Da questa procedura sono risultate le variabili in tabella:

	Estimate ($\hat{\beta}$)	Std. Error	t-value	Pr($>\|t\|$)	
`(Intercept)`	1.654	0.169	9.781	<0.001	***
`Distanza_Tee`	0.005	0.000	15.298	<0.001	***
`Green.in.Regulation1`	−1.001	0.074	−13.573	<0.001	***
`Putt`	0.968	0.048	20.118	<0.001	***
`Fairway.in.Regulation1`	−0.235	0.052	−4.486	<0.001	***
`Fairway.in.Regulationunknown`	−0.273	0.099	−2.768	0.0059	**
`HCP`	0.012	0.005	2.503	0.0127	*
`Distance.to.green.in.m`	−0.001	0.001	−2.193	0.0288	*
`Distance.Putt.in.m`	−0.008	0.006	−1.411	0.1590	

Residual standard error: 0.4686 on 468 degrees of freedom
Multiple R-squared: 0.801, Adjusted R-squared: 0.7976
F-statistic: 235.5 on 8 and 468 DF, p-value: < 2.2e-16

Tabella 4.1: Risultati del modello di regressione lineare

La Tabella 4.1 presenta i risultati del modello lineare mentre la Figura 4.2 mostra gli effetti parziali dei predittori sulla variabile risposta.

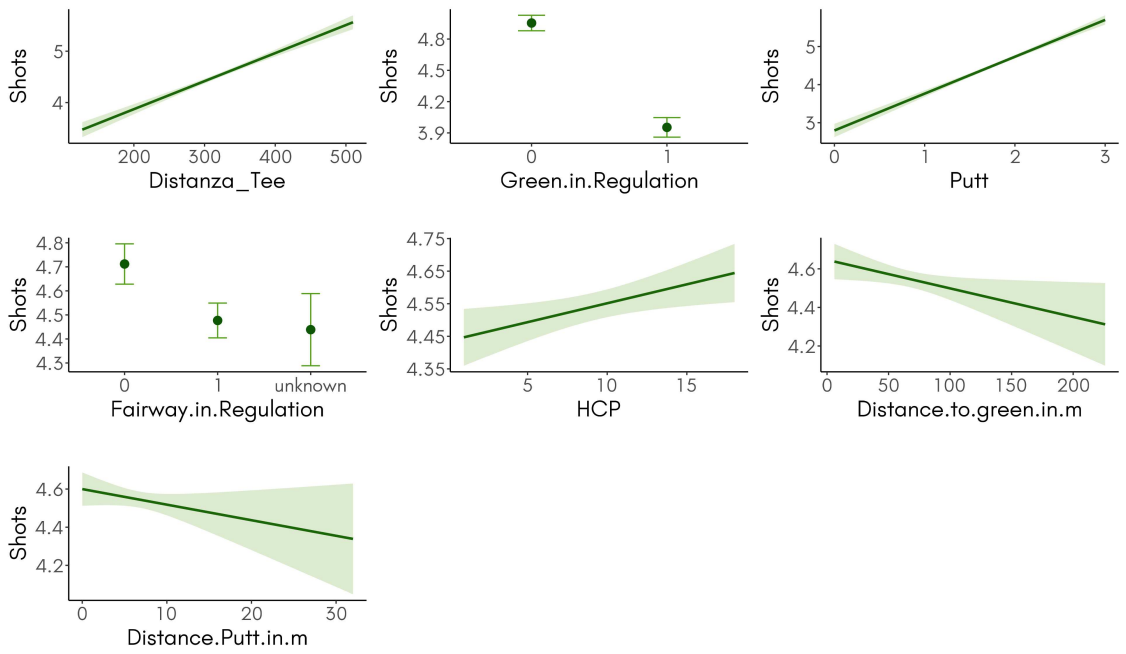


Figura 4.2: Effetti parziali dei predittori sul numero di colpi

La variabile che presenta un risultato contro intuitivo è `HCP`, principalmente perché all'interno di questa variabile manca l'informazione sul *par* della buca, perciò potrebbe essere che le 4 buche più difficili siano ad esempio *par* 3, che quindi inevitabilmente impiegano meno colpi per essere completate. La soluzione potrebbe essere quella di considerare il *net score*, ma questo porta problemi di interpretazione delle altre variabili.

Interessante anche il caso della variabile `Distance.to.green.in.m`, ovvero la distanza dalla quale si tira il colpo prima di arrivare in *green*: se esso avviene da una distanza ravvicinata al *green* è dovuto al fatto che molto probabilmente non è il colpo per la *Green in regulation*, e quindi durante la buca sono stati effettuati colpi aggiuntivi. Al contrario, colpi al *green* da 80, 100, 150 metri possono indicare che sono colpi che hanno portato la *Green in Regulation*.

Il modello prevede il numero di colpi con un errore di 0.4686, mentre l'indice di bontà di adattamento è pari a 0.801.

Spline di `Distanza_Tee`

Successivamente si è provato a modificare il predittore relativo alla variabile `Distanza_Tee`, provando a usare un lisciatore non parametrico definito da una *base spline*. Una spiegazione maggiormente dettagliata sulle *splines* verrà fornite nella sezione 4.2.

Il processo di selezione *stepwise* delle variabili ha portato ai seguenti risultati:

	Estimate ($\hat{\beta}$)	Std. Error	t-value	Pr(> $\lvert t \rvert$)	
`(Intercept)`	3.420	0.328	10.413	<0.001	***
`bs(Distanza_Tee, df = 4)1`	0.625	0.279	2.243	0.0254	*
`bs(Distanza_Tee, df = 4)2`	−0.678	0.517	−1.312	0.1901	
`bs(Distanza_Tee, df = 4)3`	1.028	0.297	3.463	0.0006	***
`bs(Distanza_Tee, df = 4)4`	1.279	0.361	3.539	0.0004	***
`Green.in.Regulation1`	−1.041	0.071	−14.661	<0.001	***
`Putt`	0.939	0.042	22.480	<0.001	***
`Fairway.in.Regulation1`	−0.243	0.052	−4.676	<0.001	***
`Fairway.in.Regulationunknown`	−1.203	0.285	−4.224	<0.001	***
`Distance.to.green.in.m`	−0.001	0.001	−2.337	0.0199	*

Residual standard error: 0.4675 on 467 degrees of freedom
Multiple R-squared: 0.8103, Adjusted R-squared: 0.8066
F-statistic: 221.6 on 9 and 467 DF, p-value: < 2.2e-16

Tabella 4.2: Risultati del modello di regressione lineare con *spline* sui dati della distanza

I nodi significativi della *spline* sono stati posizionati nei punti della distanza del *tee* pari a 127 metri, 341 metri e 510 metri.

Il numero di variabili è diminuito drasticamente, in quanto sono state escluse dal modello le variabili relative alla difficoltà della buca (HCP) la cui informazione è stata catturata dalla *spline* di `Distanza_Tee` e la distanza del primo colpo del *putt* dalla buca.

La Figura 4.3 presenta gli effetti parziali dei predittori sul numero di colpi per i modello lineare con questa modifica.

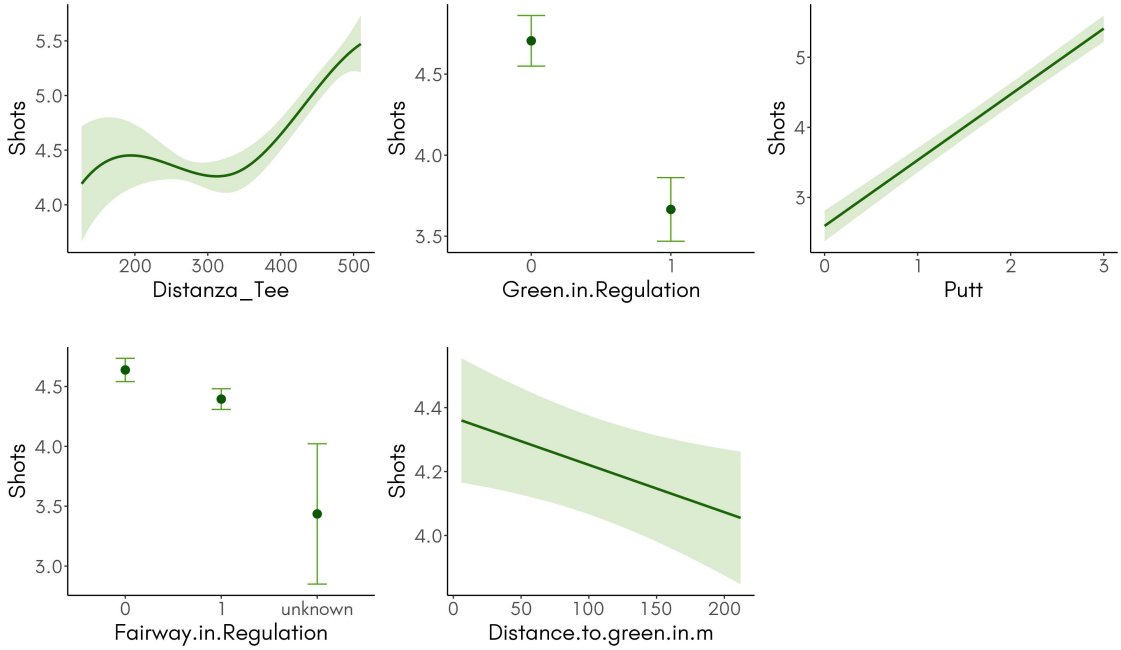


Figura 4.3: Effetti parziali dei predittori sul numero di colpi

Ovviamente, arrivare in *green* con `Green in Regulation` ha un effetto significativo sulla diminuzione del numero di colpi, come anche colpire il *fairway* con il *driver*. La modalità *unknown* della variabile `Fairway.in.Regulation` fa riferimento a quelle buche in cui non c'è il *fairway*, motivo per il quale il valore è così basso rispetto alle altre modalità.

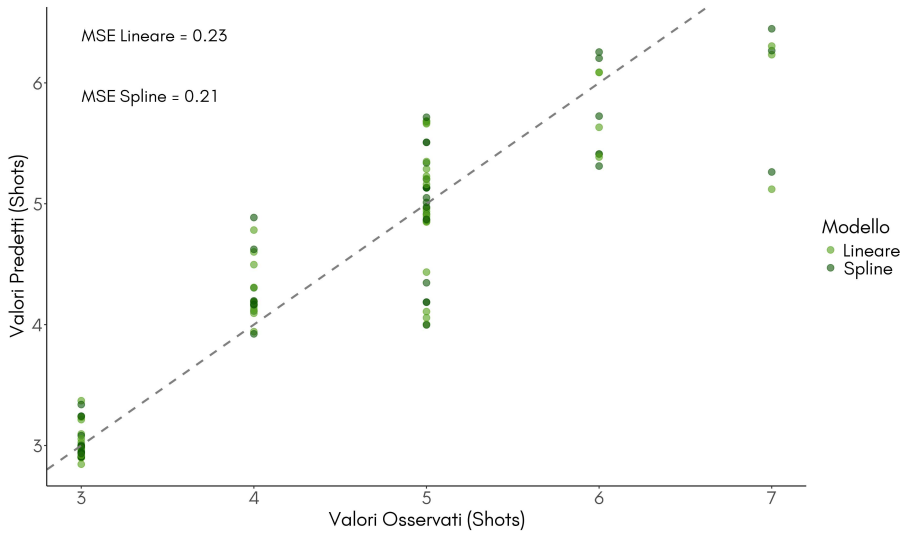


Figura 4.4: Grafico Valori predetti vs Valori osservati

La Figura 4.4 mostra il grafico dei valori predetti confrontati con i valori osservati, mettendo in risalto la differenza fra i due modelli. Il modello che presenta la *spline* su `Distanza_Tee` ha un valore di *Mean Squared Error* inferiore a quello lineare.

4.2 *Approach shot benchmark*

In questa sezione verrà implementato un modello diverso per i colpi di approccio al *green*, per i quali abbiamo i dati, prima introducendo poi utilizzando i GAM (*Generalized Additive Models*).

4.2.1 Generalized Additive Models (GAM)

I modelli additivi generalizzati (**GAM**, dall'inglese *Generalized Additive Models*) sono una classe di modelli statistici flessibili che estendono i modelli lineari generalizzati (**GLM**). Essi permettono di modellare la relazione tra la variabile dipendente e le variabili indipendenti utilizzando funzioni additive non lineari, offrendo così una maggiore flessibilità nel catturare andamenti complessi nei dati.

In un GAM, la relazione tra la variabile di risposta Y e le variabili predittive X_i è espressa come:

$$g(\mathbb{E}[Y]) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

dove:

- $g(\cdot)$ è la funzione di link che collega la media della distribuzione della risposta alla scala lineare dei predittori.
- β_0 è l'intercetta.
- $f_i(\cdot)$ sono funzioni lisce non parametriche da stimare dai dati.

I GAM consentono di modellare relazioni non lineari tra le variabili indipendenti e la variabile dipendente, senza specificare una forma funzionale predeterminata. La loro struttura additiva facilita l'interpretazione degli effetti individuali delle variabili predittive. Le funzioni f_i sono stimate utilizzando metodi di *smoothing*, come le *splines* penalizzate, il LOESS o la regressione locale, che controllano quanto sono lisce le funzioni stimate.

La stima dei parametri in un GAM coinvolge:

- **Selezione del parametro di *smoothing***: determina il grado di flessibilità delle funzioni f_i, bilanciando tra adattamento ai dati e *overfitting*.

- **Penalizzazione**: l'aggiunta di un termine di penalizzazione nella funzione di verosimiglianza riduce la complessità delle funzioni stimate.

- **Metodi di ottimizzazione**: utilizzo di algoritmi numerici, come il metodo di *backfitting*, per stimare iterativamente le funzioni lisce.

Splines

Le *splines* sono funzioni utilizzate in analisi numerica e statistica per approssimare funzioni complesse o dati osservati. Sono particolarmente utili per modellare relazioni non lineari tra variabili. Una *spline* è una funzione a tratti composta da polinomi definiti su intervalli adiacenti, uniti in modo tale da garantire una certa continuità fino a un determinato ordine di derivata.

Formalmente, data una sequenza di nodi (*knots*)

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b \tag{4.4}$$

una funzione $S(x)$ è una *spline* di grado k se:

1. Su ogni intervallo $[x_i, x_{i+1}]$, $S(x)$ è un polinomio di grado al più k.

2. $S(x)$ è $k-1$ volte continuamente derivabile su $[a, b]$.

Le *splines* sono utilizzate in regressione per modellare relazioni non lineari tra una variabile dipendente e una o più variabili indipendenti. In un modello di regressione *spline*, la relazione tra la variabile dipendente Y e la variabile indipendente X è modellata come:

$$Y(x, \beta) = \sum_{j=1}^{K+4} \hat{\beta}_j h_j(x)$$

dopo aver definito delle funzioni di base , $h_1(x) = 1$, $h_2(x) = x$, $h_3(x) = x^2$, $h_4(x) = x^3$.

Le *splines* sono uno strumento potente per la modellizzazione di dati non lineari, offrendo un equilibrio tra flessibilità e controllo della complessità del modello.

4.2.2 Applicazione ai dati

Le variabili che si è deciso di inserire all'interno del modello additivo sono la distanza del *putt* in metri (ovvero quanto è distante dalla buca il punto di arrivo del colpo di approccio), la distanza dalla quale viene eseguito l'approccio (ovvero la distanza dalla buca prima di eseguire il colpo), l'inclinazione del *green* in direzione della buca (`Slope`) e la tipologia di errore per la quale non si è raggiunto il *Green in Regulation*.

Inoltre, vista la natura dei dati, si è deciso di creare due variabili ad hoc, di cui una diventa la nuova variabile risposta. In primis si è suddivisa la distanza dal *green* in degli intervalli non regolari, successivamente si è calcolato il numero di colpi medi per ogni intervallo di distanza. Il numero di colpi medi diventa così la nuova variabile risposta.

	Estimate	Std. Error	t-value	Pr(> |t|)	
`(Intercept)`	2.346	0.014	171.746	<0.001	***
`SlopeP`	0.010	0.013	0.786	0.4323	
`SlopeU`	0.031	0.014	2.203	0.0281	*
`Fairway.in.Regulation1`	0.437	0.010	42.615	<0.001	***
`Fairway.in.Regulationunknown`	0.213	0.007	32.383	<0.001	***
`FWR.ErrDx`	0.422	0.013	31.857	<0.001	***
`FWR.ErrLg`	0.440	0.022	19.604	<0.001	***
`FWR.ErrSh`	0.410	0.037	11.053	<0.001	***
`FWR.ErrSx`	0.426	0.014	29.888	<0.001	***
`FWR.Errunknown`	0.213	0.007	32.383	<0.001	***
`GIR.ErrDx`	0.006	0.022	0.289	0.7727	
`GIR.ErrLg`	−0.020	0.025	−0.791	0.4292	
`GIR.ErrSh`	0.015	0.015	0.999	0.3186	
`GIR.ErrSx`	−0.007	0.021	−0.349	0.7273	

Approximate significance of smooth terms:
`s(Distance.to.green.in.m)` edf = 8.061, Ref.df = 8.766, F = 98.38, p-value = < 0.001 ***
Rank: 21/23
R-sq.(adj) = 0.832, Deviance explained = 83.9%
GCV = 0.0094335, Scale est. = 0.0089872, n = 424

Tabella 4.3: Risultati del modello GAM con *spline* sui dati della distanza

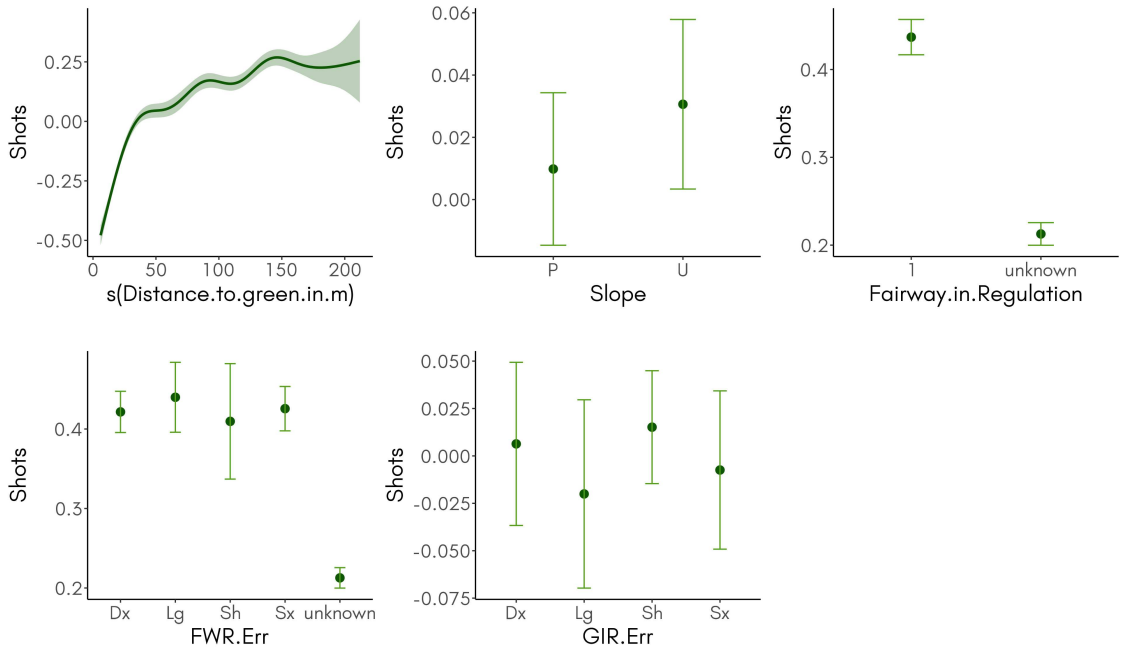


Figura 4.5: Effetti parziali del modello GAM dei predittori sul numero medio di colpi per categoria di distanza

La Tabella 4.3 mostra i risultati del modello GAM, mentre la Figura 4.5 gli effetti parziali dei predittori sulla variabile risposta.

La distanza gioca un ruolo fondamentale sul numero di colpi per terminare la buca, in quanto più vicini si fa l'approccio al *green* e il numero di colpi per terminare la buca diminuisce. Il numero di gradi di libertà stimati della *spline* è pari a 8. La variabile relativa alla tipologia di errore non è significativa sul numero di colpi, quindi è indifferente se la palla non raggiunge il *green* in regolazione a sinistra, destra, corto o lungo. Significativa è la variabile relativa alla *Fairway in Regulation*, portando a dire che tirare il *driver* non in *fairway*, ha conseguenze anche sul colpo di approccio in *green*.

L'indice di bontà di adattamento aggiustato calcolato rispetto al *dataset* di verifica R^2_{adj} è pari a 0.8686.

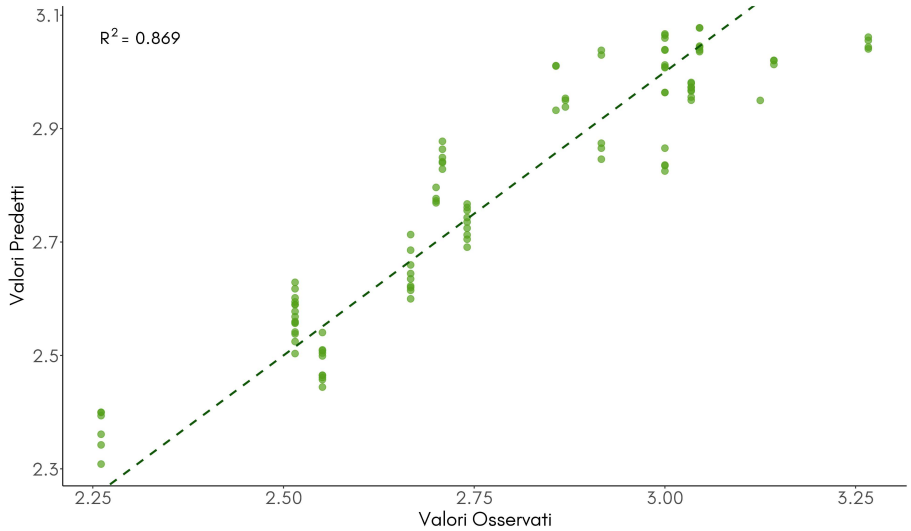


Figura 4.6: Grafico Valori predetti vs Valori osservati

4.3 *Putting shot benchmark*

Seguendo la stessa logica proposta nella sezione precedente, si va a costruire una nuova variabile risposta ad hoc, ovvero i colpi medi per terminare la buca per categoria di distanza, con le categorie costruite dividendo in intervalli regolari di lunghezza pari a 1 metro la variabile `Distance Putt in m`.

Le variabili che si vanno a considerare in questo modello sono quelle che si riferiscono solo al *green*, quindi l'inclinazione del *green* nella direzione dalla pallina alla buca (`Slope`). Purtroppo il *dataset* non presenta altre variabili utili per il calcolo del *benchmark* sul *putt*.

	Estimate	Std. Error	t-value	Pr(> |t|)	
`(Intercept)`	1.803	0.012	145.758	<0.001	***
`SlopeP`	−0.017	0.015	−1.108	0.2685	
`SlopeU`	0.028	0.017	1.697	0.0905	.

Approximate significance of smooth terms:
`s(Distance.to.green.in.m)` edf = 2.994, Ref.df = 3, F = 1072, p-value = < 0.001 ***
R-sq.(adj) = 0.897, Deviance explained = 89.8%
GCV = 0.014239, Scale est. = 0.014038, n = 424

Tabella 4.4: Risultati del modello GAM con *spline* sui dati del *putt*

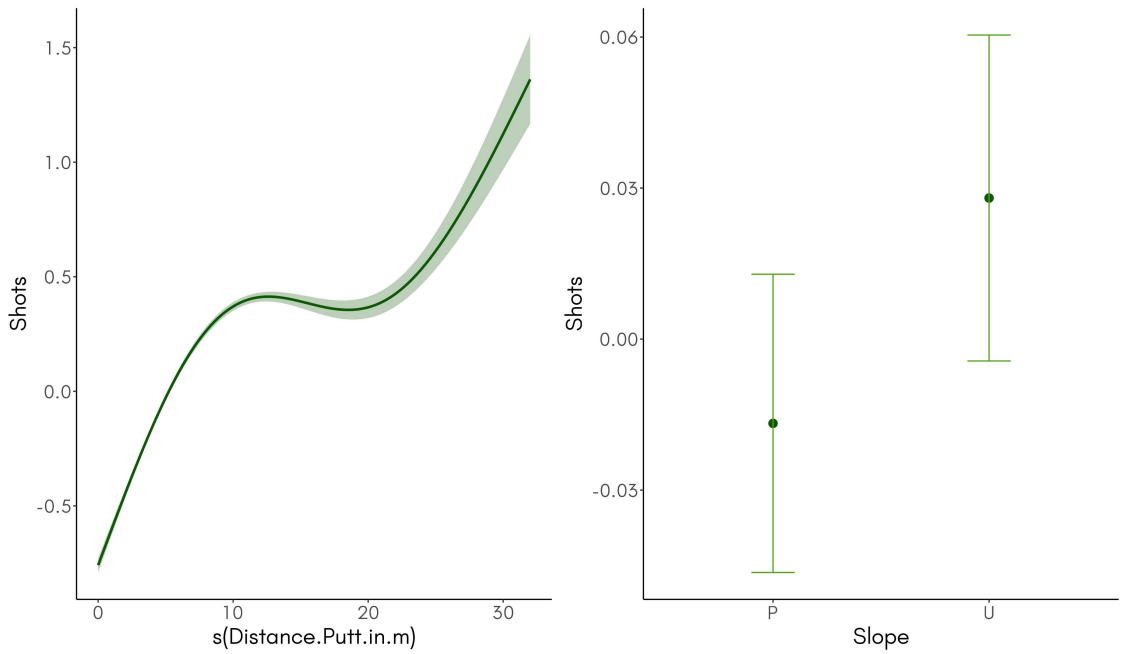


Figura 4.7: Effetti parziali del modello GAM dei predittori sul numero medio di colpi per categoria di distanza

La Tabella 4.4 mostra i risultati del modello GAM adattato ai dati sul *putt*. La variabile `Slope` fornisce una differenziazione per quanto riguarda la difficoltà del terreno in fase di *putting*, con un terreno in salita (*uphill)* si tende a fare più colpi rispetto a un terreno pianeggiante (*plain*) e al valore di riferimento (discesa: *downhill*).

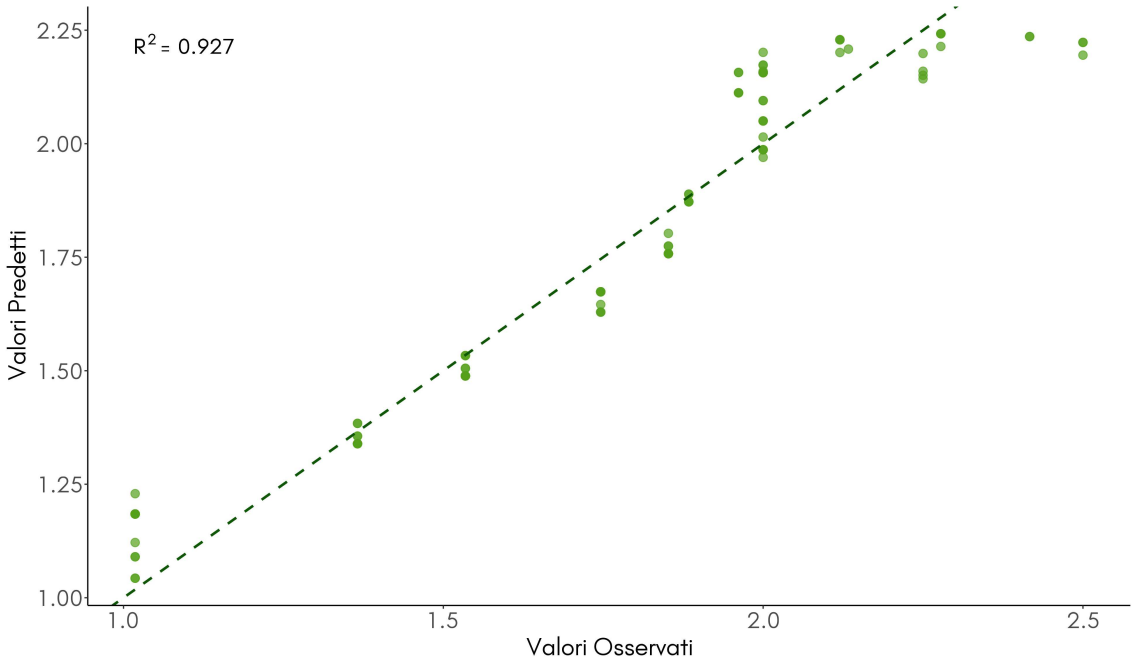


Figura 4.8: Grafico Valori predetti vs Valori osservati

La Figura 4.8 mostra il grafico dei valori predetti confrontati con i valori osservati, utilizzando i dati del *dataset* di verifica.

4.4 Esempio di applicazione completa

A scopo esemplificativo, si procede effettuando delle previsioni puntuali per determinate condizioni eseguendo 4 colpi nella buca 15.

I dettagli dei colpi sono presentati nella tabella sottostante:

Tabella 4.5: Dettagli dei colpi nella buca 15

Colpo	**Dettagli**	$J(\cdot)$	***Strokes Gained***
1	Distanza tee: 258 m Putt (media): `media_putt_h15` Distance to green: 85 m Fairway in Regulation: 1	$J(\cdot) = 3.74$ $J(\cdot) = 2.95$	$S_1 = -0.21$
2	Distanza dal Green: 85 m Slope: P Fairway in Regulation: 1 FWR Error: 0 GIR Error: 0	$J(\cdot) = 2.95$ $J(\cdot) = 1.97$	$S_2 = -0.02$
3	Slope: P Distanza Putt: 7 m	$J(\cdot) = 1.97$ $J(\cdot) = 1.18$	$S_3 = -0.21$
4	Slope: P Distanza Putt: 1 m	$J(\cdot) = 1.18$ $J(\cdot) = 0$	$S_4 = 0.18$

Per `media_putt_h15` si è deciso di prendere la media dei *putt* eseguiti nella buca 15, a indicare il numero di *putt* medio che si esegue in questa buca.

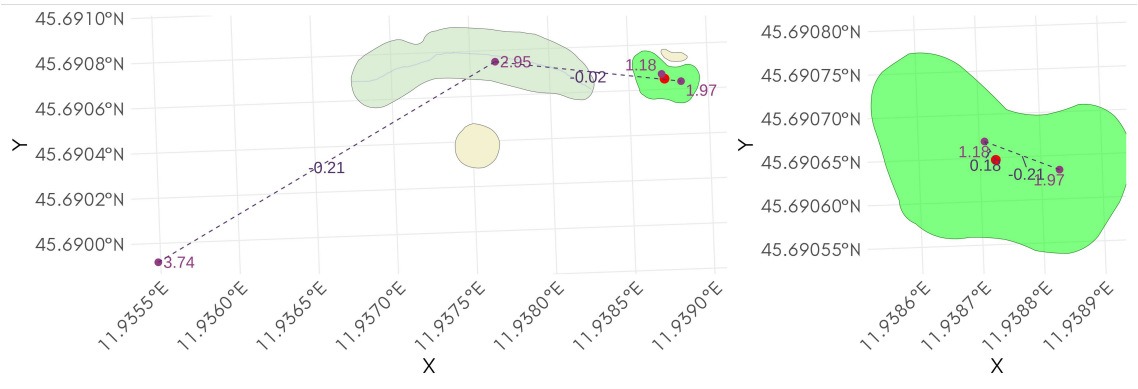


Figura 4.9: Esempio di buca giocata con calcolo *strokes gained*

Capitolo 5

Modellazione: dati 2024

In questo capitolo verranno proposti alcuni modelli per il calcolo degli *strokes gained*, verranno presentati singolarmente preceduti da una breve introduzione teorica e poi verranno analizzate le prestazioni nel complesso. La differenza rispetto al capitolo precedente è che qui, avendo i dati "colpo per colpo", verrà considerata come risposta la variabile `Remaning Shots` (colpi rimanenti) con valori che possono andare da 0 a 12 (numero massimo di colpi registrato per una buca), in modo tale che il problema che ci si trovi davanti sia un problema di **multiclassificazione**.

Rispetto al capitolo precedente, qui si suggerisce l'utilizzo di un solo modello per ogni tipologia di colpo.

Verranno testati vari modelli e, per decidere il migliore, si andranno a valutare alcune metriche come ad esempio l'*accuracy*, che rappresenta la proporzione di unità correttamente classificati rispetto al totale delle unità all'interno della matrice di confusione, generata dal confronto fra valori osservati e valori predetti.

In altre parole, l'*accuracy* misura quanto frequentemente il modello prevede correttamente la classe vera rispetto al numero totale di predizioni effettuate.

Formalmente, essa può essere espressa con la seguente formula:

$$accuracy = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i = \hat{y}_i) \tag{5.1}$$

Dove:

- N è il numero totale delle unità
- y_i è la classe osservata per l'i-esima unità
- $\hat{y}_i$ è la classe predetta per l'i-esima unità

- $\mathbb{I}(\cdot)$ è una funzione indicatrice che vale 1 se la condizione all'interno è vera, altrimenti vale 0.

L'*accuracy* assume valori compresi tra 0 e 1, dove 1 indica che tutte le predizioni del modello sono corrette.

Un altro strumento importate è il **tasso di errata classificazione**, una metrica complementare all'*accuracy* che misura la proporzione di unità che sono state classificate in modo errato rispetto al totale delle unità considerate. In altre parole essa quantifica la frequenza con cui il modello prevede una classe diversa da quella reale.

Formalmente, il tasso di errata classificazione può essere espresso come:

$$\textit{Error Rate} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i \neq \hat{y}_i) \tag{5.2}$$

Essendo complementare all'*accuracy*, il tasso di errata classificazione può anche essere calcolato come:

$$\textit{Error Rate} = 1 - \text{Accuracy} \tag{5.3}$$

Il tasso di errata classificazione assume valori compresi tra 0 e 1, dove 0 indica che tutte le predizioni del modello sono corrette (quindi nessuna unità è stata classificata erroneamente), mentre valori più vicini a 1 indicano un numero maggiore di errori di classificazione.

Ogni modello, in fase di validazione, è stato stimato su partizioni diverse dello stesso *dataset*, mediante la procedura della **convalida incrociata** per scegliere la partizione migliore.

Verrà utilizzato il tasso di errata classificazione come strumento di valutazione delle partizioni del dataset (*folds*) all'interno della convalida incrociata e verrà utilizzata l'*accuracy* per descrivere le prestazioni dei modelli.

Il punto cruciale sarà andare a determinare, per ogni colpo, la distribuzione di probabilità di terminare la buca entro X colpi.

Per ogni tipologia di modello presentato ne verranno calcolati 2 diversi, ciascuno per le due tipologie di giocatore che sono state presentate nel capitolo 3.

5.1 Classificazione logistica ordinale

Considerando la natura della variabile risposta categoriale, una classificazione logistica ordinale è il modello più adatto per analizzare questi dati. Questo tipo di modello è utilizzato quando la variabile dipendente è categoriale ordinale, ovvero le categorie hanno un ordine intrinseco.

Il modello di classificazione logistica ordinale si basa sull'idea di una variabile latente continua non osservata, che segue una distribuzione specifica (ad esempio, logistica, normale, di Gumbel o di Cauchy) con una scala unitaria, e una media che dipende linearmente dalle variabili esplicative.

Supponiamo di avere una variabile di risposta ordinale Y con K categorie ordinate $(1, 2, \ldots, K)$ e un insieme di variabili esplicative $X = (X_1, X_2, \ldots, X_p)$. Il modello può essere espresso come:

$$\log\left(\frac{P(Y \leq k \mid X)}{P(Y > k \mid X)}\right) = \theta_k - \beta^T X, \quad \text{per } k = 1, 2, \ldots, K-1$$

dove:

- $P(Y \leq k \mid X)$ è la probabilità cumulativa che la risposta sia nella categoria k o inferiore, dato X
- θ_k è la soglia (intercetta) associata alla categoria k, che determina i punti di separazione tra le categorie
- β è il vettore dei coefficienti associati alle variabili esplicative X.

Il lato sinistro dell'equazione rappresenta il logaritmo degli *odds* cumulativi di appartenere alla categoria k o inferiore rispetto alle categorie superiori. Un aspetto chiave del modello è l'assunzione delle ***odds* proporzionali**, che implica che gli effetti delle variabili esplicative siano costanti attraverso tutte le categorie della risposta. In altre parole, i coefficienti β non dipendono dalla categoria k.

I coefficienti β_j indicano l'effetto di un incremento unitario in X_j sulla *log-odds* di essere in una categoria inferiore o uguale, mantenendo costanti le altre variabili. Un coefficiente positivo suggerisce che un aumento in X_j aumenta la probabilità di appartenere a una categoria più bassa, mentre un coefficiente negativo indica il contrario.

Il modello può essere interpretato attraverso l'esistenza di una variabile latente continua Y^*, che rappresenta una propensione sottostante o un punteggio non osservato. Questa variabile latente è espressa come:

$$Y^* = \beta^T X + \varepsilon$$

dove ε è un termine di errore che segue una distribuzione specifica (ad esempio, logistica standard per il *link logit*, normale standard per il *link probit*).

Le categorie osservate di Y sono determinate dalla suddivisione di Y^* in intervalli definiti dalle soglie θ_k:

$$Y = k \quad \text{se} \quad \theta_{k-1} < Y^* \leq \theta_k$$

con $\theta_0 = -\infty$ e $\theta_K = +\infty$.

La funzione di *link* $g(\cdot)$ collega la media della distribuzione della risposta alla combinazione lineare delle variabili esplicative. Oltre al *link logit*, possono essere utilizzati altri *link*, come:

- ***Probit***: utilizza la funzione di distribuzione cumulativa della normale standard
- **Complementary Log-Log (cloglog)**: adatto per dati asimmetrici, utilizza la funzione $g(p) = \log(-\log(1-p))$
- **Log-Log**: utilizza la funzione $g(p) = -\log(-\log(p))$
- **Cauchit**: basato sulla distribuzione di Cauchy.

I parametri del modello, cioè le soglie θ_k e i coefficienti β, sono stimati tramite il metodo della massima verosimiglianza. Questo processo coinvolge l'ottimizzazione numerica della funzione di verosimiglianza, spesso utilizzando algoritmi come Newton-Raphson o l'algoritmo di Fisher scoring.

Le variabili che verranno incluse nei modelli sono la distanza dalla buca (`target.distance.before`), una *spline* bidimensionale per le informazioni sulle coordinate geografiche, il numero di *putt* eseguiti (`putts`), se si è finiti nel *bunker* oppure no (`sand.shots`), il `lie` della pallina, ovvero com'è posizionata e infine se si è raggiunto il *fairway* con il *driver*, con l'informazione anche sulla direzione dell'errore.

5.1.1 Risultati del modello HCP 5/15

	Value	Std. Error	t-value
`target.distance.before`	0.024	0.001	17.665
`ts(X, Y)`	0.000	0.000	0.756
`putts`	2.649	0.171	15.488
`sand.shots1`	0.497	0.232	2.141
`lieGreen`	−5.588	0.376	−14.845
`lieRecovery`	2.391	0.490	4.883
`lieRough`	0.049	0.236	0.207
`lieSand`	−0.043	0.495	−0.086
`lieTee`	−0.231	0.275	−0.842
`fairwayleft`	0.275	0.194	1.416
`fairwayright`	0.313	0.179	1.745
`fairwayunknown`	0.297	0.198	1.501
Intercepts:			
`0|1`	0.063	0.404	0.156
`1|2`	4.259	0.344	12.376
`2|3`	8.632	0.449	19.223
`3|4`	12.104	0.561	21.558
`4|5`	15.898	0.706	22.522
`5|6`	18.045	0.829	21.765
`6|7`	19.873	1.234	16.105

Residual Deviance: 1702.627
AIC: 1740.627

Tabella 5.1: Risultati del modello logistico ordinale

Un aumento di un'unità nella distanza dal bersaglio aumenta le *log-odds* di appartenere a una categoria superiore di colpi rimanenti (quindi fare un colpo in più) di 0.0236, mantenendo costanti le altre variabili.

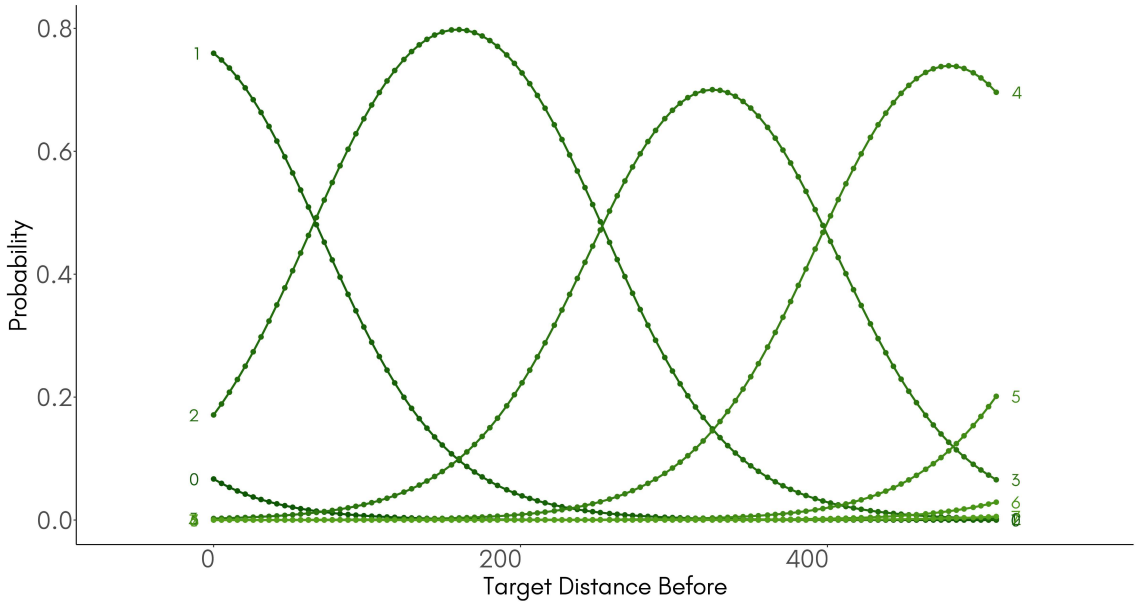


Figura 5.1: Effetto parziale della variabile della distanza dal bersaglio sul numero di colpi rimanenti

Per quanto riguarda le altre variabili, la *spline* bidimensionale non ci fornisce informazioni utili sul numero di colpi rimanenti, mentre un fattore logicamente essenziale è se il terreno dove si trova la palla è il *green* oppure no.

Tirare il *driver* fuori dal *fairway*, fa sì che le *log-odds* di terminare la buca di un colpo in più aumentino di 0.275 se si tira la palla a sinistra e 0.313 se la palla finisce a destra.

La Figura 5.1 presenta l'effetto parziale della variabile della distanza dal bersaglio sulla variabile risposta tenendo fissate le altre variabili. Con distanza dal bersaglio pari a 0 c'è quasi l'80% di probabilità di trovarsi con 1 colpo rimanente, mentre a distanza dal bersaglio massima c'è il 70% di probabilità di rimanere con 4 colpi per terminare la buca e il 20% di impiegare 5 colpi.

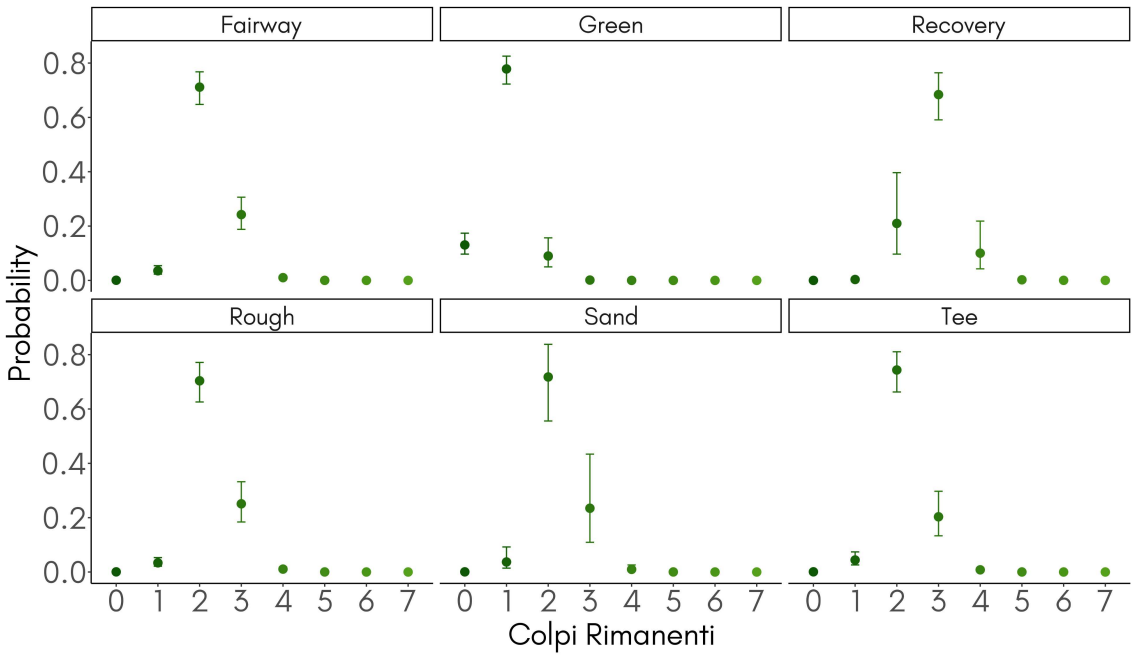


Figura 5.2: Effetto parziale della variabile *lie* sul numero di colpi rimanenti

La Figura 5.2 mostra l'effetto della variabile *lie* sui colpi rimanenti, mantenendo costanti le altre variabili.

5.1.2 Risultati del modello HCP 15/30

	Value	Std. Error	t-value
`target.distance.before`	0.023	0.001	19.241
`ts(X, Y)`	0.000	0.001	0.591
`putts`	2.019	0.123	16.412
`lieGreen`	−4.966	0.333	−14.930
`liePenalty`	1.416	1.200	1.180
`lieRough`	0.357	0.203	1.758
`lieTee`	−0.531	0.238	−2.228
`fairwayleft`	0.153	0.188	0.812
`fairwayright`	−0.379	0.206	−1.840
`fairwayunknown`	−0.038	0.193	−0.198
Intercepts:			
`0|1`	−0.418	0.358	−1.166
`1|2`	2.160	0.344	6.287
`2|3`	5.728	0.336	17.068
`3|4`	8.212	0.395	20.813
`4|5`	10.239	0.453	22.624
`5|6`	12.044	0.510	23.596
`6|7`	13.753	0.573	24.001
`7|8`	15.589	0.659	23.674
`8|9`	16.735	0.742	22.559
`9|10`	17.462	0.847	20.628
`10|11`	18.591	1.176	15.805

Residual Deviance: 2085.706
AIC: 2127.706
(18 osservazioni eliminate a causa di valori mancanti)

Tabella 5.2: Risultati del modello logistico ordinale aggiornato

La Tabella 5.2 presenta i risultati per il modello adattato al livello di gioco 15-30 HCP, il quale presenta valori diversi rispetto al modello visto precedentemente, il coefficiente riguardante il numero di colpi rimanenti quando la palla è sul *green* è passato da -5.588 a -4.966.

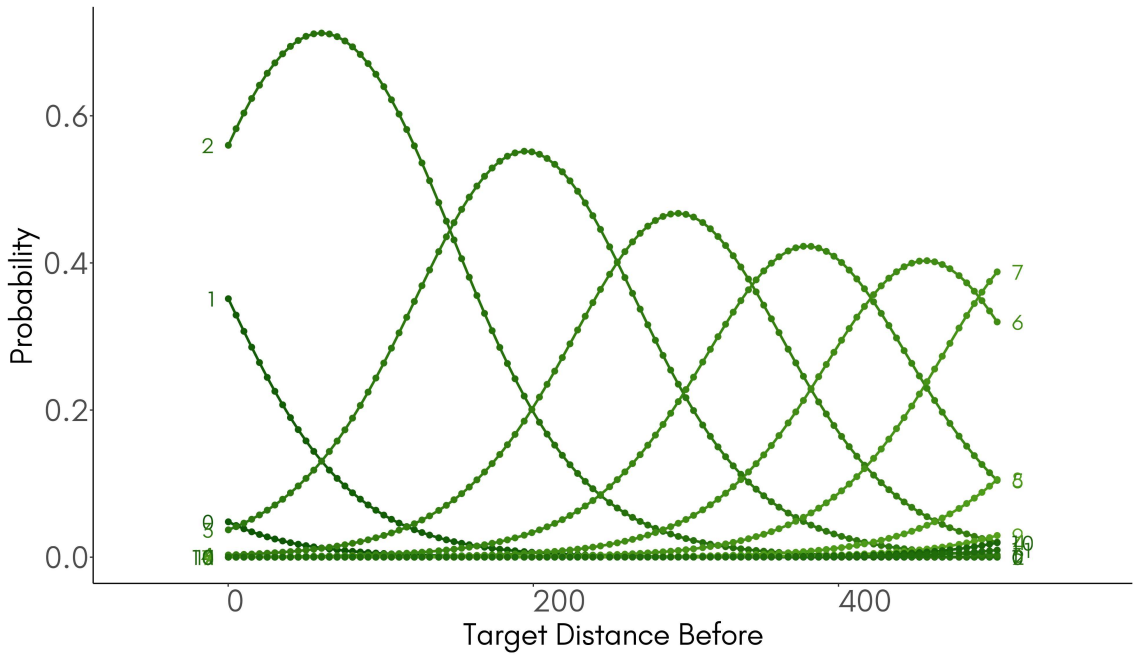


Figura 5.3: Effetto parziale della variabile della distanza dal bersaglio sul numero di colpi rimanenti

Rispetto alla Figura 5.1 risulta evidente la differenza di livello, una minima variazione della distanza va a modificare drasticamente la probabilità di finire la buca entro un certo numero di colpi. Anche il punto di partenza quando la distanza dal bersaglio sottolinea questa differenze, con una distanza massima di 515 metri, il modello dice che la probabilità più alta è di impiegare 7 colpi per finire la buca al non variare delle altre variabili, mentre per il modello con 5/15 di HCP la probabilità più alta è legata a 4 colpi.

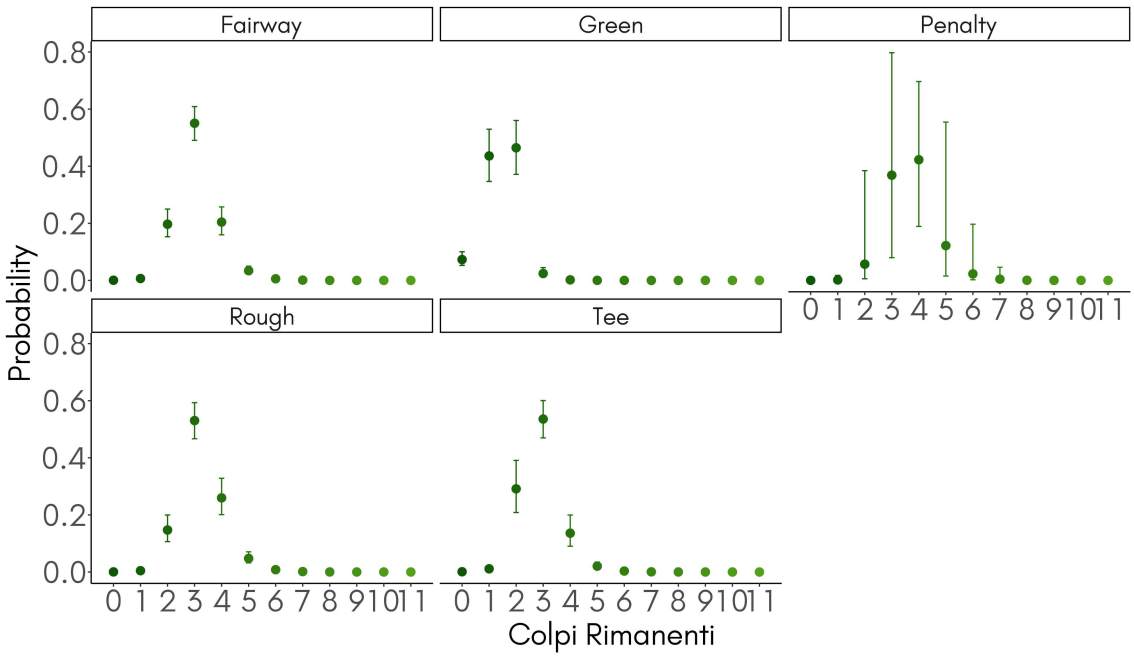


Figura 5.4: Effetto parziale della variabile `lie` sul numero di colpi rimanenti

5.2 Classificazione logistica multinomiale

Il modello di classificazione multinomiale è utilizzato per analizzare dati in cui la variabile dipendente è categoriale con più di due categorie nominali, ovvero senza un ordine intrinseco tra le categorie. A differenza della regressione logistica multinomiale, che utilizza la funzione *logit* per modellare le probabilità, il modello di regressione multinomiale può essere formulato attraverso modelli log-lineari o utilizzando altri *link* funzionali come il *probit* multinomiale.

Supponiamo di avere una variabile di risposta categoriale Y che può assumere K categorie nominali $(1, 2, \ldots, K)$ e un insieme di variabili esplicative $X = (X_1, X_2, \ldots, X_p)$. Il modello di regressione multinomiale esprime le probabilità di appartenenza a ciascuna categoria in funzione delle variabili esplicative.

Una formulazione comune del modello è attraverso le probabilità dirette:

$$P(Y = k \mid X) = f_k(X), \quad \text{per } k = 1, 2, \ldots, K$$

dove le funzioni $f_k(X)$ sono specificate in modo tale che $0 \leq f_k(X) \leq 1$ e $\sum_{k=1}^{K} f_k(X) = 1$.

Un approccio per modellare le probabilità è utilizzare un modello log-lineare, che rappresenta le log-probabilità come combinazioni lineari delle variabili esplicative:

$$\log P(Y = k \mid X) = \beta_{0k} + \beta_k^T X - \log Z(X)$$

dove:

- β_{0k} è l'intercetta specifica per la categoria k
- β_k è il vettore dei coefficienti delle variabili esplicative per la categoria k
- $Z(X)$ è una funzione di normalizzazione data da $Z(X) = \sum_{j=1}^{K} \exp(\beta_{0j} + \beta_j^T X)$.

Questo porta alla formulazione delle probabilità come:

$$P(Y = k \mid X) = \frac{\exp(\beta_{0k} + \beta_k^T X)}{\sum_{j=1}^{K} \exp(\beta_{0j} + \beta_j^T X)}$$

Un'alternativa al modello *logit* è il **modello multinomiale *probit***, che utilizza la funzione di distribuzione cumulativa normale per modellare le probabilità. In questo caso, si assume l'esistenza di variabili latenti associate a ciascuna categoria:

$$Y_k^* = \beta_{0k} + \beta_k^T X + \varepsilon_k$$

dove ε_k sono termini di errore normalmente distribuiti con media zero e matrice di varianza-covarianza Σ. L'osservazione Y appartiene alla categoria con il valore massimo tra le variabili latenti Y_k^*.

Il modello di regressione multinomiale si basa sulle seguenti assunzioni:

- **Indipendenza delle osservazioni**: le osservazioni sono indipendenti tra loro
- **Non multicollinearità**: le variabili esplicative non devono essere altamente correlate
- **Distribuzione delle risposte**: la distribuzione delle categorie della variabile dipendente segue una distribuzione multinomiale.

La stima dei parametri nel modello multinomiale può essere effettuata tramite il metodo della massima verosimiglianza. La funzione di verosimiglianza per il modello è data da:

$$L(\beta) = \prod_{i=1}^{N} \prod_{k=1}^{K} [P(Y_i = k \mid X_i)]^{I(Y_i=k)}$$

dove $I(\cdot)$ è la funzione indicatrice che vale 1 se l'argomento è verificato e 0 altrimenti.

L'ottimizzazione della funzione di verosimiglianza può essere complessa a causa della presenza di più categorie e della normalizzazione richiesta. Spesso si utilizzano algoritmi numerici come la massimizzazione condizionale o metodi basati sull'algoritmo EM (*Expectation-Maximization*).

I coefficienti stimati β_{jk} rappresentano l'effetto di un incremento unitario in X_j sulla *log-odds* della categoria k rispetto alla categoria di riferimento. Un coefficiente positivo indica che un aumento in X_j aumenta la probabilità di appartenere alla categoria k, mentre diminuisce la probabilità di appartenere alla categoria di riferimento.

5.2.1 Risultati del modello HCP 5/15

Colpi Rimanenti →	1	2	3	4	5	6	7
`(Intercept)`	1.870	−7.118	−15.325	−22.540	−29.604	−20.946	−62.591
`target.distance.before`	0.883	0.924	0.951	0.965	0.971	0.955	1.001
`ts(X, Y)`	0.003	0.002	0.002	0.001	0.007	0.003	0.017
`putts`	3.346	8.977	11.078	12.408	13.774	11.486	17.506
`sand.shots1`	−0.139	0.056	1.231	2.185	2.627	2.346	12.613
`lieGreen`	−11.921	−20.487	−51.397	−56.342	−71.399	−54.162	−28.774
`lieRecovery`	7.465	9.408	13.137	13.735	14.878	−0.650	15.136
`lieRough`	2.043	2.016	2.692	2.537	3.250	−12.989	14.672
`lieSand`	4.232	4.869	5.466	−21.931	0.600	−3.108	10.423
`lieTee`	−37.253	−38.324	−38.845	−38.557	−38.446	−39.483	−35.731
`fairwayleft`	0.318	1.213	0.908	1.478	1.914	2.322	−10.624
`fairwayright`	0.395	0.982	0.563	1.066	1.681	1.738	−6.916
`fairwayunknown`	0.424	0.500	0.696	1.538	2.424	2.370	0.454

Residual Deviance: 1280.707
AIC: 1462.707

Tabella 5.3: Coefficienti nel modello di classificazione multinomiale

La Tabella 5.3 mostra i coefficienti (*log-odds*) di appartenere alla specifica categoria di colpi rimanenti rispetto alla categoria di riferimento (0).

Ad esempio, il valore del *log-odds* di impiegare 1 colpo in più per terminare la buca, tenendo fisse le altre variabili, per la variabile relativa alla distanza dal bersaglio, è pari a 0.883, questo significa che con ogni unità di aumento in `target.distance.before`, le probabilità di appartenere alla categoria 1 rispetto alla categoria di riferimento (0) aumentano di circa $e^{0.883} = 2.418$ volte.

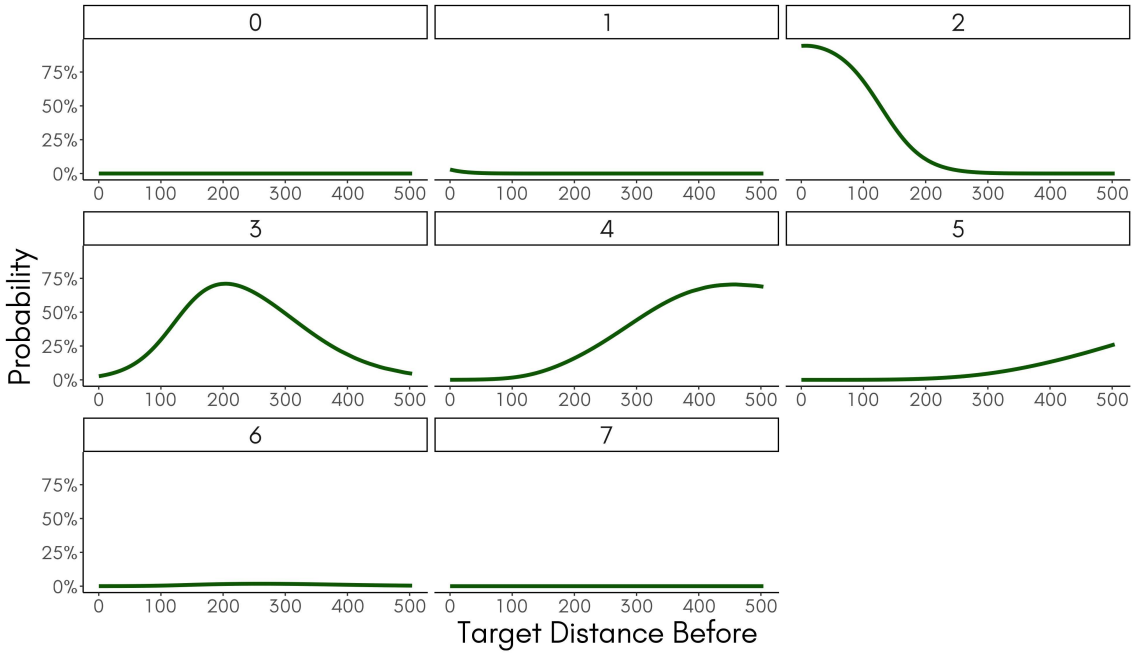


Figura 5.5: Effetto parziale della variabile `target.distance.before` sul numero di colpi rimanenti

La Figura 5.5 mostra l'effetto parziale della variabile relativa alla distanza sul numero di colpi.

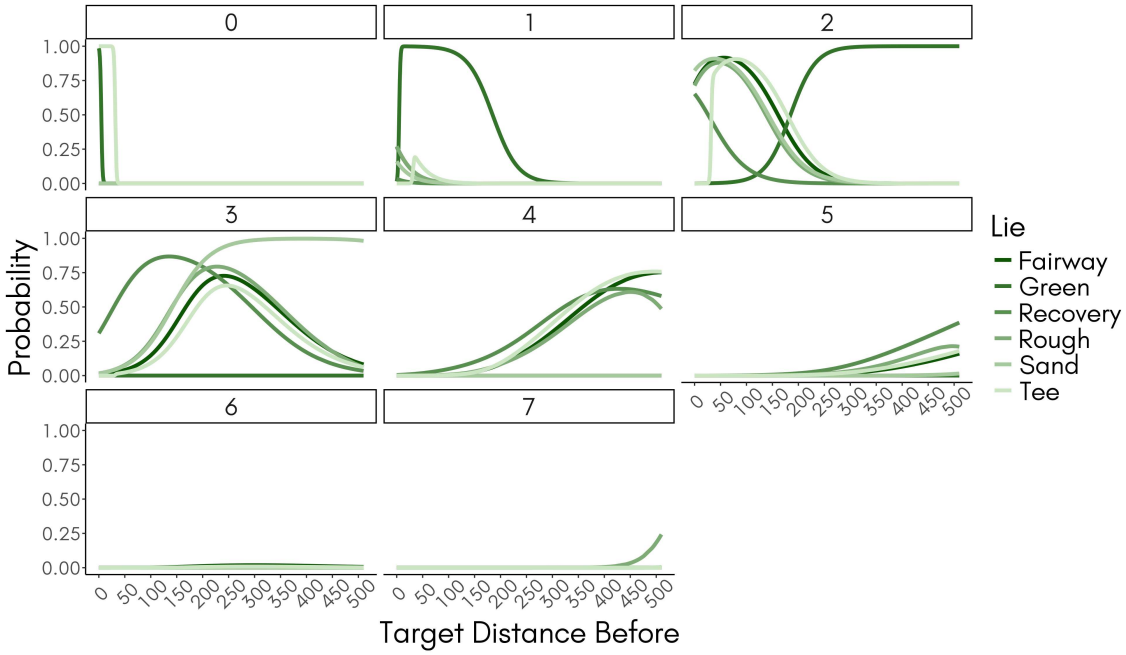


Figura 5.6: Effetto parziale della variabile `target.distance.before` e della variabile `lie` sul numero di colpi rimanenti

Espandendo l'analisi dell'effetto parziale anche in base alla variabile relativa al *lie*, è possibile fare considerazioni generali sull'andamento della probabilità in base ad alcune condizioni, ad esempio con una distanza di 182 metri e la palla in *fairway*, la probabilità di avere 1 *shot* rimanente è praticamente 0%, mentre si alza quella di averne 2 o 3.

5.2.2 Risultati del modello HCP 15/30

Colpi Rimanenti →	1	2	3	4	5	6	7	8	9	10	11
`(Intercept)`	0.312	−6.925	−15.789	−21.419	−26.368	−33.002	−39.073	−58.683	−58.893	−45.716	−54.848
`target.distance.before`	0.433	0.479	0.510	0.526	0.537	0.547	0.556	0.579	0.595	0.571	0.564
`ts(X, Y)`	0.001	0.000	−0.001	−0.002	−0.001	0.002	0.001	0.018	0.029	0.020	0.003
`putts`	0.719	4.942	7.976	9.147	10.033	11.310	12.285	12.137	13.594	12.067	15.024
`sand.shots1`	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
`lieGreen`	−3.113	−8.983	−13.351	−23.085	−27.514	−25.189	−25.790	−4.144	−17.439	−21.613	−13.454
`liePenalty`	−3.965	−11.578	8.841	10.446	−4.435	−2.016	−0.098	0.481	0.712	0.561	0.329
`lieRough`	1.584	2.927	3.513	3.733	4.218	4.196	4.445	−3.394	9.294	−5.304	0.575
`lieTee`	−28.958	−10.735	−9.672	−10.011	−9.872	−10.531	−10.363	−1.801	−13.623	−11.888	−7.805
`fairwayleft`	−0.040	−0.155	−0.092	0.136	0.299	0.725	0.682	1.898	−10.114	−3.223	−3.058
`fairwayright`	−0.368	−0.325	−0.513	−0.689	−1.129	−1.617	−2.209	−6.314	−5.570	−4.848	−5.977
`fairwayunknown`	0.040	−0.165	−0.262	−0.564	0.723	1.459	0.479	3.712	−4.248	−3.658	−2.422

Residual Deviance: 1801.753
AIC: 2043.753

Tabella 5.4: Coefficienti nel modello di classificazione multinomiale

Il valore del *log-odds* di impiegare 1 colpo in più per terminare la buca, tenendo fisse le altre variabili, per la variabile relativa alla distanza dal bersaglio, è pari a 0.443, questo significa che con ogni unità di aumento in `target.distance.before`, le probabilità di appartenere alla categoria 1 rispetto alla categoria di riferimento (0) aumentano di circa $e^{0.443} = 1.542$ volte.

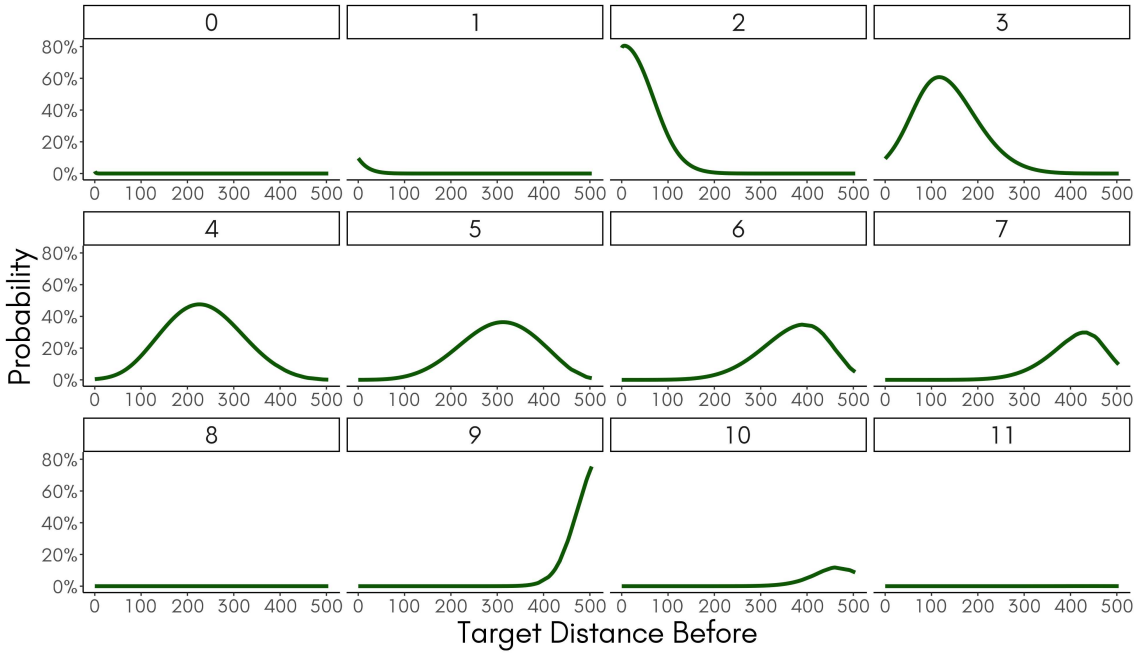


Figura 5.7: Effetto parziale della variabile `target.distance.before` sul numero di colpi rimanenti

La Figura 5.7 mostra l'effetto parziale della variabile relativa alla distanza sul numero di colpi.

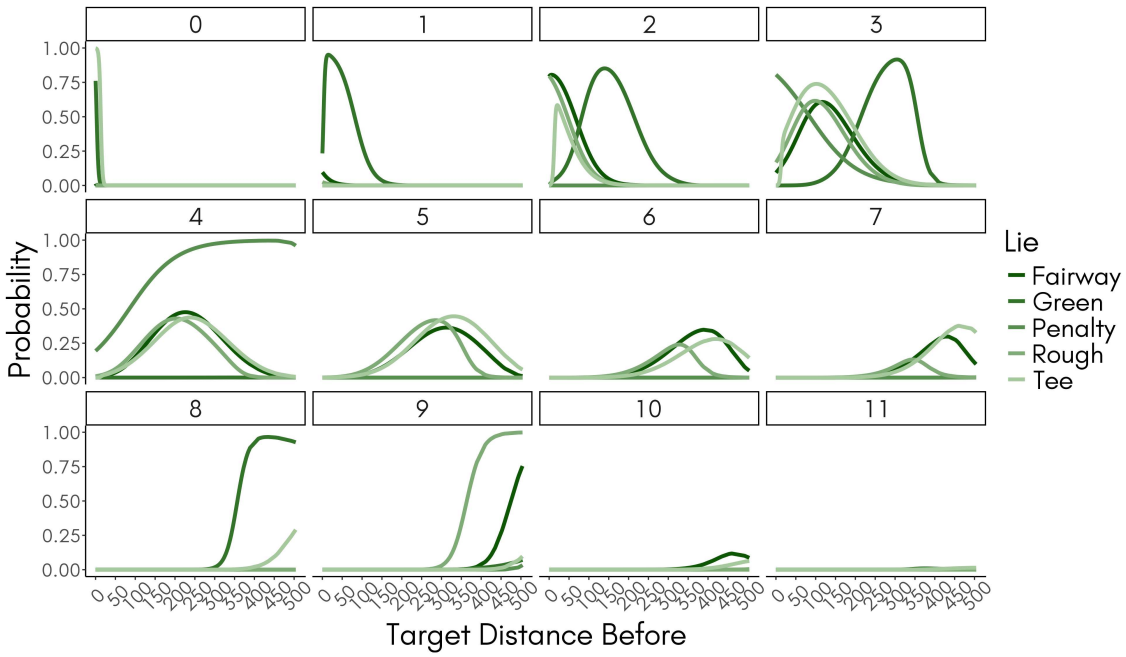


Figura 5.8: Effetto parziale della variabile `target.distance.before` e della variabile `lie` sul numero di colpi rimanenti

5.3 MARS (Multivariate Adaptive Regression Splines)

I modelli MARS (*Multivariate Adaptive Regression Splines*) sono una tecnica di regressione non parametrica - in questo caso applicate alla classificazione - che combina la flessibilità dei metodi di *smoothing* con la capacità interpretativa dei modelli lineari. Sviluppati da Jerome Friedman nel 1991, i MARS sono progettati per catturare relazioni non lineari e interazioni tra le variabili esplicative senza richiedere una specifica forma funzionale a priori.

I modelli MARS costruiscono una relazione tra la variabile dipendente Y e le variabili esplicative $X = (X_1, X_2, \ldots, X_p)$ utilizzando una combinazione di **funzioni base a pezzi** (*piecewise basis functions*), in particolare le **funzioni *hinge*** (*hinge functions*), che permettono di modellare cambiamenti nella pendenza della funzione di regressione.

Una funzione *hinge* è definita come:

$$h(X, t) = \begin{cases} (X - t) & \text{se } X > t \\ 0 & \text{altrimenti} \end{cases}$$

dove t è un punto di nodo (*knot*) nel dominio della variabile X.

Il modello MARS può essere espresso come una somma di funzioni base:

$$\hat{Y} = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(X)$$

dove:

- β_0 è l'intercetta del modello
- β_m sono i coefficienti da stimare
- $B_m(X)$ sono le funzioni base, che possono essere funzioni *hinge* delle variabili X o prodotti di esse per catturare interazioni
- M è il numero totale di funzioni base nel modello.

La costruzione di un modello MARS avviene in due fasi principali:

1. **Fase Forward**: in questa fase, il modello viene costruito in modo incrementale. Si aggiungono iterativamente funzioni base che riducono l'errore residuo, selezionando i punti di nodo t e le variabili X che migliorano maggiormente l'adattamento
2. **Fase *Backward Pruning***: poiché il modello iniziale potrebbe essere sovradimensionato, si applica una potatura (*pruning*) per rimuovere le funzioni base meno significative. Questo processo utilizza criteri come il **Generalized Cross-Validation** (GCV) per bilanciare l'errore di adattamento e la complessità del modello

L'uso del GCV durante la fase di potatura aiuta a prevenire l'*overfitting* e a selezionare un modello con una buona capacità di generalizzazione.

5.3.1 Risultati del modello HCP 5/15

Nell'introduzione al capitolo è stato descritto il processo di convalida incrociata per andare a definire la partizione del *dataset* migliore al fine di calcolare il modello. Per il MARS andiamo a includere nell'algoritmo la selezione di due iper-parametri necessari per questo modello, ovvero *degree* che rappresenta il grado di interazione fra le funzioni *hinge* e *prune number*, ovvero il numero di termini presenti nel modello finale. Questi

due parametri sono stati fatti variare in una griglia equispaziata in cui il valore di *degree* assumeva i valori 1 e 2, mentre *prune number* assumeva i valori da 2 a 20.

I valori selezionati sono stati rispettivamente 2 e 12.

Le funzioni di base che sono venute fuori dal modello 12 su 19 termini, e sono stati usati solo 3 predittori su 12 totali.

Funzioni Hinge
`(Intercept)`
`lieGreen`
`h(target.distance.before-45)`
`h(target.distance.before-211)`
`h(putts-2)`
`h(target.distance.before-6) * lieGreen`
`h(45-target.distance.before) * lieGreen`
`h(2-putts) * lieGreen`
`h(target.distance.before-7) * h(putts-2)`
`h(45-target.distance.before) * h(putts-2)`
`h(45-target.distance.before) * h(2-putts)`
`h(target.distance.before-45) * h(putts-2)`

GCV: 0.3781
RSS: 402.6097

Tabella 5.5: Coefficiente delle funzioni *hinge* nel modello MARS

In questo caso i nodi sono nei punti 6, 7, 45 e 211 per la variabile `target.distance.before`. È possibile associare questi punti a cambiamenti sostanziali nella scelta della mazza con cui eseguire il colpo. Sopra i 45 metri iniziano i colpi in cui si utilizza un ferro, quindi con lo *swing* a pieno movimento, fino a 211 metri, distanza media del *driver*. Variabili importanti per determinare il numero di colpi rimanenti sono anche se la pallina si trova o meno sul *green* e il numero di *putt*.

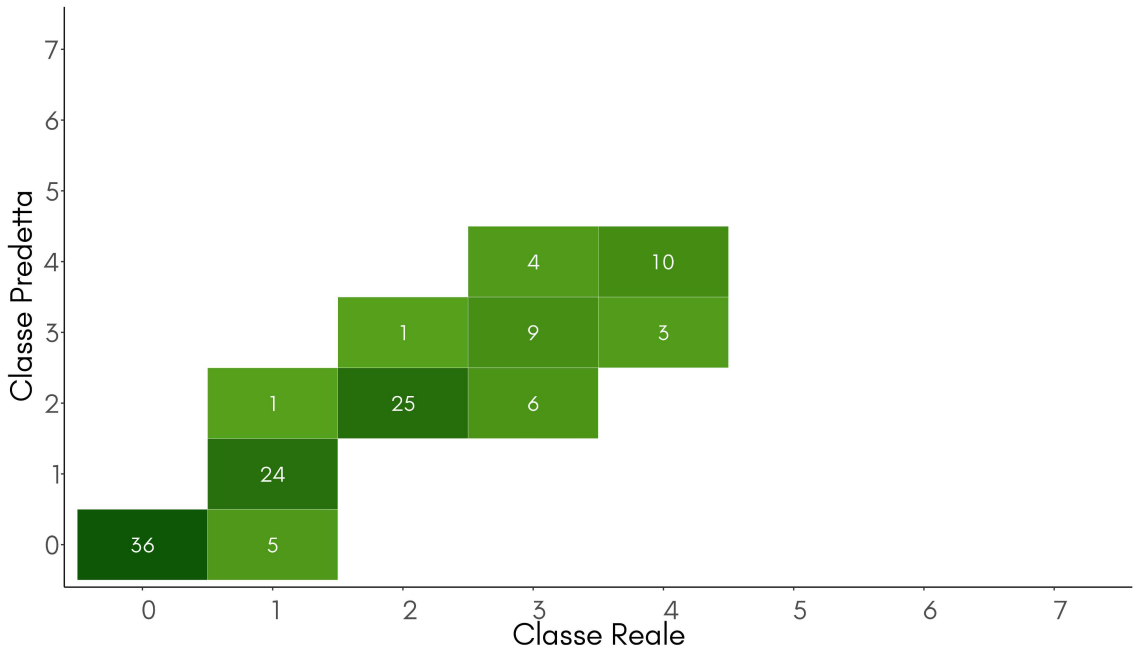


Figura 5.9: Matrice di confusione del modello MARS sul dataset di verifica

La Figura 5.9 rappresenta la matrice di confusione del modello MARS applicato. Questa matrice consente di valutare le prestazioni del modello confrontando le classi predette con le classi reali. Ogni cella della matrice indica il numero di osservazioni appartenenti a una specifica combinazione di classe reale e classe predetta. I colori delle celle variano da bianco a verde scuro, con un gradiente che evidenzia l'intensità delle frequenze: celle con frequenza pari a zero sono colorate di bianco, mentre frequenze più elevate assumono tonalità di verde più intense.

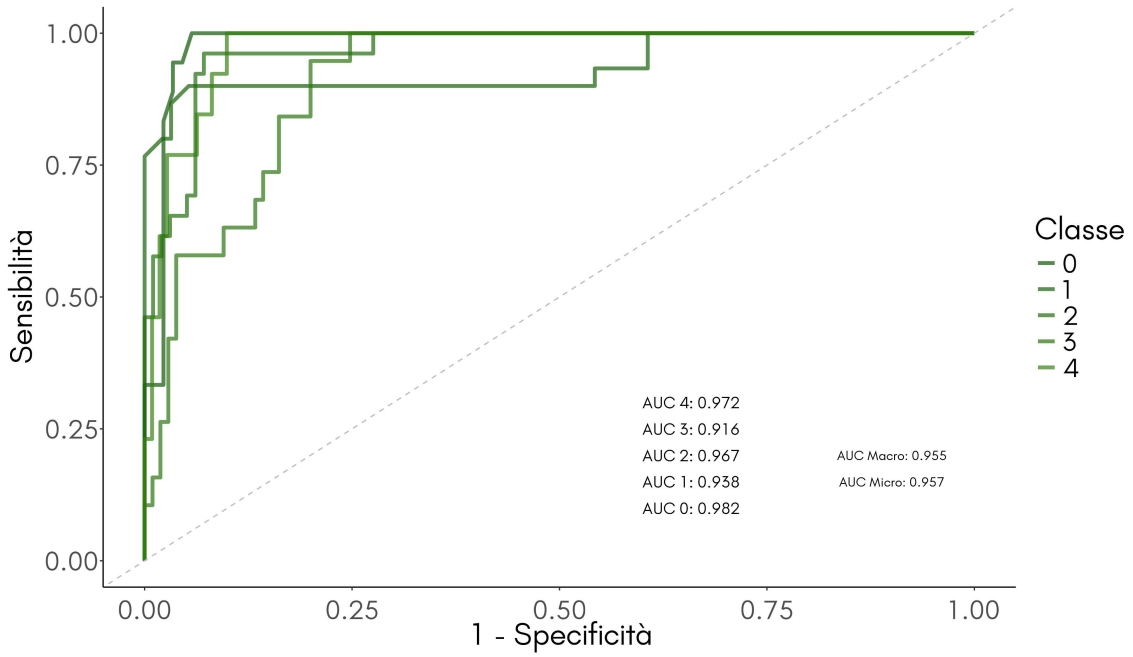


Figura 5.10: Curva ROC e AUC per il modello MARS

La Figura 5.10 illustra le curve ROC (*Receiver Operating Characteristic*) per ciascuna delle classi del problema di classificazione multiclasse. Ogni curva rappresenta la sensibilità (*True Positive Rate*), definita come $\frac{\sum \mathbb{I}(y_i=\hat{y}_i)}{\sum \mathbb{I}(y_i=\hat{y}_i)+\sum \mathbb{I}(y_i\neq\hat{y}_i)}$, rispetto a 1-specificità (*False Positive Rate*), definita come $1-\frac{\sum \mathbb{I}(y_i\neq\hat{y}_i)}{\sum \mathbb{I}(y_i\neq\hat{y}_i)+\sum \mathbb{I}(y_i=\hat{y}_i)}$, per una determinata classe, permettendo di valutare la capacità del modello di distinguere tra le diverse categorie. La linea diagonale tratteggiata indica il comportamento di un classificatore casuale, fungendo da riferimento per valutare le prestazioni del modello. Le curve ROC che si discostano maggiormente dalla diagonale indicano una migliore discriminazione del modello per quella specifica classe.

5.3.2 Risultati del modello HCP 15/30

Per questo dataset i valori selezionati per gli iper-parametri *degree* e *prune number* sono rispettivamente 2 e 13.

Le funzioni di base che sono venute fuori dal modello sono quindi 13 su 19 termini, e sono stati usati solo 3 predittori su 11 totali.

Funzioni Hinge
(Intercept)
lieGreen
h(target.distance.before-41)
h(target.distance.before-160)
h(putts-2)
h(6-target.distance.before) * lieGreen
h(2-putts) * lieGreen
h(putts-2) * lieGreen
h(target.distance.before-7) * h(putts-2)
h(41-target.distance.before) * h(putts-2)
h(41-target.distance.before) * h(2-putts)
h(target.distance.before-41) * h(putts-2)
h(target.distance.before-160) * h(putts-2)

GCV: 0.5687
RSS: 505.0381

Tabella 5.6: Coefficiente delle funzioni hinge nel modello MARS

La Tabella 5.6 fa riferimento alle funzioni di base calcolate dal modello MARS, in questo caso i nodi sono nei punti 6, 7, 41 e 160. Provando a dare un'interpretazione sulla stessa linea precedente, questi punti indicano un cambiamento sostanziale nella tipologia di colpo che si va ad effettuare, sotto i 41 metri si parla di un colpo di approccio, sotto i 6 o 7 metri si utilizza sostanzialmente il *putter*, mentre sopra i 160 metri non è più il colpo verso al *green* ma un colpo per fare più distanza possibile.

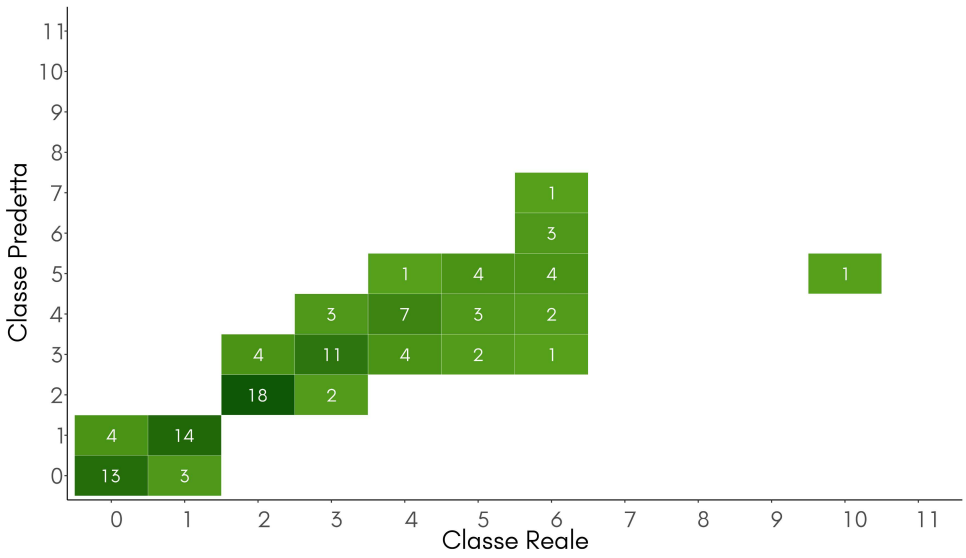


Figura 5.11: Matrice di confusione del modello MARS sul dataset di verifica

Questa matrice di confusione risulta più sparsa della precedente in quanto il livello di gioco essendo più basso porta ad un numero maggiore di colpi. C'è una classificazione *outlier* in quanto un'osservazione è stata classificata con 5 colpi rimanenti ma in realtà da quel punto specifico ne mancavano 10.

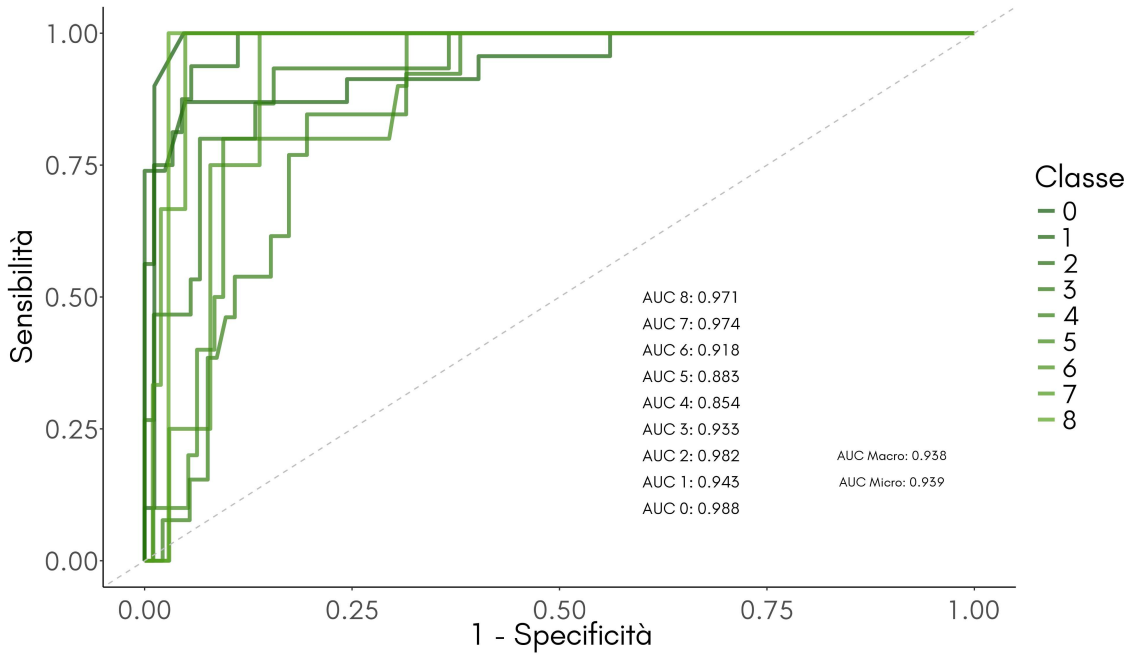


Figura 5.12: Curva ROC e AUC per il modello MARS

Essendo il numero di classi più alto, la curva ROC presenta un numero maggiore di linee. I risultati mostrano un valore di AUC alto per tutte le classi, dove ha un maggiore significato per le classi 0, 1, 2 in quanto sono presenti in maggior numero nel dataset rispetto alla classe 7 o 8. La classificazione peggiore avviene attorno alle classi 4 e 5, con un valore di AUC comunque alto (0.854 per la classe 4 e 0.883 per la classe 5), dove i singoli colpi tendono ad avere le stesse caratteristiche.

5.4 Alberi di classificazione con splines naturali

Gli alberi di classificazione sono modelli predittivi non parametrici utilizzati per assegnare osservazioni a categorie predefinite della variabile dipendente categoriale. Essi suddividono iterativamente lo spazio delle variabili esplicative in regioni omogenee rispetto alla variabile di risposta,

creando una struttura ad albero decisionale che facilita l'interpretazione dei risultati.

In alcuni casi, le relazioni tra le variabili esplicative e la variabile di risposta possono essere altamente non lineari. Per catturare queste relazioni complesse, è possibile applicare trasformazioni alle variabili esplicative, come le ***splines* naturali**, prima di costruire l'albero di classificazione. Questo approccio combina la flessibilità delle *splines* con la capacità interpretativa degli alberi di decisione.

Un albero di classificazione si costruisce seguendo questi passaggi:

1. **Suddivisione ricorsiva**: il *dataset* viene suddiviso in nodi interni basati sui valori delle variabili indipendenti che massimizzano l'omogeneità della variabile di risposta all'interno dei nodi figli.

2. **Criterio di suddivisione**: si utilizza tipicamente l'indice di impurità di Gini, l'entropia o il tasso di misclassificazione per determinare il punto ottimale di *split*.

3. **Crescita dell'albero**: la suddivisione continua in modo ricorsivo fino al raggiungimento di un criterio di arresto, come una dimensione minima dei nodi o una profondità massima dell'albero.

4. **Potatura dell'albero**: per evitare l'*overfitting*, l'albero può essere potato riducendo il numero di nodi terminali, basandosi su criteri come la complessità o la validazione incrociata.

Il criterio di suddivisione mira a trovare la variabile X_j e il valore di soglia s che massimizzano la riduzione dell'impurità. Ad esempio, utilizzando l'indice di Gini, la funzione di costo da minimizzare è:

$$\min_{X_j,s} [N_L \cdot Gini(L) + N_R \cdot Gini(R)]$$

dove:

- N_L e N_R sono il numero di osservazioni nei nodi figli sinistro (*left*) e destro (*right*), rispettivamente

- $Gini(L)$ e $Gini(R)$ sono gli indici di Gini dei nodi figli

- x_{ij} è il valore della variabile X_j per l'osservazione i.

L'indice di Gini per un nodo è definito come:

$$Gini = 1 - \sum_{k=1}^{K} p_k^2$$

dove p_k è la proporzione di osservazioni nel nodo appartenenti alla classe k, e K è il numero totale di classi.

Le ***splines* naturali** sono funzioni lisce a tratti che possono essere utilizzate per modellare relazioni non lineari tra le variabili esplicative e la variabile di risposta. Incorporando le *splines* naturali in un albero di classificazione, è possibile catturare andamenti più complessi nelle variabili esplicative, migliorando la capacità predittiva del modello.

Una *spline* naturale di grado k con nodi $\{\xi_1, \xi_2, \ldots, \xi_m\}$ è una funzione *spline* che è un polinomio di grado k in ogni intervallo tra i nodi e che soddisfa condizioni di continuità fino alla derivata di ordine $k-1$. Inoltre, le *splines* naturali impongono che la funzione sia lineare oltre i nodi esterni, rendendo il modello più stabile ai margini.

Le *splines* naturali possono essere rappresentate come una combinazione lineare di funzioni base:

$$s(X_j) = \sum_{h=1}^{K} \theta_h B_h(X_j)$$

dove:

- $s(X_j)$ è la funzione spline naturale applicata alla variabile X_j
- θ_h sono i coefficienti da stimare
- $B_h(X_j)$ sono le funzioni base *spline* naturali
- K è il numero di funzioni base.

Per integrare le *splines* naturali negli alberi di classificazione, le variabili esplicative selezionate vengono trasformate utilizzando le funzioni base delle *splines*, e queste trasformazioni vengono utilizzate come predittori nel modello.

Costruzione del modello

Il modello di albero di classificazione con *splines* naturali può essere espresso come:

$$\hat{Y} = \text{Classifica}\left(\sum_{\ell=1}^{L} c_\ell \cdot I(X \in R_\ell)\right)$$

dove:

- L è il numero di nodi terminali dell'albero
- c_ℓ è la classe predominante nel nodo terminale ℓ
- $I(\cdot)$ è la funzione indicatrice che vale 1 se la condizione è vera e 0 altrimenti
- R_ℓ è la regione dello spazio delle variabili esplicative definita dal percorso nell'albero fino al nodo terminale ℓ.

Le variabili X includono sia le variabili originali sia le trasformazioni *spline*:

$$X = (X_1, X_2, \ldots, X_p, s_{j_1}(X_{j_1}), s_{j_2}(X_{j_2}), \ldots)$$

dove $s_{j_k}(X_{j_k})$ sono le funzioni *spline* naturali applicate alle variabili selezionate X_{j_k}.

Il criterio di suddivisione viene applicato anche alle variabili *spline* trasformate. Questo permette all'albero di utilizzare le relazioni non lineari catturate dalle *splines* per effettuare le suddivisioni. La funzione di costo rimane la stessa, ad esempio per l'indice di Gini:

$$\min_{X_j, s} \left[N_L \cdot Gini(L) + N_R \cdot Gini(R)\right]$$

dove ora X_j può essere una delle funzioni base *spline* $B_h(X_{j_k})$.

Nella costruzione di questo modello, si è deciso di applicare la convalida incrociata e di scegliere i parametri di *tuning* del modello attraverso la minimizzazione del tasso di errata classificazione. I parametri principali nella costruzione di un albero di classificazione sono *minsize* e *mindev* che rappresentano rispettivamente la dimensione minima permessa dai nodi e il valore minimo che deve essere la devianza entro i nodi per permettere la suddivisione in foglie. Questi due parametri sono stati fatti variare su una griglia equi spaziata che assume rispettivamente i valori interi da 1 a 10, e i valori discreti da 0.001 a 0.01 con incremento successivo di 0.001.

5.4.1 Risultati per il modello HCP 5/15

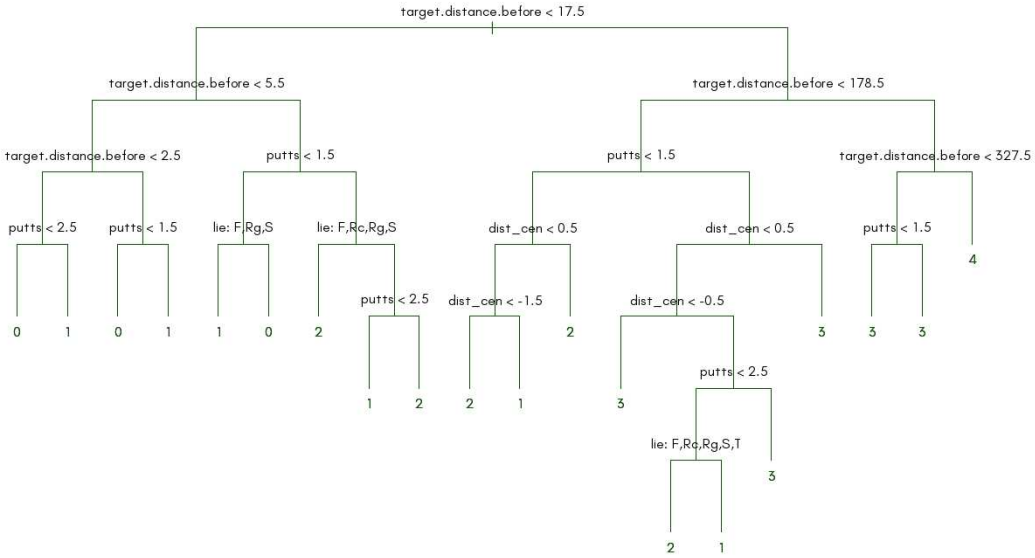


Figura 5.13: Risultati dell'albero di regressione con *splines*

Leaf_Node	Predicted_Class	0	1	2	3	4	5	6	7
16	0	0.931	0.069	·	·	·	·	·	·
17	1	0.400	0.533	0.067	·	·	·	·	·
18	0	0.933	0.067	·	·	·	·	·	·
19	1	0.250	0.750	·	·	·	·	·	·
20	1	·	1.000	·	·	·	·	·	·
21	0	0.750	0.250	·	·	·	·	·	·
22	2	·	0.063	0.875	0.063	·	·	·	·
46	1	·	0.989	0.011	·	·	·	·	·
47	2	·	0.111	0.889	·	·	·	·	·
48	2	·	·	0.950	0.050	·	·	·	·
49	1	·	0.879	0.091	0.030	·	·	·	·
25	2	·	·	0.875	0.100	0.025	·	·	·
52	3	·	·	0.040	0.800	0.080	0.040	0.040	·
212	2	·	·	0.946	0.018	0.036	·	·	·
213	1	·	1.000	·	·	·	·	·	·
107	3	·	·	·	1.000	·	·	·	·
27	3	·	·	0.033	0.833	0.100	0.033	·	·
28	3	·	·	0.219	0.656	0.063	0.031	0.031	·
29	3	·	·	0.010	0.608	0.330	0.041	0.010	·
15	4	·	·	·	0.131	0.667	0.162	0.020	0.020

Tabella 5.7: Probabilità delle foglie del modello

Come illustra la Figura 5.13, le variabili più impattanti in questo modello sono la distanza dal bersaglio prima del colpo, la distanza dal centro dal *fairway*, e il *lie* Il primo *split* può essere interpretato come la scelta del bastone di riferimento, sotto i 17.5 metri tendenzialmente si preferisce usare il *putter* dove, se il bersaglio è inferiore ai 5.5 metri allora è più probabile avere zero colpi rimanenti (con il bersaglio sotto i 2.5 metri facendo 2 *putt* c'è il 93% di probabilità di lasciarsi 0 colpi rimanenti (ovvero aver terminato la buca), mentre con 3 *putt* la probabilità lasciarsi 0 colpi è pari a 0.4). Se il bersaglio è compreso fra 2.5 e 5.5 il numero di *putt* per terminare la buca è determinante, sotto i 1.5 allora c'è il 93% di probabilità di aver terminato la buca, altrimenti con due *putt* la probabilità di imbucare è pari al 25%.

La parte destra della Figura 5.13 permette la stessa interpretazione sulla scelta del bastone, se la distanza dalla buca è inferiore a 178.5 metri verrà scelto un ferro, dove poi entra in gioco la distanza dal centro del *fairway*, essere a destra o a sinistra è fondamentale nel calcolo del numero di colpi rimanenti.

Dai valori degli *split* si evince come sia preferibile tirare alla sinistra del *fairway* rispetto che a destra, sebbene la numerosità della foglia 213 (tabella 5.7) sia pari a 1.

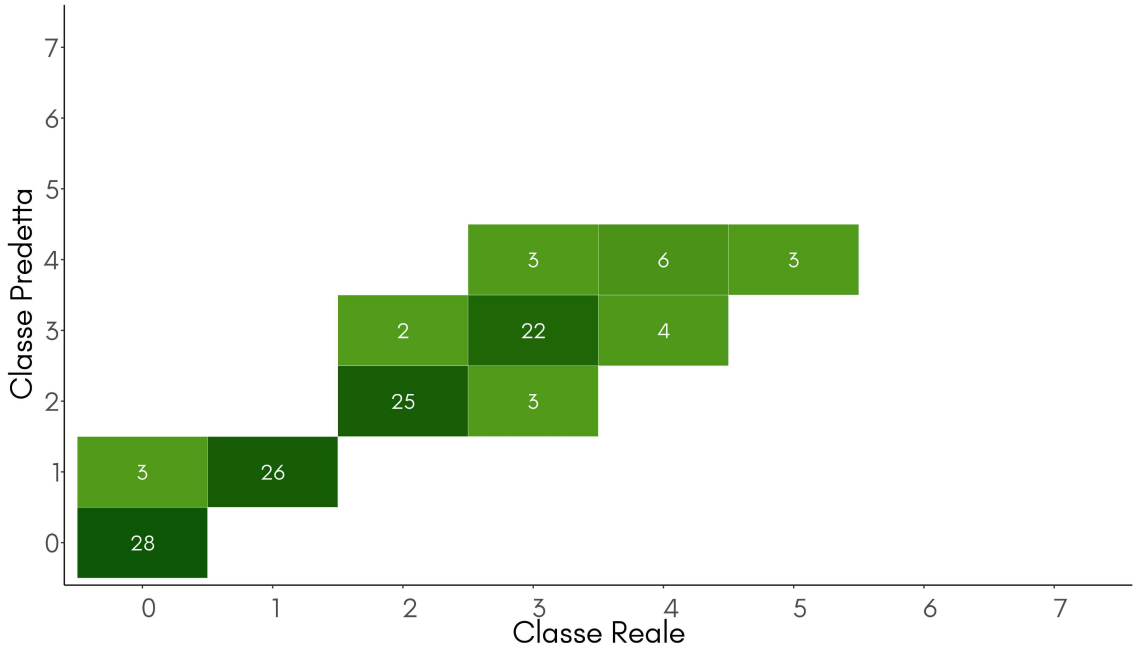


Figura 5.14: Matrice di confusione del modello di classificazione mediante albero nel *dataset* di verifica

La matrice di confusione illustra un ottima *performance* del modello nel classificare correttamente il numero di colpi rimasti, con solamente 24 unità classificate erroneamente rispetto alle 125 del *dataset* di verifica.

5.4.2 Risultati per il modello HCP 15/30

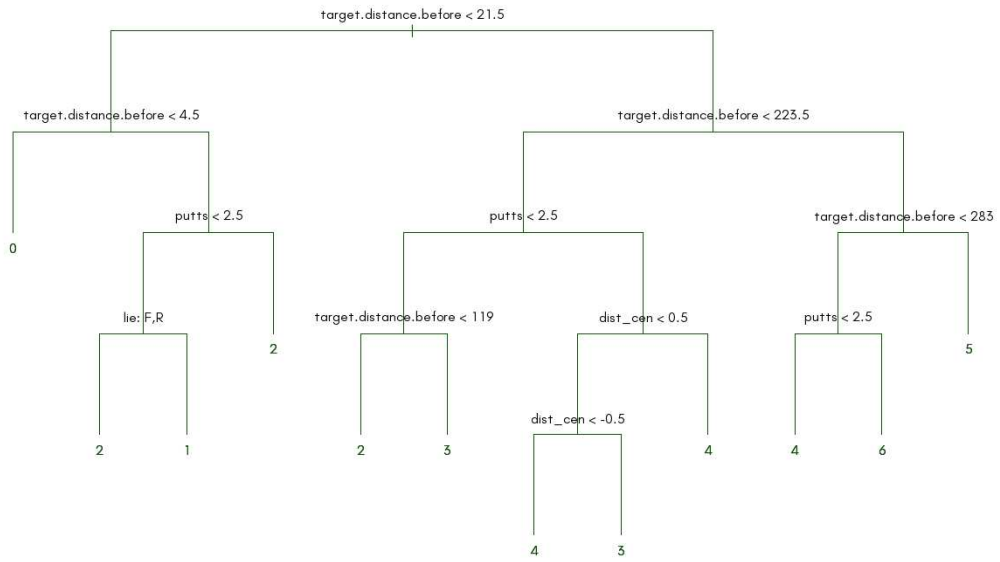


Figura 5.15: Risultati dell'albero di regressione con *splines*

Leaf_Node	Predicted_Class	0	1	2	3	4	5	6	7
4	0	0.609	0.353	0.039	·	·	·	·	·
20	2	·	0.355	0.613	0.032	·	·	·	·
21	1	0.067	0.933	·	·	·	·	·	·
11	2	0.033	0.100	0.700	0.167	·	·	·	·
24	2	·	0.051	0.663	0.235	0.041	0.010	·	·
25	3	·	0.010	0.112	0.561	0.204	0.092	0.020	·
52	4	·	·	·	·	0.600	0.333	·	0.067
53	3	·	·	0.071	0.714	0.071	0.095	0.048	·
27	4	·	·	·	·	0.515	0.273	0.182	0.030
28	4	·	·	0.020	0.224	0.531	0.224	·	·
29	6	·	·	·	·	0.125	0.250	0.500	0.063
15	5	·	·	·	0.030	0.180	0.300	0.240	0.160

Tabella 5.8: Probabilità delle foglie del modello

Nella figura 5.15 possiamo andare a distinguere due macro-zone divise dal nodo `target.distance.before < 21.5`. A sinistra sono racchiusi i

colpi di approccio breve e *putt*; se la buca è sotto i 4.5 metri di distanza allora c'è il 60.9% di probabilità che il colpo venga imbucato e il 35.3% di probabilità che serva un'altro colpo per terminare la buca. Sotto i 21.5 metri se il *lie* della palla è *fairway* o *rough* allora c'è il 61.3% di probabilità che vi siano 2 colpi ulteriori per terminare la buca, altrimenti c'è il 93.3% di probabilità che vi sia 1 solo colpo rimanente, sotto intendendo che il *lie* della pallina sia il *green*.

La parte a destra del primo *split* si basa sostanzialmente sulla divisione in base alla distanza dal bersaglio: più aumenta la distanza dalla buca e più aumenta il numero di colpi.

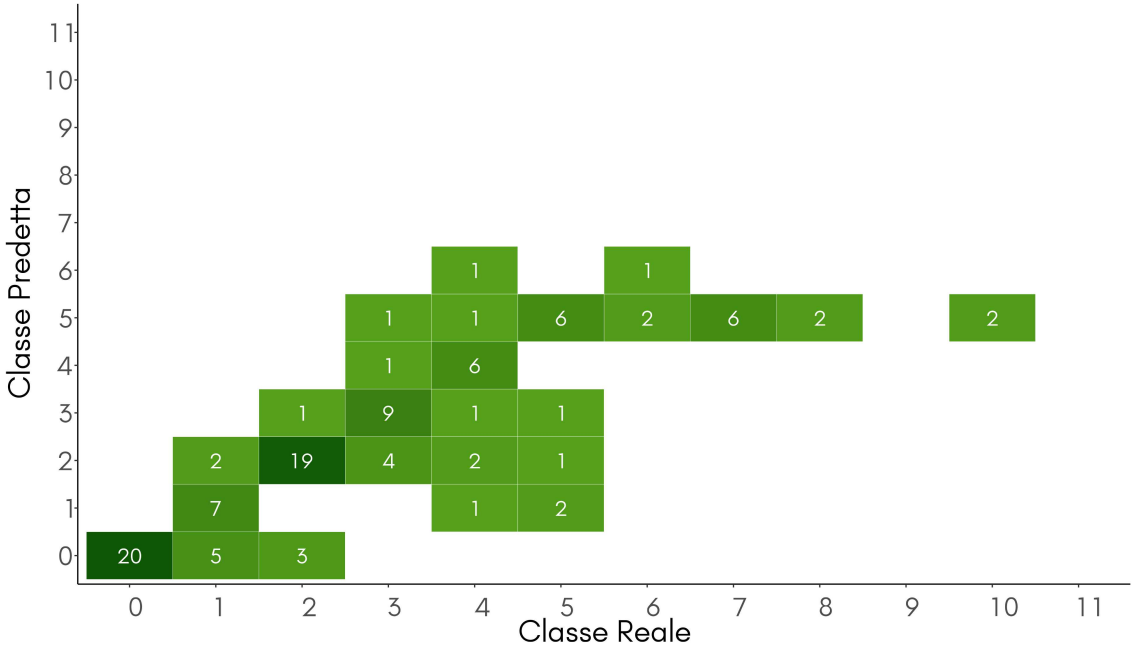


Figura 5.16: Matrice di confusione del modello di classificazione mediante albero nel *dataset* di verifica

Anche in questo modello come il modello MARS, la classificazione è meno precisa e più sparsa, con un totale del 36.5% di unità classificate male nel dataset di verifica.

5.5 *Random Forest*

I *Random Forest* (foreste casuali) sono una potente tecnica di apprendimento automatico basata su alberi decisionali, introdotta da Leo Breiman nel 2001. Essi appartengono alla classe dei metodi di ***ensemble learning***, che combinano le previsioni di molti modelli di base (in questo caso,

alberi decisionali) per migliorare le prestazioni predittive e la robustezza rispetto all'uso di un singolo albero.

I *Random Forest* costruiscono una collezione (foresta) di alberi decisionali su sottocampioni casuali del dataset originale. Le previsioni finali sono ottenute aggregando le previsioni di tutti gli alberi dove la previsione finale è la classe con il maggior numero di voti tra gli alberi (**voto di maggioranza**).

La costruzione di una *Random Forest* coinvolge i seguenti passaggi:

1. ***Bootstrap sampling***:
   - Per ciascun albero, si crea un campione *bootstrap* prelevando casualmente con ripetizione dal *dataset* di *training* originale
   - Questo introduce diversità tra gli alberi, poiché ognuno viene addestrato su un diverso sotto campione dei dati.
2. **Costruzione degli alberi**:
   - Durante la costruzione di ogni albero, ad ogni nodo di *split*, si considera un sottoinsieme casuale di variabili predittive (*feature subset*).
   - Questo ulteriore livello di casualità riduce la correlazione tra gli alberi e migliora la capacità di generalizzazione del modello.
3. **Aggregazione**:
   - Le previsioni dei singoli alberi vengono combinate per produrre la previsione finale dell'*ensemble*.

Grazie al processo di aggregazione e alla casualità introdotta, i *Random Forest* tendono a sovradattarsi meno rispetto ai singoli alberi decisionali, possono gestire un gran numero di variabili predittive senza richiedere una selezione preliminare delle stesse.

Forniscono inoltre stime dell'importanza delle variabili, utili per interpretare il modello e selezionare le feature più rilevanti e sono meno sensibili a valori mancanti e anomalie nei dati.

A differenza dei singoli alberi decisionali però, i *Random Forest* sono modelli di *"black box"* e risultano meno interpretabili, possono essere computazionalmente intensivi, specialmente con grandi *dataset* e un elevato numero di alberi e, sebbene meno propensi all'*overfitting* rispetto ai singoli alberi, in presenza di dati estremamente rumorosi possono comunque sovradattarsi.

I *Random Forest* forniscono una stima interna dell'errore, chiamata *Out-of-Bag* (OOB) *error*, senza necessità di un *set* di validazione separato. Questo è possibile perché, in media, per ogni albero, circa un terzo dei dati non viene utilizzato nel campione *bootstrap* e può essere usato per testare le prestazioni dell'albero.

Anche per questa tipologia di modello, si è scelto il parametro di regolazione (numero di variabili da includere nel modello) mediante convalida incrociata.

5.5.1 Risultati per il modello HCP 5/15

Il numero di variabili da includere a ogni nodo, che minimizza il tasso di errata classificazione, è pari a 4.

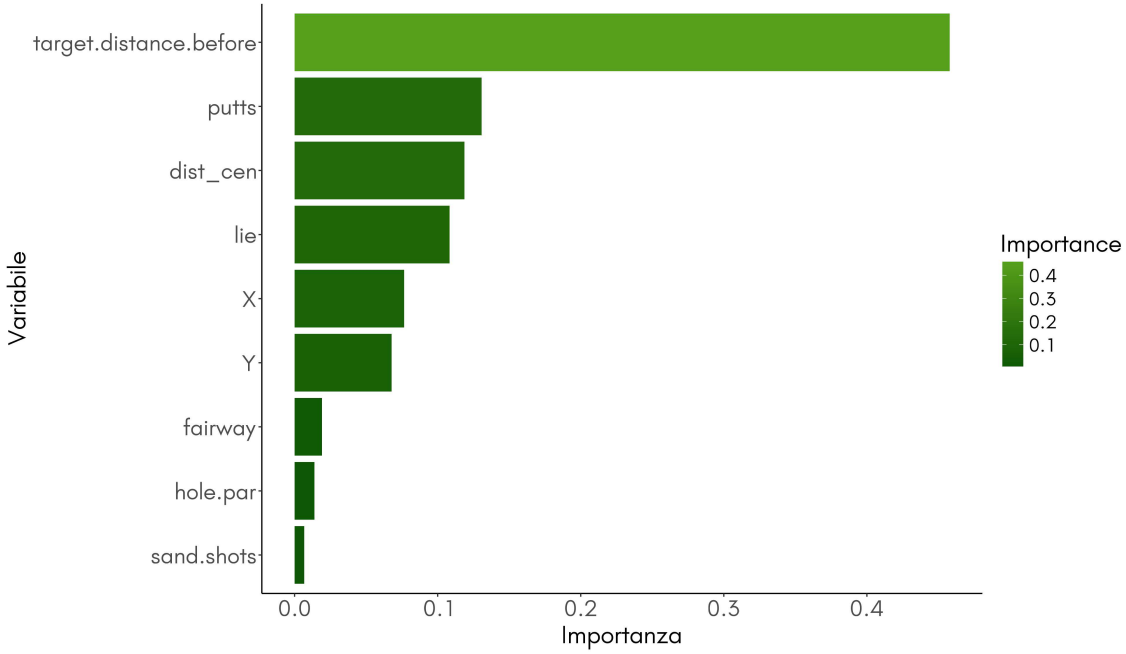


Figura 5.17: Grafico dell'importanza delle variabili nel modello *Random Forest*

La Figura 5.17 mostra l'importanza delle variabili nel modello di *Random Forest*. La variabile più importante è la distanza dal bersaglio con un'importanza del 45.88%, seguita dal numero di *putt* per terminare la buca con il 13.22%.

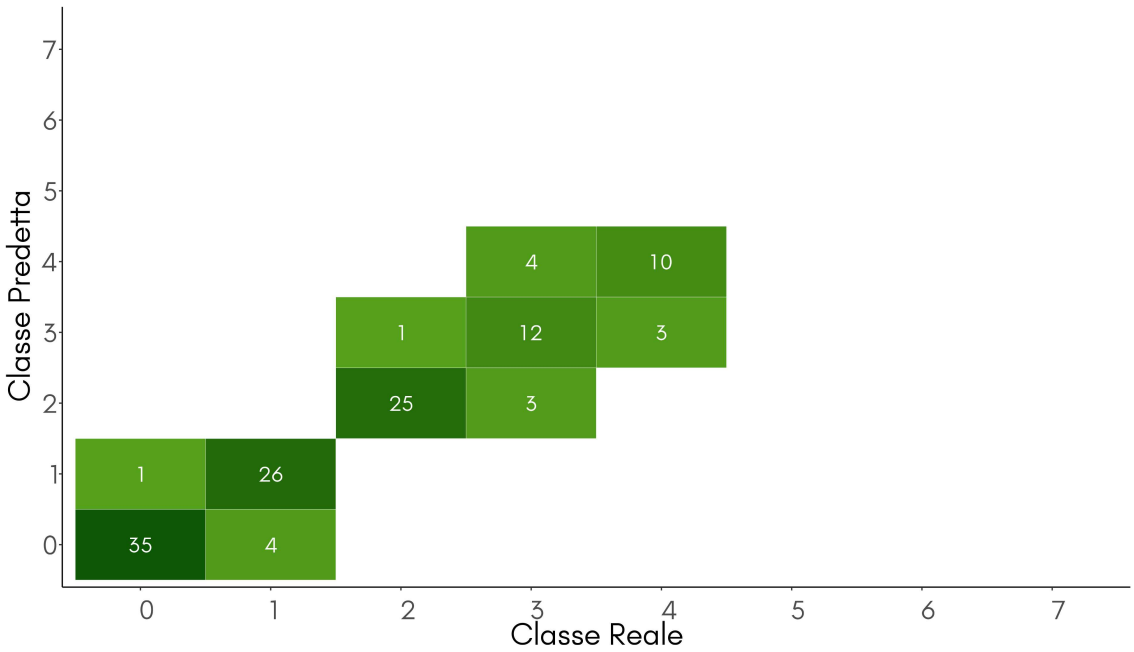


Figura 5.18: Matrice di confusione del modello *Random Forest* nel dataset di verifica

Questa tipologia di modello essendo di sua natura *casuale* non lascia spazio a molte interpretazioni in sé per sè, quindi l'attenzione si sposta sulle prestazioni del modello, come mostrato dalla matrice di confusione nella Figura 5.18.

Il numero di unità classificate correttamente è pari a 108 su 125, che porta a considerare il modello *Random Forest* come il migliore per classificare il numero di colpi rimanenti.

5.5.2 Risultati per il modello HCP 15/30

Il numero di variabili da includere ad ogni nodo che minimizza il tasso di errata classificazione è pari a 7.

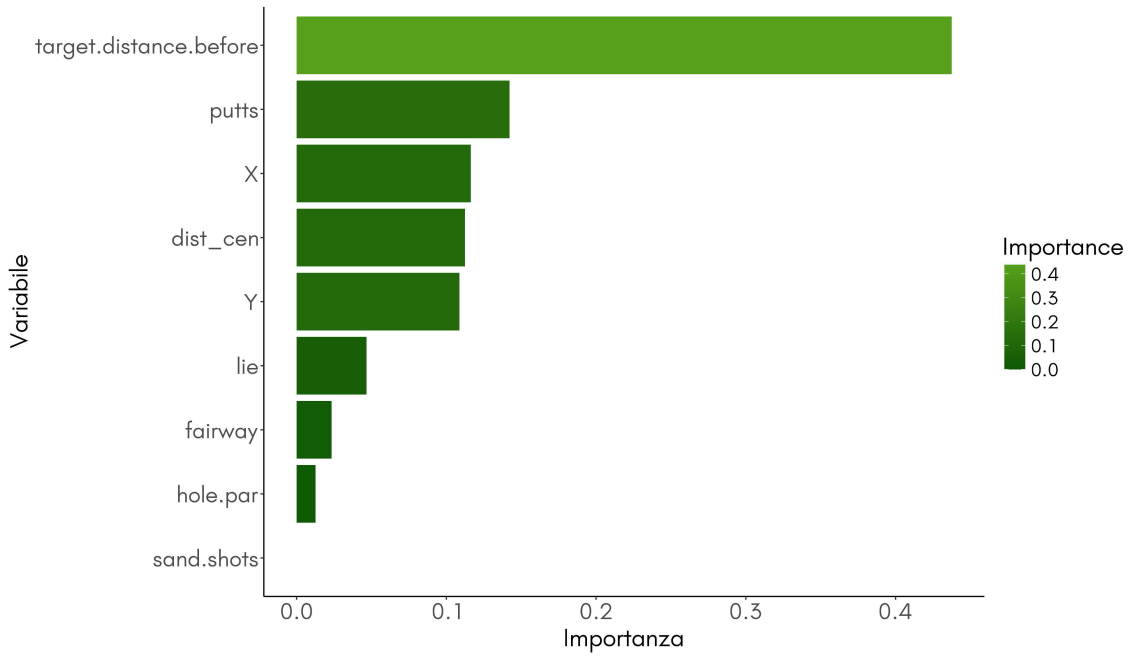


Figura 5.19: Grafico dell'importanza delle variabili nel modello *Random Forest*

Anche con questo *dataset* la variabile più importante risulta essere la distanza dal bersaglio, con un'importanza del 43.74%. Con questo *dataset* le variabili spaziali come X e Y assumono maggiore importanza.

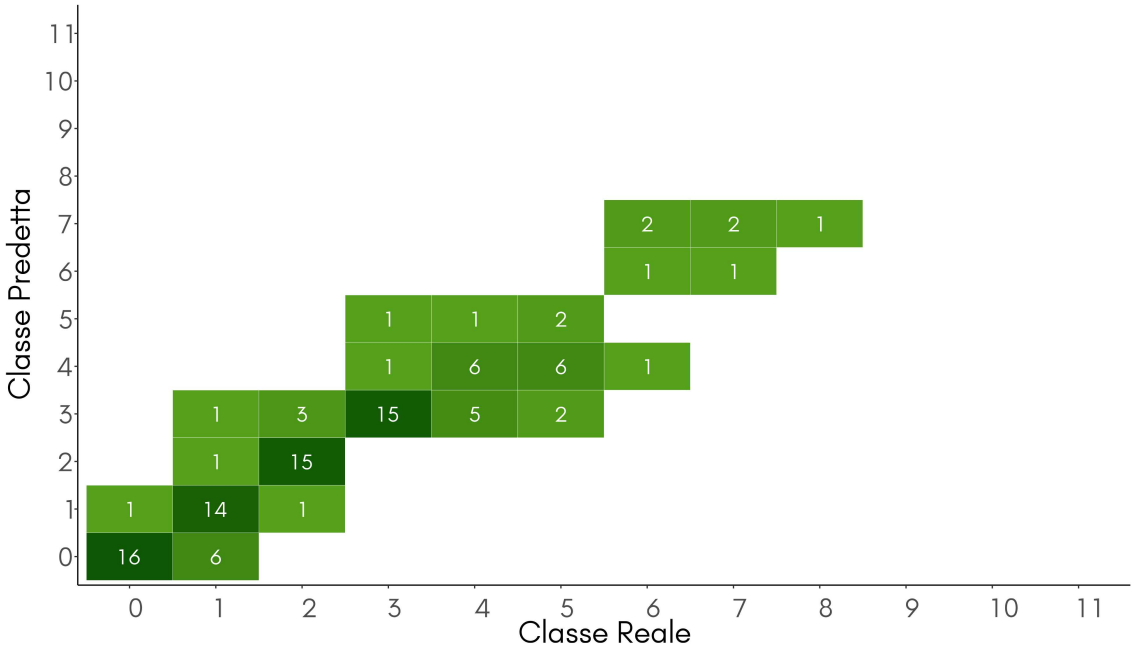


Figura 5.20: Matrice di confusione del modello Random Forest nel dataset di verifica

Questo *dataset* in tutte le analisi è risultato essere più difficile da modellare, in quanto presenta modalità della variabile risposta più sparse ed una numerosità leggermente inferiore. Il *Random Forest* classifica molto bene per le prime 5 classi, mentre perde potenza per le classi 5, 6, 7 ed 8 anche a causa delle poche osservazioni.

5.6 Strokes gained

In questa sezione, si seleziona il modello con la prestazione migliore e si procede con la definizione degli *strokes gained*.

Modello	**Accuracy HCP 5/15**	**Accuracy HCP 5/15**
Logistica Ordinale	0.6532	0.4952
MARS	0.8388	0.6952
Logistica Multinomiale	0.7177	0.5143
Random Forest	0.8710	0.7048
Albero di Classificazione	0.8710	0.6449

Tabella 5.9: Accuratezza dei modelli

La Tabella 5.9 mostra i risultati generali di *accuracy* per ogni modello.

Il modello migliore dal punto di vista di interpretazione sarebbe l'albero di classificazione, ma dal punto di vista delle prestazioni è preferibile il modello *Random Forest*, che ci restituisce, per ogni modalità della variaible `remaining_shots` una distribuzione di probabilità, dalla quale per voto di maggioranza si va a scegliere la classe rappresentativa come previsione.

Però è possibile utilizzare questa distribuzione per andare a definire in primis il valore atteso della variabile risposta per quel determinato colpo.

Supponiamo di definire la modalità di risposta dell'i-esima come Y_{ik} e la corrispondente distribuzione di probabilità stimata come $\hat{p}_{i\cdot}$. Possiamo quindi a creare il *benchmark* per il colpo i-esimo come:

$$J(f(\mathbf{X}_i)) = \mathbb{E}[Y_{i\cdot}] \tag{5.4}$$

$$= \sum_{k=0}^{K} Y_{ik}\hat{p}_{ik} \tag{5.5}$$

Ad esempio se prendessimo la prima riga del *dataset* del giocatore con HCP 5/15, si riferisce al primo colpo della prima buca, che in quella specifica data ha terminato in 6 colpi, il *benchmark* per quel colpo risulta essere 5.4104.

Utilizzando la formula per il calcolo degli *strokes gained* (1.1) è possibile prendere le prime 6 osservazioni del *dataset*, che rappresentano la prima buca giocata e calcolare questa metrica per ogni colpo:

$$S_i = J(f(\mathbf{X}_i)) - J(f(\mathbf{X}_{i+1})) - 1 \tag{5.6}$$

Colpo	**Posizione**	$J(f(\mathbf{X}_i))$	***Strokes Gained***
1	Da: 466m, Tee A: 291m, Fairway	$J(\cdot) = 5.41$ $J(\cdot) = 4.51$	$S_1 = -0.10$
2	Da: 291m, Fairway A: 154m, Fairway	$J(\cdot) = 4.51$ $J(\cdot) = 3.98$	$S_2 = -0.47$
3	Da: 154m, Fairway A: 31m, Fairway	$J(\cdot) = 3.98$ $J(\cdot) = 3.00$	$S_3 = -0.02$
4	Da: 31m, Fairway A: 6m, Green	$J(\cdot) = 3.00$ $J(\cdot) = 1.96$	$S_4 = 0.04$
5	Da: 6m, Green A: 5m, Green	$J(\cdot) = 1.96$ $J(\cdot) = 1.38$	$S_5 = -0.42$
6	Da: 5m, Green A: In Buca	$J(\cdot) = 1.38$ $J(\cdot) = 0$	$S_6 = 0.38$

Tabella 5.10: Calcolo degli *strokes gained* per le prime 6 osservazioni del *dataset* del giocatore con HCP 5/15

Per questo campione di osservazioni i colpi che sono stati valutati peggio, basandosi sullo storico del giocatore (*benchmark*), sono il numero 2 (un colpo "sicuro" per tenersi il terzo tiro in *green*) e il numero 5 (un brutto *putt* che ha percorso solo un metro di distanza).

5.6.1 *Danger zone*

I modelli sono stati costruiti in base alla concezione che nel campo non vi siano ostacoli in altezza come alberi, e che quindi il giocatore sia portato a tirare sempre verso la buca. Però, come nel caso della Tabella 5.10 e della conformazione della buca 1, tirare il secondo colpo direttamente verso la buca potrebbe portare il giocatore a colpire gli alberi, motivo per il quale sono state tracciate per tutto il campo delle *Danger nones* in corrispondenza degli alberi.

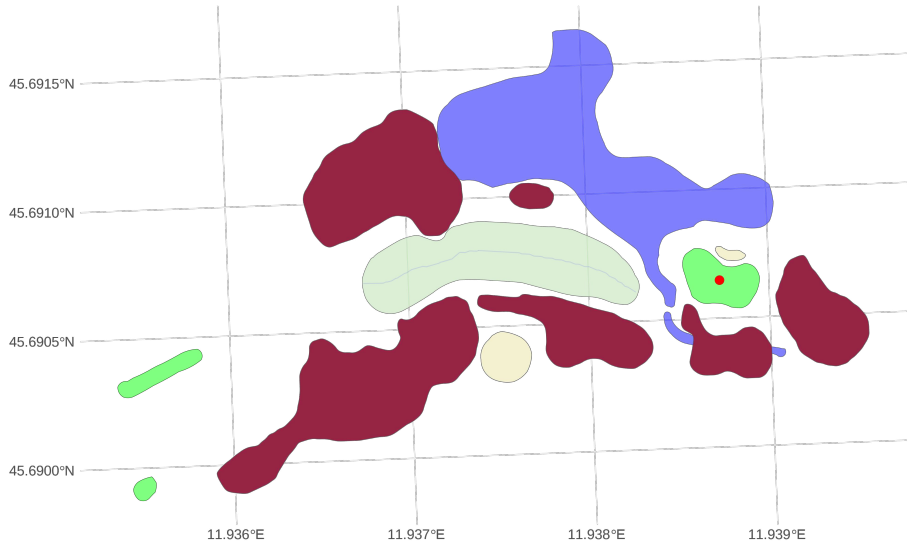


Figura 5.21: Mappa della buca 15 con la presenza delle *danger zones*

L'idea, quindi, è quella di modificare la distribuzione di probabilità ottenuta dal modello basandosi sulla percentuale di *danger zone* che la linea ideale del tiro dovrebbe attraversare.

La modifica della distribuzione avviene mediante l'operazione di **convoluzione**, ovvero la somma di due distribuzioni di probabilità, fra la distribuzione ottenuta dal modello e una funzione di spostamento generata appositamente. La convoluzione $Q(z)$ della distribuzione originale $P(x)$ con la distribuzione di spostamento $S(y)$ si calcola come:

$$Q(z) = \sum_{x=0}^{7} P(x) \cdot S(z - x)$$

Questa funzione parte dall'idea che il giocatore, una volta entrato nella *danger zone* debba ragionare in maniera diversa rispetto a quando si trova senza ostacoli, ovvero cercando di fare prima un colpo di recupero per mettere la palla in sicurezza e poi successivamente poter tornare a giocare la buca normalmente.

Concettualmente, se un giocatore ha davanti a sé il 100% di *danger zone*, si trova costretto a fare un colpo in direzione del *fairway* più vicino e tirare poi il colpo successivo, perdendo così 1 colpo. La funzione di spostamento che definiamo con $S(y, d_r)$ andrà a sommare 1 a tutta la distribuzione di probabilità se la percentuale di *danger zone* è pari a 100, mentre andrà a sommare 0 se è pari a 0.

Per poter validare l'idea di una modifica *ex-post* alla distribuzione è necessario raccogliere una moltitudine di dati in più, in quanto generalmente la probabilità di finire in queste zone è bassa. A tale motivo si

presentano 4 possibili funzioni di spostamento che andrebbero validate in presenza di molti più dati.

$$S(y; d_r) = \begin{cases} 1 - d_r & y = 0 \\ d_r(1 - d_r) & y = 1 \\ d_r^2 & y = 2 \end{cases} \tag{5.7}$$

$$S(y; d_r) = \begin{cases} 1 - d_r & y = 0 \\ d_r(1 - \omega d_r) & y = 1 \\ \omega d_r^2 & y = 2 \end{cases} \tag{5.8}$$

$$S(y; d_r) = \begin{cases} 1 - d_r & y = 0 \\ \alpha d_r(1 - d_r) & y = 1 \\ d_r[1 - \alpha(1 - d_r)] & y = 2 \end{cases} \tag{5.9}$$

$$S(y; d_r) = \begin{cases} 1 - {d_r}^{\frac{1}{2}} & y = 0 \\ {d_r}^{\frac{1}{2}} & y = 1 \end{cases} \tag{5.10}$$

Le quattro equazioni illustrate sopra hanno sviluppi diversi, la 5.7 considera che se un giocatore si trova completamente all'interno della *danger zone* impiega 2 colpi in più rispetto al normale, nella 5.8 viene inserito il parametro ω che è la probabilità di impiegare 2 colpi a uscire dalla zona rispetto che 1. Nella 5.9 il parametro α regola la velocità di spostamento fra $y = 1$ e $y = 2$ mentre nella 5.10 si considera che il massimo spostamento sia di un colpo, ma l'effetto segue l'andamento di una funzione a radice quadrata.

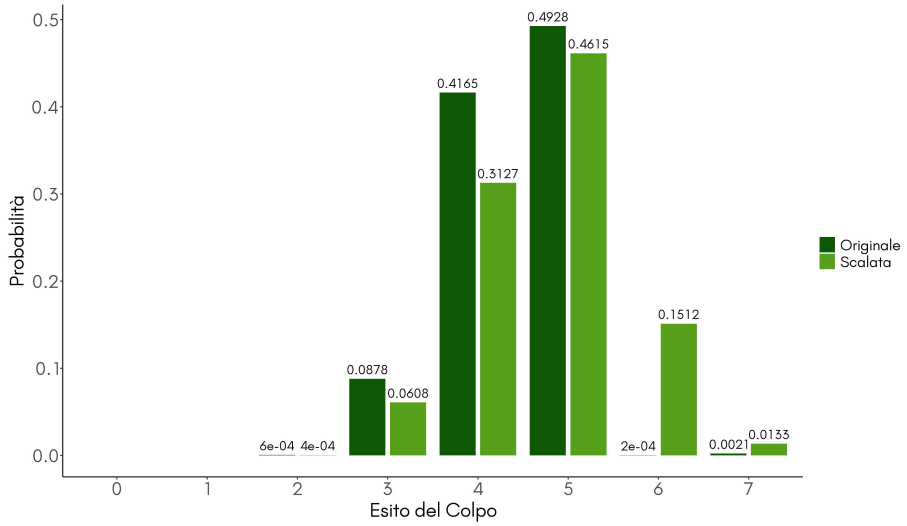


Figura 5.22: Differenza fra distribuzione di probabilità originale e scalata con la convoluzione

La Figura 5.22 mostra la differenza fra la distribuzione di probabilità del primo colpo (Tabella 5.10) originale, e quella riscalata con la convoluzione (5.8), considerando una percentuale di *danger zone* di esempio del 31% e un valore di ω di 0.28.

Capitolo 6

Conclusioni

Questa ricerca ha presentato approcci diversi e più completi per il calcolo degli *strokes gained*, aggiungendo alcune variabili di gioco che aiutano a fornire una panoramica più dettagliata sulla prestazione. La variabile più significativa in tutti i modelli è risultata essere la distanza dalla buca, che gioca un ruolo fondamentale molto più del trovarsi a sinistra o destra del *fairway* o trovarsi nella sabbia.

Il modello con la prestazione migliore con i dati del 2024 è stato il *random forest*, con un'accuratezza di 0.8710. Questo modello non lascia spazio a interpretazioni in quanto consiste in una combinazione di classificatori e funziona come una *black-box*. La variabile *putts* è stata considerata nei modelli sebbene sia una forma di informazione *leaker* per la risposta. Difatti, gioca un ruolo secondario rispetto alla distanza dal bersaglio nell'addestramento dei modelli. Una possibile soluzione è quella di inserire la media dei *putt* per quella buca fatti dal giocatore come storico in fase di previsione puntuale.

La suddivisione dei *dataset* in base all'abilità dei giocatori tramite HCP, si è rivelata fondamentale per differenziare le *performance* e le strategie in base al livello di competenza. Questo approccio ha evidenziato come i giocatori con HCP più basso tendano a sviluppare colpi più lunghi e consistenti, mentre quelli con HCP più alto necessitano di adattamenti per raggiungere buone *performance*.

L'adozione di modelli come i *Random Forest*, sebbene computazionalmente intensivi, potrebbe aprire prospettive interessanti per analisi in tempo reale delle *performance*, offrendo ai giocatori suggerimenti immediati sulle migliori strategie. Una possibile estensione futura potrebbe integrare tali modelli con dispositivi di monitoraggio portatili, in grado di raccogliere e analizzare dati colpo per colpo, migliorando la personalizzazione delle raccomandazioni e riducendo i margini di errore.

Poter utilizzare strumenti avanzati di raccolta dati come il *FlightScope Mevo Launch Monitor* - un dispositivo portatile che immagazzina tutti i dati fisici del colpo a partire dalla velocità del bastone e della pallina (*club speed* e *ball speed*), angolo di attacco (*attack angle*), *loft* dinamico (*dynamic loft*) e rotazione della pallina (*spin rate*) - permetterebbe di personalizzare ulteriormente il calcolo degli *strokes gained*, rendendolo più adatto ai singoli giocatori e alle loro caratteristiche.

Questo studio dimostra come un approccio *data-driven* possa trasformare l'allenamento golfistico, offrendo *feedback* dettagliati e mirati che i metodi tradizionali di osservazione non potrebbero fornire. Analizzare le aree di forza e debolezza consente infatti una personalizzazione dell'allenamento, in modo da poter ottimizzare sia il gioco lungo che le fasi di approccio e *putting*. Questo tipo di analisi offre ai *coach* uno strumento potente per identificare trend di lungo periodo e per risolvere inefficienze nel gioco.

L'analisi geografica effettuata tramite QGIS ha permesso di rilevare come la struttura del campo e la posizione dei colpi influiscano sulle *performance* complessive, mostrando differenze significative tra colpi eseguiti in zone centrali rispetto a quelli in aree periferiche o con ostacoli. Questa dimensione spaziale evidenzia il potenziale dell'uso di strumenti GIS per prevedere i risultati basati sulla posizione geografica e sulle condizioni ambientali specifiche di ogni colpo.

Una fra le problematiche affrontate è stata la qualità dei dati forniti dall'applicazione per la rilevazione degli stessi, che non permette di registrare molte altre variabili. Inoltre, l'uniformità dei soggetti nella raccolta dati ha rappresentato una sfida: essendo lo strumento di raccolta un *software* utilizzabile in molteplici modi, nonostante le indicazioni fornite dal manuale, non è sempre intuitivo seguirle, portando a distorsioni dei dati.

C'è stato anche il problema della gestione dei dati particolari, come i colpi legati alle situazioni di penalità o ai *drop* senza penalità in luoghi diversi da dove è atterrata la pallina. La scarsa numerosità del *dataset* completo ha influenzato negativamente l'indagine, producendo risultati non molto soddisfacenti specialmente per la categoria 5/15 HCP, che ha presentato dati più dispersi.

Un possibile punto futuro per ampliare questa ricerca sarebbe la costante raccolta dati anche dalle situazioni di pericolo, come presentato nella sezione 5.6, e la raccolta dati da parte di un numero maggiore di soggetti. Anche l'estensione a campi diversi, con la conseguente mappatura di un numero superiore di circoli, arricchirebbe questa ricerca

in maniera completa, potendo diversificare così le fonti dati.

La potenzialità nello sviluppo di un'applicazione *mobile*, con una UI (*User Interface*) facile da interpretare e accessibile a tutte le età e tutti i livelli di gioco, all'interno della quale è possibile avere una visione tridimensionale dell'area di atterraggio della pallina, permetterebbe di poter aiutare i giocatori con le proprie scelte strategiche, magari con dettagli a schermo che permettano di vedere le possibili traiettorie del colpo da fare, con le conseguenti distribuzioni di probabilità per il numero di colpi rimanenti per finire la buca, oppure evidenziando con colori diversi le zone che hanno una maggiore probabilità di andare a incrementare il numero di colpi rimanenti.

In fase di stesura di questa ricerca si è pensato di usare il *Deep Reinforcement Learning*, un ramo del *machine learning* che combina *Reinforcement Learning* e *Deep Learning*, con l'obiettivo di insegnare ad un agente a completare una buca da golf nel minor numero di colpi possibili. Il *Deep Learning* utilizza una rete neurale per trasformare un set di *input* in un insieme di *output* attraverso una rete neurale. Il *Reinforcement Learning* è un processo per il quale un'agente (*agent*) impara a muoversi e prendere decisioni in un'ambiente (*enviroment*) attraverso prove, errori ed un sistema di ricompense basato su essi. L'idea è quella di addestrarlo sui dati raccolti sul campo e introdurre penalizzazioni in base alla distanza dalla buca o al terreno in cui si trova, con lo scopo finale di ottenere un modello che, avendo in input un qualsiasi punto all'interno del campo, consenta di mostrare al giocatore il percorso migliore per arrivare alla buca (Omstedt, 2020).

Lo sviluppo odierno di modelli basati su intelligenza artificiale, permette a questo sport di andare in una direzione dove le scelte e l'approccio al gioco sono sempre più *data driven*, favorendo lo sviluppo anche di nuove figure professionali che affiancherebbero i giocatori professionisti durante gli allenamenti e le gare ufficiali.

Appendice A

Descrizione delle Variabili

Dataset `rounds`

Tabella A.1: Descrizione delle variabili del *dataset* `rounds`

Variabile	**Descrizione**
`player name`	Nome del giocatore che ha effettuato il *round*.
`date`	Data in cui è stato giocato il *round*.
`start time`	Orario di inizio del *round*.
`course name`	Nome del campo da golf in cui si è svolto il *round*.
`course holes`	Numero totale di buche presenti nel campo.
`tee name`	Nome del tee di partenza utilizzato (ad esempio, "*White*", "*Yellow*", "*Red*").
`rating`	Valutazione del campo, che rappresenta la difficoltà standard per un giocatore *scratch*.
`slope`	Valore che indica la difficoltà relativa del campo per un giocatore medio rispetto a un giocatore *scratch*.
`course handicap`	*Handicap* del campo per il giocatore, calcolato in base al suo *handicap* personale e alle caratteristiche del campo.
`scoring format`	Formato di punteggio utilizzato nel round (ad esempio, "*Stroke Play*", "*Stableford*").
`completed holes`	Numero di buche completate dal giocatore durante il *round*.
`gross score`	Punteggio lordo totale ottenuto dal giocatore.
`gross score over par`	Differenza tra il punteggio lordo e il *par* totale del campo.

Continua nella pagina successiva

Tabella A.1 – continua dalla pagina precedente

Variabile	Descrizione
`net score or points`	Punteggio netto o punti ottenuti dal giocatore, tenendo conto dell'*handicap*.
`putts`	Numero totale di putt effettuati durante il round.
`penalties`	Numero totale di colpi di penalità ricevuti durante il *round*.
`GIRs`	Numero di "*Green in Regulation*" ottenuti durante il *round*.
`fairways`	Numero di "*Fairway in Regulation*" colpiti dal giocatore dal *tee* di partenza.
`sand shots`	Numero di colpi effettuati dal *bunker* durante il *round*.
`URL`	Indirizzo *web* associato al *round* per ulteriori dettagli.
`finish time`	Orario di fine del *round*.

Dataset `shots`

Tabella A.2: Descrizione delle variabili del dataset `shots`

Variabile	Descrizione
`date`	Data in cui è stato effettuato il colpo.
`course name`	Nome del campo da golf dove è stato effettuato il colpo.
`hole number`	Numero della buca in cui il colpo è stato giocato (da 1 a 18).
`shot number`	Numero progressivo del colpo all'interno della buca.
`lie`	Posizione della palla prima del colpo (ad esempio, *Tee*, *Fairway*, *Rough*, *Bunker*, *Green*).
`club`	Tipo di bastone utilizzato per il colpo (ad esempio, *Driver*, *Ferro 7*, *Putter*).
`club details`	Dettagli specifici sul bastone, come marca, modello o altre caratteristiche.
`shot length meters`	Distanza percorsa dal colpo in metri.

Continua nella pagina successiva

Tabella A.2 – continua dalla pagina precedente

Variabile	**Descrizione**
`target distance before`	Distanza dal bersaglio prima del colpo, in metri.
`target distance after`	Distanza dal bersaglio dopo il colpo, in metri.
`outcome`	Superficie dove atterra il colpo effettuato (ad esempio, *Fairway*, *Rough*, *Green*).
`included in distance stats`	Indicatore (*yes*/*no*) se il colpo è incluso nelle statistiche di distanza.
`strokes gained`	Vedi 1.2 con formula di Broadie (2012)
`distance from center of fairway`	Distanza laterale in metri dalla posizione finale della palla al centro del *fairway*.
`time`	Orario in cui il colpo è stato effettuato.

Dataset `holes`

Tabella A.3: Descrizione delle variabili del *dataset* `holes`

Variabile	**Descrizione**
`player name`	Nome del giocatore che ha completato la buca.
`date`	Data in cui la buca è stata giocata.
`course name`	Nome del campo da golf.
`tee name`	Nome del *tee* di partenza utilizzato per la buca specifica.
`hole number`	Numero della buca giocata (da 1 a 18).
`hole par`	*Par* della buca, che rappresenta il numero standard di colpi in cui un giocatore esperto dovrebbe completare la buca.
`total strokes`	Numero totale di colpi effettuati dal giocatore per completare la buca.
`putts`	Numero di *putt* effettuati sul *green* per la buca.
`penalties`	Numero di colpi di penalità ricevuti nella buca.
`sand shots`	Numero di colpi eseguiti dal *bunker* nella buca.

Continua nella pagina successiva

Tabella A.3 – continua dalla pagina precedente

Variabile	**Descrizione**
`fairway`	Indicatore (*hit*/*right*/*left*) se il giocatore ha centrato il *fairway* con il colpo di partenza nelle buche *par* 4 e *par* 5.
`GIR`	Indicatore (*sì*/*no*) se il giocatore ha raggiunto il *green* in regolazione per la buca.

Appendice B

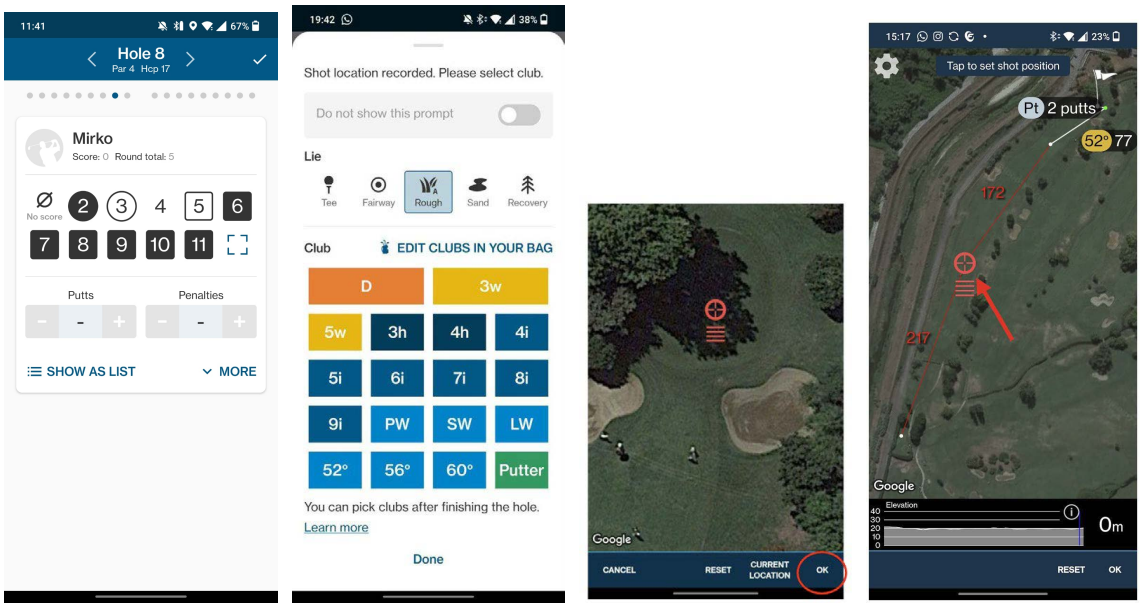
Metodo di raccolta

B.1 Tecniche di raccolta

In questa sezione viene descritto nel dettaglio com'è avvenuta la raccolta dati dei soggetti che si sono prestati a questa ricerca, ai quali è stato chiesto di:

- Inserire all'interno dello spazio riservato il numero di colpi e putt eseguiti. (B.1a)
- Successivamente, modificare la lista dei colpi inserendo manualmente la superficie di partenza (*Tee*, *Fairway*, ecc.) e la mazza. (B.1b)
- Inserire manualmente la posizione della bandiera (*Flag Position*). (B.1c)
- Per ogni colpo, modificare esattamente la posizione geografica di partenza. (B.1d)

Ripetendo questa procedura per tutte le 18 buche, il *software* riesce a ottenere quante più informazioni possibili, come descritto nel paragrafo dedicato alla descrizione delle variabili.



(a) Inserimento del numero di colpi e putt

(b) Modifica della lista dei colpi

(c) Inserimento della posizione della bandiera

(d) Modifica della posizione geografica di partenza

Figura B.1: Procedura di raccolta dei dati

Una volta esportati i dati, l'applicazione ci offre tre tipi di *dataset* diversi: `shots`, `rounds` e `holes`.

La descrizione delle variabili che compongono questo *dataset* è presente nell'appendice A.

B.2 *Shiny Dashboard*

Ai partecipanti dello studio successivamente è stata data la possibilità di monitorare le proprie statistiche, mediante la creazione di un sito *web* dedicato. [1]

Ai partecipanti è stato prima sottoposto un questionario con domande di profilazione generali (età, circolo di appartenenza, HCP) e più specifiche per il gioco del golf (*"Da quanto tempo pratichi questa disciplina"*, *"Quanto spesso ti alleni a settimana"*).

Il questionario è stato creato rispettando tutte le norme vicenti in termini di protezione dei dati persionali e conforme al GDPR.

Una volta concluso il questionario viene fornito il *link* al sito *web*, dove gli utenti sono invitati a registrarsi mediante *username* e *password*, le quali informazioni sono salvate all'interno di un *database* SQL che si

[1] https://ztr0o8-mirko0gabriel-briglia.shinyapps.io/dashboard/

appoggia ad AWS (*Amazon Web Services*) ed encriptate attravero un sistema di *password hashing*.

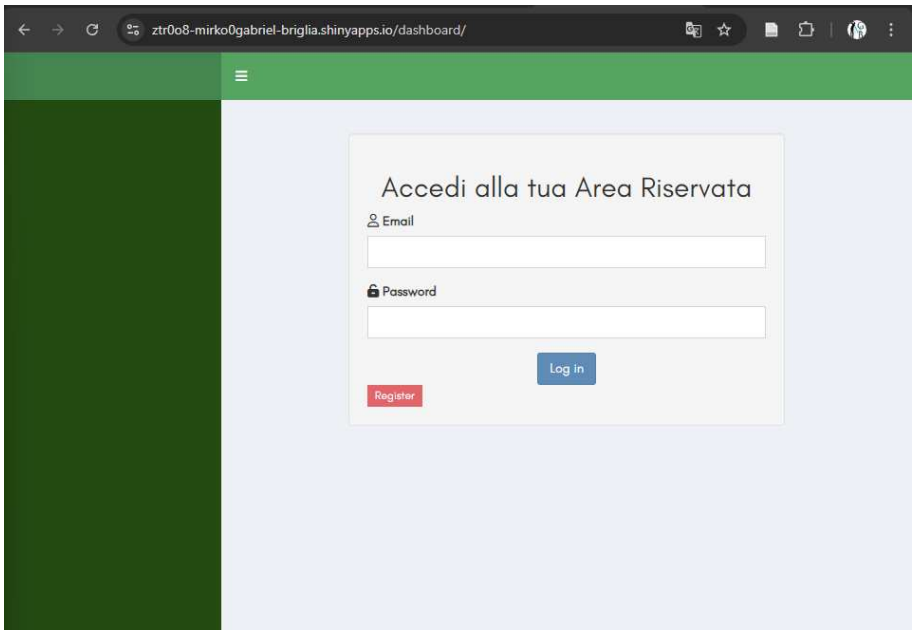


Figura B.2: *Screenshot* del sito web

L'intera struttura del sito web è stata creata attraverso il linguaggio di programmazione `R` con l'utilizzo di particolari librerie aggiuntive come `shiny` e `DBI` [2].

[2]Link per vedere il codice: https://github.com/MirkoGB/LaureaMagistrale

Bibliografia

[1] Mark Broadie. "Assessing Golfer Performance on the PGA TOUR". en. In: *Interfaces* 42.2 (apr. 2012), pp. 146–165. ISSN: 0092-2102, 1526-551X. DOI: 10.1287/inte.1120.0626. URL: https://pubsonline.informs.org/doi/10.1287/inte.1120.0626 (visitato il giorno 04/09/2024).

[2] Mark Broadie. "Assessing Golfer Performance Using Golfmetrics". en. In: Science and Golf V: Proceedings of the 2008 World Scientific Congress of Golf (2008), pp. 253–262.

[3] Mark Broadie. "Measuring Putting on the PGA TOUR". en. In: (2011).

[4] Mark Broadie e William J. Hurley. "Golf Analytics: Developments in Performance Measurement and Handicapping". In: *Handbook of Statistical Methods and Analyses in Sports*. Num Pages: 20. Chapman e Hall/CRC, 2016. ISBN: 978-1-315-16607-0.

[5] Mark Broadie e Soonmin Ko. "A simulation model to analyze the impact of distance and direction on golf scores". en. In: *Proceedings of the 2009 Winter Simulation Conference (WSC)*. Austin, TX, USA: IEEE, dic. 2009, pp. 3109–3120. ISBN: 978-1-4244-5770-0. DOI: 10.1109/WSC.2009.5429280. URL: http://ieeexplore.ieee.org/document/5429280/ (visitato il giorno 04/09/2024).

[6] Sangit Chatterjee, Frederick Wiseman e Robert Perez. "Studying improved performance in golf". In: *Journal of Applied Statistics* 29.8 (nov. 2002). Publisher: Routledge, pp. 1219–1227. ISSN: 02664763. DOI: 10.1080/0266476022000011283. URL: https://search.ebscohost.com/login.aspx?direct=true&db=bsu&AN=7437381&site=ehost-live (visitato il giorno 23/09/2024).

[7] Robert Connolly e Richard Rendleman. "Skill, luck, and streaky play on the PGA Tour". In: *Journal of the American Statistical Association* 103 (feb. 2008), pp. 74–88. DOI: 10.1198/016214507000000310.

[8] Robert A. Connolly e Richard J. Rendleman. "Going for the Green: A Simulation Study of Qualifying Success Probabilities in Professional Golf". en. In: *Journal of Quantitative Analysis in Sports* 7.4 (gen. 2011). ISSN: 1559-0410. DOI: 10.2202/1559-0410.1308. URL: https://www.degruyter.com/document/doi/10.2202/1559-0410.1308/html (visitato il giorno 18/10/2024).

[9] Robert A. Connolly e Richard J. Rendleman. "Going for the Green: A Simulation Study of Qualifying Success Probabilities in Professional Golf". eng. In: *Journal of quantitative analysis in sports* 7.4 (2011). Publisher: De Gruyter. ISSN: 1559-0410. DOI: 10.2202/1559-0410.1308.

[10] Tom Cunneff. *Is the Tour's ShotLink data merely for Strokes Gained geeks? Don't bet on it.* eng. ISSN: 1056-5493 Issue: 9 Num Pages: 78- Place: New York Publication Title: Golf magazine (1991) Volume: 61. 2019.

[11] Christian Drappi e Lance Co Ting Keh. "Predicting golf scores at the shot level". en. In: *Journal of Sports Analytics* 5.2 (apr. 2019), pp. 65–73. ISSN: 2215020X, 22150218. DOI: 10.3233/JSA-170273. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JSA-170273 (visitato il giorno 18/10/2024).

[12] Markus Dumke. "Reinforcement Learning in R". eng. mast. 2017. DOI: 10.5282/ubm/epub.59093. URL: https://epub.ub.uni-muenchen.de/59093/?template=user_area (visitato il giorno 04/09/2024).

[13] Douglas Fearing, Jason Acimovic e Stephen C. Graves. "How to Catch a Tiger: Understanding Putting Performance on the PGA Tour". en. In: *SSRN Electronic Journal* (2010). ISSN: 1556-5068. DOI: 10.2139/ssrn.1538300. URL: http://www.ssrn.com/abstract=1538300 (visitato il giorno 04/09/2024).

[14] Patrik Fisher. "Objective Analysis of Golf". In: (1998), pp. 41–42.

[15] Mark Heinsman. *From the Fairway to the Spreadsheet: The Rise of Golf Analytics and Data-Driven Strategies.* en-US. Mag. 2023. URL: https://www.alteryx.com/blog/from-the-fairway-to-the-spreadsheet-the-rise-of-golf-analytics-and-data-driven-strategies (visitato il giorno 23/09/2024).

[16] Erik L. Heiny e Robert Lowell Heiny. "Stochastic model of the 2012 PGA Tour season". en. In: *Journal of Quantitative Analysis in Sports* 10.4 (gen. 2014). ISSN: 1559-0410, 2194-6388. DOI: 10.1515/jqas-2014-0043. URL: https://www.degruyter.com/document/doi/10.1515/jqas-2014-0043/html (visitato il giorno 18/10/2024).

[17] Erik L. Heiny e Robert Lowell Heiny. "Stochastic model of the 2012 PGA Tour season". eng. In: *Journal of quantitative analysis in sports* 10.4 (2014). Publisher: De Gruyter, pp. 367–379. ISSN: 2194-6388. DOI: 10.1515/jqas-2014-0043.

[18] Andrew Hoegh. "Defining the Performance Coefficient in Golf: A Case Study at the 2009 Masters". eng. In: *Journal of quantitative analysis in sports* 7.2 (2011). Publisher: De Gruyter, pp. 1331–1331. ISSN: 1559-0410. DOI: 10.2202/1559-0410.1331.

[19] Soonmin Ko. "A Simulation Model to Analyze the Impact of Golf Skills and a Scenario-based Approach to Options Portfolio Optimization". en. In: ().

[20] Roland Minton. *Golf by the Numbers: How Stats, Math, and Physics Affect Your Game*. en. Google-Books-ID: ARQSbyrZIEMC. JHU Press, set. 2012. ISBN: 978-1-4214-0452-3.

[21] Aminu Hamza Nababa (Al'amin). *Understanding Q-Learning in Reinforcement Learning*. en. Lug. 2023. URL: https://medium.com/@alaminhnab4/understanding-q-learning-in-reinforcement-learning-3b0e10223ae5 (visitato il giorno 04/09/2024).

[22] Fredrik Omstedt. *A deep reinforcement learning approach to the problem of golf using an agent limited by human data*. eng. 2020. URL: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-277832 (visitato il giorno 04/09/2024).

[23] Richard J. Rendleman Jr. "The relative roles of skill and luck within 11 different golfer populations". eng. In: *Journal of quantitative analysis in sports* 16.3 (2020). Publisher: De Gruyter, pp. 237–254. ISSN: 2194-6388. DOI: 10.1515/jqas-2019-0028.

[24] Ryuzaburo SHIOJI. "A Study on Analyzing Golf Data Using Q-learning". jpn. In: *JSAI Technical Report, Type 2 SIG* 2014.DOCMAS-007 (2014). Publisher: The Japanese Society for Artificial Intelligence, pp. 04–. ISSN: 2436-5556. DOI: 10.11517/jsaisigtwo.2014.DOCMAS-007_04.

[25] Michael Stöckl, Peter F Lamb e Martin Lames. "The ISOPAR Method: A New Approach to Performance Analysis in Golf". en. In: *Journal of Quantitative Analysis in Sports* 7.1 (gen. 2011). ISSN: 1559-0410. DOI: 10.2202/1559-0410.1289. URL: https://www.degruyter.com/document/doi/10.2202/1559-0410.1289/html (visitato il giorno 18/10/2024).

[26] Michael Stöckl, Peter F. Lamb e Martin Lames. "The ISOPAR Method: A New Approach to Performance Analysis in Golf". eng. In: *Journal of quantitative analysis in sports* 7.1 (2011). Publisher: De Gruyter, pp. 1289–1289. ISSN: 1559-0410. DOI: 10.2202/1559-0410.1289.

[27] Takeru Suzuki et al. "Investigating factors that improve golf scores by comparing statistics of amateur golfers in repeat scramble strokes and one-ball conditions". en. In: *Journal of Human Sport and Exercise* 16.4 (2020). ISSN: 1988-5202. DOI: 10.14198/jhse.2021.164.09. URL: http://hdl.handle.net/10045/107171 (visitato il giorno 04/09/2024).

[28] *The Increasing Presence Of Data Analytics In Golf*. en-GB. Gen. 2020. URL: https://www.sportperformanceanalysis.com/article/increasing-presence-of-data-analytics-in-golf (visitato il giorno 23/09/2024).

[29] *The reinforcement learning problem*. URL: https://cran.r-project.org/web/packages/ReinforcementLearning/vignettes/ReinforcementLearning.html (visitato il giorno 04/09/2024).

[30] Kasra Yousefi e Tim B. Swartz. "Advanced putting metrics in golf". en. In: *Journal of Quantitative Analysis in Sports* 9.3 (gen. 2013). ISSN: 1559-0410, 2194-6388. DOI: 10.1515/jqas-2013-0010. URL: https://www.degruyter.com/document/doi/10.1515/jqas-2013-0010/html (visitato il giorno 18/10/2024).

[31] Kasra Yousefi e Tim B. Swartz. "Advanced putting metrics in golf". eng. In: *Journal of quantitative analysis in sports* 9.3 (2013). Publisher: De Gruyter, pp. 239–248. ISSN: 2194-6388. DOI: 10.1515/jqas-2013-0010.

Vorrei ringraziare di cuore il mio relatore Francesco Lisi e il mio correlatore Federico Cacciafesta per la loro professionalità e disponibilità dimostratami in questi mesi di lavoro, il Golf Club Ca'Amata per avermi accolto come socio a braccia aperte e in particolar modo il maestro Gabriele Heinrich per i suoi insegnamenti e per essere sempre stato disponibile per seguirmi in tutti i passi di questo progetto.

I would also like to thank Dr.Fredrik Omstedt, with whom I corresponded and who provided me with excellent insights for working on Deep Reinforcement Learning.