UNIVERSITA' DEGLI STUDI DI BRESCIA

FACOLTA' DI ECONOMIA

CORSO DI LAUREA MAGISTRALE IN

AZIENDA, MERCATO E INFORMAZIONE

TESI DI LAUREA

"Quando la palla non è solo rotonda: tecniche di Data Mining applicate al mondo del calcio."

RELATORE:

CHIAR.MA PROF.SSA PAOLA ZUCCOLOTTO

CORRELATORE:

CHIAR.MO PROF. MAURIZIO CARPITA

LAUREANDA:

ROBERTA TENGATTINI

MATRICOLA N. 76327

ANNO ACCADEMICO 2010/2011

“Il senso del calcio è che vinca il migliore in campo, indipendentemente dalla storia, dal prestigio e dal budget.”

Johan Cruijff

INDICE

INTRODUZIONE.

Il calcio, come tutti sanno, è uno sport di squadra, nel quale si affrontano due squadre composte ciascuna da undici giocatori, che usando un pallone sferico all'interno di un campo di gioco rettangolare con due porte, cercano di segnare più punti rispetto all'avversario, facendolo passare fra i pali della porta di quest'ultimo.

È diventato uno sport popolarissimo e praticato in tutto il mondo principalmente perché non richiede attrezzature speciali ed inoltre è estremamente adattabile ad ogni situazione.

Nell'antichità sappiamo che ci sono stati numerosi giochi simili a questo, in particolare in Cina, Giappone, Grecia ed Italia, ma ognuno di essi, pur avendo alcune caratteristiche in comune con quello che poi sarebbe diventato lo sport più seguito in tutto il mondo, riscontrava anche profonde differenze. Ad esempio, in alcuni era consentito anche l'utilizzo delle mani, mentre in altri la dimensione delle porte era veramente irrisoria (30-40 centimetri). Date queste incongruenze con le regole del calcio praticate ai giorni nostri, possiamo affermare che la vera patria del calcio moderno fu l'Inghilterra, o meglio, i college britannici. Questo sport nacque infatti come sport d'élite in quanto il football fu inizialmente praticato dai giovani delle scuole più ricche e delle università. Le classi erano sempre composte da dieci alunni, e a questi si aggiungeva il maestro che giocava sempre insieme a loro; da qui nacque la consuetudine di giocare in undici. Il "capitano" di ogni squadra era una sorta di discendente del maestro che, in quanto tale, dirigeva la sua classe di alunni. Le diverse scuole britanniche giocavano ognuna secondo le proprie regole, spesso basilarmente diverse. Nel 1848, per ovviare a questo problema, i rappresentanti delle varie scuole e dei club inglesi si riunirono per trovare un punto d'incontro. La riunione durò otto ore e produsse un importante risultato; vennero infatti stilate le prime basilari regole del calcio, dette anche Regole di Cambridge.

Il calcio si espanse a macchia d'olio. In Inghilterra ben presto divenne lo sport per eccellenza della classe lavorativa e non solo di quella benestante, dato che uno sport divertente, semplice e stancante era l'ideale per sfogarsi dopo una settimana lavorativa. Dall'Inghilterra, il calcio moderno venne esportato prima nelle vicine Scozia (1873), Galles (1876) e Irlanda del Nord (1880) e successivamente in tutta Europa, o per opera degli emigrati di ritorno dall'Inghilterra stessa (che furono tra i primi a conoscere il football), o su iniziativa degli stessi inglesi che si trovavano all'estero. Fu proprio così, che il calcio si diffuse anche nella nostra Italia a partire dalla fine del XIX secolo.

Le prime città italiane ad avvicinarsi a questo sport furono Torino, Genova e Milano. Non a caso furono proprio le tre potenze industriali dell'epoca a fondare quello che di li a poco divenne sport nazionale. Queste città, proprio grazie agli scambi commerciali con mercanti e marinai Inglesi incominciarono a conoscere il gioco del football e così decisero di formare delle squadre che si sarebbero poi affrontate in una sorta di campionato ad eliminazione diretta sul modello della Coppa d'Inghilterra. Il Genoa fu indiscutibilmente la prima Grande squadra della nostra penisola, aggiudicandosi i primi tre tornei rispettivamente nel 1888, 1889 e nel 1890, fu poi il Milan, nel 1891, a fermare la corsa dei fenomenali genovesi.

Nel 1895 ci fu un'importantissima riforma del campionato che sostituì alle gare secche, ad eliminazione diretta, una serie di gruppi preliminari, i così detti Gironi Eliminatori Regionali, propedeutici al Girone Finale Nazionale. Vennero introdotte allo stesso modo le partite di andata e ritorno. Fu proprio in questa occasione che la Juventus riuscì a conquistare il suo primo trionfo.

Col passare degli anni anche altre squadre riuscirono ad inserirsi nel Girone che di li a poco avrebbe coinvolto squadre provenienti da tutte le regioni italiane; arrivando così alla formazione del Girone di Serie A, così come lo conosciamo noi oggi. Il numero delle squadre presenti nel girone ha subito leggere variazioni nel corso degli anni, ma dal 2004 è stato stabilito che il numero di squadre ammesse è 20. Queste, si affrontano nel girone di andata e in quello di ritorno; per ciascuna partita vengono assegnati tre punti alla squadra vincitrice e zero a quella sconfitta; mentre in caso di pareggio viene conferito un punto ad entrambe. Alla fine della stagione, la squadra prima classificata vince lo scudetto e viene premiata con la Coppa Campioni d'Italia.

Il calcio fin dagli albori è stato uno sport che ha sempre movimentato moltissimi soldi e anche in periodi di crisi non ne ha mai subito gli effetti. Si ha sempre avuto la convinzione che una squadra per essere forte doveva possedere i migliori giocatori al mondo, i così detti "fenomeni", e per ottenere questo risultato i vari club tuttora spendono cifre da capogiro. Pertanto possiamo affermare che, il calcio è sempre stato uno sport prettamente di atleti, nel quale non si trovava spazio per i così detti "secchioni" che, con i loro studi, cercavano di proporsi alle grandi squadre per analizzare i dati che potevano essere rilevati alla fine di ogni match e che sarebbero risultati molto utili per valutare i maggiori punti di forza e di debolezza di ciascun giocatore. I dirigenti e gli allenatori, tuttavia, hanno sempre snobbato questo tipo di "aiuto", pensando che tali informazioni fossero del tutto inutili, e che le uniche variabili che

incidessero realmente sul risultato delle partite fossero le reti segnate dalla propria squadra rispetto all'avversaria. Questo ragionamento, naturalmente ovvio e banale, cominciò a vacillare solamente alcuni anni fa, quando incominciarono a comparire i primi dati statistici riferiti alle partite di campionato. Apparvero timidamente sulle pagine rosa della Gazzetta dello Sport, quasi per completare il filo del discorso e poi si instaurarono prepotentemente all'interno della vita vera e propria del calcio.

Oggi qualsiasi trasmissione sportiva cita in continuazione dati statistici e alla fine degli incontri viene mostrato un prospetto riepilogativo relativo ad alcune fra le maggiori variabili rilevate durante ogni match. Questi sono solo piccoli esempi di come la statistica sia quasi diventata indispensabile e vitale per tutti. Sembra proprio che ora il calcio non possa più fare a meno della statistica e viceversa. Gli allenatori e i giocatori non appena finisco un incontro corrono dagli statistici per ottenere i dati in merito alla partita appena disputata e raffrontarli con quelli delle partite precedenti. Gli statistici, dal canto loro, raccolgono sempre più informazioni per ogni match disputato. Non sappiamo di preciso quale sia la ragione che ha spinto statistici e matematici ad avvicinarsi al mondo del calcio, forse è stata la passione sfrenata di milioni di persone ad incuriosirli e a portarli alla decisione di compiere analisi statistiche anche molto sofisticate su questi dati. Di per certo sappiamo che la statistica sta dando un fortissimo aiuto al calcio e che molto probabilmente nel corso degli anni diventerà sempre più fondamentale, anzi, oserei dire parte integrante.

Di recente è stato presentato alla Mostra del Cinema di Venezia il film *Money Ball: L'arte di vincere*, il quale sottolinea l'importanza della statistica nel mondo dello sport. Il film parla del mondo del baseball americano, e della abissale differenza di budget per l'acquisto dei giocatori tra le diverse squadre, tale differenza non lascia spazio a piccole squadre come quella dell'Oakland Athletics, protagonista del film, le quali si vedono allontanare i propri fuoriclasse, attirati dai grandi contratti delle squadre più potenti. La svolta avviene quando il coach degli Oakland Athletics, Billy Beane incontra un matematico di Yale, Peter Brand il quale, tramite formule matematiche e statistiche dimostrerà, prima ad Oakland e successivamente a tutto il mondo, che non sempre il talento di un giocatore corrisponde al suo costo. Una nuova mentalità che sicuramente, all'inizio, non viene accolta a braccia aperte; ma che successivamente dimostrerà la sua forza e credibilità trasformando gli Oakland in una squadra in grado, con pochi soldi, di affrontare squadre a cinque stelle e raggiungere risultati invidiabili. L'idea di fondo del film è che una grande squadra debba essere o divenire grande non grazie ai soldi, ma grazie al talento, cuore e testa di tutti coloro che la compongono.

Pertanto il film cerca di focalizzare l'attenzione sul rapporto sport denaro e sulle difficoltà che le squadre medio-piccole in qualsiasi sport di squadra si trovano ad affrontare per riuscire a sopravvivere contro lo strapotere dei grandi club che basano la loro strategia sullo spendere montagne di denaro in modo da riuscirne a guadagnare altrettanto. Il film punta a far riscoprire l'autentico valore dello sport fatto di competizione, sudore, risultati e record da migliorare; sottolineando allo stesso modo come le ormai avanzate tecniche statistiche possano aiutare in tale direzione.

Proprio partendo da queste provocazioni ha preso spunto l'idea di tesi che mi presto a sviluppare nelle pagine seguenti.

Nel recente periodo di inchiesta sul calcio scommesse, ognuno di noi, tifoso o meno, si è interessato a questo strano mondo del calcio che, per colpa di pochi soggetti, risulta essere sempre più malato. Alcuni di noi si sono interessati maggiormente all'ambito morale della questione, mentre altri hanno cercato di comprendere realmente il problema alla radice.

Il calcio scommesse muove ingenti somme di denaro e grazie alle ricevitorie e a numerosi siti internet, è diventato sempre più facile giocare. Le quotazioni riferite alle varie partite, tuttavia, sono calcolate attraverso calcoli statistici molto avanzati che vanno a recuperare in enormi database lo storico riferito alle squadre che da li a poco si scontreranno. Questo fatto ha sollevato un po' di curiosità in merito a che tipo di variabili vengano rilevate per ciascuna partita. Così, navigando sul web siamo giunti nel sito della Panini Digital, azienda leader in Italia nella raccolta statistica e nella fornitura dei servizi informatici a supporto delle società calcistiche e del mondo dei media.

Scorrendo il sito non si può che rimanere sorpresi dall'enormità di dati contenuti e dalla perfetta organizzazione di questi a seconda delle varie giornate di campionato. Non avrei mai pensato che per una partita venissero prese in considerazione un numero così elevato di variabili. Da qui l'idea di analizzare se esiste un qualche tipo di correlazione tra determinate variabili e il risultato finale delle partite. Obbiettivo forse leggermente ambizioso ma allettante, che ci ha portato a rivolgerci direttamente presso la Panini Digital per ottenere, per ogni partita del campionato di serie A 2010/2011, i file relativi alle variabili rilevate.

Per sviluppare questo progetto di tesi ho deciso di seguire scrupolosamente le fasi principali tipiche di ogni progetto di data mining, le quali verranno illustrate in maniera sintetica nel capitolo 1. Successivamente vi sarà una breve introduzione sulle principali tecniche statistiche

che verranno implementate nel nostro modello per l'ottenimento dei risultati e seguiranno alcuni brevissimi cenni in merito al software open source "R", utilizzato per la stesura dei codici. La vera e propria applicazione pratica verrà sviluppata nel terzo capitolo, prendendo in considerazione i dati del campionato Italiano di Serie A 2010/2011. Sulla base dei risultati ottenuti dai vari modelli statistici, verranno tratte le opportune considerazioni.

Per alleggerire la lettura sono state create numerose appendici nelle quali è possibile trovare nel dettaglio alcuni passi fondamentali per l'elaborazione dei dati e la loro comprensione.

CAPITOLO 1: IL DATA MINING E LE FASI PRINCIPALI DEI PROGETTI.

Prima di illustrare le sei fasi principali di un progetto di data mining facciamo un passo indietro e chiediamoci cosa intendiamo con il termine data mining. La recente letteratura offre diverse ed utili definizioni del concetto di data mining, alcune delle quali sono riportate di seguito.

"Data Mining is the efficient discovery of valuable, non obvious information from a large collection of data. (...) The idea is that the raw material is the business data, and the data mining algorithm is the excavator, sifting through the vast quantity of raw data looking for the valuable nuggets of business information."

Joseph P. Bigus

"Il Data Mining è una strategia di apprendimento di natura induttiva che costruisce modelli per identificare pattern nascosti nei dati. Un modello creato da un algoritmo di Data Mining è una generalizzazione concettuale dei dati. Tale generalizzazione può assumere la forma di un albero, di una rete, di un'equazione o di un insieme di regole."

R. J. Roiger, M. W. Geatz

"Processo di esplorazione e analisi, in modo totalmente o parzialmente automatizzato, di una grande quantità di dati al fine di individuare schemi e regole significativi non noti a priori."

M.J.A. Berry, G.S. Linoff

"*Il Data Mining è un processo atto a scoprire correlazioni, relazioni e tendenze nuove e significative, setacciando grandi quantità di dati immagazzinati nei repository, usando tecniche di riconoscimento delle relazioni e tecniche statistiche e matematiche.*"

Gurtner Group

"The promise of Data Mining is to find the interesting patterns lurking in all these billions and trillions of bytes. Merely finding patterns is not enough. You must respond to the patterns and act on them, ultimately turning data into information, information into action, and action into value."

M.J.A. Berry, G.S. Linoff

Il termine data mining è basato sull'analogia delle operazioni che vengono svolte dai minatori, i quali scavano all'interno delle miniere grandi quantità di materiale di poco valore per riuscire a trovare l'oro. Nel nostro ambito, l'oro è l'informazione precedentemente

sconosciuta; mentre il materiale di poco valore sono i dati e le operazioni di scavo non sono altro che le tecniche di esplorazione dei dati.

È importante sottolineare che il data mining non è legato a tecniche specifiche, anzi molto spesso, i migliori risultati si ottengono combinando una serie di tecniche distinte.

La cosa più importante da tenere in considerazione quando si parla di questo argomento è che il data mining non prende decisioni, ma fornisce ai decision-maker (coloro che sono chiamati a prendere decisioni sull'argomento studiato attraverso il data mining) le informazioni necessarie a fronteggiare le difficoltà dei mercati competitivi, o più in generale l'incertezza della vita reale. Pertanto i veri fattori critici di successo di un progetto di data mining sono la conoscenza dell'argomento posto in analisi e l'esperienza maturata nel tempo del soggetto che si appresta a prendere le decisioni. Queste conoscenze, insieme alle utili informazioni tratte dalle analisi di data mining creano un forte processo sinergico che porta alla presa di decisioni brillanti e veloci.

Pertanto possiamo dire che i caratteri distintivi del data mining sono:

- analizzare grandi quantità di dati all'interno delle quali il Data Miner troverà qualcosa di interessante; è cruciale per la buona riuscita del modello controllare che i risultati ottenuti non siano errati;
- la nuova informazione estrapolata dai dati deve portare un vantaggio al business;
- l'obiettivo per il Data Miner è trovare qualcosa di non intuitivo, infatti più l'informazione si discosta dall'ovvio e più è grande il suo valore potenziale.

Per ottenere questo valore, il Data Miner, deve seguire un certo numero di fasi che vanno dalla definizione degli obiettivi alla valutazione dei risultati.

In letteratura esistono numerose varianti per la schematizzazione del processo di data mining, ma la cosa più importante da sottolineare è che benché il numero di fasi possa variare notevolmente, i concetti fondamentali differiscono di poco.

Ora verrà presentato il modello noto come Cross Industry Standard Process for Data Mining (CRISP-DM). Questo modello è, in realtà, un progetto finanziato dalla Commissione Europea, volto a definire un approccio standard ai progetti di data mining indipendentemente dalla tipologia di business. Il CRISP-DM divide il ciclo di vita di un progetto di data mining in sei fasi principali, tuttavia la sequenza di queste non è rigorosa; infatti spesso è necessario andare avanti ed indietro tra le diverse fasi.

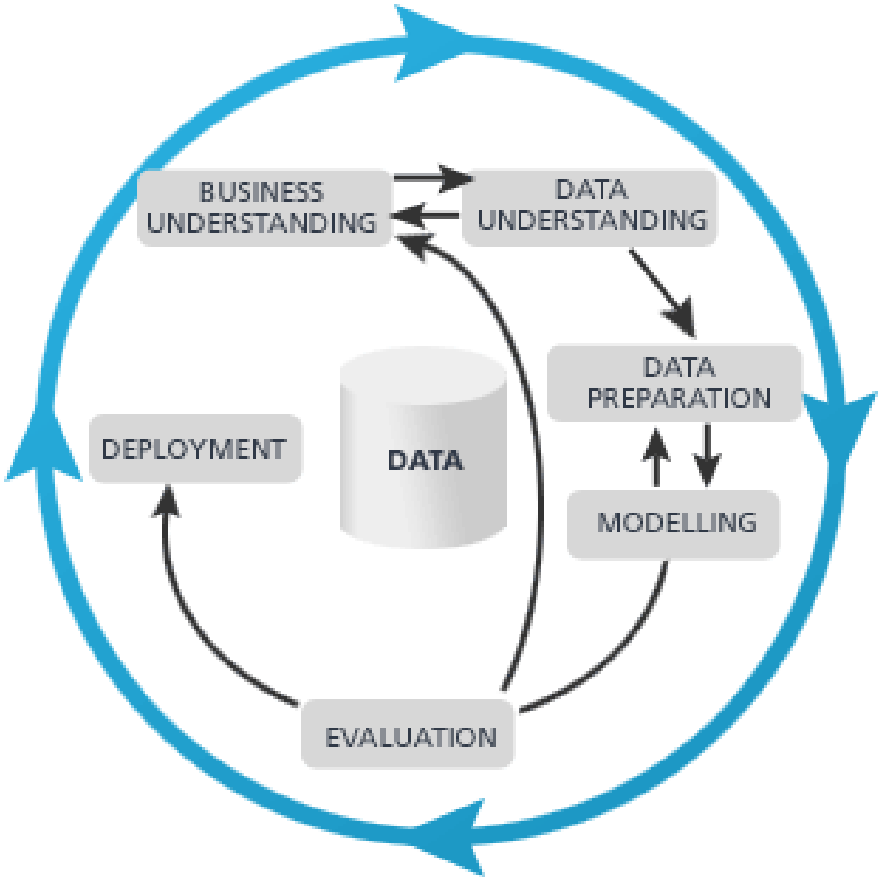


Figura 1: Cross Industry Standard Process for Data Mining CRISP-DM.

Le frecce nel diagramma indicano le interdipendenze più importanti e frequenti che possono avvenire tra le fasi, mentre il cerchio esterno al diagramma simboleggia la natura iterativa del processo di data mining; infatti, quasi sempre, il processo continua anche dopo che una soluzione è stata distribuita, così che i processi futuri possano beneficiare di quelli precedenti e trovare soluzioni sempre migliori.

Fase 1 – Comprensione del business.

L'oggetto della prima fase è l'individuazione degli obiettivi e la definizione chiara di ciò che deve essere portato a termine. Per fare ciò è importante capire, quali dati saranno necessari per riuscire a compire l'analisi e dove reperirli.

Sempre all'interno di questa fase, bisognerebbe stimare il costo del progetto e il ricavo che ci si aspetta di ottenere dopo aver compiuto quest'analisi. In altre parole, si dovrebbe valutare quanto vale in termini monetari questa informazione, tenuto conto che affinché un progetto sia efficiente i ricavi devono sempre superare i costi. Questo aspetto, tuttavia, non verrà trattato in questo elaborato in quanto questa analisi viene effettuata solo a scopo didattico.

Fase 2 – Comprensione dei dati.

La seconda fase consiste nella raccolta dei dati da utilizzare nel processo di analisi, nella comprensione delle variabili a disposizione e nella creazione di nuove variabili che potrebbero risultare utili per raggiungere i risultati desiderati.

I dati possono essere di due tipi:

- *Qualitativi*: assumono valori discreti non ordinabili; servono per identificare una categoria e distinguerla dalle altre. A questo tipo di variabile, appartiene una fattispecie particolare, le *variabili dicotomiche o dummy*, che assumono solo due modalità, solitamente 0 e 1, le quali indicano rispettivamente l'assenza o la presenza dell'elemento indicato con la variabile.
- *Quantitativi*: quando vengono attribuiti valori discreti enumerabili oppure valori continui. Questi valori assumono un pieno significato numerico e pertanto sono ordinabili.

Fase 3 – Preparazione dei dati.

Prima di effettuare le analisi statistiche e quindi l'estrazione delle informazioni utili, occorre procedere ad un accurato controllo dei dati a nostra disposizione, in modo da verificare se essi presentino le caratteristiche necessarie a renderli idonei alle successive elaborazioni. Queste operazioni, chiamate pre-processing, si collocano a monte delle analisi statistiche, ciò significa che è necessario attuarle prima di creare il modello vero e proprio e importarvi i dati all'interno. Sintetizzando possiamo affermare che la pulizia dei dati (data cleaning) è un processo in grado di garantire, con una certa soglia di affidabilità, la correttezza di un insieme di dati.

Attraverso questa fase, quindi, vogliamo assicurare la qualità dei dati scelti, e per farlo è necessario effettuare un trattamento dei dati anomali e mancanti.

I dati anomali, sono quelli che differiscono sensibilmente dai restanti. Per definizione, infatti, sono osservazioni che essendo atipiche o erronee, si discostano decisamente dal comportamento degli altri dati, con riferimento al tipo di analisi considerata. Questa definizione è molto importante in quanto pone l'accento sul tipo di analisi che viene effettuata, infatti molto spesso accade che i valori non risultano anomali se le variabili vengono esaminate singolarmente, mentre lo diventano quando vengono considerate congiuntamente. La loro presenza può essere dovuta ad errori umani tipicamente nelle

operazioni di trascrizione, ad errori sistematici nella raccolta dei dati, oppure ad errori dovuti semplicemente al caso, che ha fatto si che nella raccolta dei dati alcune osservazioni abbiano prodotto dati molto lontani dalla media del campione (outlier), tuttavia, la loro identificazione può avvenire facilmente stabilendo un valore soglia.

In questo contesto è importante localizzare record doppi o meglio chiamati dati ridondanti che possono essere individuati facilmente attraverso l'analisi di correlazione. Queste correlazioni, infatti, potrebbero compromettere i risultati del modello in modo imprevedibile. Per effettuare questa operazione, siccome generalmente le basi dati sono molto grandi, vengono utilizzati calcoli statistici come media e varianza, in modo da riuscire ad ottenere le informazioni utili per la valutazione delle variabili stesse.

Infine, i dati mancanti o missing values rappresentano, informazioni perse e possono dipendere da:

- malfunzionamenti nei sistemi di raccolta dati,
- inconsistenza con valori di altri attributi del set di dati (ad esempio quando lo stesso campo ha valori differenti in diverse tabelle, molto spesso questo è causato da aggiornamenti effettuati in maniera non corretta),
- i dati non sono stati inseriti a causa di incomprensioni,
- alcuni dati possono non essere considerati importanti al momento dell'inserimento,
- mancata registrazione dei cambiamenti nei dati.

Si può procedere in diversi modi, come ad esempio l'eliminazione dei record contenenti dati mancanti, o la sostituzione del dato mancante con la media della classe o con valori rilevati per osservazioni simili. È importante tenere in considerazione due aspetti prima di effettuare qualsiasi operazione su questi dati:

- l'utilizzo di queste congetture può incidere notevolmente sui risultati, ed
- è necessario valutare attentamente le tecniche che si intende utilizzare, in quanto alcune di queste sono in grado di riconoscere e sviluppare i dati mancanti, mentre altre richiedono che tutti i valori siano presenti.

Una volta fatto tutto questo è necessario passare alla trasformazione dei dati, infatti, a seconda della tecnica utilizzata, le variabili potrebbero dover essere revisionate per rispondere a determinate caratteristiche del modello.

Con il termine "trasformazione di una variabile" intendiamo la derivazione di nuove variabili attraverso l'applicazione di funzioni a quelle originarie.

$$Y' = f(Y)$$

Un metodo di trasformazione molto comune è la normalizzazione, la quale è in grado di modificare i valori in modo che cadano tutti all'interno di un determinato range, o meglio, questo tipo di trasformazione fa si che le grandezze siano "riscalate" secondo range definiti. Questa operazione permette di mettere a confronto distribuzioni diverse.

I metodi di normalizzazione più importanti sono:

- *Normalizzazione min-max*: questa trasformazione consente di ottenere valori compresi nell'intervallo [0, 1].

  $$Y' = \frac{(Y - \min_A)}{(max_A - min_A)}$$

  Dove *min* rappresenta il valore minimo assunto originariamente dalla variabile e *max* il suo valore massimo.
- *Normalizzazione z-score*: è il metodo di normalizzazione più utilizzato ed effettua la conversione del valore attraverso l'utilizzo della media e della deviazione standard.

  $$Y' = \frac{Y - E[Y]}{\sqrt{Var[Y]}}$$
- *Normalizzazione tramite decimal scaling*: consiste nella divisione di ogni valore di uno stesso range per la base 10 del range stesso.

  $$Y' = \frac{Y}{10^j}$$

  Si sposta il punto decimale di Y di j posizioni, in modo che j è il numero minimo di posizioni spostate che fa si che il massimo valore assoluto cada tra [0 … 1].
- *Normalizzazione logaritmica*: viene fatta prevalentemente per riscalare i range senza perdere informazioni e viene effettuata sostituendo ai valori il rispettivo logaritmo.

  $$Y' = \ln(Y)$$

Un altro aspetto molto importante da valutare in questa fase è la dimensione dei dati a nostra disposizione, molto spesso infatti i database possono contenere terabytes di dati e quindi le tecniche di data mining complesse potrebbero richiedere lunghi tempi di elaborazione. Per ovviare a questo problema, si effettua la riduzione dei dati, che ha l'obiettivo di ridurre la rappresentazione del set di dati iniziale in un set di dati di minore dimensione tale però da

produrre gli stessi risultati (o quasi). Tuttavia, se si ha a disposizione un database grande, ma non eccessivo, consiglio sempre di provare a lanciare il modello con tutti i dati a nostra disposizione (naturalmente puliti dai problemi trattati nelle righe precedenti), in quanto negli ultimi anni i software si sono sviluppati notevolmente e permettono di effettuare moltissime operazioni in tempi estremamente ridotti. Se poi in fase di esecuzione ci si rende conto che i tempi sono insostenibili, si può sempre interrompere il processo di esecuzione e ridimensionare i dati. Buona norma, sarebbe inoltre, non avere un numero di variabili di molto superiore al numero di osservazioni; cosa che invece, accade abbastanza spesso nei problemi di data mining caratterizzati frequentemente da database a rettangolo.

Fase 4 – Modellazione.

In questa fase l'obiettivo principale è l'applicazione di una o più tecniche di data mining, attraverso le quali è possibile costruire dei modelli in grado di fornire informazioni efficaci rispetto agli obiettivi della ricerca. Alcune tecniche per essere applicate richiedono che i dati soddisfino determinate caratteristiche, per questo, molto spesso, risulta necessario fare ritorno alla fase di preparazione dei dati per adattare il database iniziale alle nuove esigenze.

Perché questa fase abbia successo è necessario conoscere in maniera approfondita le diverse tecniche statistiche e i diversi ambiti nei quali applicarle; inoltre per ogni algoritmo è importante conoscere il tipo di variabili da inserire in input e il tipo di variabili che ci si aspetta di ottenere come output.

Grazie al repentino sviluppo delle nuove tecnologie, oggi, è possibile sviluppare modelli contenenti svariate tipologie di variabili.

Fase 5 – Valutazione.

La fase di valutazione è necessaria per mettere a confronto i risultati di alcuni modelli presi a confronto. Solitamente questi confronti vengono effettuati su problematiche di tipo supervisionato per le quali, quindi, si conoscono i risultati. Se nessuna tecnica mostrasse performance soddisfacenti, potrebbe risultare necessario tornare alle fasi precedenti e rivedere alcuni passaggi.

Per ottenere buoni risultati è indispensabile avere dati di qualità, infatti è impensabile riuscire a trarre conoscenza o per lo meno riuscire ad ottenere risultati soddisfacenti se nel database a nostra disposizione non sono contenute le informazioni che stiamo cercando; infatti per quanto un modello sia corretto ed affinato, non potrà mai sopperire alla scarsa correttezza

(distorsione) delle informazioni fornite in input al modello. L'ambiente dati a monte deve essere quanto più possibile robusto ed affidabile.

Fase 6 – Implementazione.

Una volta che è stato selezionato un modello e ne è stata verificata la correttezza, i risultati ottenuti possono essere integrati nei processi decisionali.

È importante definire le aree applicative nelle quali la conoscenza prodotta, attraverso il modello, può apportare benefici effettivi.

CAPITOLO 2: METODOLOGIE E STRUMENTI DI ANALISI.

Ora analizzeremo le principali tecniche statistiche che verranno utilizzate per affrontare la risoluzione del nostro progetto e successivamente verranno inseriti alcuni cenni in merito al software open source che verrà utilizzato per implementare i codici necessari all'ottenimento dei risultati, sui quali poi soffermarsi per trarre le dovute conclusioni.

2.1 GLI ALBERI DECISIONALI.

Gli alberi decisionali sono una particolare tecnica statistica utilizzata per risolvere problemi di classificazione e di regressione. L'obiettivo di tale metodologia è di ottenere una segmentazione gerarchica di un insieme di unità mediante l'individuazione di regole. Lo scopo principale, infatti, è proprio quello di individuare la regola di decisione ottimale che, dato un insieme di variabili rilevate, consente di prevedere al meglio la classe di appartenenza delle singole unità. Per raggiungere questo obiettivo le classi devono essere note a priori.

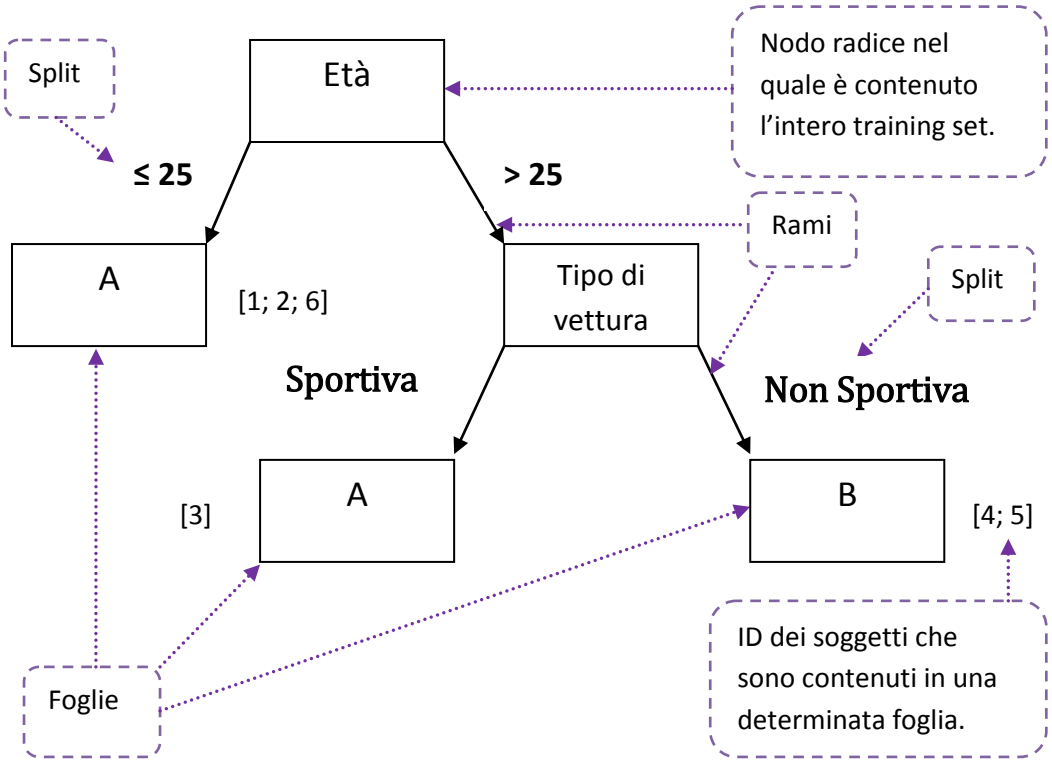
L'output grafico ottenuto da questa procedura è una struttura gerarchica ad albero contenente nodi, rami e foglie. I nodi rappresentano i gruppi di unità nei diversi stadi del processo di segmentazione, i rami determinano le condizioni che conducono alle suddivisioni e le foglie sono i nodi terminali per i quali non è ritenuta utile nessuna suddivisione e che contengono le principali informazioni comunicate dal modello. Un nodo viene chiamato genitore rispetto ai nodi che esso genera, mentre viene denominato figlio rispetto al nodo da cui discende. Ogni foglia mostra una chiara legge di classificazione delle osservazioni, che può essere letta seguendo il percorso che collega il nodo iniziale ad ognuna di esse.

Introduciamo ora un esempio per capire in maniera più approfondita il funzionamento di questa particolare tecnica statistica.

Si supponga che una compagnia di assicurazioni voglia identificare il legame che associa le classi di rischio, in cui vengono suddivisi i clienti, con l'età anagrafica e il tipo di vettura posseduta. Lo studio viene effettuato su un gruppo di clienti per i quali si conosce la corrispettiva classe di appartenenza. Di seguito verrà rappresentato il training set.

ID	Età	Tipo di auto	Rischio
1	23	Berlina	A
2	18	Sportiva	A
3	43	Sportiva	A
4	68	Berlina	B
5	32	Furgone	B
6	20	Berlina	A

Un possibile albero di decisione potrebbe essere il seguente:



Da questo semplice grafico possiamo dedurre che tutti i ragazzi con meno di 25 anni appartengono alla classe di rischio "A", indipendentemente dal tipo di macchina posseduta. Per i ragazzi con più di 25 anni invece, la situazione è diversa, i possessori di una vettura non sportiva con più di 25 anni verranno assegnati alla classe di rischio "A", mentre i possessori di vettura sportiva con più di 25 anni verranno assegnati alla classe di rischio "B".

Il nostro training set di 6 soggetti è stato opportunamente partizionato nei tre nodi foglia che si sono venuti a creare. Accanto ad ognuno di essi sono indicati, tra parentesi quadre, i codici identificativi dei soggetti che sono contenuti al loro interno.

2.1.1 PRINCIPALI CARATTERISTICHE DEGLI ALBERI DECISIONALI.

Gli alberi decisionali vengono suddivisi in due tipologie, gli "alberi di classificazione", in cui la variabile dipendente è di tipo categoriale e gli "alberi di regressione", nei quali la variabile dipendente è di tipo quantitativo.

Entrambe queste tipologie, appartengono alla classe delle tecniche di classificazione supervisionata, in quanto la segmentazione può trarre vantaggio dall'informazione supplementare sul gruppo di appartenenza che risulta noto per un numero ristretto di unità. L'insieme di unità su cui è determinato l'albero viene chiamato training set o training sample.

Questa tecnica statistica non pone tutte le variabili disponibili sul medesimo piano logico, ma attribuisce ad una variabile il ruolo di variabile dipendente mentre le altre vengono considerate come variabili esplicative.

Le procedure ad albero sono flessibili e non richiedono il soddisfacimento di stringenti ipotesi, adattandosi molto bene all'analisi di data set di grandi dimensioni. Vengono spesso utilizzati anche per la loro capacità di ordinare le variabili indipendenti in base all'ordine di priorità, cioè per scegliere le variabili più importanti in grado di spiegare un certo problema ed eventualmente eliminare quelle meno significative in modo da semplificare l'analisi.

La segmentazione gerarchica ottenuta mediante un albero decisionale può essere definita come una procedura "per passi" (stepwise), attraverso la quale l'insieme delle unità statistiche è suddiviso progressivamente, secondo un criterio di ottimizzazione, in una serie di sottogruppi disgiunti e che presentano al loro interno un grado di omogeneità maggiore rispetto all'insieme iniziale. La segmentazione fornisce pertanto una successione gerarchica di partizioni (partizionamento ricorsivo) attraverso la quale ad ogni passo del processo, l'eterogeneità nei gruppi si riduce rispetto al passo precedente; le foglie dell'albero presentano un grado di omogeneità tale da poterle attribuire ad una delle classi di partenza.

Per la costruzione dell'albero è necessario selezionare le variabili di splitting, individuando quelle che spiegano maggiormente le osservazioni. Ad ogni nodo, quindi, l'algoritmo deve cercare l'attributo che produce la suddivisione migliore rispetto alle osservazioni disponibili sul nodo corrente; bisogna, tuttavia, tenere in considerazione il fatto che il criterio utilizzato per la ricerca di tale attributo non sempre è in grado di trovare la soluzione ottimale in senso assoluto, ma spesso individua la migliore solo localmente in quanto non tiene conto delle suddivisioni future. Dopo il processo di generazione dell'albero è prevista la fase di potatura (pruning) dello stesso. Questa fase è indispensabile in quanto un modello molto complesso

potrebbe fornire buoni risultati sui dati di training, ma non consentire l'ottenimento di un output soddisfacente a fronte di nuovi input. In questo caso l'albero sarebbe in grado di descrivere alla perfezione i dati su cui è stato costruito, ma difficilmente potrebbe essere generalizzato per un nuovo campione di input, questa situazione si verifica quando il modello impara "a memoria" le caratteristiche dei dati sui quali è stato generato e non apprende i concetti generali alla base di questi. Si cade pertanto in un problema di overfitting (sovra adattamento). Un'efficace modo per evitarlo consiste nell'utilizzare le tecniche di pruning che prevedono la manipolazione dell'albero al fine di migliorarne la capacità di generalizzazione.

Gli alberi decisionali vengono costruiti "apprendendo dall'esperienza", tuttavia, il ruolo dell'esperto non risulta irrilevante e neppure minimale, in quanto, l'utente deve interpretare i risultati per riuscire ad ottenere il corretto output attraverso successivi raffinamenti del modello.

2.1.2 ALGORITMI PER LA COSTRUZIONE E LA POTATURA DI ALBERI DECISIONALI: I CART.

Esistono numerosi algoritmi per la costruzione di alberi di decisione, che si differenziano in particolare per il numero di ramificazioni ammesse ad ogni livello. La metodologia CART (Classification e Regression Trees) è riconosciuta come uno degli algoritmi più diffusi nei processi di data mining. Questa tecnica, proposta da Breiman nel 1984, consente solo partizioni binarie ed ha rappresentato una svolta rispetto alle tecniche di segmentazione note in precedenza. Tra le caratteristiche più salienti vi è il fatto che:

- la variabile dipendente può essere sia qualitativa che quantitativa;
- è possibile considerare congiuntamente variabili esplicative qualitative e quantitative;
- si possono eseguire split considerando come predittori combinazioni lineari di variabili quantitative;
- gli split vengono definiti in base al criterio di riduzione dell'impurità;
- il dimensionamento ottimale degli alberi viene proposto attraverso la procedura di potatura (pruning).

Il processo di crescita dell'albero inizia con la scelta del campo che meglio di tutti separa i record in gruppi e che offre quindi il massimo guadagno d'informazione. Gli split, come detto in precedenza, vengono definiti in base all'indice di impurità calcolato sulla variabile dipendente, il quale fornisce la misura di quanto i records sono equamente divisi tra le classi; l'eterogeneità (impurità) è nulla se le n unità statistiche presentano tutte la medesima modalità, mentre è massima se le unità statistiche sono uniformemente ripartite fra le r

modalità, cosicché ciascuna modalità presenta la medesima frequenza relativa *1/r*. In altre parole, questo indice viene interpretato come la probabilità che, due elementi qualsiasi di una popolazione, scelti casualmente, appartengano a due classi differenti.

Per valutare l'eterogeneità si può fare riferimento a diversi criteri. Tipicamente si utilizza l'indice di Gini o l'entropia si Shannon se la variabile dipendente (Y) è categoriale, la varianza o lo scarto quadratico medio se la Y è quantitativa. Riportiamo di seguito la formula per il calcolo dell'impurità attraverso l'indice Gini:

$$G = 1 - \sum_{i=1}^{r} f_i^2$$

dove f_i è la frequenza relativa della modalità i-esima d'un fenomeno qualitativo che può assumere r modalità. L'indice G assume valore minimo (pari a 0) nel caso di massima omogeneità (cioè di eterogeneità nulla) e valore massimo $\frac{(r-1)}{r}$ nel caso di massima eterogeneità (Sergio Zani e Andrea Cerioli, 2007).

Come abbiamo anticipato in precedenza, uno dei tratti più originali della metodologia CART consiste nella proposta d'un metodo per la validazione dell'albero.

Tale criterio rappresenta una regola di blocco nella procedura di costruzione dell'albero di classificazione. Un criterio di arresto ragionevole, può essere quello di fissare una soglia minima β per il decremento di impurità dell'albero, passando da uno stadio a quello successivo, al di sotto della quale la procedura si ferma. Tuttavia, la scelta soggettiva della soglia influenza pesantemente i risultati; infatti se β è troppo piccolo è probabile ottenere un albero finale profondo (con molte foglie) con conseguenti difficoltà interpretative; mentre se β è troppo elevato un nodo *t* può essere dichiarato terminale escludendo la possibilità che i suoi nodi discendenti ammettano un decremento di impurità $\geq \beta$. Una considerazione ulteriore riguarda la stima di $R(T)$, dove T è un generico albero di classificazione. La stima per ricostituzione $\hat{R}(T)$ è inversamente proporzionale al numero di foglie dell'albero. Pertanto, l'accuratezza di $\hat{R}(T)$ decresce al crescere delle dimensioni dello stesso e la scelta in merito alla dimensione, basata esclusivamente su tale stima, porta alla selezione di classificatori con numerosi *split*. La stima basata sul campione test $\hat{R}_{ts}(T)$ ha, invece, prima un andamento decrescente e poi crescente all'aumentare del numero di foglie, oltre una certa soglia.

Le considerazioni appena effettuate in relazione alla scelta del criterio di arresto e alla stima del tasso di errata classificazione hanno portato all'introduzione di una metodologia di validazione degli alberi detta pruning (potatura) le cui fasi possono essere così riassunte:

- creazione dell'albero massimo T_{max} che si ottiene fissando $\beta = 0$, per cui le foglie sono costituite da casi appartenenti alla stessa classe o al limite da un solo caso;
- selezione dei sottoalberi che si possono ottenere tagliando T_{max} in determinati punti e stima dell'errore di previsione dei diversi sottoalberi mediante uno stimatore appropriato di $R(T)$, dove T è un generico albero di classificazione. Tale fase, costituisce il nucleo del pruning, poiché l'albero viene potato eliminando alcuni rami secondari;
- scelta del sottoalbero che fornisce la migliore stima di $R(T)$.

Per ovviare al problema che il numero di possibili sottoalberi possa essere molto elevato anche quando l'albero massimo ha un numero limitato di foglie si utilizza una procedura di pruning selettivo, la quale consente di individuare una sequenza di sottoalberi di dimensione decrescente. Ogni sottoalbero appartenente alla sequenza ottimale è il migliore rispetto ai sottoalberi appartenenti alla stessa classe, o meglio, rispetto ai sottoalberi aventi il medesimo numero di foglie. Allo scopo di individuare la sequenza ottimale si definisce, per ogni albero $T \leq T_{max}$, una misura $R_\alpha(T)$ detta funzione di costo-complessità, la quale mira alla riduzione dell'errore di previsione tenendo conto di una penalità collegata alla dimensione dell'albero:

$$R_\alpha(T) = \hat{R}(T) + \alpha\left|\tilde{T}\right|$$

dove $\left|\tilde{T}\right|$ è il numero di foglie dell'albero T, $\hat{R}(T)$ è la stima per ricostituzione del tasso di errata classificazione e α è un numero reale non negativo detto parametro di complessità.

Questo parametro α può essere visto come una penalità connessa agli alberi di grandi dimensioni. Nel caso in cui fossero a confronto due alberi, per i quali il valore di $\hat{R}(T)$ è il medesimo, si dovrebbe selezionare quello che contiene il minor numero di foglie. Il vero valore che penalizza gli alberi con un elevato numero di foglie è, quindi, α; per cui al crescere di questo parametro si avrà un numero sempre minore di foglie contenute nel sottoalbero ottimale (Sergio Zani e Andrea Cerioli, 2007).

2.1.3 METODOLOGIE PER IL CONTROLLO DELL'OVERFITTING.

Attraverso questi cenni di teoria abbiamo capito che la scelta finale è un compromesso tra accuratezza e previsione, in quanto, producendo alberi "impuri" si cerca di dare spazio alla capacità stessa dell'albero di generalizzare, aumentando così, da un lato, il rischio di commet-

tere errori di previsione, ma riducendo allo stesso tempo il rischio di overfitting e quindi di sovra adattamento del modello ai dati di cui dispone. È provato, infatti, che il tasso di errore calcolato su un insieme di dati diverso dal set di addestramento, inizialmente decresce all'aumentare del numero di nodi, ma superata una certa soglia, quantificata sempre in numero di nodi, il tasso di errore registra un'impennata. Questo, è provocato dal fatto che il modello sta andando ad adattarsi all'osservazione dei dati piuttosto che al modello generatore degli stessi.

Vi sono diversi metodi per controllare l'overfitting, tra cui i più importanti sono il metodo del validation set, o più in generale la cross-validation. Con il metodo validation set il campione viene diviso in due parti, il training set con cui si opera l'addestramento ed il validation set su cui si controlla l'errore. In genere all'aumentare del numero di nodi di un albero l'errore che si commette sul validation set decresce inizialmente, per poi inverte la tendenza a causa del manifestarsi dell'overfitting. In questo modo si può intuire che il numero ottimale di nodi si ottiene in corrispondenza del minimo. Il metodo appena menzionato, tuttavia, può risentire della particolare scelta effettuata nella definizione del validation set. Infatti potrebbe risultare che, suddividendo il campione in due parti differenti, si ottengano risultati diversi. Per ovviare a questo problema si può utilizzare il metodo della cross-validation con la quale il campione è suddiviso in K sottogruppi e l'albero viene addestrato K volte, ogni volta escludendo dal training set uno dei K sottogruppi, ed utilizzandolo per calcolare l'errore. L'errore finale viene ottenuto dalla media dei K errori parziali.

Tramite questi due metodi, il modello viene valutato sulla base dell'errore di generalizzazione su un "nuovo" campione di dati, diverso dai dati utilizzati per l'addestramento, rendendo tale validazione maggiormente significativa.

2.1.4 PRO E CONTRO IN MERITO ALLA METODOLOGIA DEGLI ALBERI DECISIONALI.

Le tecniche di segmentazione basate sugli alberi decisionali vengono spesso utilizzate perché sono capaci di generare regole chiare e comprensibili, fornendo al contempo un'immagine sintetica e intuitiva delle relazioni tra le variabili esplicative. Tuttavia possono presentare alcune limitazioni; tra le quali quella maggiormente problematica è rappresentata dalla struttura intrinsecamente instabile dell'albero e pertanto piccole variazioni nella matrice dei dati possono condurre a grandi differenze nelle regole di classificazione e nella corrispondente segmentazione delle unità.

Per ovviare a questo problema e quindi limitare l'instabilità di un albero decisionale, è buona norma rinunciare in partenza all'idea di farlo crescere fino al livello massimo possibile; questo obiettivo è solitamente raggiunto fissando un limite minimo per il numero di unità assegnate a ciascun gruppo e/o potando in modo opportuno l'albero di dimensione massima (come è stato spiegato in precedenza).

Spesso, inoltre, il modello viene criticato in quanto non tiene conto dell'influenza che la scelta di un campo divisore potrebbe avere in futuro, dato che, tale scelta non viene mai riconsiderata in seguito; infatti tutti i campi divisori risultano dipendenti dai precedenti. A seguire un altro problema è insito nella procedura di potatura che, come abbiamo visto in precedenza, risente di un certo grado di soggettività e non garantisce completamente dalla naturale tendenza degli alberi a cadere in situazioni di overfitting.

Un punto a favore degli alberi decisionali è rappresentato dal fatto che questi non sono sensibili a differenze di scala, a valori fuori scala o a distribuzioni asimmetriche ed offrono buone risposte anche a fronte di dati rumorosi.

Concludendo possiamo affermare che il modello degli alberi decisionali rappresenta una buona scelta quando l'obiettivo è classificare record o scegliere le variabili più importanti per prevedere un determinato risultato, in quanto racchiude notevoli informazioni utili per la classificazione.

2.2 LEARNING ENSEMBLES.

Negli ultimi anni la ricerca ha appreso che la realtà è un complesso e imperscrutabile insieme di meccanismi, impossibili da descrivere con i classici modelli semplificatori tradizionalmente utilizzati nelle analisi statistiche. Può essere necessario utilizzare una nuova classe di modelli in grado di gestire la presenza di un gran numero di variabili interagenti secondo meccanismi complicati, i cosiddetti modelli algoritmici. Alcuni esempi di questi modelli possono essere le reti neurali o i learning ensembles che si basano sull'aggregazione di un gran numero di funzioni di previsioni distinte (ensemble members) e sulla loro combinazione lineare per ottenere una previsione unica. I learning ensembles possono essere costruiti impiegando diversi tipi di funzioni, tuttavia quella maggiormente utilizzata è quella che si basa sull'utilizzo di alberi decisionali CART (Classification And Regression Trees). I CART come detto precedentemente sono uno strumento utile sia per la classificazione che per la regressione. Utilizzano le variabili esogene per partizionare ripetutamente il campione in due sottocampioni disgiunti, all'interno dei quali la variabile Y (da prevedere) presenti minore

eterogeneità. Come già sottolineato nel precedente capitolo il vantaggio principale riferibile alla metodologia dei CART è di essere in grado di modellare anche relazioni di tipo non lineare tra la variabile Y e i predittori X, purtroppo, però, presentano anche due grandi limiti. Il primo è legato all'instabilità, il che comporta che a causa di leggere perturbazioni nel database si potrebbero ottenere alberi completamente diversi, spesso vicini in termini di errore, ma molto distanti in termini di significato. Il secondo riguarda la procedura di potatura, la quale risente di un certo grado di soggettività che non assicura l'assenza di situazioni di overfitting. Tuttavia, è stato dimostrato che, l'aggregazione di molteplici CART in una learning ensemble consente di superare entrambi questi problemi, traendo al contempo vantaggio dai principali punti di forza di questa metodologia, i quali derivano principalmente dal meccanismo di bipartizione del campione.

2.2.1 RANDOM FOREST.

Le Random Forest (RF), introdotte da Breiman nel 2001, hanno lo scopo di individuare le variabili che maggiormente influiscono sulla determinazione della variabile Y e allo stesso modo consentono di studiare il legame esistente tra una variabile dipendente e più predittori. Vengono costruite attraverso l'aggregazione di uno svariato numero di CART diversi, ottenuti attraverso l'introduzione di una componente di casualità nel meccanismo della loro costruzione. Infatti, ad ogni nodo di ogni albero, viene selezionato casualmente (random features) un piccolo sottoinsieme di predittori (X) e solo tra essi avviene la "competizione" in modo da far emergere quello che tra questi assicura la migliore bipartizione. Una differenza importante rispetto agli alberi di decisione è che, attraverso questo procedimento, gli alberi vengono sviluppati fino alla loro dimensione massima senza essere potati.

Molto spesso, a questa procedura, viene affiancata la tecnica bagging, introdotta da Breiman nel 1996, al fine di superare la necessità di suddividere il campione in training set e test set. Si è notato che l'utilizzo congiunto di questa tecnica, con la sopra citata random features, fa sì che aumenti la precisione del modello e al contempo permette di stimare l'errore di generalizzazione, che verrà trattato in maniera più approfondita di seguito. Questa tecnica prevede che per la creazione di ogni albero venga utilizzato un diverso sottoinsieme di dati, provenienti dal dataset originale, selezionato casualmente. Dal momento che ogni albero è ottenuto utilizzando solo n_1 elementi provenienti dal campione originale di numerosità n, i rimanenti $n_2 = n - n_1$ elementi non danno nessun contributo alla sua costruzione e questo permette il cosiddetto *out-of-bag estimation* (OOB) dell'errore di generalizzazione. Questa tecnica, come si può facilmente intuire, viene applicata per valutare la capacità di

generalizzazione del modello, la quale rappresenta la capacità di fornire previsioni accurate su di un insieme di dati che non è stato utilizzato per l'addestramento. La stima dell'errore di generalizzazione out-of-bag viene ottenuta calcolando la previsione della Random Forest per ogni soggetto del campione, mediando le previsioni dei soli alberi che sono stati ottenuti con un dataset che non conteneva il soggetto in questione. Il confronto tra i valori previsti, calcolati in questo modo, e i valori osservati fornisce una misura della capacità di generalizzazione delle RF. L'errore di previsione out-of-bag nel caso di Y quantitativa può essere calcolato come stima dello Sum Squared Error (SSE) che viene ottenuto come segue:

$$SSE = \sum_{i=1}^{N} (y_i - \widehat{y_i})^2$$

Nel caso di Y qualitativa, invece, vengono utilizzate misure come il tasso di errata classificazione.

Riassumendo, i passi per la costruzione di un insieme k di classificatori casuali, attraverso la metodologia CART, sono i seguenti (Sandri e Zuccolotto, 2004):

1) dato un campione casuale $S = \{(y_1, X_1); (y_2, X_2); \dots; (y_n, X_n)\}$, k sottocampioni di n_1 elementi $(n_1 < n)$, $S_1, S_2, \dots, S_k$, sono scelti casualmente e vanno a sostituirsi a S.
2) Ogni sottocampione S_i viene usato per far crescere un albero di classificazione con la metodologia CART, con la differenza che ad ogni nodo la divisione (split) è determinata considerando solo il piccolo gruppo di variabili esogene, selezionate casualmente da X.
3) Gli alberi sono sviluppati fino alla loro dimensione massima e non potati.

La previsione finale ottenuta con le RF viene calcolata attraverso la media delle previsioni dei singoli alberi che compongono la foresta casuale. La filosofia alla base di tale procedura è di cercare di perturbare al massimo la costruzione dei CART in modo da consentire la massima espressione della loro instabilità, la quale dovrebbe poi essere neutralizzata dalla combinazione di tutte le previsioni nella loro media. Un complesso apparato probabilistico sottostante a questa semplice idea di base ha consentito di dimostrare alcuni importanti teoremi che assicurano il funzionamento di tale procedura, che risulta inoltre esente da problemi di overfitting (Breiman, 2001).

Lo svantaggio principale connesso all'utilizzo dei modelli di tipo algoritmici è legato al loro scarso potere interpretativo della natura dei fenomeni. Sono delle black box con le quali si

possono ottenere anche accurate previsioni, ma dalle quali non è possibile comprendere il meccanismo generatore delle stesse, e più nello specifico, risulta impossibile distinguere il contributo dei singoli predittori alla definizione del risultato. Per le Random Forest, questo problema è ancora più cruciale, in quanto il metodo funziona molto bene specialmente in presenza di un piccolo numero di predittori informativi nascosti tra un grande numero di variabili "rumore". Per ovviare, in parte, a questo problema, sono state introdotte misure di Variable Importance le quali permettono di estrapolare da queste black box alcune informazioni aggiuntive utili all'interpretazione del fenomeno stesso.

Nel caso delle Random Forest, vengono calcolate alcune grandezze che misurano l'importanza delle diverse variabili esogene nella previsione di Y (Variable Importance Measures, VIM). Le due fondamentali VIM di una generica variabile esogena X_h sono ottenute nel modo seguente (Breiman, 2002):

- M1 – Predictive Importance: vengono casualmente permutati i valori assunti da X_h e calcolate le previsioni con questo nuovo dataset, in cui l'eventuale associazione tra X_h e Y è completamente distrutta. L'incremento dell'errore di previsione ottenuto con il nuovo dataset misura l'importanza di X_h nella previsione di Y.

$$M1(X_h) = OOB\ SSE\left(\widetilde{X_h}\right) - OOB\ SSE$$

- M2 – Constructive Importance: vengono sommate le diminuzioni di eterogeneità ottenute lungo tutti i nodi di tutti gli alberi che sono stati bipartiti in base alla variabile X_h.

$$I_h(T) = \sum_{j \in J} [d(h,j) I(h,j)]$$

dove *I(h,j)* è l'indicatore della funzione, il quale assume valore pari a 1 quando la variabile *h* è usata per dividere il nodo *j* e 0 altrimenti.

Alcuni recenti studi hanno mostrato che la misura M2 è affetta da una distorsione che tende ad attribuire maggiore importanza alle variabili che consento per loro natura un maggior numero di possibili bipartizioni del campione o che sono affette da un maggior numero di valori mancanti. Pertanto, il suo utilizzo impone una preventiva depurazione di tale distorsione (Sandri e Zuccolotto, 2007).

La possibilità di calcolare le Variable Importance Measures fa sì che le Random Forest possano essere utilizzate per la variable selection, ossia, l'identificazione dei predittori più importanti di un dato fenomeno. Si è dimostrato che le Random Forest, in quest'ambito,

mostrano notevoli potenzialità, in particolare in quei casi in cui pochi importanti predittori si trovano "nascosti" in mezzo ad un gran numero di variabili "di disturbo" (Sandri e Zuccolotto, 2006). Questo fattore è di notevole importanza tenuto conto del fatto che, negli ultimi anni, data la crescente disponibilità di set di dati di grandi dimensioni con all'interno un gran numero di variabili osservate, ha fatto si che la selezione delle variabili sia diventata un problema critico, a prescindere dal tipo di tecnica che si intende utilizzare per modellare il fenomeno.

2.3 SOFTWARE STATISTICO OPEN SOURCE "R".

A questo punto analizziamo le caratteristiche e la storia del software statistico open source che verrà utilizzato per la stesura del codice necessario all'implementazione dei modelli statistici sopra analizzati.

R, ormai da anni, costituisce, nell'ambito dei software di tipo statistico, una valida alternativa ai più diffusi ambienti statistici. La sua licenza viene distribuita gratuitamente in Internet in quanto si tratta di un software open source che viene costantemente sviluppato da un team di ricercatori in ambito statistico e informatico di fama mondiale. Essendo un software open source il suo codice sorgente viene distribuito gratuitamente e liberamente, costituendo così un sistema aperto a chiunque voglia aumentarne le possibilità di utilizzo e di calcolo.

R fu inizialmente creato da Ross Ihaka e Robert Gentelman, del Dipartimento di Statistica dell'Università di Auckland, Nuova Zelanda. Successivamente un enorme gruppo di persone cominciò a dare il proprio contributo, fondando l'R Core Team che dal 1997 si occupa dei codici sorgenti di R. Fin da subito questo software è stato considerato un clone o un dialetto di S, linguaggio di programmazione statistico sviluppato nel 1980 dai Bell Labs, noti tra l'altro per lo sviluppo del sistema operativo UNIX e del linguaggio C. S è stato sviluppato da un gruppo di ricercatori guidati da John Chambers, che oggi fa parte anche dell'R Core Team, e che nel 1998 ha ricevuto il premio ACM Software System Award proprio per il linguaggio S.

Esistono notevoli versioni di R che permettono di usare questo software su svariate piattaforme e sistemi informativi. R essendo un vero e proprio ambiente di programmazione, permette una elevatissima flessibilità nell'implementazione di funzioni di calcolo e di rappresentazione grafica statistica. In sintesi possiamo affermare che R rappresenta un insieme integrato di risorse software le quali consentono la manipolazione di dati, il calcolo e la visualizzazione di grafici (Angelo M. Mineo).

CAPITOLO 3: ELABORAZIONE DATI CAMPIONATO ITALIANO SERIE A.

FASE 1 - COMPRENSIONE DEL BUSINESS.

Prima di iniziare a mettere a fuoco quali sono gli obiettivi che si vogliono raggiungere attraverso queste analisi è opportuno fare un passo indietro e spiegare, per i non addetti ai lavori, il funzionamento del Campionato Italiano di Serie A.

Questo Campionato si divide in due parti: il girone di andata e quello di ritorno. La sua durata totale è di 38 giornate; 19 per ciascun girone. Essendo venti le squadre a competere per il titolo verranno disputate ben 380 partite al termine delle quali verrà premiata la prima classificata con il titolo di "Campione d'Italia". A questa squadra verrà conferito lo scudetto e l'opportunità di partecipare alla Champions League. L'UEFA Champions League (letteralmente Lega dei Campioni), o Coppa dei Campioni d'Europa rappresenta il massimo torneo calcistico europeo per squadre di club maschili e viene considerato il primo al mondo per livello calcistico.

L'opportunità di partecipare alla Champions League, generalmente per la nazione Italia, oltre ad essere conferita alla prima squadra classificata all'interno del campionato nazionale viene anche conferita alla seconda e alla terza, anche se, il numero preciso di squadre ammesse per nazione dipende da un particolare tipo di rating calcolato dalla UEFA per misurare le performance calcistiche di ciascuna di esse. In base a questo indicatore le tre nazioni con il rating più elevato presentano in Champions le prime quattro squadre classificate nel proprio campionato, le successive tre nazioni più forti ne porteranno tre, a seguire, per le nazioni dalla settima alla quindicesima sarà concesso presentarne solo due ed infine, oltre la sedicesima nazione ne sarà ammessa solo una. Tuttavia, non tutte le squadre presentate da ogni nazione entrano direttamente a competere nei gironi, infatti ciascuna nazione è costretta ad iscrivere una squadra ai cosiddetti preliminari di Champions League; naturalmente questa squadra sarà quella che, in base alla classifica all'interno del proprio campionato, ha ottenuto il risultato peggiore raffrontato alle altre squadre nazionali iscritte a questa competizione.

Tornando al nostro Campionato Italiano di Serie A è opportuno sottolineare che le venti squadre facenti parte dello stesso non rimangono fisse, infatti il loro permanere all'interno del campionato dipende dal loro posizionamento finale. Le ultime tre classificate, per l'appunto, sono costrette a retrocedere alla categoria inferiore la Serie B e allo stesso modo, le prime tre classificate nel Campionato di Serie B entrano a far parte della Serie A. Questo accade ogni anno alla fine del campionato.

Pertanto è necessario analizzare quali squadre facevano parte del Campionato Italiano di Serie A per l'anno 2010/2011, in modo da riuscire a comprendere le analisi e le considerazioni che verranno effettuate nelle pagine seguenti.


Bari


Bologna


Brescia


Cagliari


Catania


Cesena


Chievo


Fiorentina


Genoa


Inter


Juventus


Lazio


Lecce


Milan


Napoli


Palermo


Parma


Roma


Sampdoria


Udinese

Dopo questa breve introduzione sul funzionamento del Campionato Italiano, possiamo incominciare a mettere a fuoco quali sono gli obiettivi che vogliamo raggiungere alla fine di questa tesi.

Come tutti sanno il risultato di una partita è determinato da due elementi: il primo, che sicuramente possiamo considerare il più predominante è caratterizzato dal gioco e quindi dalla bravura atletica di una squadra rispetto alla sua avversaria, mentre il secondo, altrettanto importante, è l'elemento dell'aleatorietà o meglio della fortuna. Molto spesso quest'ultimo fattore incide pesantemente sul risultato finale delle partite e accade che, questa casualità, venga di sovente associata alla sfericità del pallone. Proprio da quest'associazione ha preso spunto il titolo di questo elaborato.

Come accennato nella fase introduttiva, proprio a causa del recente fermento per quanto riguarda il mondo del calcio e il recente scandalo scommesse, che ha visto coinvolto giocatori ed ex giocatori più o meno noti di varie società, anche in questo caso più o meno di primo piano, tra le quali le più conosciute sono sicuramente Atalanta, Siena e Chievo; abbiamo deciso di analizzare tutti i dati delle partite di Serie A 2010/2011 per riuscire a scorporare

ciascun incontro dall'effetto casualità prima citato e osservare, attraverso analisi statistiche sofisticate, se alcune tra le variabili rilevate durante gli incontri, possono essere considerate maggiormente critiche per la definizione del risultato finale della partita stessa. Sarebbe inoltre interessante redigere una nuova classifica sulla base proprio di queste variabili, per esaminare come sarebbe andato a finire il campionato se l'elemento fortuna non esistesse nella vita reale. Una volta stilata questa nuova classifica, naturalmente bisognerà confrontarla con la classifica reale e verificare di quanto le due si discostano. Ovviamente non ci aspettiamo la perfetta coincidenza fra le due, ma d'altro canto non ci aspettiamo nemmeno una completa inversione delle stesse. Sulla base di questa nuova classifica si potranno poi analizzare le singole partite del campionato e verificare, per ciascuna di esse, il livello di probabilità assegnato dal modello al risultato effettivamente realizzato dalle due squadre.

Guardando il lato pratico di questi obiettivi e tenendo presente che ciascun progetto di data mining deve produrre valore, possiamo affermare che il modello potrebbe risultare utile agli allenatori, i quali potrebbero spingere i giocatori della propria squadra a prestare maggiore attenzione alle variabili considerate più critiche ai fini del risultato finale. Allo stesso modo, potrebbero essere interessate a questo tipo di analisi anche le società di scommesse in quanto, attraverso i risultati ottenuti, potrebbero meglio definire le varie quotazioni da attribuire ad ogni match.

I dati necessari per compiere questo tipo di analisi sono le variabili che vengono rilevate durante una partita di campionato.

Scorrendo vari siti internet di dati statistici riferiti al mondo del calcio, abbiamo deciso di rivolgerci presso la Panini Digital, per ottenere tutti i dati necessari allo sviluppo di questo progetto. Ci siamo rivolti direttamente a questa società in quanto è considerata l'azienda leader in Italia nella raccolta statistica di dati e nella fornitura dei servizi informatici a supporto delle società calcistiche e del mondo dei media.

FASE 2 - COMPRENSIONE DEI DATI.

La Panini Digital, come dicevo, ci ha fornito tutti i dati da noi richiesti, confidandoci che nessuno mai le aveva chiesto una quantità di dati così elevata per lo sviluppo di una tesi.

Abbiamo ricevuto i file di ogni partita in formato elettronico, per un totale di 380 file; questo perché le squadre che compongono il campionato Italiano di Serie A sono 20, pertanto vengono giocate 10 partite per giornata, il numero di giornate è 38 e quindi il totale di partite disputate durante il campionato è 380.

Una volta ottenuti i file è stato necessario organizzarli in un unico database.

Fin da subito, durante la fase di importazione degli stessi nel database unico, abbiamo notato che alcuni file contenevano un numero diverso di variabili rispetto agli altri. In particolare, tutte le partite della 1° e 2° giornata e le prime sei riferite alla 3° contengono 394 variabili rilevate per squadra a partita; mentre le partite di Bari – Roma, Juventus – Chievo e Sampdoria – Palermo rispettivamente della 35°, 36° e 37° giornata hanno 399 variabili, sempre per squadra a partita. I restanti file ne contengono 398.

Per riuscire a tenere in considerazione il maggior numero di informazioni possibili abbiamo deciso di considerare per ogni squadra il numero più elevato di variabili, 399, e dove alcune di queste non fossero state rilevate di lasciare la cella vuota. Questi dati, saranno poi considerati dal modello come missing values.

Se ci fossimo limitati a spiegare le variabili contenute nei vari file forniteci avremmo avuto un database di 798 variabili, tuttavia, seppure sembrino molte, abbiamo deciso di inserirne molte altre. Alcune dettate dalla necessità come il nome delle squadre, la giornata di campionato e il risultato delle partite, mentre altre sono state inserite per curiosità come il meteo durante la partita e la differenza di ciascuna variabile riferita alla squadra di casa con la rispettiva variabile riferita alla squadra ospite. Naturalmente siamo perfettamente consapevoli che alcune di queste variabili "differenza" non hanno senso, ma di questo aspetto ce ne occuperemo nella fase di preparazione dei dati.

Il nostro database è quindi formato da 1.204 variabili, che possono essere ripartite in 6 macro categorie: *dati relativi alla giornata*, *dati generali*, *dati sui portieri*, *fase difensiva*, *fase offensiva* e *calci piazzati battuti*.

All'interno dell'Appendice A potrete trovare l'elenco completo di tutte le variabili, suddivise nelle 6 macro categorie appena citate, insieme ad una breve descrizione di ciascuna di esse e al nome identificativo assegnato alle stesse all'interno del database.

FASE 3 – PREPARAZIONE DEI DATI.

Passiamo ora alla fase vera e propria di preparazione dei dati, e consideriamo solamente le variabili del database che devono essere modificate, in particolare per renderle confrontabili con le altre variabili contenute nello stesso, o addirittura eliminate per vari errori commessi in fase di travaso dati o in fase di rilevazione degli stessi. È necessario eliminare anche le informazioni ridondanti in quanto non portano a nessuna informazione aggiuntiva.

3.3.1 VARIABILI DUMMY.

Iniziamo con la variabile meteo. Questa variabile presenta 10 modalità, e più precisamente sono:

- coperto;
- nebbia;
- neve;
- nubi sparse;
- pioggia;
- pioggia debole;
- pioggia e schiarite;
- poco nuvoloso;
- sereno;
- temporale.

Ai fini della nostra analisi, abbiamo deciso di creare due file, uno contenente tutte queste modalità e l'altro nel quale ci interessa sapere solo se il tempo meteorologico durante i vari incontri può essere considerato bello o brutto. Per creare questo secondo file dobbiamo trasformare la variabile meteo in una variabile dicotomica contenente solo due modalità: 0 se il tempo può essere considerato buono ed 1 se la partita si gioca con condizioni atmosferiche avverse. Consideriamo condizioni atmosferiche avverse il temporale, la pioggia, la neve e la nebbia.

Facciamo tutto questo in modo da individuare se esistono cambiamenti notevoli nei valori delle altre variabili al variare del tempo atmosferico ed al variare della modalità in cui questo viene espresso.

3.3.2 MISSING VALUES.

Nel database vi sono dei dati mancanti (missing values) e sono riferiti alle variabili:

- Duelli aerei (D_DUELLI)
- Percentuale di duelli aerei vinti (D_%DUELLI1)
- Duelli aerei persi (D_DUELLI2)
- Percentuale di duelli aerei persi (D_%DUELLI2)
- Precisione nel tiro (O_PREC_TIRO).

Per le prime quattro variabili sono presenti 51 missing values, questo significa che per 51 incontri su 380 queste variabili non sono state rilevate. Pertanto, solo per il 13,42% delle partite dell'intero campionato, il dato è mancante. Essendo una percentuale limitata, ho deciso di mantenerli comunque nel database, considerando anche il fatto che gli alberi decisionali e le Random Forest sono in grado di analizzare i dati anche in presenza di questi valori mancanti. Abbiamo deciso di rimediare a questa mancanza compilando i campi del database rimasti vuoti.

Osservando attentamente i dati abbiamo verificato che la variabile duelli aerei vinti è stata rilevata per ogni partita, pertanto avendo a disposizione sia il numero di duelli aerei vinti dalla squadra di casa, sia quelli vinti dalla squadra ospite possiamo ottenere anche tutti gli altri valori che inizialmente non erano presenti nel database. Infatti sommando le due variabili appena citate possiamo ottenere: il numero totale di duelli aerei, calcolare le percentuali di duelli aerei vinti e persi e di conseguenza il numero effettivo di duelli aerei persi. Queste variabili, come vedremo nelle righe seguenti, sono in realtà espresse in funzione di altre variabili e quindi sarà necessario valutarle attentamente e decidere quali mantenere all'interno del dataset.

La variabile "precisione nel tiro", purtroppo, è stata rilevata solo in 3 partite su 380 dell'intero campionato. Essendo un numero veramente irrisorio, suppongo che questa variabile non potrà essere considerata tra le più importanti per la definizione del risultato di una partita. Pertanto penso che sia opportuno eliminarla; questa operazione, comporta l'eliminazione di tre variabili: due riferite alla precisione nel tiro rispettivamente per la squadra di casa e quella ospite, mentre l'altra riferita alla differenza tra il valore assunto da questa variabile per la squadra di casa e quello riferito alla squadra ospite.

3.3.3 SELEZIONE DELLE VARIABILI DA UTILIZZARE PER IL MODELLO.

A questo punto è stato necessario ridimensionare il nostro database principalmente per tre motivi: e più precisamente in quanto alcune variabili sarebbero risultate inutili ai fini della nostra analisi perché troppo specifiche su elementi tecnici di secondo piano; altre variabili risultavano essere ridondanti, infatti sarebbero potute comunque essere calcolate come funzioni di altre variabili ed infine l'ultima serie di variabili è stata eliminata in quanto rappresentava variabili "troppo" significative per il nostro modello. Queste variabili sono ad esempio il numero di reti segnate da una squadra in una determinata modalità, i punti assegnati alla fine dell'incontro, i rigori trasformati e così via. Abbiamo deciso di eliminare queste variabili dal database perché, se le avessimo lasciate, sicuramente il modello ci avrebbe

restituito che proprio queste potevano essere considerate quelle che incidevano maggiormente sul risultato di una partita. Ovviamente il fattore più determinante per la buona riuscita di una partita sono proprio le maggiori reti segnate da una squadra rispetto all'avversaria, ma non è questo che ci interessa; infatti vogliamo capire quali sono le variabili che determinano la vittoria di una squadra nascondendo al modello i risultati delle partite e quindi anche il numero di reti.

Attraverso questa scrematura siamo riusciti ad ottenere un database con 729 variabili e 380 osservazioni, che sarà inserito nel modello per essere elaborato e di conseguenza per far sì che si possano trarre le dovute considerazioni.

Per il dettaglio delle variabili eliminate nelle varie fasi è possibile consultare l'Appendice B.

FASE 4: MODELLAZIONE.

3.4.1 RANDOM FOREST.

Abbiamo applicato la metodologia delle Random Forest ai dati a nostra disposizione utilizzando il software di programmazione statistica R. Nell'Appendice C è possibile visionare il codice attraverso il quale siamo giunti alla determinazione dei risultati.

Inizialmente ci eravamo prefissati tre variabili obiettivo da determinare attraverso questo modello, e più precisamente:

- Y1: variabile che considera sullo stesso piano le partite perse e pareggiate dalla squadra di casa, raggruppandole in un'unica previsione, considerando separatamente solo le partite vinte da questa.
- Y2: la quale prende in considerazione solo le partite vinte e perse dalla squadra di casa, escludendo dall'analisi tutte quelle osservazioni che si riferivano ad incontri conclusi con un pareggio.
- Y3: considera in maniera separata le tre possibili modalità con le quali si può concludere una partita di campionato; vittoria, sconfitta e pareggio.

Sono state attuate molte analisi su queste variabili per verificare la bontà dei risultati ottenuti. Questi studi sono stati effettuati per entrambi i database a nostra disposizione; e più precisamente per quello contenente la variabile meteo espressa nelle sue dieci modalità originarie e per quello con questa variabile espressa in maniera dicotomica.

Ci siamo resi conto che i risultati da noi ottenuti non variavano significativamente al variare del database utilizzato per l'analisi. Molto probabilmente perché la variabile meteo non è mai risultata tra le variabili più significative per la determinazione del risultato finale di una partita. Pertanto risultava inutile mantenere il doppio database per effettuare le analisi successive, ed allora ho deciso di considerare solo il database contenente le dieci modalità riferite alla variabile meteo, in modo da avere un'informazione più dettagliata, per quanto riguarda questa variabile, in ogni partita del campionato.

Di seguito verranno mostrati alcuni dei risultati di previsione ottenuti per le tre variabili obiettivo sopra citate.

VARIABILE Y_1.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,  $
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 26

        OOB estimate of  error rate: 25%
Confusion matrix:
                     sconfitta o pareggio vittoria class.error
sconfitta o pareggio                  158       42   0.2100000
vittoria                               53      127   0.2944444
```

Figura 2: previsioni ottenute attraverso la Random Forest per la prima variabile obiettivo.

Per la variabile Y_1 notiamo che il nostro modello va abbastanza bene sia a livello generale che per singole modalità. Infatti, sbaglia nel 25% dei casi, e più precisamente quando una partita dovrebbe essere classificata come *sconfitta o pareggio* il modello classifica la partita in maniera errata nel 21% dei casi; mentre per quanto riguarda le partite che dovrebbero essere classificate come *vittorie*, il modello sbaglia nel 29,44% dei casi.

Anche ripetendo per un numero elevato di volte queste analisi, i risultati variano di poco e lo stesso vale anche per le variabili che risultano maggiormente significative ai fini della determinazione dei risultati.

VARIABILE Y_2.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,  $
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 26

        OOB estimate of  error rate: 19.72%
Confusion matrix:
          sconfitta vittoria class.error
sconfitta        58       46  0.44230769
vittoria         10      170  0.05555556
```

Figura 3: previsioni ottenute attraverso la Random Forest per la seconda variabile obiettivo

Per la variabile Y_2 apparentemente il modello funziona bene, infatti a livello generale si nota che classifica le partite in maniera errata solo nel 19,72% dei casi. Tuttavia analizzando più approfonditamente i risultati notiamo che il modello tende a classificare maggiormente le partite come vinte, a scapito della precisione sulla modalità *sconfitta*. Tale discrepanza di risultati si riscontra in ogni iterazione del modello.

Proprio a causa di questa anomalia e della perdita di numerose osservazioni dovute al fatto che le partite disputate si fossero concluse con un pareggio, abbiamo deciso di eliminare la variabile Y_2 dalle nostre variabili obiettivo.

VARIABILE Y_3.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,  $
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 26

        OOB estimate of  error rate: 41.32%
Confusion matrix:
          vittoria sconfitta pareggio class.error
vittoria       161        10        9   0.1055556
sconfitta       40        53       11   0.4903846
pareggio        66        21        9   0.9062500
```

Figura 4: previsioni ottenute attraverso la Random Forest per la terza variabile obiettivo

Infine per la variabile Y_3 otteniamo, rispetto alle precedenti variabili obiettivo, il peggior risultato a livello generale. Notiamo che come sempre il modello classifica bene la modalità *vittoria*, mentre sbaglia quasi nel 50% dei casi quando si tratta di prevedere la modalità *sconfitta* e lo possiamo quasi definire disastroso quando il match si conclude con un *pareggio*, sbagliando nel 90,62% dei casi.

Pur presentando risultati a primo acchito abbastanza demotivanti, abbiamo deciso di mantenere questa variabile nel modello per vedere se, attraverso opportune modifiche all'interno dello stesso, possiamo migliorare i risultati ottenuti. Inoltre, valutando attentamente la situazione possiamo dire che con il modello si ha una maggiore possibilità di individuare la modalità esatta rispetto al caso, infatti se ci affidassimo a quest'ultimo, in media, riusciremmo ad individuare la modalità corretta solo nel 33% dei casi, mentre con il modello riusciamo a farlo quasi per il 60%.

Naturalmente qualora i risultati successivi non dovessero ancora essere appaganti procederemo ad eliminare anche questa variabile obiettivo.

3.4.2 ANALISI DEGLI INDICI.

Passiamo ora ad analizzare le variabili più significative per il modello; attraverso le quali quest'ultimo si è addestrato ed ha permesso di produrre i risultati sopra citati.

Analizzeremo quindi i classici grafici a barre sia per l'indice Mean Decrease in Accuracy sia per quello sulla Gini Impurity e successivamente individueremo le variabili più importanti al fine della previsione del modello, aiutandoci con il grafico di dispersione (scatter plot), ottenuto per ciascuna variabile obiettivo, sulla base dell'andamento degli indici sopra citati.

VARIABILE Y_1.

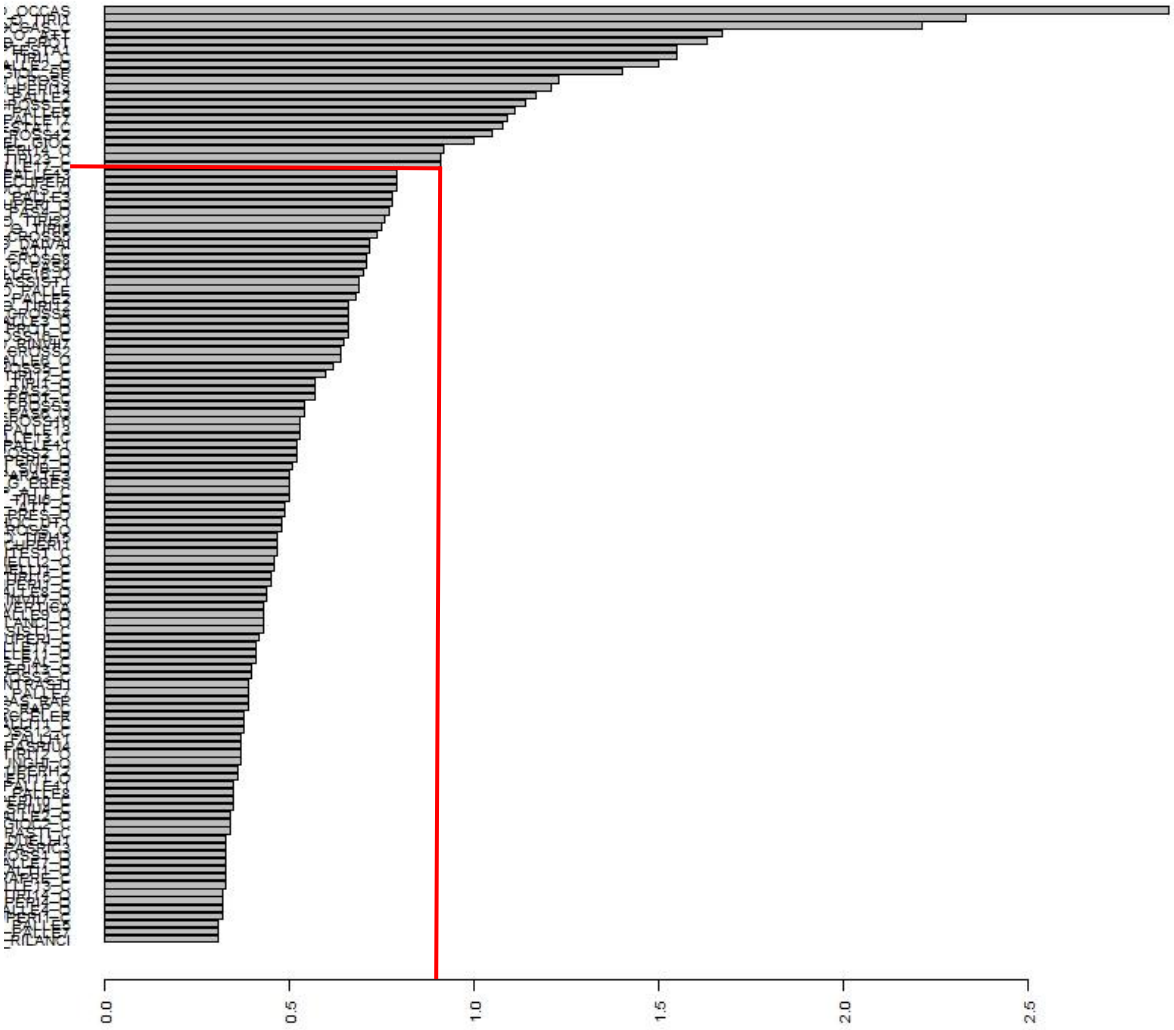


Figura 5: diagramma a barre riferito alla prima variabile obiettivo per l'indice Mean Decrease in Accuracy

In questo grafico sono state illustrate le prime 120 variabili più importanti secondo l'indice Mean Decrease in Accuracy. Sul lato sinistro degli assi cartesiani possiamo notare i nomi corrispondenti a ciascuna variabile raffigurata all'interno dello stesso. Tuttavia, essendo molto elevato il numero di variabili, risulta difficile identificarle basandosi unicamente sul disegno e pertanto è stato necessario implementare una stringa di codice in modo da ottenere un elenco ordinato di tutte le variabili con a fianco il valore assunto da ciascuna di esse per ognuno dei due indici di Variable Importance. Tornando al grafico, possiamo notare in corrispondenza delle linee rosse che l'importanza in merito alle variabili precipita vorticosamente facendo formare la classica curvatura a gomito. Questo significa che le variabili con valori di importanza inferiori a questa linea contribuiscono a spiegare la variabile obiettivo in maniera

sempre minore. Notiamo inoltre, che questo fenomeno si verifica in corrispondenza del valore 0,9. Pertanto è stata ripetuta l'iterazione per un numero cospicuo di volte tenendo sempre in considerazione solo le variabili con valori superiori a 0,9 in riferimento a questo indice. Abbiamo potuto fare questo in quanto osservando questo tipo di grafico per ogni iterazione inerente a questo indice, abbiamo riscontrato che sempre in prossimità del valore 0.9 si verifica questa curvatura a gomito. La soglia, come si è potuto capire, è stata scelta in maniera soggettiva, tuttavia questa decisione non deve essere considerata cruciale ai fini della buona riuscita del modello in quanto lo scostamento effettivo da questo valore, per ciascuna analisi, avrebbe consentito di includere o escludere solo 3 o 4 variabili, rispettivamente in più o in meno, rispetto a quanto è stato ottenuto con questa decisione.

Successivamente abbiamo osservato l'andamento a seconda dell'indice di Gini impurity.

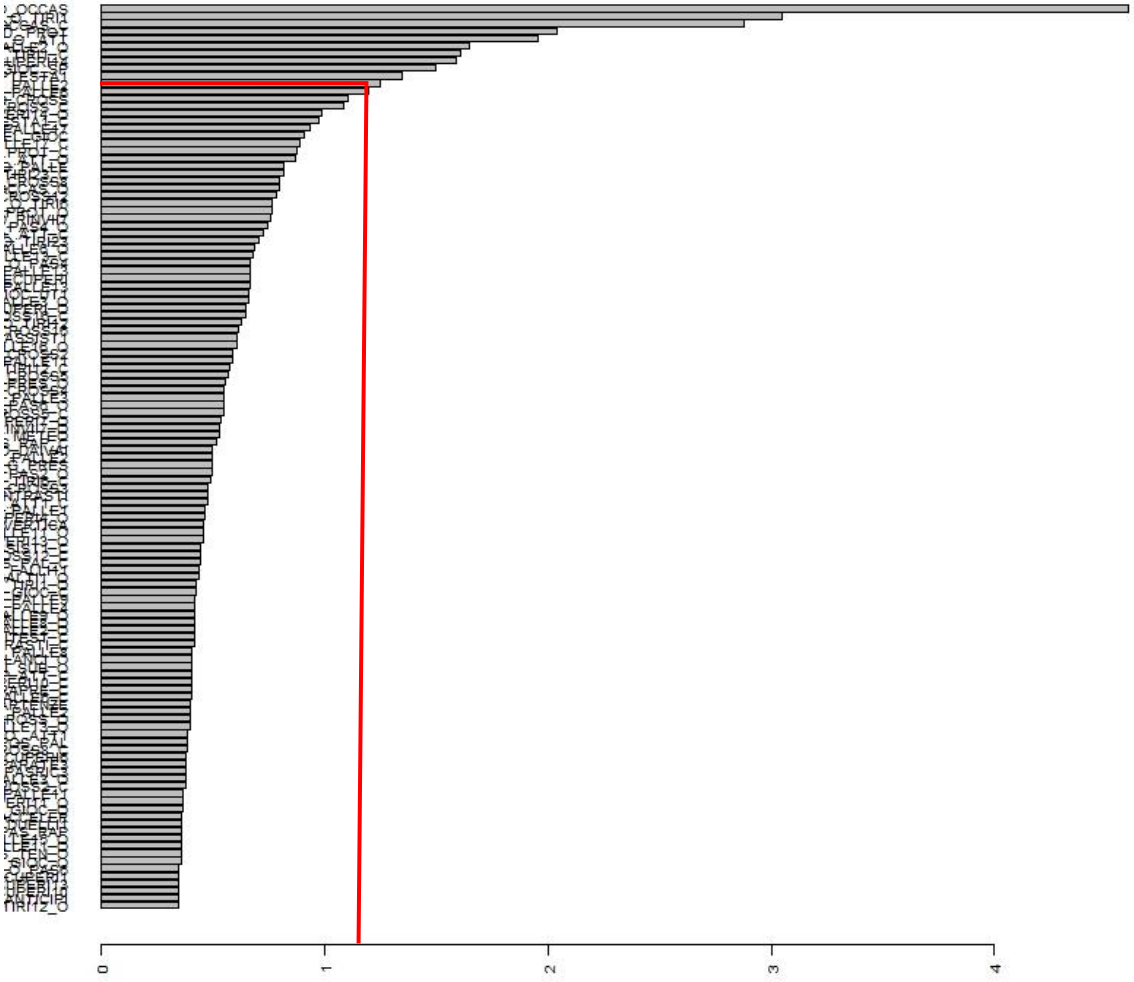


Figura 6: diagramma delle variabili più significative in riferimento alla prima variabile obiettivo, calcolate attraverso l'indice Gini Impurity

Osservando questo grafico notiamo immediatamente che rispetto al precedente la curvatura a gomito si manifesta molto prima. Tuttavia, essendo nostra intenzione tenere in considerazione solo le variabili più importanti che vengono considerate tali da entrambi gli indici abbiamo deciso di tenere in considerazione per ogni iterazione tutte quelle variabili che assumono, in riferimento a questo indice, valori superiori a 1,05. Abbiamo stabilito questa soglia in quanto delimita l'area nella quale poi si registra la curvatura a gomito che avevamo preso come riferimento anche per il precedente indice.

Come visto in precedenza, anche in questo caso, la procedura è stata ripetuta per un numero cospicuo di volte; tenendo sempre in considerazione solo le variabili con valori più elevati rispetto alla soglia appena identificata. Naturalmente, essendo abbastanza elevato questo valore soglia abbiamo registrato una variabilità abbastanza significativa in merito alle variabili maggiormente determinanti per la definizione del risultato e pertanto è stato necessario procedere con diverse analisi in modo da registrare quali tra queste risultassero essere, con maggior frequenza, quelle più significative per il modello in questione.

Una volta definite le due soglie è stato possibile generare il grafico di dispersione o meglio chiamato Scatter Plot, che maniera intuitiva, ci permette di apprendere quali sono le variabili prese in considerazione per il raggiungimento dei nostri obiettivi.

Il grafico, riportato nella pagina successiva, è stato costruito ponendo rispettivamente sull'asse delle ascisse l'indice Mean Decrease in Accuracy e su quello delle ordinate il Gini Impurity.

Osservando l'andamento della nube di punti notiamo che le variabili posizionate nell'angolo in alto a destra sono quelle maggiormente significative al fine della determinazione del risultato finale delle partite. Pertanto, abbiamo deciso di considerare, come già argomentato in precedenza, solo le variabili che abbiano, per quanto riguarda l'indice Mean Decrease in Accuracy un valore superiore o uguale a 0,9; mentre per quanto riguarda l'indice di Gini Impurity un valore superiore o uguale a 1,05. La scelta di questo valore soglia è stata attuata, come anticipato poco sopra, in quanto rappresenta il miglior compromesso tra il numero di variabili selezionate e la significatività di queste ultime.

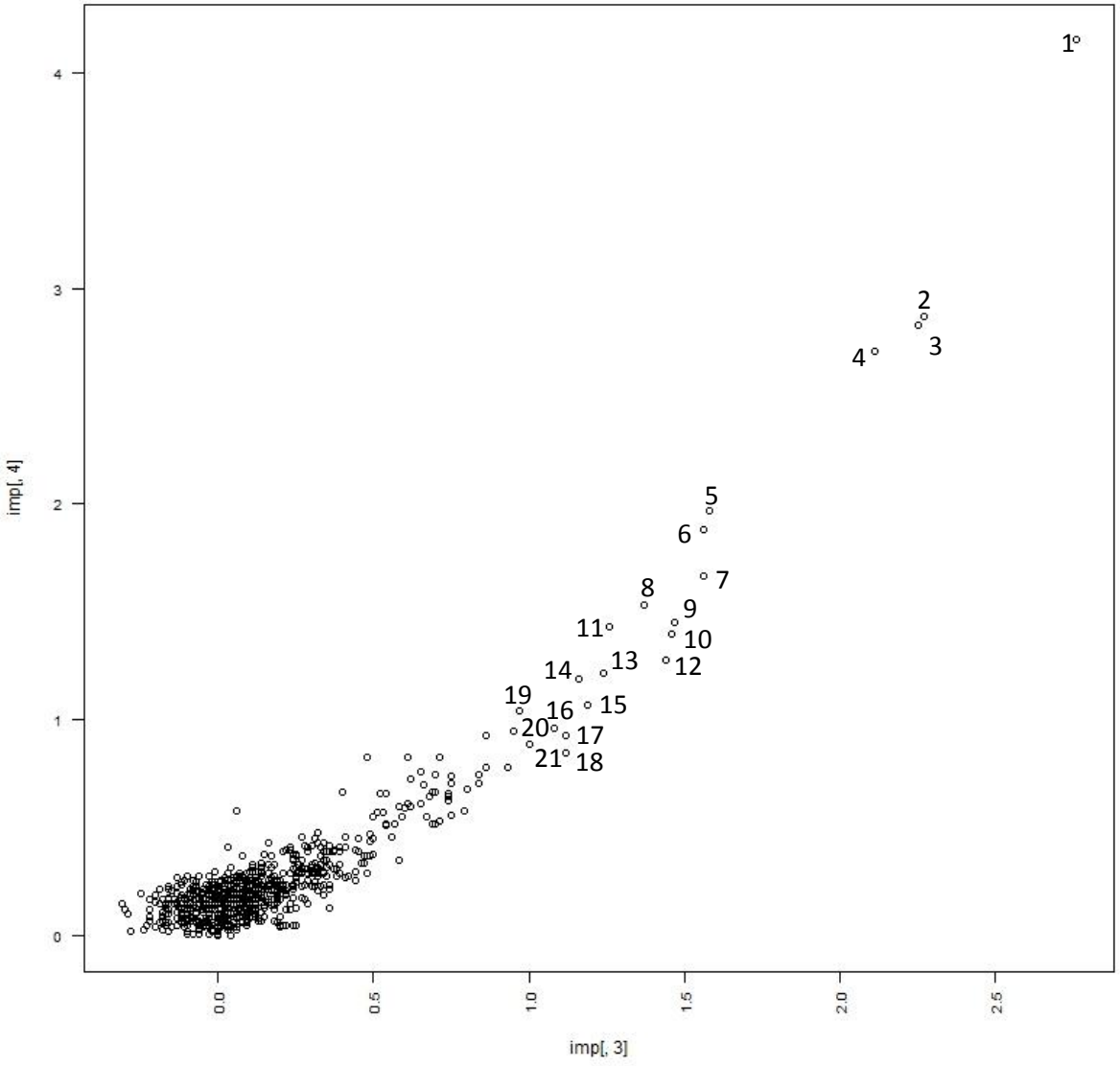


Figura 7: Scatter Plot in merito agli indici Mean Decrease in Accuracy e Gini Impurity per quanto riguarda la prima variabile obiettivo.

Procedendo con l'osservazione del grafico notiamo una specie di "nuvola nera", la quale rappresenta un grandissimo numero di variabili che assumono, per gli indici sopra citati, valori pressoché simili. Pertanto, nella rappresentazione grafica tendono a sovrapporsi in maniera tale da creare una sorta di zona di "caos" nella quale sono contenute variabili ininfluenti per le analisi che ci apprestiamo a compiere. L'insieme di queste variabili può essere etichettato con il nome "variabili rumore".

Di seguito verranno riportate le variabili che sono risultate più importanti per la determinazione della variabile obiettivo Y_1.

1. DIF_O_OCCAS: differenza di occasioni tra la squadra di casa e la squadra ospite;

2. DIF_O_TIRI1: differenza di tiri dentro tra le due squadre;
3. O_OCCAS_C: numero di occasioni create dalla squadra di casa;
4. DIF_O_%ATT: differenza di percentuale di attacco alla porta tra le due squadre;
5. DIF_D_%PROT: differenza di percentuale di protezione dell'area;
6. DIF_G_COLPTESTA1: differenza di numero di colpi di testa effettuati in area avversaria;
7. O_TIRI1_C: numero di tiri dentro effettuati dalla squadra di casa;
8. O_PALLE2_O: numero di palle a scavalcare il centrocampo calciate dalla squadra ospite;
9. DIF_G_GIOC_SP: differenza in merito al numero di giocate spettacolari effettuate da entrambe le squadre;
10. DIF_O_CROSS: differenza di numero di cross effettuati dalle due squadre;
11. DIF_D_RECUPERI14: differenza di recuperi aerei effettuati in area;
12. DIF_O_PALLE2: differenza di palle a scavalcare il centro campo;
13. O_CROSS_C: numero di cross effettuati dalla squadre di casa;
14. DIF_G_PALLE6: differenza di palle giocate sulla fascia centrale in difesa;
15. DIF_D_PALLE17: differenza di palle perse effettive in attacco;
16. G_COLPTESTA1_C: numero di colpi di testa effettuati in area avversaria dalla squadra ospite;
17. DIF_O_CROSS12: differenza di cross su azione;
18. DIF_G_VEL_GIOC: differenza di velocità di gioco fra le due squadre che disputano la partita;
19. D_RECUPERI14_O: numero di recuperi aerei in area effettuati dalla squadra ospite;
20. D_PALLE17_C: numero di palle perse effettive in attacco dalla squadra di casa;
21. O_OCCAS_O: numero di occasioni create della squadra ospite durante lo svolgimento della partita.

Prima di effettuare analisi più approfondite su queste variabili passiamo ad esaminare i grafici ottenuti con riferimento alla terza variabile obiettivo (Y_3).

VARIABILE Y_3.

Per quanto riguarda la terza variabile obiettivo, che prevede l'effettivo risultato di una partita considerando separatamente anche le partite che si sono concluse con un pareggio, possiamo osservare i seguenti grafici:

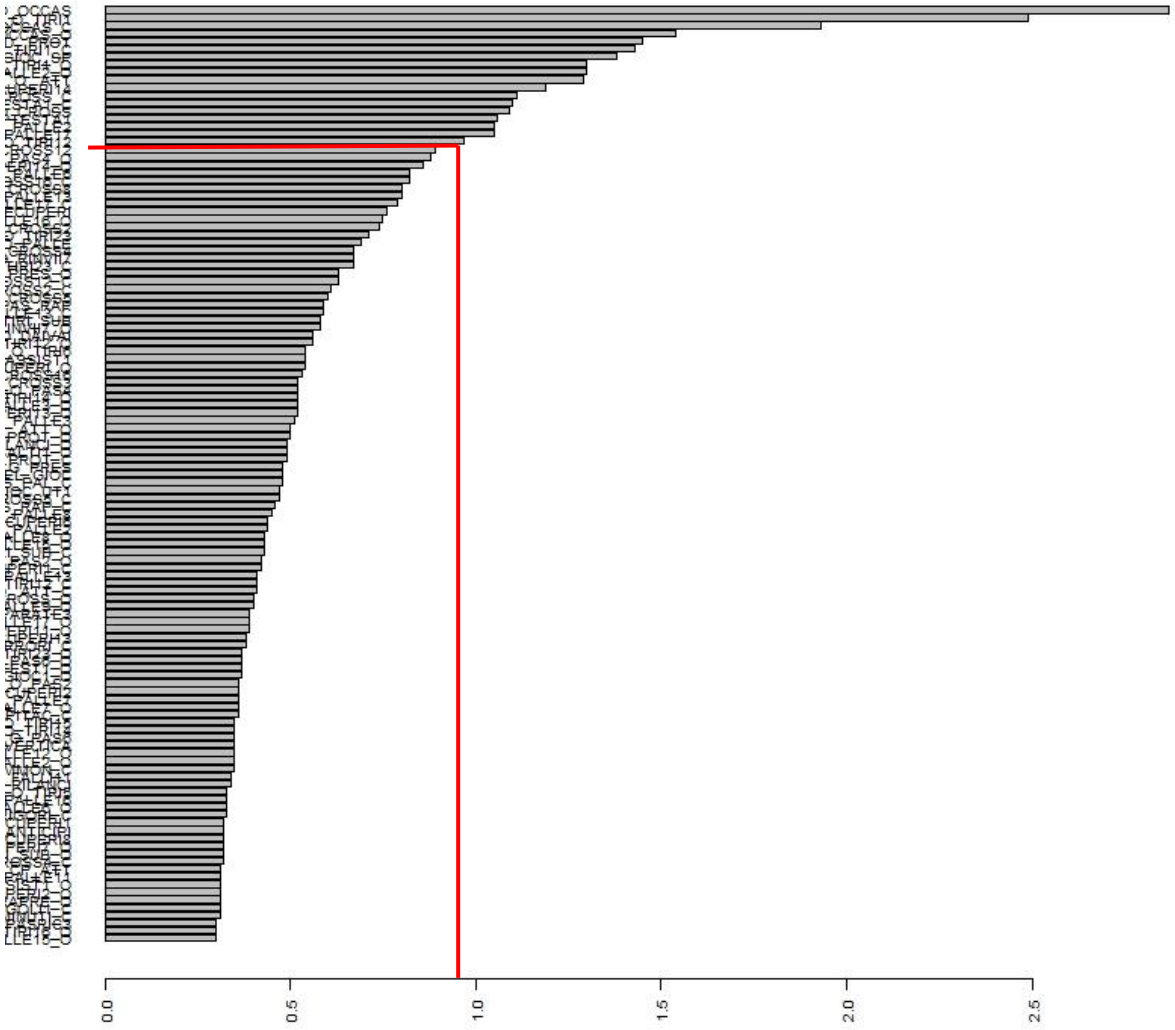


Figura 8: diagramma a barre riferito alla terza variabile obiettivo per l'indice Mean Decrease in Accuracy.

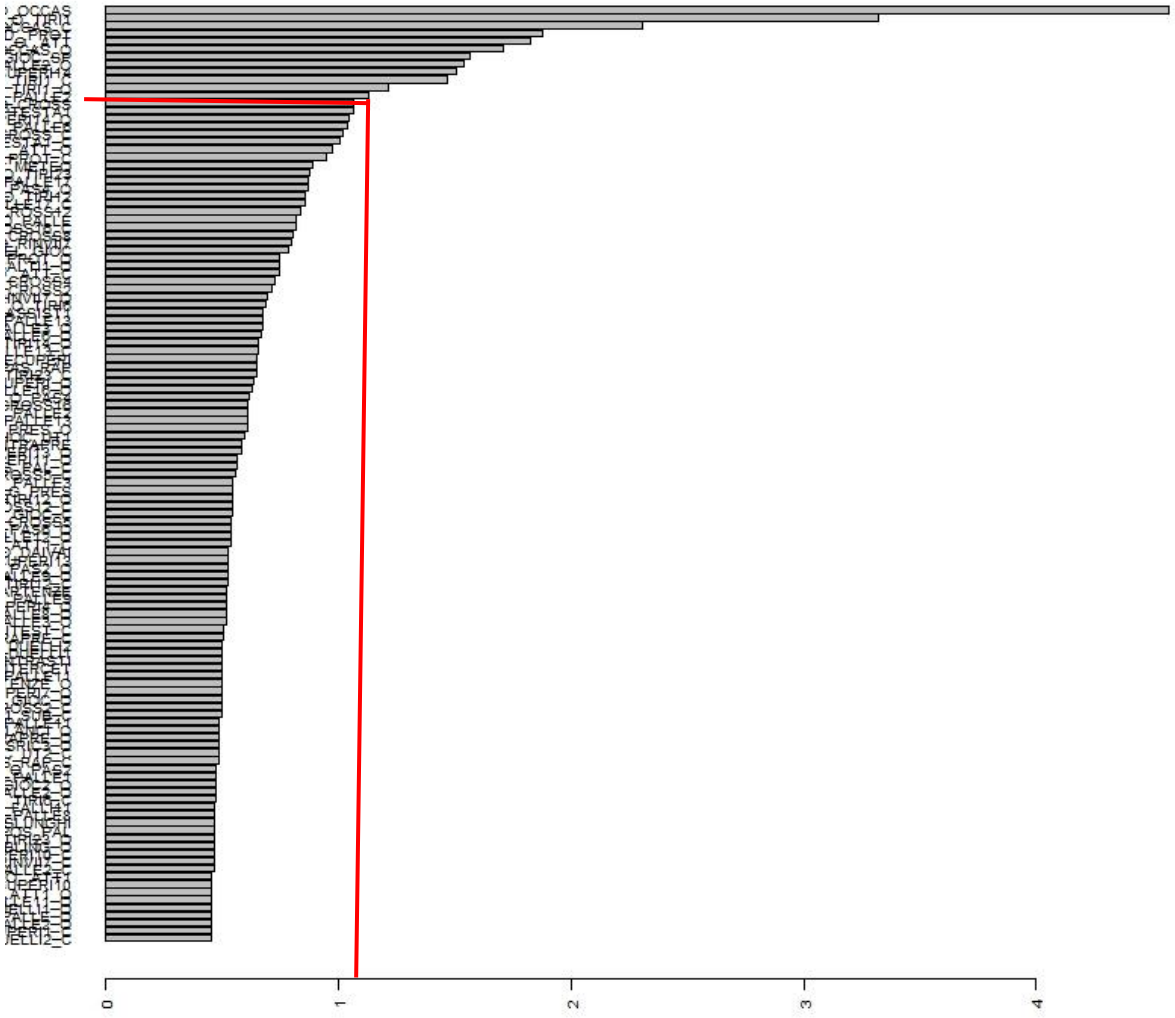


Figura 9: diagramma delle variabili più significative in riferimento alla terza variabile obiettivo calcolate attraverso l'indice Gini Impurity.

Osservando i due grafici notiamo immediatamente che ci si ripropone una situazione molto simile a quella ottenuta per la prima variabile decisionale; infatti il valore soglia ottimale per l'indice di Mean Decrease in Accuracy è sempre intorno al valore 0,9; mentre per l'indice Gini Impurity si aggira intorno a 1,05.

Fatte queste semplici considerazioni passiamo ad analizzare anche lo Scatter Plot in riferimento agli indici appena menzionati. Ci aspettiamo che tale grafico sia molto simile a quello visto in precedenza per la variabile Y_1 in quanto i singoli grafici in merito agli indici oggetto di analisi risultano essere molto simili a quelli osservati per la prima variabile obiettivo.

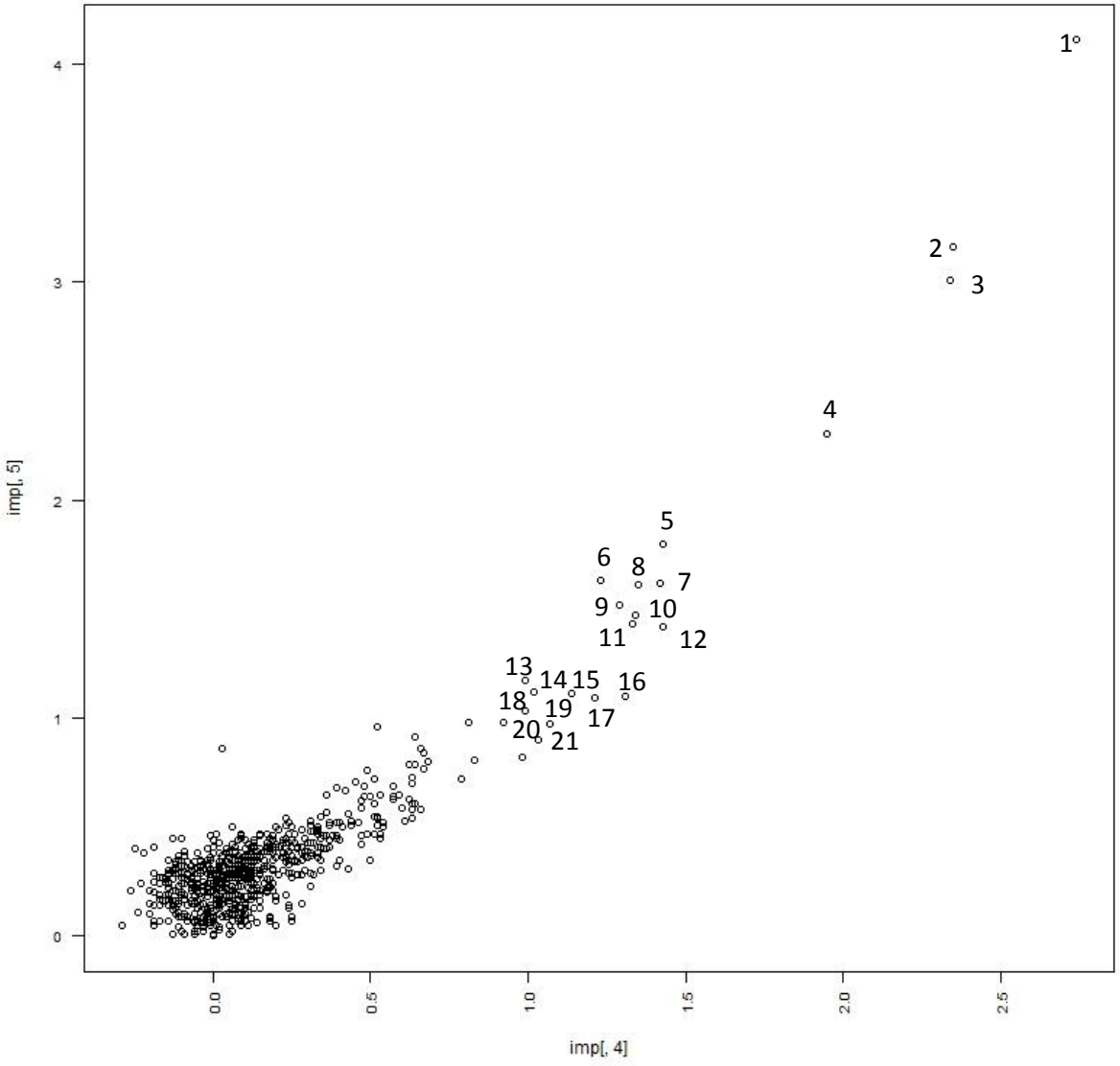


Figura 10: Scatter Plot in merito agli indici Mean Decrease in Accuracy e Gini Impurity per quanto riguarda la terza variabile obiettivo.

Possiamo affermare che le nostre aspettative si sono concretizzate, infatti l'andamento dei punti è molto simile. Naturalmente possiamo notare una leggera differenza di posizionamento delle variabili che si trovano in alto a destra. Questa differenza è giustificabile in quanto dovendo prevedere una modalità di risultato in più, rispetto alla precedente variabile decisionale, ci aspettiamo che l'importanza delle variabili, intesa come valori assunti dalle stesse, cambi e che allo stesso modo alcune di queste differiscano dalle precedenti. Tuttavia, lavorando sul medesimo dataset e tenendo in considerazione che la variabile obiettivo attuale riprende i medesimi valori soglia assunti dalla variabile obiettivo precedente, ci aspettiamo che le variabili non considerate come importanti nella determinazione della prima variabile

obiettivo non siano in numero eccessivamente elevato rispetto a quelle ottenute per la determinazione di Y_3.

Passiamo quindi ad analizzare le variabili risultate maggiormente apprezzabili per la determinazione della variabile obiettivo presa in considerazione:

1. DIF_O_OCCAS: differenza di occasioni tra la squadra di casa e la squadra ospite;
2. DIF_O_TIRI1: differenza di tiri dentro tra le due squadre;
3. O_OCCAS_C: numero di occasioni create dalla squadra di casa;
4. DIF_D_%PROT: differenza di percentuale di protezione dell'area;
5. O_TIRI1_C: numero di tiri dentro effettuati dalla squadra di casa;
6. DIF_G_GIOC_SP: differenza in merito al numero di giocate spettacolari effettuate da entrambe le squadre;
7. O_TIRI1_O: numero di tiri dentro effettuati dalla squadra ospite;
8. O_PALLE2_O: numero di palle a scavalcare il centrocampo calciate dalla squadra ospite;
9. DIF_O_%ATT: differenza di percentuale di attacco alla porta tra le due squadre;
10. DIF_D_RECUPERI14: differenza di recuperi aerei effettuati in area;
11. O_CROSS_C: numero di cross effettuati dalla squadra di casa;
12. G_COLPTESTA1_C: numero di colpi di testa effettuati in area avversaria dalla squadra ospite;
13. DIF_O_CROSS: differenza di numero di cross effettuati dalle due squadre;
14. DIF_G_COLPTESTA1: differenza di numero di colpi di testa effettuati in area avversaria;
15. DIF_O_PALLE2: differenza di palle a scavalcare il centro campo;
16. DIF_D_PALLE17: differenza di palle perse effettive in attacco;
17. DIF_O_TIRI12: differenza di tiri di piedi effettuati con palla bassa dalle due squadre;
18. O_OCCAS_O: numero di occasioni create della squadra ospite durante lo svolgimento della partita;
19. DIF_O_CROSS12: differenza di cross su azione;
20. DIF_G_PALLE6: differenza di palle giocate sulla fascia centrale in difesa;
21. D_RECUPERI14_O: numero di recuperi aerei in area effettuati dalla squadra ospite.

Proprio come ci aspettavamo le variabili grosso modo sono rimaste le stesse, ne è variato naturalmente l'ordine, ma le uniche che non erano considerate come importanti per la determinazione della variabile Y_1 sono O_TIRI1_O e DIF_O_TIRI12; allo stesso modo, per

quanto riguarda la determinazione della variabile Y_3 non sono risultate come importanti DIF_G_VEL_GIOC e D_PALLE17_C.

Grazie al fatto che sono solo due le variabili che si discostano passando da una variabile obiettivo all'altra, abbiamo deciso di creare un nuovo database contenente tutte le osservazioni e i relativi valori per le variabili identificate come più importanti emerse dalle analisi appena effettuate. In questo modo abbiamo un database di 380 osservazioni e 23 variabili.

3.4.3 ANALISI DELLE VARIABILI PIU' RILEVANTI PER LA DETERMINAZIONE DEI RISULTATI FINALI DELLE PARTITE.

Passiamo ora ad analizzare in maniera più approfondita le variabili di cui disponiamo.

Ad un primo sguardo notiamo che le variabili *occasioni* e *tiri dentro* sono presenti in tutte e tre le modalità disponibili per ciascuna variabile, e quindi sono citate sia per la squadra di casa, sia per la squadra ospite che per differenza di valori tra la squadra di casa e la squadra ospite. Altre variabili, come, i *colpi di testa in area avversaria*, i *cross* e le *palle perse effettive in attacco*, sono presenti in riferimento alla squadra di casa e come differenza di valori tra le due squadre; mentre le variabili *palle a scavalcare il centrocampo* e i *recuperi aerei in area* sono presenti per la squadra ospite e come differenza. Infine, le variabili presenti con una sola modalità sono tutte riferite al valore differenza tra la modalità assunta dalla variabile riferita alla squadra di casa e quella assunta dalla squadra ospite; più precisamente, queste variabili, sono riferite alla *percentuale di attacco alla porta*, alla *percentuale di protezione dell'area*, alle *giocate spettacolari*, alle *palle giocate sulla fascia centrale in difesa*, ai *cross su azione*, alla *velocità di gioco* e ai *tiri di piede con palla bassa*.

Osservando i dati, emerge che la maggior parte di queste variabili (13/23) sono riferite alla fase offensiva (prefisso O_), mentre le restanti si dividono in maniera eguale tra le variabili riferite alla fase difensiva (prefisso D_) e quelle riferite ai dati generali (prefisso G_).

Passiamo ora ad analizzare il comportamento della Random Forest sulla base di questo nuovo database.

VARIABILE Y_1.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 21.58%
Confusion matrix:
                     sconfitta o pareggio vittoria class.error
sconfitta o pareggio                  163       37       0.185
vittoria                               45      135       0.250
```

Figura 11: previsione della Random Forest per la prima variabile obiettivo ottenuta attraverso l'utilizzo del database contenente solo le variabili più significative.

Confrontando i risultati appena ottenuti, con quelli prima descritti, possiamo notare un significativo miglioramento nella previsione del modello; infatti osserviamo che l'errore di previsione diminuisce di quasi 4 punti percentuali, ed allo stesso tempo anche la previsione in merito alle due modalità presenti al suo interno migliora rispettivamente di circa 3 e 4 punti percentuali.

In generale, possiamo ritenerci molto soddisfatti, in quanto questo database ci ha permesso di ridurre significativamente l'errore di previsione in merito alla variabile obiettivo appena descritta. Questo miglioramento, può essere imputato al fatto che, riducendo il database e prendendo in considerazione solo le variabili maggiormente significative, si riduce anche il "rumore" presente nei dati e quindi, come avevamo anticipato nella parte teorica in merito alle Random Forest, solitamente a fronte di una riduzione di rumore all'interno del database si riscontra un miglioramento in previsione del modello.

Per dare un giudizio definitivo sull'efficienza del nuovo database, è necessario valutare anche il comportamento del modello applicato alla terza variabile obiettivo, la quale era stata mantenuta nelle nostre analisi seppur non presentasse risultati pienamente soddisfacenti.

VARIABILE Y_3.

Notiamo con grande piacere che anche la previsione in merito alla terza variabile obiettivo è significativamente migliorata, guadagnando quasi 8 punti percentuali rispetto all'errore di classificazione nel suo complesso.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 33.95%
Confusion matrix:
          vittoria sconfitta pareggio class.error
vittoria       155        11       14   0.1388889
sconfitta       14        69       21   0.3365385
pareggio        40        29       27   0.7187500
```

Figura 12: previsione della Random Forest per la terza variabile obiettivo, ottenuta attraverso l'utilizzo del database contenente solo le variabili più significative.

Osservando attentamente le singole modalità riferite a questa variabile obiettivo osserviamo un netto miglioramento sia della modalità pareggio, che precedentemente poteva essere considerata disastrosa con un errore di previsione pari al 90%, sia della modalità sconfitta, che, sempre in precedenza, si aggirava intorno ad un errore del 50%.

A questo punto, possiamo affermare che anche la variabile Y_3 può essere considerata interessante ai fini della nostra analisi, visto e considerato che l'errore di previsione generale è migliorato significativamente in corrispondenza della selezione delle variabili esplicative. Inoltre ora, anche i dati relativi alle previsioni delle singole modalità hanno un senso logico; più precisamente, per la modalità vittoria, quasi tutte le partite che dovevano essere classificate in questo modo vengono riconosciute come tali e solo un piccolissimo numero di queste viene classificato in maniera errata, dividendosi quasi equamente tra le restanti due modalità. Anche per quanto riguarda la modalità sconfitta, la maggior parte delle partite viene classificata in maniera corretta. L'unica modalità che ha un comportamento leggermente in controtendenza rispetto alle precedenti è il pareggio, infatti in questo caso la maggior parte delle partite viene classificata in maniera errata, tuttavia questa modalità, se ci pensiamo attentamente, è la più critica in assoluto anche se non venisse considerata sotto l'aspetto statistico in quanto molto vicina ad entrambi i risultati (vittoria e sconfitta).

Molto spesso, come sanno perfettamente gli appassionati di questo sport, accade che vi sia la predominanza netta di una squadra rispetto all'avversaria, eppure la partita si conclude con un pareggio solo perché la seconda squadra è stata molto più fortunata in quanto, a fronte di un numero decisamente inferiore di occasioni, è riuscita a segnare lo stesso numero di reti. È il "bello" del calcio. Le componenti di casualità e di fortuna rendono difficile la previsione dei risultati attraverso modelli matematici, i quali cercano di trovare delle regole meticolose per riuscire a classificare le varie osservazioni (partite) in maniera corretta. Forse, è proprio

questa componente aleatoria che tiene incollati milioni di telespettatori allo schermo durante le dirette, molto spesso facendo innervosire un gran numero di donne che preferirebbero fare qualcosa di diverso la domenica pomeriggio.

Tornando a noi, possiamo concludere che, a fronte di queste considerazioni possiamo ritenerci pienamente soddisfatti, sia per quanto riguarda il modello utilizzato che per il database selezionato e quindi proseguire con altre analisi.

3.4.4 ANALISI DELLE COMPONENTI PRINCIPALI.

La nostra prossima "sfida" è quella di cercare di ridurre nuovamente il numero di variabili contenute nel database, principalmente sostituendo ad esse dei fattori che possono essere considerati delle sintesi delle variabili precedenti. Per fare questo tipo di operazione ci affidiamo al metodo dell'analisi delle componenti principali, introdotta da H. Hotelling 1953, la quale consente, partendo da una matrice dei dati di dimensioni $n \times p$ (con tutte variabili quantitative), di sostituire alle p variabili tra loro correlate un nuovo insieme di variabili (chiamate componenti principali) che godono di due importantissime proprietà:

1) sono tra loro incorrelate (o meglio ortogonali);
2) sono elencate in ordine decrescente rispetto alla loro varianza.

La prima componente principale rappresenta la combinazione lineare delle p variabili di partenza avente massima varianza; la seconda componente principale è la combinazione lineare delle p variabili con varianza immediatamente inferiore, soggetta al vincolo di essere ortogonale alla componente precedente e così via.

Se le p variabili sono fortemente correlate, un numero k di componenti principali ($k < p$) tiene conto di una quota elevata di varianza totale, per cui ci si può limitare a considerare solo tali componenti, trascurando le restanti $(p - k)$. Possiamo quindi affermare che l'interesse operativo dell'analisi delle componenti principali si manifesta nei casi in cui poche componenti principali sono in grado di "spiegare" una percentuale elevata di varianza totale, poiché in tale circostanza la sostituzione delle componenti principali alle variabili di partenza soddisfa un criterio di parsimonia, con una perdita di informazioni limitata e con un deciso miglioramento nell'interpretabilità del *pattern* dei dati.

Ogni componente principale, come si è detto, è una combinazione lineare di tutte le variabili di partenza. Tuttavia, se in una certa analisi si sono estratte più componenti, l'interpretazione delle stesse può essere resa più agevole se ogni componente principale è effettivamente

collegata solo ad un sottoinsieme di variabili, mentre le restanti hanno un'importanza trascurabile. Per riuscire ad avvicinarsi a questa situazione "ideale" è possibile ruotare gli assi di riferimento, in modo che ciascun autovettore abbia alcuni elementi prossimi a zero. La rotazione ortogonale delle prime *k* componenti principali lascia immutata la quota di varianza totale complessivamente spiegata dalle medesime, ma la prima componente ruotata non è più la combinazione lineare delle variabili di partenza avente massima varianza, e così pure per le successive componenti principali. In altri termini, le quote di varianza spiegata dalle componenti principali ruotate non coincidono con quelle prima della rotazione, pur essendo identica la loro somma. Questa tecnica è fortemente consigliata quando si procede all'analisi dei fattori (Sergio Zani e Andrea Cerioli, 2007).

Dopo questi brevi cenni teorici torniamo alla fase applicativa della nostra tesi cercando di suddividere le variabili che compongono il nostro database in "blocchi" a seconda del loro significato, in modo da trovare delle variabili di "sintesi" che abbiano un senso logico. Considerando le variabili a nostra disposizione siamo riusciti ad individuare quattro blocchi in cui queste possono essere classificate. Più precisamente, vi sono alcune variabili in merito alla pericolosità della squadra di casa, altre riferite alla pericolosità della squadra ospite, altre ancora alla difesa della squadra ospite ed infine le variabili differenza, le quali non possono essere classificate in nessuna delle precedenti in quanto questa prerogativa dipenderebbe dal valore assunto, per ogni osservazione, da ognuna di esse. Infatti quando assumono valore positivo significa che la squadra di casa ha un punteggio più elevato e quindi può essere considerata migliore rispetto alla ospite per quanto riguarda quella determinata variabile; viceversa, invece quando il valore è negativo.

Di seguito verranno riportate le variabili contenute in ogni "blocco" da noi individuato, per facilitare la comprensione delle analisi successive.

1) Pericolosità della squadra di casa:
   - O_OCCAS_C: numero di occasioni della squadra di casa;
   - O_TIRI1_C: tiri dentro;
   - G_COLPTESTA1_C: colpi di testa in area avversaria;
   - O_CROSS_C: numero totale di cross effettuati dalla squadra.
2) Pericolosità della squadra ospite:
   - O_PALLE2_O: palle a scavalcare il centrocampo;
   - O_OCCAS_O: numero di occasioni della squadra ospite;
   - O_TIRI1_O: tiri dentro.

3) Difesa della squadra ospite:
   - D_RECUPERI14_O: recuperi aerei in area effettuati dalla squadra ospite;
   - D_PALLE17_C: palle perse effettive in attacco dalla squadra di casa.
4) Variabili differenza:
   - DIF_O_OCCAS: differenza di occasioni tra la squadra di casa e quella ospite;
   - DIF_O_TIRI1: differenza di tiri dentro tra le due squadre;
   - DIF_O_%ATT: differenza di percentuale di attacco alla porta;
   - DIF_D_%PROT: differenza di percentuale di protezione dell'area;
   - DIF_G_COLPTESTA1: differenza di colpi di testa in area avversaria;
   - DIF_G_GIOC_SP: differenza di giocate spettacolari;
   - DIF_O_CROSS: differenza di numero di cross;
   - DIF_D_RECUPERI14: differenza di recuperi aerei in area;
   - DIF_O_PALLE2: differenza di numero di palle a scavalcare il centrocampo;
   - DIF_G_PALLE6:differenza di numero di palle giocate sulla fascia centrale in difesa;
   - DIF_D_PALLE17: differenza di palle perse effettive in attacco;
   - DIF_O_CROSS12: differenza di cross su azione;
   - DIF_G_VEL_GIOC: differenza di velocità di gioco;
   - DIF_O_TIRI12: differenza di tiri di piede con palla bassa.

Una volta effettuate queste distinzioni in relazione alle variabili rimaste nel nostro database, abbiamo deciso di applicare l'analisi delle componenti principali separatamente ad ognuno di questi blocchi, in modo da ottenere dei fattori che sintetizzino le variabili al proprio interno.

Partiamo con le variabili che stanno ad indicare la pericolosità della squadra di casa.

Da entrambi i grafici sottostanti notiamo che il numero migliore di fattori necessari per sintetizzare le quattro variabili in merito a questo sottoinsieme sono due. E più precisamente:

Fattore 1:	Fattore 2:
- O_OCCAS_C (+)	- G_COLPTESTA1_C (+)
- O_TIRI1_C (+)	- O_CROSS_C (+)

Tenuto conto delle variabili contenute in ciascuno dei due fattori, possiamo rinominarli nel seguente modo: *fase d'attacco della squadra di casa* e *indice di pericolosità aerea della squadra di casa*.

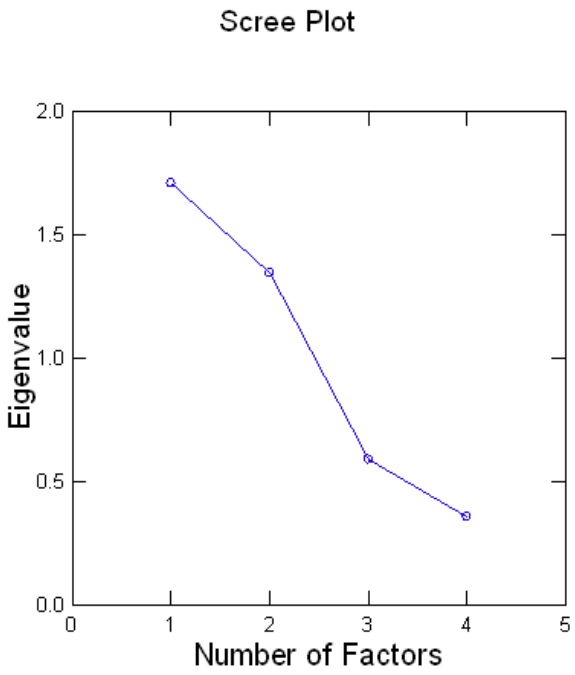


Figura 13: analisi delle componenti principali per le variabili relative alla pericolosità della squadra di casa.

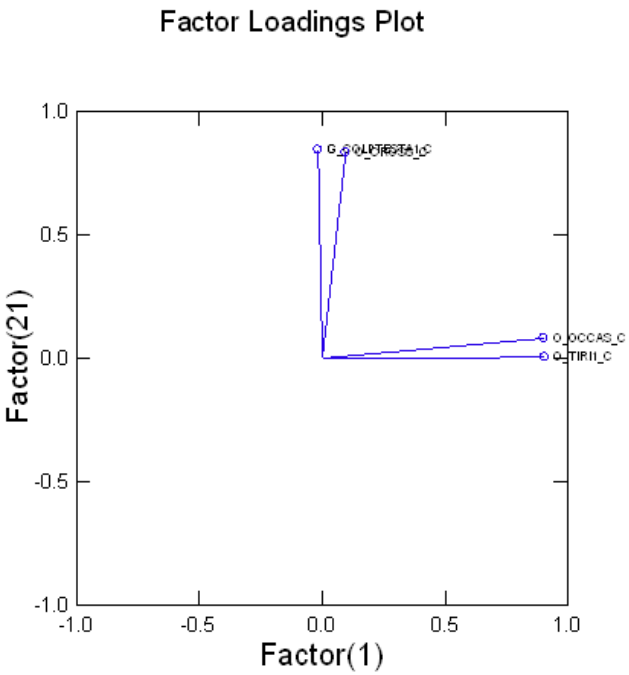


Figura 14: Factor Loadings Plot che evidenzia che sono necessari due fattori per sintetizzare le variabili.

I due fattori ottenuti spiegano più del 76% della varianza totale e questo è un ottimo risultato.

Percent of Total Variance Explained	
Fattore 1	**Fattore 2**
41.016	35.314

Le singole variabili all'interno di ciascun fattore hanno tutte segno positivo e questo sta ad indicare il fatto che, al crescere di una di esse, aumentano simultaneamente anche le altre (correlazione positiva).

Passiamo ora ad analizzare le variabili relative alla pericolosità della squadra ospite.

Component Loadings	
	Fattore 1
O_PALLE2_O	-0.387
O_OCCAS_O	0.895
O_TIRI1_O	0.904

Percent of Total Variance Explained
58.932

Le variabili contenute all'interno di questo sottoinsieme sono state riassunte in un unico fattore; che verrà nominato come *indice di pericolosità della squadra ospite*.

Tale fattore sintetizza ben tre variabili riuscendo ad ottenere una percentuale di varianza totale spiegata di quasi il 59%. Rispetto ai fattori delle precedenti variabili, sopra analizzate, notiamo che, per questo fattore, le variabili al suo interno non sono tutte correlate in maniera positiva (e quindi tutte con segno +), ma la variabile riferita alle *palle a scavalcare il*

centrocampo è correlata in maniera negativa rispetto alle altre due. Questa correlazione inversa sta ad indicare che al crescere di questa variabile, le altre due diminuiscono e viceversa.

Inizialmente, questa considerazione, ci ha lasciati leggermente sorpresi, ma poi analizzando attentamente il significato di queste variabili è emerso che una squadra detiene un numero elevato di *palle a scavalcare il centrocampo* quando si trova in una sorta di difficoltà, quasi sempre dovuta ad una maggiore forza a centrocampo della squadra avversaria. Data questa difficoltà riscontrata dalla squadra ospite, per proseguire l'attacco, molto spesso vengono effettuati lanci lunghi che vanno a superare la zona di centrocampo nella quale gli avversari sono "più forti", in modo da poter proseguire con l'attacco. Naturalmente immaginiamo che al crescere del numero di *occasioni* create da una squadra crescano anche il numero di *tiri dentro* effettuati dalla stessa, ma allo stesso modo diminuisca il numero di *palle a scavalcare il centrocampo* in quanto, questa squadra, non si trova più nella situazione di difficoltà prima illustrata.

Anche per le variabili che rientrano nella difesa della squadra ospite si ripropone una situazione simile alla precedente. Vengono classificate tutte all'interno di un unico fattore che verrà rinominato come *fase di difesa della squadra ospite*. Questo fattore riesce a spiegare più dell'83% della varianza totale e le variabili contenute al suo interno sono entrambe correlate in maniera positiva.

Component Loadings

	Fattore 1
D_RECUPERI14_O	0.913
D_PALLE17_C	0.913

Percent of Total Variance Explained
83.419

Ora passiamo al blocco più consistente riguardante le 14 variabili espresse come differenza di valori tra quanto rilevato per la squadra di casa in relazione alle performance della squadra ospite.

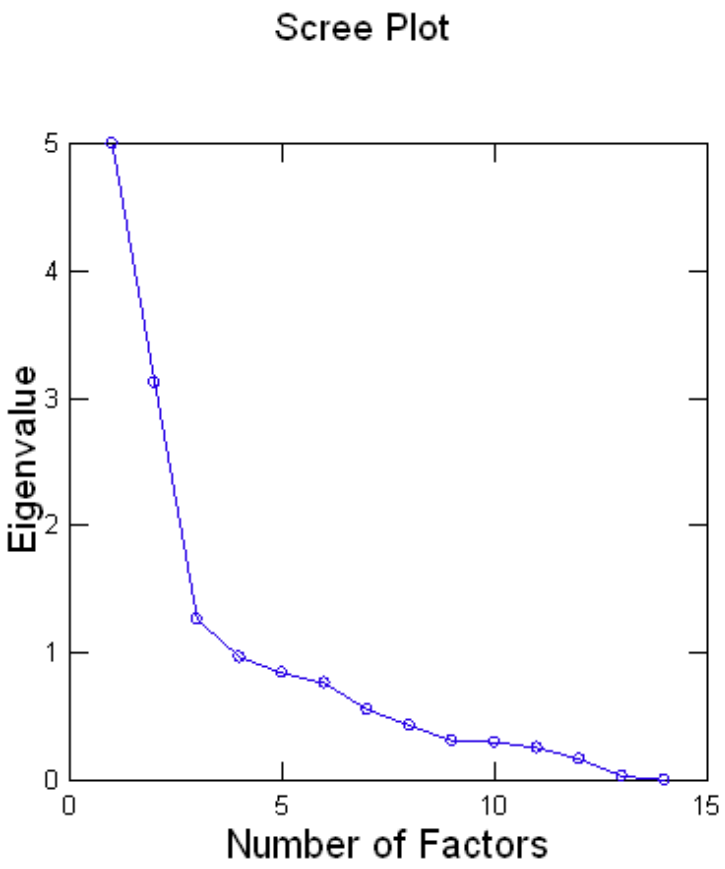


Figura 15: Scree Plot in merito al numero di fattori necessari per sintetizzare le 14 variabili differenza.

Osservando lo Scree Plot fornito, notiamo subito che senza ombra di dubbio il numero ottimale di fattori da utilizzare è 3. Infatti constatiamo che dopo il terzo pallino vie è un cambio netto di pendenza e pertanto i primi tre fattori riescono a sintetizzare in maniera più che buona le quattordici variabili oggetto della nostra attuale analisi. Questa tesi è avvalorata dal fatto che congiuntamente questi tre fattori riescono a spiegare più del 66% della varianza totale.

Percent of Total Variance Explained		
Fattore 1	**Fattore 2**	**Fattore 3**
33.732	24.082	9.196

Analizziamo ora le variabili contenute all'interno di ciascuno di questi fattori:

Fattore 1
- DIF_G_COLPTESTA1 (+)
- DIF_G_PALLE6 (-)
- DIF_D_RECUPERI14 (-)
- DIF_D_PALLE17 (+)
- DIF_O_CROSS12 (+)
- DIF_O_PALLE2 (-)
- DIF_O_CROSS (+)

Fattore 2
- DIF_D_%PROT (+)
- DIF_O_TIRI1 (+)
- DIF_O_%ATT (+)
- DIF_O_OCCAS (+)
- DIF_O_TIRI12 (+)

Fattore 3
- DIF_G_VEL_GIOC (+)
- DIF_G_GIOC_SP (+)

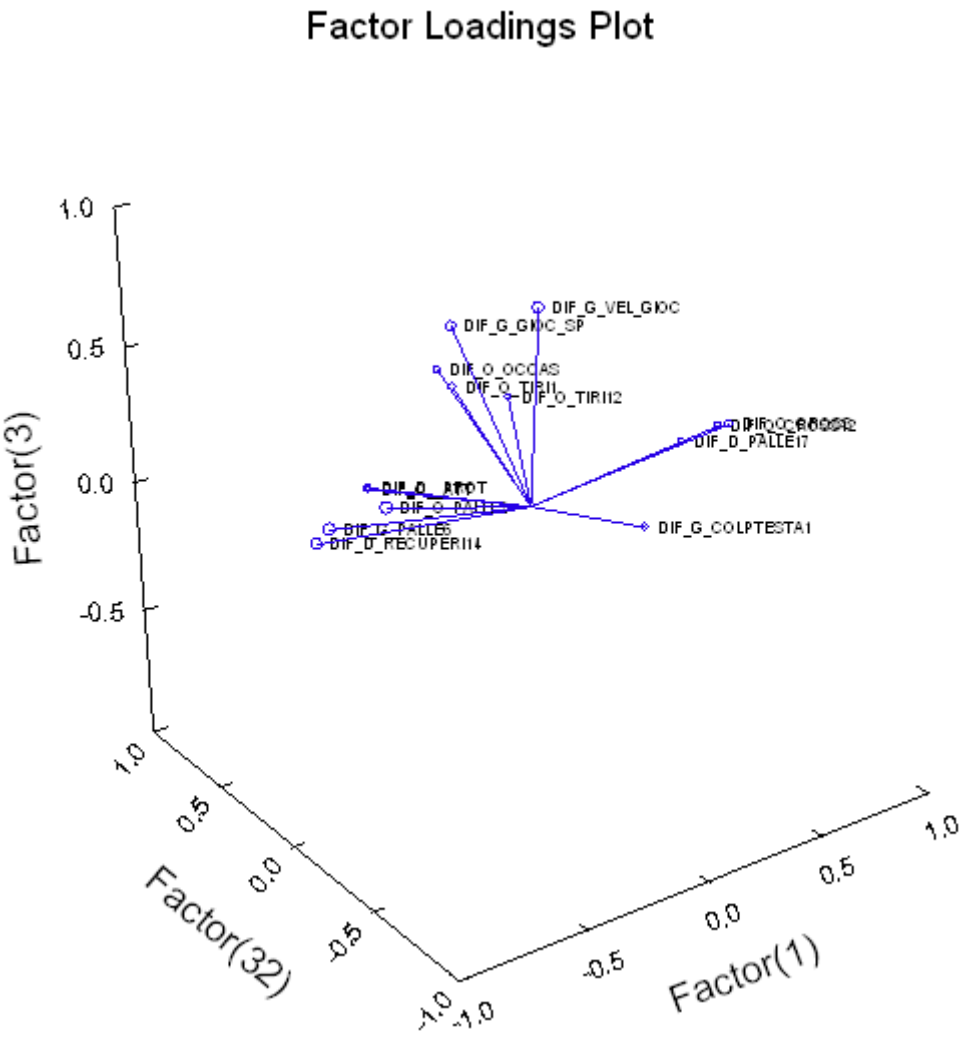


Figura 16: Factor Loadings Plot dei tre fattori riferiti alle variabili "differenza".

Questa volta è più difficile individuare esattamente la divisione delle variabili nei vari fattori basandoci esclusivamente sull'osservazione del grafico sopra illustrato. Trattandosi di uno spazio tridimensionale, sarebbe stato necessario riuscire a ruotare il grafico per osservare esattamente quali variabili potevano essere considerate in merito ad un fattore e quali invece no. In questo caso, abbiamo osservato la matrice dei fattori ruotati e attraverso questa siamo riusciti ad attribuire a ciascun fattore le variabili che presentavano il valore assoluto più elevato.

A questo punto è necessario rinominare in maniera significativa i tre fattori creati sulla base delle *variabili differenza*, basandoci sulle variabili contenute al loro interno. Rispetto alle altre classificazioni, questa è quella che presenta maggiori difficoltà interpretative, in quanto a seconda della partita può indicare la predominanza della squadra di casa rispetto all'avversaria o viceversa. Dopo lunghe riflessioni su questi fattori siamo riusciti ad attribuire un "titolo" ad ognuno di essi. Il primo fattore sarà l'*indice di sviluppo di azione in fase*

offensiva della squadra di casa in relazione alla squadra ospite, il secondo fattore rappresenterà l'*indice di efficacia del gioco* della squadra di casa in relazione alla squadra ospite ed infine il terzo fattore indicherà la *briosità del gioco* della squadra di casa sempre in relazione alla squadra ospite.

Arrivati a questo punto abbiamo creato un nuovo database contenente solo i sette fattori risultati dall'analisi delle componenti principali, in modo da poter riapplicare il metodo delle Random Forest e verificare se vi è stato un brusco od un leggero calo di precisione, o perché no, magari anche un leggero miglioramento. Quest'ultima possibilità è abbastanza remota considerando il fatto che, tramite questo tipo di analisi, immancabilmente si perde una parte d'informazione (come prima anticipato nei cenni di teoria).

Ricapitolando le uniche variabili all'interno del database al momento sono:

1) fase d'attacco della squadra di casa;
2) indice di pericolosità aerea della squadra di casa;
3) indice di pericolosità della squadra ospite;
4) fase di difesa della squadra ospite;
5) indice di sviluppo d'azione in fase offensiva della squadra di casa in relazione alla squadra ospite;
6) indice di efficacia del gioco della squadra di casa in relazione alla squadra ospite;
7) briosità del gioco della squadra di casa in relazione alla squadra ospite.

RISULTATI IN MERITO ALLA VARIABILE Y_1.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 20%
Confusion matrix:
                     sconfitta o pareggio vittoria class.error
sconfitta o pareggio                  163       37   0.1850000
vittoria                               39      141   0.2166667
```

Figura 17: previsioni ottenute, per la prima variabile obiettivo, attraverso l'utilizzo del database contenente le 7 componenti principali precedentemente calcolate.

Osservando i dati notiamo che a dispetto delle nostre aspettative l'indice OOB riferito a questa variabile obiettivo è migliorato di 1,5 punti percentuali. Questo miglioramento è legato alla maggior precisione nella previsione della modalità vittoria, la quale guadagna quasi 4

punti percentuali. Tale risultato è decisamente incoraggiante, in quanto passando da 23 variabili a 7 siamo riusciti addirittura ad ottenere un miglioramento nella precisione di previsione.

Naturalmente prima di dare il nostro giudizio definitivo su questo database dobbiamo osservare come questo si comporta prevedendo la terza variabile obiettivo.

RISULTATI IN MERITO ALLA VARIABILE Y_3.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 34.74%
Confusion matrix:
          vittoria sconfitta pareggio class.error
vittoria       152        14       14   0.1555556
sconfitta       14        63       27   0.3942308
pareggio        36        27       33   0.6562500
```

Figura 18: previsioni ottenute, per la terza variabile obiettivo, attraverso l'utilizzo del database contenente i 7 fattori ottenuti dall'analisi delle componenti principali.

Per quanto riguarda questa variabile obiettivo notiamo una leggera diminuzione nell'indice OOB (meno di un punto percentuale), legata al calo di precisione delle modalità vittoria e sconfitta. Tuttavia la modalità prima considerata "critica" (*pareggio*) migliora di ben 6 punti percentuali. Questo è un ottimo traguardo per la nostra capacità previsionale.

In conclusione possiamo ritenerci pienamente soddisfatti di questo nuovo database in quanto la capacità previsionale è quasi ai medesimi livelli di quella precedente contenente ben 23 variabili contro le attuali 7. Tuttavia l'interpretazione di queste sette variabili non è stata semplicissima, in particolare per quanto riguarda il blocco contenente tutte le *variabili differenza*. Per questo motivo abbiamo deciso di ripetere tutto il procedimento visto fin ora, escludendo, fin dalle prime fasi, tutte le *variabili differenza*, in modo da osservare se è possibile migliorare il modello sotto l'aspetto dell'interpretabilità dei dati, senza peggiorare di molto l'aspetto dell'accuratezza nella previsione.

3.4.5 RIPETIZIONE DELLE ANALISI SUL DATABASE NEL QUALE SONO STATE ELIMINATE TUTTE LE VARIABILI DIFFERENZA.

Iniziamo con il valutare l'accuratezza previsionale del modello applicato all'intero database escludendo le variabili differenza.

VARIABILE Y_1.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 21

        OOB estimate of  error rate: 28.95%
Confusion matrix:
                     sconfitta o pareggio vittoria class.error
sconfitta o pareggio                  155       45   0.2250000
vittoria                               65      115   0.3611111
```

Figura 19: applicazione della Random Forest alla prima variabile obiettivo in merito al database depurato da tutte le variabili differenza.

Osservando i risultati a nostra disposizione notiamo immediatamente che il nostro tasso di errore di stima peggiora di circa 4-5 punti percentuali. La maggior parte di questo peggioramento è legato all'accuratezza previsionale della modalità vittoria, la quale subisce quasi totalmente gli effetti dovuti a questo cambiamento di database.

Siamo disposti a tollerare questo lieve peggioramento in quanto, siamo passati da un database contenente 726 variabili ad uno di 487. Pertanto a fronte di una diminuzione di ben 239 variabili siamo disposti ad accettare un peggioramento di circa 5 punti percentuali.

Tutte queste considerazioni sono valide, naturalmente se alla fine di tutta questa analisi riusciremo ad ottenere un significativo miglioramento interpretativo degli indici costruiti sulla base di una sintesi delle variabili rimaste all'interno del nostro database, riapplicando la procedura appena illustrata per quanto riguardava l'intero database.

Riproducendo fedelmente la procedura prima affrontata per l'anali dei risultati ottenuti, passiamo ora ad analizzare nel dettaglio i diagrammi a barre in merito agli indici Mean Decrease in Accuracy e Gini Impurity. L'obiettivo è quello di riuscire ad individuare le variabili più significative per questo modello riapplicandolo successivamente alla selezione, in modo da riuscire a valutare i miglioramenti o peggioramenti previsionali legati a questa brusca riduzione dimensionale.

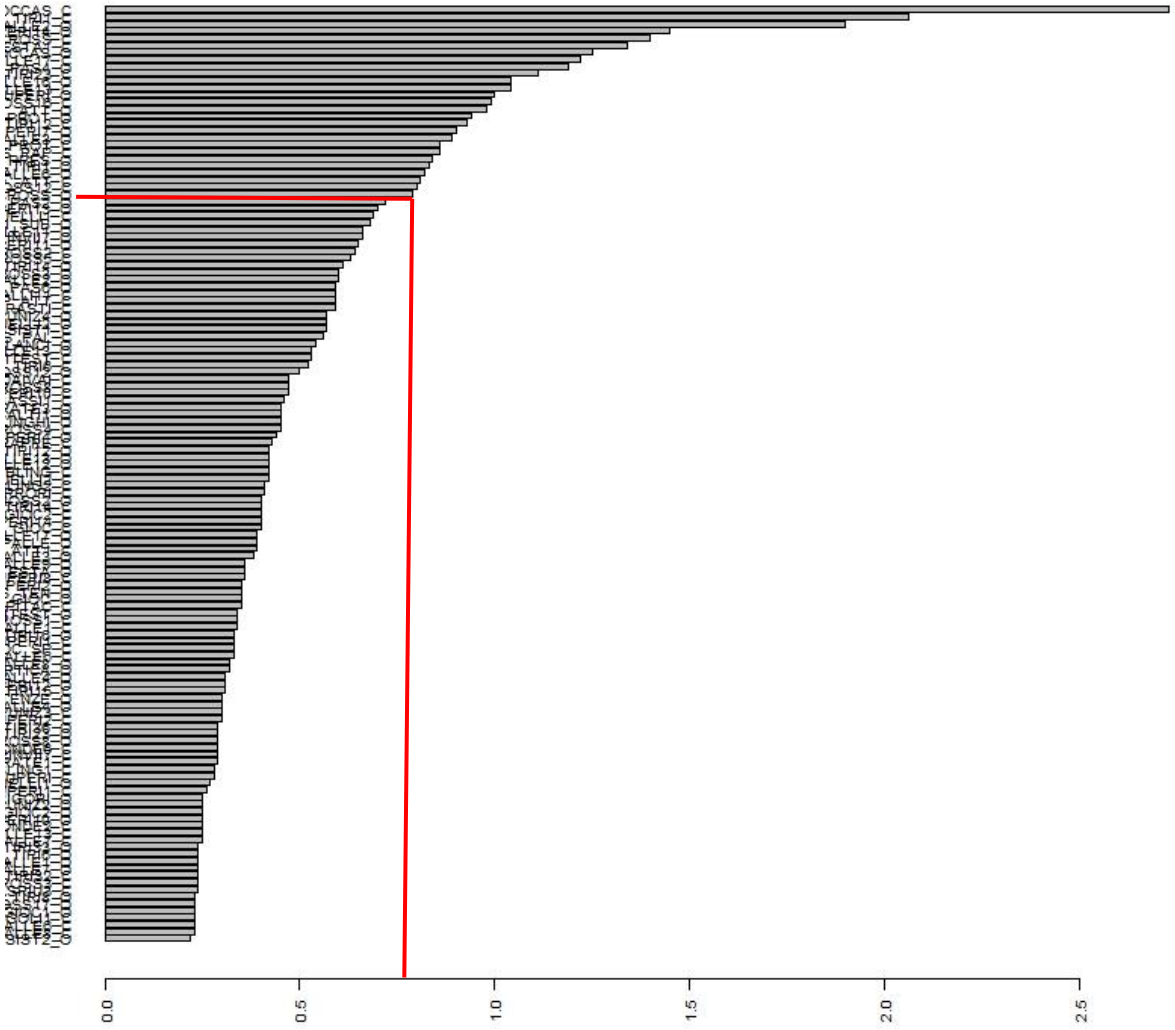


Figura 20: diagramma a barre riferito alle prime 120 variabili più significative per la determinazione della prima variabile obiettivo calcolata attraverso l'indice Mean Decrease in Accuracy legato al database depurato dalle variabili "differenza".

Come già appreso in precedenza, possiamo notare che in corrispondenza delle linee rosse l'importanza in merito alle variabili precipita formando la classica curvatura a gomito. Questo significa che le variabili successive a queste linee contribuiscono a spiegare il fenomeno in maniera sempre minore. Per quanto riguarda la prima variabile decisionale, la curvatura a gomito si verifica in corrispondenza del valore 0,75. Ripetendo l'analisi per un numero cospicuo di volte e osservando i grafici corrispondenti, abbiamo riscontrato che sempre in prossimità del valore 0,75 si verifica tale curvatura.

Successivamente abbiamo osservato l'andamento per quanto riguarda l'indice di Gini Impurity.

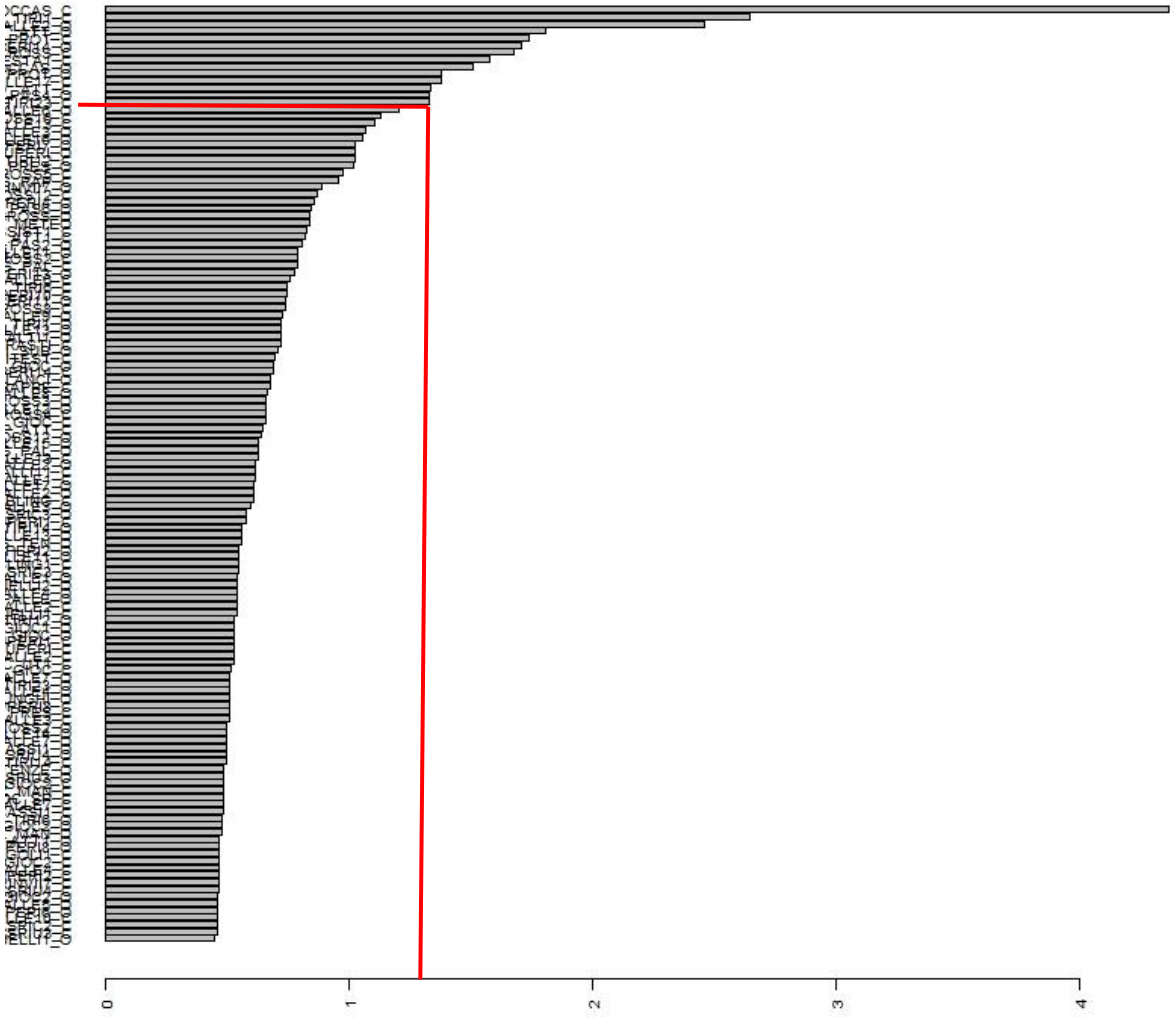


Figura 21: diagramma a barre delle prime 120 variabili calcolate attraverso l'indice Gini Impurity ottenuto sulla base della prima variabile obbiettivo in merito al database depurato da tutte le variabili "differenza".

Anche per questo database, le variabili calcolate attraverso l'indice Gini Impurity decrescono molto più velocemente rispetto al precendente indice Mean Descrease in Accuracy facendo si che la curvatura a gomito si manifesti molto prima.

Essendo nostra intenzione tenere in considerazione solo le variabili più importanti che vengono considerate tali da entrambi gli indici abbiamo deciso di tenere in considerazione per ogni iterazione tutte quelle variabili che assumono per questo indice valori maggiori a 1,3. Abbiamo stabilito questa soglia in quanto delimita l'area nella quale poi si registra la curvatura a gomito che avevamo preso come riferimento anche per il precedente indice. A questo punto, la procedura è stata ripetuta per un numero cospicuo di volte; tenendo sempre in

considerazione solo le variabili con valori superiori alla soglia appena identificata. Naturalmente, essendo abbastanza elevato questo valore abbiamo registrato una varianza abbastanza significativa in merito alle variabili più rilevanti, e pertanto è stato necessario procedere con diverse analisi in modo da registrare, in media, quali tra queste fossero quelle di maggior interesse per il modello in questione.

Tenendo in considerazione solo le variabili risultate maggiormente rilevanti per entrambi gli indici, definiamo le più significative per la determinazione di Y_1:

1. O_OCCAS_C: numero di occasioni della squadra di casa;
2. O_TIRI1_C: tiri dentro effettuati dalla squadra di casa;
3. O_PALLE2_O: numero di palle a scavalcare il centrocampo calciate dalla squadra ospite;
4. G_COLPTESTA1_C: colpi di testa in area avversaria effettuati dalla squadra di casa;
5. O_CROSS_C: numero di cross effettuati dalla squadra di casa;
6. O_OCCAS_O: numero di occasioni della squadra ospite;
7. D_RECUPERI14_O: numero di recuperi aerei in area effettuati dalla squadra ospite;
8. O_TIRI23_C: tiri di piede su palla bassa da azione effettuati dalla squadra di casa;
9. D_PALLE17_C: palle perse effettive in attacco dalla squadra di casa;
10. O_PAS4_O: passaggi lunghi utili da dietro effettuati dalla squadra ospite;
11. D_RECUPERI_O: numero di recuperi effettuati dalla squadra ospite in difesa;
12. D_RECUPERI7_O: recuperi in zona area effettuati dalla squadra ospite;
13. O_CROSS12_C: numero di cross su azione effettuati dalla squadra di casa;
14. O_CROSS16_C: numero di cross su calcio piazzato effettuati dalla squadra di casa;
15. D_%PROT_C: percentuale di protezione dell'area in merito alla squadra di casa;
16. O_PALLE3_O: numero di palle a scavalcare il centrocampo utili calciate dalla squadra ospite;
17. O_%ATT_O: percentuale di attacco alla porta effettuata dalla squadra ospite;
18. D_%PROT_O: percentuale di protezione dell'area in merito alla squadra ospite;
19. G_PALLE6_O: palle giocate dalla squadra ospite sulla fascia centrale in difesa;
20. O_%ATT_C: percentuale di attacco alla porta effettuata dalla squadra di casa;
21. O_TIRI12_C: tiri di piede effettuati dalla squadra di casa con palla bassa;
22. D_PALLE13_C: numero di palle perse effettivamente in attacco dalla squadra di casa;
23. D_PALLE16_O: palle perse effettive a centrocampo dalla squadra ospite;
24. G_PRES_O: pressing effettuato dalla squadra ospite;

25. O_CROSS5_C: cross ad uscire effettuati dalla squadra di casa;
26. G_PAS_RAP_C: numero di passaggi rapidi effettuati dalla squadra di casa;
27. D_RINVII7_O: rinvii effettuati dalla squadra ospite da zona arretrata.

VARIABILE Y_3.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 21

        OOB estimate of  error rate: 45.26%
Confusion matrix:
          vittoria sconfitta pareggio class.error
vittoria       162         8       10   0.1000000
sconfitta       52        39       13   0.6250000
pareggio        66        23        7   0.9270833
```

Figura 22: prima applicazione della Random Forest, alla terza variabile obiettivo, in merito al database depurato da tutte le variabili differenza.

Anche per quanto riguarda la variabile Y_3 ci troviamo di fronte ad un lieve peggioramento nell'accuratezza previsionale, infatti il modello perde circa 4 punti percentuali rispetto al primo database considerato. Tale peggioramento è dovuto principalmente alla modalità sconfitta, la quale regredisce di ben 13 punti percentuali.

Pur non trovandoci in una situazione particolarmente incoraggiante continuiamo la nostra analisi, tenendo sempre presente che il nostro obiettivo è quello di migliorare l'interpretazione degli indici da noi creati sulla base della sintesi delle variabili individuate attraverso questa prima fase.

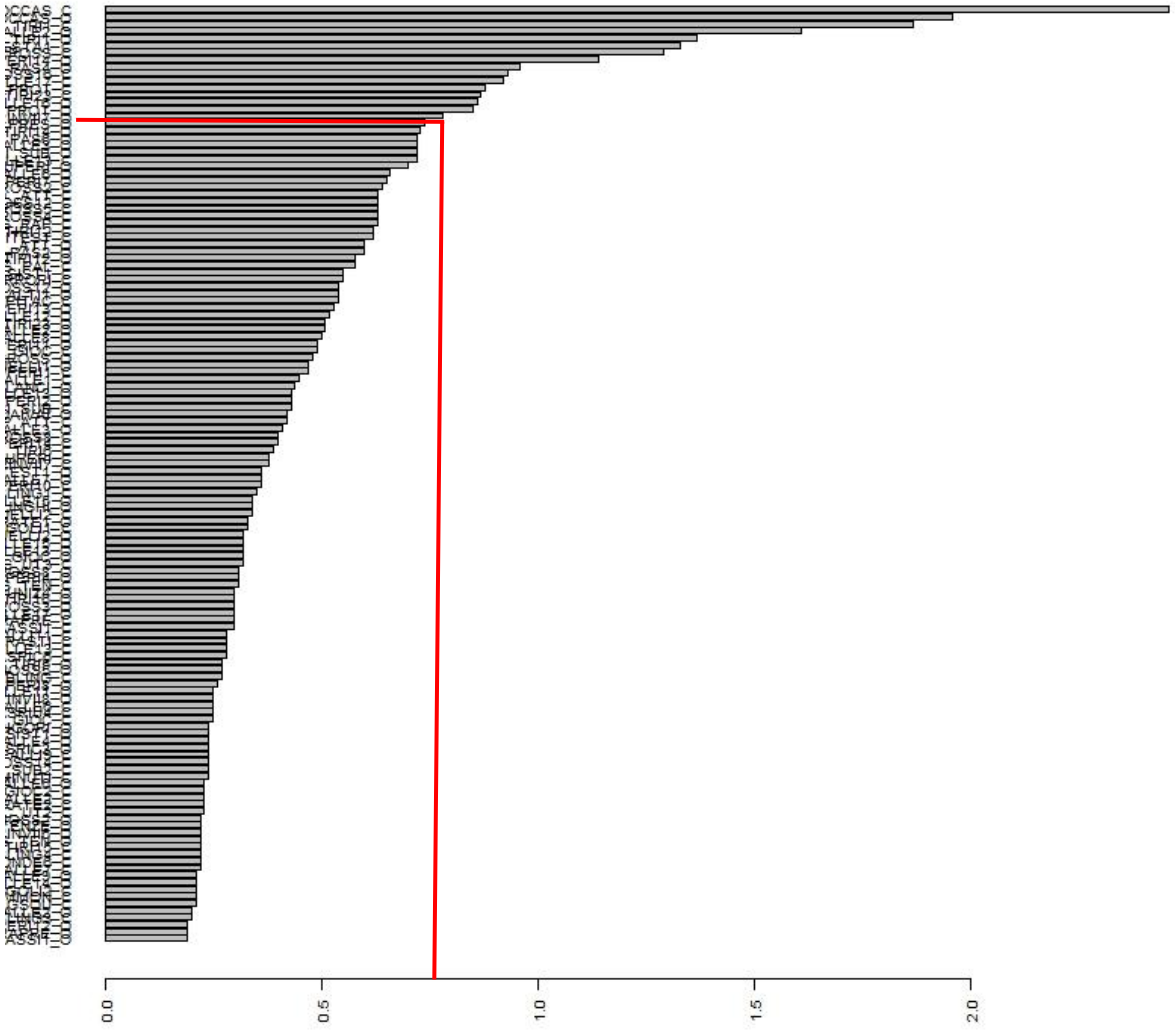


Figura 23: diagramma a barre riferito alle prime 120 variabili più significative per la determinazione della terza variabile obiettivo calcolata attraverso l'indice Mean Decrease in Accuracy legato al database depurato dalle variabili "differenza".

Come per la precedente variabile obiettivo, andiamo ad individuare il valore soglia per entrambi gli indici a nostra disposizione (Mean Decrease in Accuracy e Gini Impurity), in modo da individuare le variabili più significative fornite dal modello per la previsione della variabile Y_3.

Per quanto riguarda l'indice Mean Decrease in Accuracy notiamo che il valore soglia coincide con quello trovato per la variabile Y_1 (0,75); mentre per l'indice di Gini Impurity, sotto raffigurato, il valore soglia si aggira intorno all'1,25. Tale valore è molto vicino alla soglia sopra individuata per la prima variabile obiettivo.

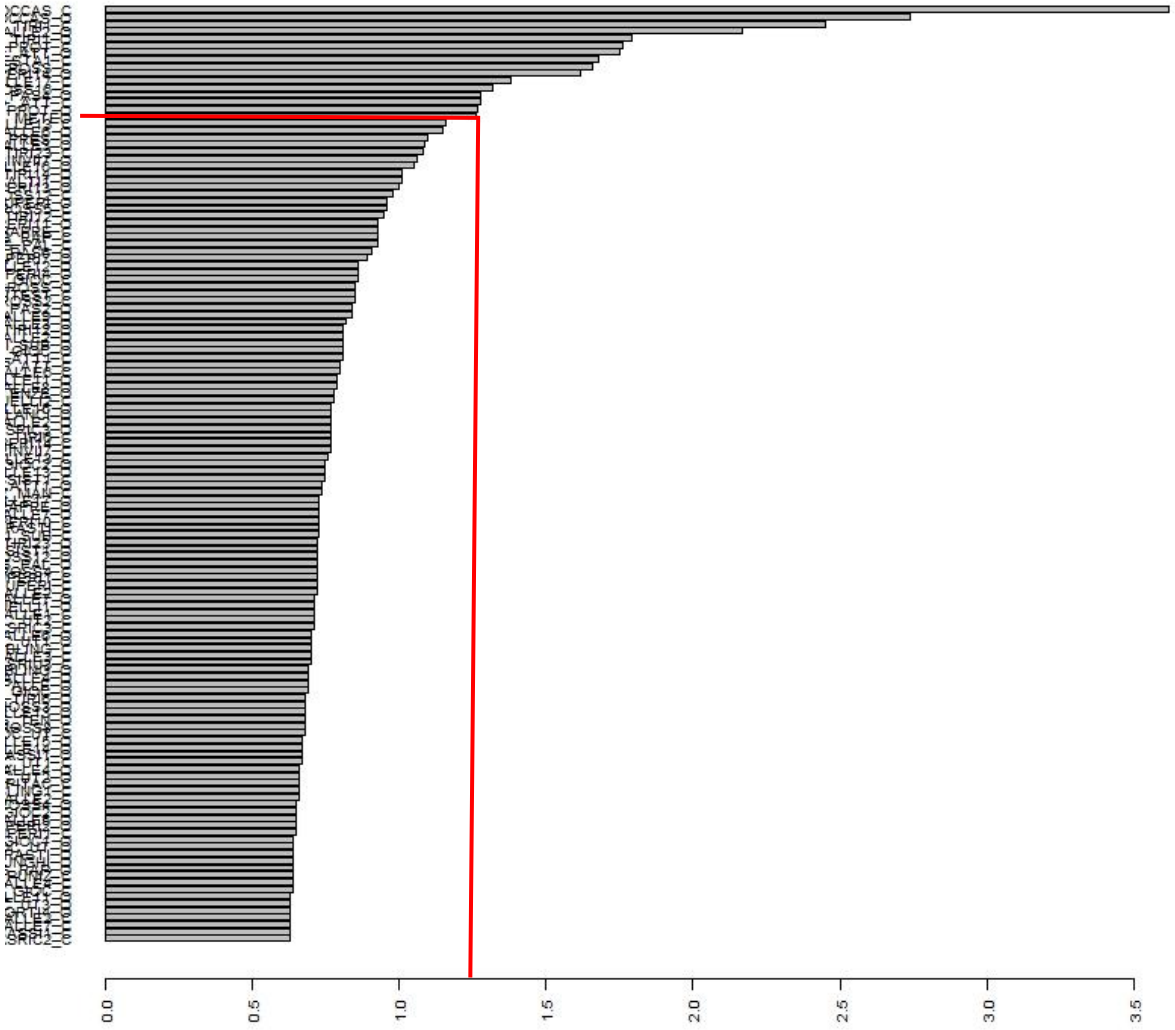


Figura 24: diagramma a barre delle prime 120 variabili calcolate attraverso l'indice Gini Impurity ottenuto sulla base della terza variabile obbiettivo in merito al database depurato da tutte le variabili "differenza".

Passiamo ora ad analizzare nel dettaglio le variabili più importanti per la determinazione della variabile $\boldsymbol{Y_3}$.

1. O_OCCAS_C: numero di occasioni create dalla squadra di casa;
2. O_OCCAS_O: numero di occasioni create dalla squadra ospite;
3. O_TIRI1_C: numero di tiri dentro calciati dalla squadra di casa;
4. O_PALLE2_O: numero di palle a scavalcare il centrocampo calciate dalla squadra ospite;
5. O_TIRI1_O: numero di tiri dentro effettuati dalla squadra ospite;
6. O_%ATT_O: percentuale di attacco alla porta in riferimento alla squadra ospite;
7. O_CROSS_C: numero di cross effettuati dalla squadra di casa;

8. D_RECUPERI14_O: numero di recuperi aerei in area effettuati dalla squadra ospite;
9. D_%PROT_C: percentuale di protezione dell'area della squadra di casa;
10. G_COLPTESTA1_C: colpi di testa della squadra di casa in area avversaria;
11. O_CROSS16_C: cross effettuati dalla squadra di casa su calcio piazzato;
12. D_PALLE17_C: palle perse effettive dalla squadra di casa in attacco;
13. O_PAS4_O: passaggi lunghi utili da dietro della squadra ospite;
14. D_%PROT_O: percentuale di protezione dell'area della squadra ospite;
15. O_%ATT_C: percentuale di attacco alla porta della squadra di casa;
16. G_PRES_O: pressing della squadra ospite;
17. G_PALLE6_O: palle giocate dalla squadra ospite sulla fascia centrale in difesa;
18. O_TIRI23_C : tiri di piede della squadra di casa su palla bassa da azione;
19. D_PALLE16_O: palle perse effettive a centrocampo dalla squadra ospite;
20. G_PASALTI1_O: numero di passaggi alti ricevuti dalla squadra ospite;
21. D_PALLE13_C: palle perse effettive in attacco dalla squadra di casa;
22. O_PALLE3_O: palle a scavalcare il centrocampo utili per la squadra ospite;
23. D_RINVII7_O: rinvii da zona arretrata effettuati dalla squadra ospite;
24. D_RECUPERI_O: numero di recuperi in difesa della squadra ospite;
25. O_CROSS5_C: cross ad uscire effettuati dalla squadra di casa.

Proprio come nella prima analisi, ci troviamo di fronte a variabili molto simili sia per quanto riguarda Y_1 che per Y_3. Più precisamente sono sei le variabili che trovano riscontro solo in una delle due analisi appena effettuate: D_RECUPERI7_O, O_CROSS12_C, O_TIRI12_C, G_PAS_RAP_C, O_TIRI1_O e G_PASALTI1_O. Le prime quattro sono contenute nell'elenco riferito alla prima variabile obiettivo, mentre le restanti due in quelle relative alla terza variabile decisionale.

Abbiamo deciso di creare un nuovo database contenente la selezione individuata per entrambe le variabili decisionali. In questo modo il nuovo database è formato da 29 variabili. Questa operazione è stata necessaria per poter riapplicare il modello alle sole variabili significative e verificare se la sua accuratezza migliora grazie alla riduzione del rumore prodotto dalle variabili meno importanti, o se ci troviamo di fronte ad un peggioramento previsionale in quanto sono state eliminate troppe variabili, alcune delle quali erano particolarmente rilevanti per la determinazione dei risultati.

Rispetto alle variabili prese in considerazione possiamo affermare che ci troviamo di fronte ad un'equa ripartizione tra variabili riguardanti la squadra di casa e quella ospite. Questo ci fa

subito capire che il modello "ragiona" in maniera appropriata in quanto considera ugualmente importanti per la determinazione del risultato finale delle partite i valori associati ad entrambe le squadre. Le variabili, rispetto al database precedente, sono molto variegate, a seconda che la squadra giochi in casa o come ospite, infatti solo quattro di queste sono state selezionate sia per l'una che per l'altra. Più precisamente sono la *percentuale di protezione dell'area*, le *occasioni* create, i *tiri dentro* effettuati e la *percentuale di attacco alla porta*.

Passiamo ora ad analizzare il comportamento della nostra Random Forest, applicata a questo nuovo database.

VARIABILE Y_1.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 5

        OOB estimate of  error rate: 24.74%
Confusion matrix:
                     sconfitta o pareggio vittoria class.error
sconfitta o pareggio                  158       42   0.2100000
vittoria                               52      128   0.2888889
```

Figura 25: seconda applicazione della Random Forest effettuata sulla selezione di variabili sopra riportata.

Riapplicando il nostro modello al database selezionato, riscontriamo un netto miglioramento nella capacità previsionale di ben 4 punti percentuali, la maggior parte dei quali dovuti alla modalità vittoria. Questa guadagna quasi 8 punti percentuali rispetto ai risultati precedentemente osservati.

VARIABILE Y_3.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 5

        OOB estimate of  error rate: 40.79%
Confusion matrix:
          vittoria sconfitta pareggio class.error
vittoria       147        14       19   0.1833333
sconfitta       24        59       21   0.4326923
pareggio        50        27       19   0.8020833
```

Figura 26: risultati della Random Forest per la terza variabile obiettivo ricavati dal database contenente la selezione di variabili.

Come per la variabile Y_1, possiamo riscontrare un netto miglioramento nell'accuratezza previsionale sia a livello generale, che per singole modalità. In particolare per le modalità sconfitta e pareggio si registra un brusco calo nell'errore di previsione, il quale porta di conseguenza ad un miglioramento generale del modello.

Proprio come nella nostra prima analisi notiamo che, addestrando il modello solo sulla base delle variabili rilevate come più importanti il modello migliora l'accuratezza previsionale. Tuttavia, sempre confrontando entrambe le analisi, notiamo che la capacità di previsione del modello è migliore utilizzando il database contenente le *variabili differenza*. Proprio per questo motivo è necessario procedere con l'Analisi delle Componenti Principali, in modo da appurare se questo peggioramento previsionale è giustificato da un miglioramento interpretativo degli indici che si verranno a creare.

Per procedere all'applicazione del modello delle PCA dobbiamo analizzare le variabili a nostra disposizione e suddividerle in blocchi a seconda del loro significato intrinseco. Queste possono essere classificate in 4 macro categorie e più precisamente:

1) Pericolosità della squadra di casa:
   - G_PAS_RAP_C: numero di passaggi rapidi effettuati dalla squadra di casa;
   - G_COLPTESTA1_C: numero di colpi di testa della squadra di casa in area avversaria;
   - O_TIRI12_C: tiri di piede su palla bassa;
   - O_CROSS12_C: numero di cross su azione effettuati dalla squadra di casa;
   - O_OCCAS_C: occasioni create dalla squadra di casa;
   - O_TIRI1_C: tiri dentro;
   - O_CROSS_C: cross totali della squadra;
   - O_CROSS16_C: cross su calcio piazzato;
   - O_%ATT_C: percentuale di attacco alla porta sempre in merito alla squadra di casa;
   - O_TIRI23_C: tiri di piede su palla bassa da azione;
   - O_CROSS5_C: cross ad uscire.
2) Pericolosità della squadra ospite:
   - G_PRES_O: pressing effettuato dalla squadra ospite;
   - G_PASALTI1_O: passaggi alti ricevuti dalla squadra ospite;
   - O_OCCAS_O: numero di occasioni create;
   - O_PALLE2_O: numero di palle a scavalcare il centrocampo;
   - O_TIRI1_O: tiri dentro;

- O_%ATT_O: percentuale di attacco alla porta sempre in merito alla squadra ospite;
- O_PAS4_O: passaggi lunghi utili da dietro;
- O_PALLE3_O: numero di palle a scavalcare il centrocampo utili.

3) Difesa della squadra di casa:
   - D_%PROT_C: percentuale di protezione dell'area della squadra di casa;
   - D_PALLE16_O: numero di palle perse effettivamente a centrocampo dalla squadra ospite.
4) Difesa della squadra ospite:
   - D_PALLE17_C: numero di palle perse effettivamente in attacco dalla squadra di casa;
   - D_PALLE13_C: numero totale di palle perse in attacco dalla squadra di casa;
   - G_PALLE6_O: palle giocate dalla squadra ospite sulla fascia centrale in difesa;
   - D_RECUPERI7_O: recuperi effettuati dalla squadra ospite in zona area;
   - D_RECUPERI14_O: recuperi aerei in area effettuati dalla squadra ospite;
   - D_%PROT_O: percentuale di protezione dell'area della squadra ospite;
   - D_RINVII7_O: rinvii da zona arretrata;
   - D_RECUPERI_O: recuperi effettuati in difesa nei confronti della squadra di casa.

Una volta effettuate queste distinzioni in relazione alle variabili rimaste nel nostro database, abbiamo deciso di applicare l'analisi delle componenti principali separatamente per ognuno di questi blocchi, in modo da ottenere dei fattori che sintetizzino le variabili al proprio interno.

Partiamo con le variabili che stanno ad indicare la pericolosità della squadra di casa.

Da entrambi i grafici sottostanti notiamo che il numero migliore di fattori necessari per sintetizzare le undici variabili in merito a questo sottoinsieme sono due. E più precisamente:

Fattore 1:
- G_COLPTESTA1_C (+)
- O_CROSS16_C (+)
- O_CROSS_C (+)
- O_CROSS12_C (+)
- O_CROSS5_C (+)

Fattore 2:
- G_PAS_RAP_C (+)
- O_TIRI1_C (+)
- O_OCCAS_C (+)
- O_%ATT_C (+)
- O_TIRI23_C (+)
- O_TIRI12_C (+)

Tenuto conto delle variabili contenute in ciascuno dei due fattori, possiamo considerarle come due diverse metodologie di gioco messe appunto dalla squadra di casa. Il primo di questi

fattori predilige le *giocate aeree* facendo ricorso prevalentemente ai cross e ai colpi di testa, mentre il secondo privilegia il *gioco pericoloso con palla bassa*, atto a costruire occasioni da gol.

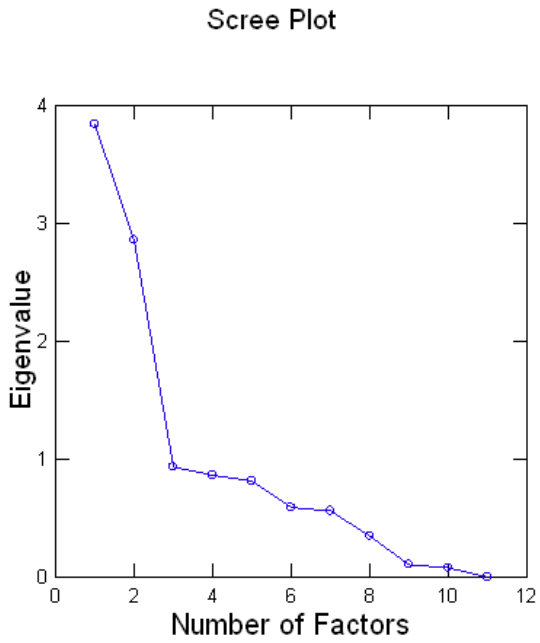


Figura 28: analisi delle componenti principali per le variabili relative alla pericolosità della squadra di casa.

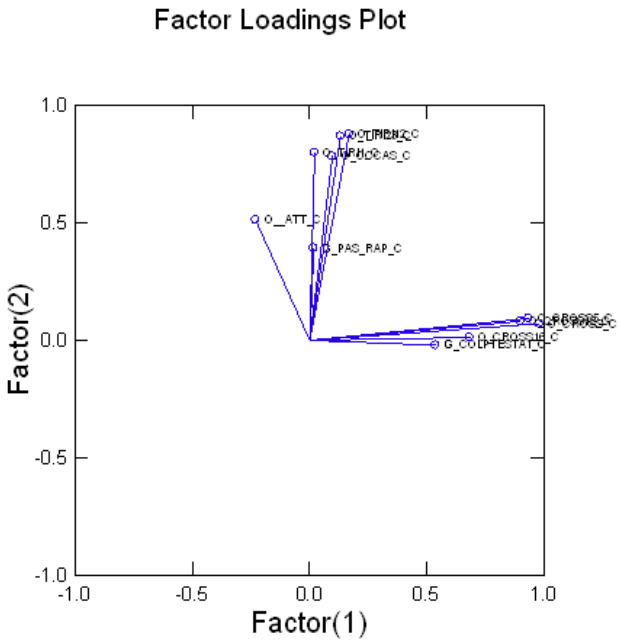


Figura 27: Factor Loadings Plot che evidenzia la necessità di due fattori per sintetizzare le variabili.

I due fattori ottenuti spiegano più del 60% della varianza totale e questo è un ottimo risultato.

Percent of Total Variance Explained	
Fattore 1	**Fattore 2**
31.808	29.119

Le singole variabili all'interno di ciascun fattore hanno tutte segno positivo e questo sta ad indicare il fatto che al crescere di una di esse aumentano simultaneamente anche le altre (correlazione positiva).

Passiamo ora ad analizzare le variabili relative alla pericolosità della squadra ospite.

Osservando lo Scree Plot, sotto riportato, notiamo subito che, senza ombra di dubbio, il numero ottimale di fattori da utilizzare è 3. Infatti possiamo osservare che dopo il terzo pallino vi è un cambio netto di pendenza e pertanto i primi tre fattori riescono a sintetizzare in maniera più che buona le otto variabili oggetto della nostra attuale analisi.

Questa tesi è avvalorata dal fatto che congiuntamente i tre fattori riescono a spiegare più del 72% della varianza totale.

Percent of Total Variance Explained		
Fattore 1	Fattore 2	Fattore 3
35.064	23.973	13.799

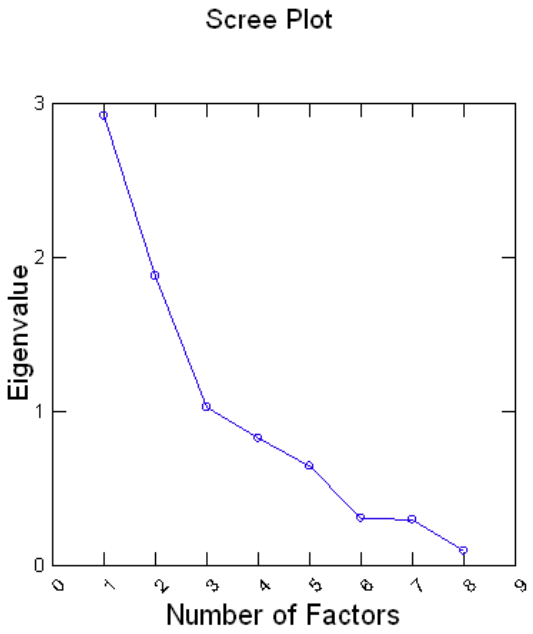


Figura 29: le otto variabili in merito alla pericolosità della squadra ospite vengono sintetizzate in tre fattori.

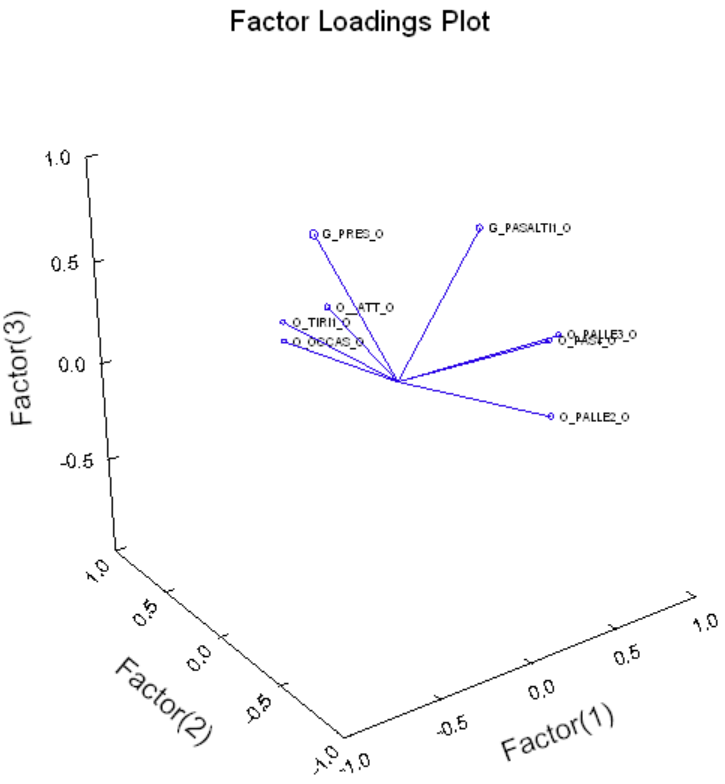


Figura 30: Factor Loadings Plot che permette di osservare come le variabili si dividono all'interno dei tre fattori individuati.

Analizziamo ora le variabili contenute all'interno di ciascuno di questi fattori:

Fattore 1	Fattore 2	Fattore 3
- O_PALLE2_O (+)	- O_OCCAS_O (+)	- G_PRES_O (+)
- O_PAS4_O (+)	- O_TIRI1_O (+)	- G_PASALTI1_O (+)
- O_PALLE3_O (+)	- O_%ATT_O (+)	

A questo punto è necessario rinominare in maniera appropriata i tre fattori ottenuti, sulla base delle variabili contenute al loro interno. Il primo fattore indica i *passaggi lunghi oltre il centrocampo*, il secondo la *pericolosità d'attacco* e il terzo i *tentativi di contropiede*, naturalmente in riferimento alla squadra ospite.

Rispetto alla prima analisi di questo elaborato, notiamo immediatamente che risulta molto più semplice attribuire un significato agli indici calcolati, sia per il significato intrinseco di ciascuna variabile contenuta al loro interno, sia per il fatto che tutte le variabili assumono il medesimo segno all'interno di ciascun fattore. Pertanto, questo significa che, sono tutte correlate in maniera positiva tra loro. Questo è un vantaggio non indifferente per le nostre analisi, in quanto ci mette nella condizione di capire effettivamente come variano i risultati al variare dei valori assunti da questi indici.

Il blocco contenente le variabili in merito alla difesa della squadra di casa, viene riassunto in un unico fattore, il quale riesce a spiegare più del 57% della varianza totale.

Component Loadings	
	Fattore 1
D__PROT_C	0.757
D_PALLE16_O	0.757

Percent of Total Variance Explained
57.314

Tale fatto non ci sorprende in quanto erano solamente due le variabili contenute all'interno di questo sottogruppo e pertanto è più che ragionevole che basti un unico fattore per sintetizzarle. Tale fattore misura l'*indice di difesa (o la fase difensiva) della squadra di casa*.

Passiamo quindi all'ultimo gruppo da analizzare: le variabili relative alla difesa della squadra ospite. In questo raggruppamento sono presenti otto variabili ed attraverso l'analisi delle componenti principali emerge chiaramente che sono necessari due fattori per sintetizzarle.

Percent of Total Variance Explained	
Fattore 1	**Fattore 2**
64.695	15.381

Come sempre siamo riusciti a fare queste deduzioni osservando l'andamento dello Scree Plot, sotto riportato, il quale evidenzia senza ombra di dubbio che sono necessari due fattori per sintetizzare al meglio le variabili in merito alla difesa della squadra ospite. Infatti, dopo il secondo fattore si registra un'inversione di tendenza, che sta ad indicare che i fattori successivi sintetizzano in maniera sempre minore le variabili.

Questo è confermato anche dall'andamento della varianza spiegata, la quale mostra che i due fattori, congiuntamente, spiegano più dell'ottanta per cento della varianza totale.

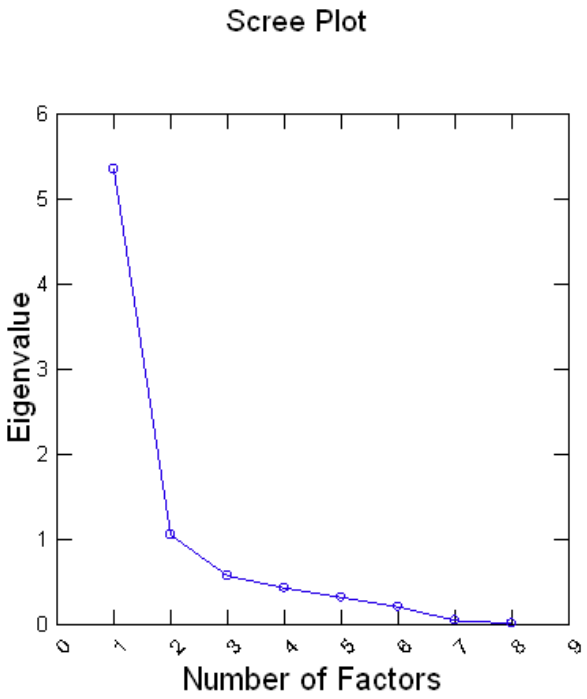


Figura 31: Scree Plot in merito alle otto variabili relative alla difesa della squadra ospite.

Osservando attentamente il grafico Factor Loadings Plot possiamo definire in maniera accurata come queste variabili si siano divise tra i fattori 1 e 2.

Fattore 1:

- D_PALLE17_C (+)
- D_PALLE13_C (+)
- G_PALLE6_O (+)
- D_RECUPERI7_O (+)
- D_RECUPERI14_O (+)
- D_RINVII7_O (+)
- D_RECUPERI_O (+)

Fattore 2:

- D_%PROT_O (+)

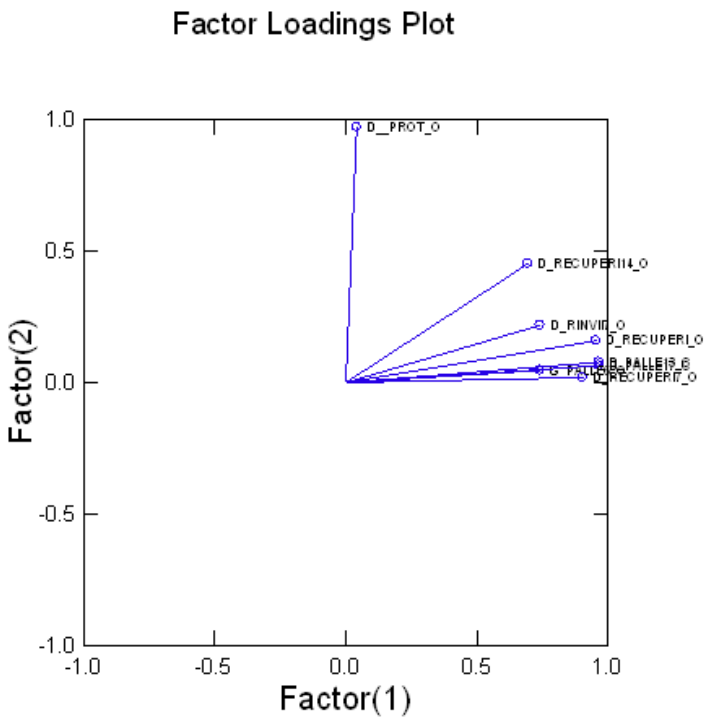


Figura 32: distinzione delle variabili contenute all'interno dei due fattori identificati per la selezione in riferimento alla difesa della squadra ospite.

A fronte di questa ripartizione delle variabili all'interno dei due fattori, possiamo rinominare quest'ultimi rispettivamente *fase difensiva protezione della porta* e *barriera difensiva a protezione area* della squadra ospite.

Arrivati a questo punto abbiamo creato un nuovo database contenente solo gli otto fattori ricavati dall'analisi delle componenti principali, in modo da poter riapplicare il metodo delle Random Forest e verificare se vi è stato un brusco od un leggero calo nella precisione delle previsioni; o perché no, magari anche un leggero miglioramento. Quest'ultima possibilità è abbastanza remota considerando il fatto che tramite questo tipo di analisi immancabilmente si perde una parte d'informazione.

Ricapitolando gli indici contenuti all'interno del database sono:

1) giocate aeree effettuate dalla squadra di casa;
2) giocate pericolose a palla bassa eseguite dalla squadra di casa;
3) passaggi lunghi oltre il centrocampo attuati dalla squadra ospite;
4) indice di pericolosità d'attacco della squadra ospite;
5) tentativi di contropiede della squadra ospite;
6) indice di difesa (o fase difensiva) della squadra di casa;
7) fase difensiva protezione della porta della squadra ospite;

8) barriera difensiva a protezione dell'area della squadra ospite.

RISULTATI IN MERITO ALLA VARIABILE Y_1.

Dai risultati previsionali ottenuti dall'applicazione del modello Random Forest al nuovo database contenente solo i fattori ottenuti attraverso l'analisi delle componenti principali, notiamo che il modello migliora di ben due punti percentuali rispetto all'applicazione sul precedente database.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 22.63%
Confusion matrix:
                     sconfitta o pareggio vittoria class.error
sconfitta o pareggio                  161       39   0.1950000
vittoria                               47      133   0.2611111
```

Figura 33: risultati previsionali per la prima variabile obiettivo, calcolata sugli 8 indici ottenuti attraverso la PCA.

RISULTATI IN MERITO ALLA VARIABILE Y_3.

```
Call:
 randomForest(x = x, y = y, ntree = nt, replace = TRUE, importance = TRUE,
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 38.42%
Confusion matrix:
          vittoria sconfitta pareggio class.error
vittoria       148        20       12   0.1777778
sconfitta       26        57       21   0.4519231
pareggio        39        28       29   0.6979167
```

Figura 34: risultati previsionali per la terza variabile obiettivo calcolata sugli 8 indici ottenuti attraverso la PCA.

Come per la variabile decisionale appena analizzata, anche l'errore di classificazione in merito alla variabile Y_3 migliora di oltre due punti percentuali, in particolare grazie ad un miglioramento sorprendente nell'accuratezza della modalità pareggio.

Questi risultati sono sorprendenti in quanto, anche a fronte di una notevole riduzione di dimensionalità, il risultato previsionale migliora di ben due punti percentuali per entrambe le variabili decisionali.

FASE 5: VALUTAZIONE.

Passiamo ora alla valutazione dei risultati ottenuti, raffrontando le due analisi oggetto di studio.

Confrontando i valori ottenuti dall'analisi Random Forest si nota che il database, contenente anche i fattori relativi alle *variabili differenza,* fornisce risultati previsionali migliori di circa due/tre punti percentuali rispetto a quanto ottenuto dalla seconda analisi. Tuttavia, è opportuno tenere in considerazione che l'interpretabilità degli indici ottenuti, attraverso quest'ultima analisi, è risultata nettamente superiore rispetto alla precedente.

A fronte di queste considerazioni abbiamo deciso di utilizzare entrambi i database, ma per scopi completamente diversi; infatti, i risultati ottenuti nella prima analisi (i quali risultano migliori per quanto riguarda l'accuratezza nelle previsioni) verranno utilizzati per stilare una classifica che permetterà di osservare come questa si discosta dalla realtà. I secondi, invece, presentano risultati migliori sotto il profilo dell'interpretazione, quindi verranno utilizzati per calcolare gli indici di pericolosità e di difesa in merito a ciascuna squadra facente parte del campionato italiano di Serie A 2010/2011.

FASE 6: IMPLEMENTAZIONE.

3.6.1 CLASSIFICA.

Le venti squadre protagoniste del campionato oggetto di studio, come già accennato in precedenza, si sono sfidate per ben 38 giornate tra partite di andata e di ritorno alla fine delle quali la capolista è risultata essere il Milan, con ben 8 punti di scarto sulla seconda classificata (Inter).

Per riuscire a stilare la classifica relativa alle previsioni ottenute attraverso la Random Forest ho iniziato ad attribuire punti alle varie squadre proprio a seconda del risultato fornitoci dal modello. In Appendice D è possibile osservare le previsioni per singola partita del campionato. Naturalmente, dato che attraverso il modello non è possibile tener conto dell'effetto aleatorietà, non possiamo assolutamente sperare che la classifica derivante da questo rispecchi fedelmente quella reale. Non a caso, una volta completata l'assegnazione dei vari punteggi e tirato le somme sui risultati finali di ciascuna squadra è emerso che secondo il modello l'Inter avrebbe dovuto vincere il campionato, mentre il Milan avrebbe dovuto classificarsi al secondo posto.

Ma confrontiamo più attentamente la classifica ottenuta dal modello con quella reale.

Classifica reale:

1)	Milan	82
2)	Inter	76
3)	Napoli	70
4)	Udinese	66
5)	Lazio	66
6)	Roma	63
7)	Juventus	58
8)	Palermo	56
9)	Fiorentina	51
10)	Genoa	51
11)	Catania	46
12)	Parma	46
13)	Chievo	46
14)	Cagliari	45
15)	Cesena	43
16)	Bologna	42
17)	Lecce	41
18)	Sampdoria	36
19)	Brescia	32
20)	Bari	24

Classifica ricavata dal modello:

1)	Inter	89
2)	Milan	77
3)	Udinese	72
4)	Juventus	65
5)	Roma	62
6)	Genoa	60
7)	Palermo	58
8)	Lazio	57
9)	Catania	56
10)	Cagliari	55
11)	Fiorentina	51
12)	Napoli	48
13)	Lecce	47
14)	Brescia	46
15)	Sampdoria	45
16)	Bari	41
17)	Cesena	39
18)	Parma	38
19)	Chievo	33
20)	Bologna	30

A livello di posizionamento in classifica non abbiamo ottenuto la corrispondenza di nessuna squadra in relazione alla classifica reale. Tuttavia, se osserviamo le squadre all'interno dei riquadri rossi possiamo notare che, seppur con diverso ordine, sei su otto sono comprese anche nella classifica appena redatta.

Anche le squadre che sarebbero dovute retrocedere in Serie B sono diverse dalla realtà, tuttavia, possiamo osservare dai riquadri blu, che queste assumono valori molto simili tra loro anche secondo le previsioni del modello, quindi questo per noi è un risultato incoraggiante.

Un risultato molto particolare lo riscontriamo all'interno dei riquadri color viola. Infatti la squadra della Fiorentina raggiunge un punteggio pari a 51, sia nella classifica reale che in quella redatta sulla base delle previsioni del modello. Questo indica che, seppur prevedendo in maniera diversa alcune partite, i punteggi sono finiti per compensarsi.

Proviamo ad analizzare le motivazioni per le quali le due classifiche si discostano in maniera così netta:

- innanzitutto nel calcio è presente una forte componente di casualità e di fortuna, da cui i modelli matematici e statistici sono in grado di depurare il risultato;
- nel corso del campionato, sono stati riscontrati errori arbitrali che hanno interferito sul risultato finale di alcune partite, avvantaggiando alcune squadre a scapito di altre. Tuttavia, questi fatti possono essere considerati sotto l'ipotesi di casualità dell'errore e associati alla componente fortuna che influenza l'andamento di ogni incontro;
- le partite concluse con un pareggio sono quelle in cui l'effetto aleatorietà è maggiormente presente, in quanto molto vicine ad entrambi gli altri due possibili risultati. In questi casi, molto spesso, la fortuna gioca un ruolo predominante;
- infine, dato tutto il clamore intorno allo scandalo calcio scommesse, potrebbe essere che alcune previsioni si discostino dai dati reali proprio perché le partite alle quali si riferiscono, sono state manipolate in modo da far si che il match si concludesse con il risultato desiderato da questi accaniti scommettitori. Tuttavia, l'inchiesta su "Calciopoli" non si è ancora conclusa, quindi non possiamo affermare con decisione questa ipotesi, anche se, siamo consapevoli che nel corso del campionato si sono verificate scorrettezze più o meno rilevanti che si sono ripercosse sui risultati delle partite.

Per analizzare in maniera più approfondita il legame esistente tra le due graduatorie a nostra disposizione ricorriamo all'indice di cograduazione. Si dice cograduazione (rank correlation) la metodologia statistica che studia le relazioni tra i posti d'ordine delle modalità di variabili quantitative oppure ordinali (Sergio Zani e Andrea Cerioli 2007).

Per misurare il livello di concordanza tra le due classifiche ricorriamo all'indice di cograduazione di Spearman:

$$_{S}I = 1 - \frac{6\sum_{i=1}^{N}(g_{i1} - g_{i2})^{2}}{N^{3} - N}$$

Dove g_1 e g_2 rappresentano rispettivamente le due graduatorie a nostra disposizione, ed N il numero di soggetti, che nel nostro caso corrisponde alle 20 squadre facenti parte del campionato.

L'indice ${}_SI$ assume valori compresi tra -1 e +1; più precisamente:

- ${}_SI = -1$ indica perfetta discordanza (contrograduazione) tra le due graduatorie (in pratica le due graduatorie sono completamente invertite: il primo è diventato l'ultimo, il secondo il penultimo e così via).
- ${}_SI = +1$ indica perfetta concordanza (cograduazione) tra le classifiche (le posizioni sono rimaste perfettamente uguali).
- Valori negativi di ${}_SI$ indicano discordanza tra le graduatorie in misura pari a ${}_SI\%$, dove la percentuale si intende rispetto al massimo teorico in caso di perfetta discordanza.
- Valori positivi di ${}_SI$ indicano concordanza tra le graduatorie, in misura pari a ${}_SI\%$ dove la percentuale si intende rispetto al valore massimo teorico in caso di perfetta corcondanza.

Squadra	Classifica reale	Classifica modello	$g_1 - g_2$	$(g_1 - g_2)^2$
Milan	1	2	-1	1
Inter	2	1	1	1
Napoli	3	12	-9	81
Udinese	4	3	1	1
Lazio	5	8	-3	9
Roma	6	5	1	1
Juventus	7	4	3	9
Palermo	8	7	1	1
Fiorentina	9	11	-2	4
Genoa	10	6	4	16
Catania	11	9	2	4
Parma	12	18	-6	36
Chievo	13	19	-6	36
Cagliari	14	10	4	16
Cesena	15	17	-2	4
Bologna	16	20	-4	16
Lecce	17	13	4	16
Sampdoria	18	15	3	9
Brescia	19	14	5	25
Bari	20	16	4	16

$\sum_{i=1}^{20}$ → **302**

$$ {}_SI = 1 - \frac{6 * 302}{8.000 - 20} = 0{,}773 $$

Tra le due graduatorie c'è una buona concordanza, pari al 77% circa del massimo teorico. Questo significa che vi è connessione tra la classifica reale e la classifica ottenuta attraverso il modello della Random Forest applicato alle componenti principali derivate dalla selezione di variabili individuate in precedenza. Questa tesi è avvalorata anche dall'indice ρ che assume valore pari a 0,755.

Sotto viene riportato un grafico che permette di visualizzare il posizionamento delle venti squadre del campionato tenendo in considerazione sia la classifica reale che quella ottenuta attraverso il modello Random Forest.

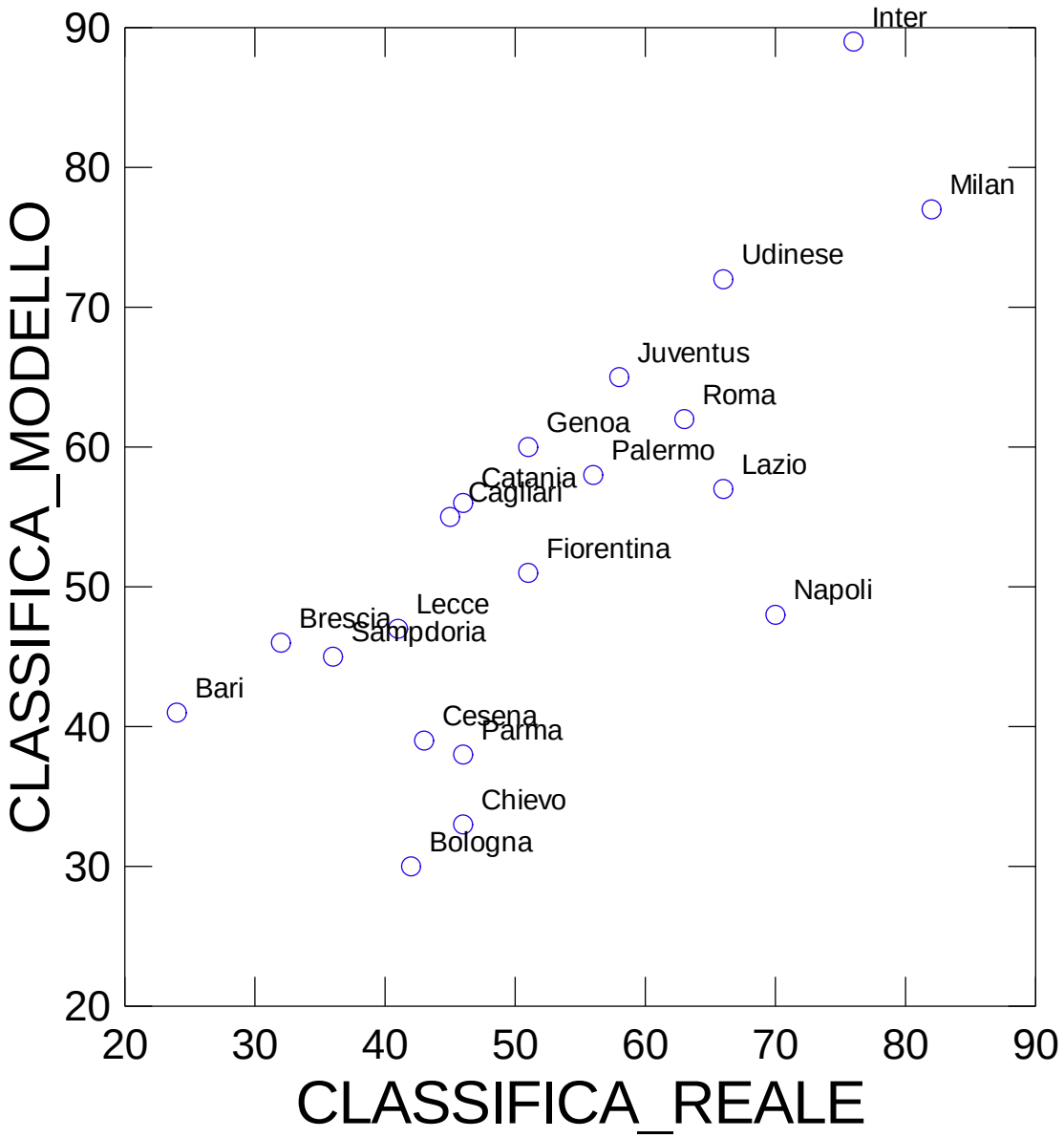


Figura 35: Scatter Plot che evidenzia il posizionamento delle squadre a seconda dei punteggi attribuititi dalle due classifiche a nostra disposizione.

3.6.2 ANALISI DELLE PREVISIONI OTTENUTE.

Il metodo della Random Forest assegna ad ogni modalità una probabilità e prevede, nel nostro caso specifico i risultati delle partite, restituendo quella a cui ha assegnato la percentuale più elevata. Pertanto costruendo una tabella, contenente per ogni singola partita la probabilità associata a ciascuna modalità e sottraendo, alla probabilità della modalità effettivamente realizzatasi, quella associata alla modalità più probabile, secondo i canoni del modello, ho costruito il grafico sottostante, ordinando i vari incontri a seconda della probabilità legata al risultato effettivo.

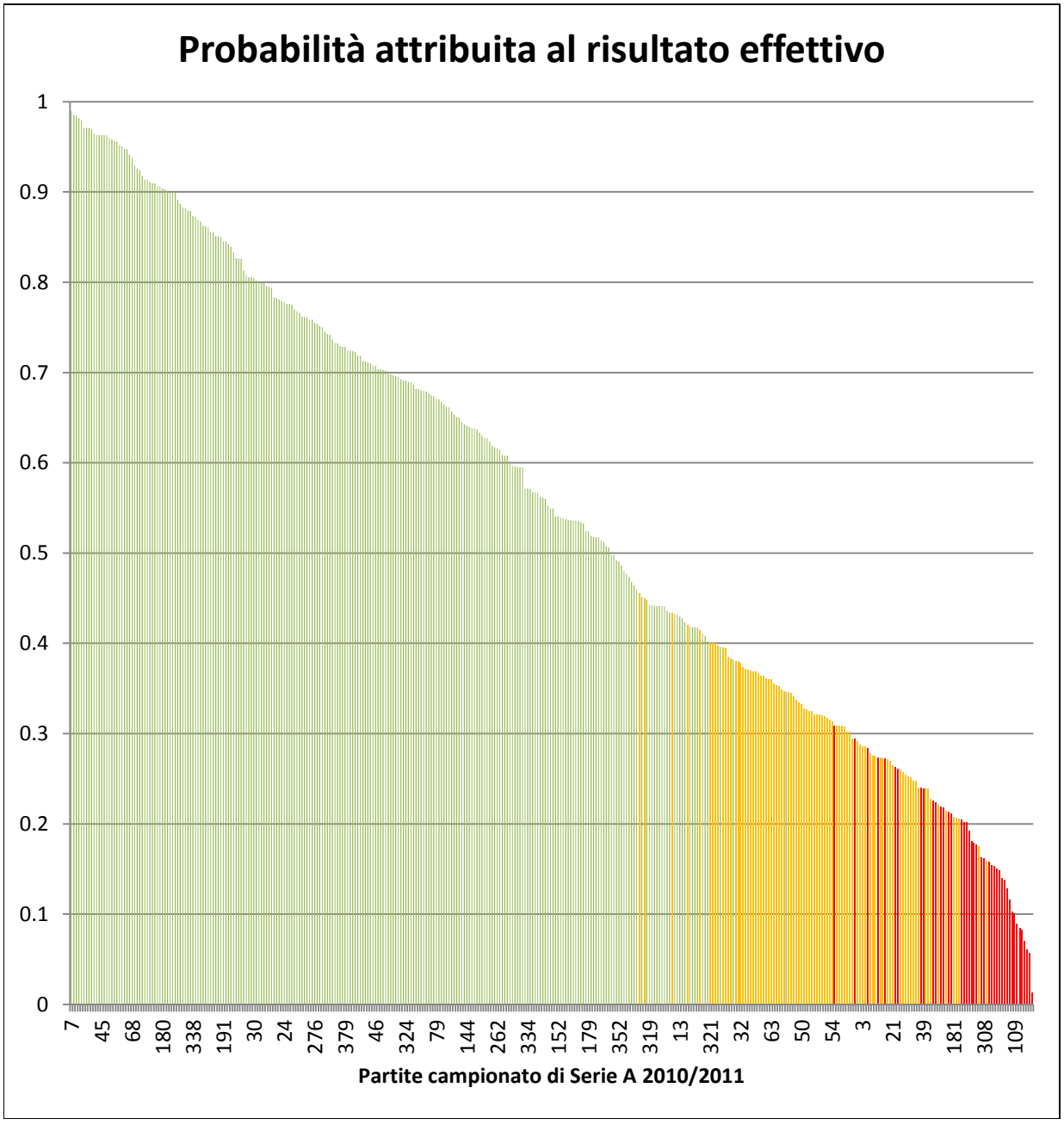


Figura 36: istogramma che evidenzia quali partite sono state previste correttamente.

Sull’asse delle ascisse possiamo individuare le varie partite disputate durante il campionato; ordinate a seconda della percentuale attribuita alla modalità che effettivamente si è manifestata. Tale probabilità viene rappresentata sull’asse delle ordinate. Il livello di dettaglio dell’asse delle ascisse, purtroppo non è dei migliori in quanto, essendo ben 380 partite, risultava impossibile riuscire a visualizzarle in maniera comprensibile su di un unico foglio.

Le linee verdi rappresentano quelle partite che sono state previste in maniera corretta dal modello, mentre quelle arancio e rosse sono quelle previste in maniera errata. Abbiamo deciso di attribuire due colori a queste ultime in quanto, riteniamo interessante valutare se il modello ha sbagliato in maniera lieve, assegnando una percentuale di poco inferiore alla modalità che effettivamente si è manifestata; oppure se a questa modalità aveva assegnato una probabilità talmente bassa che mai sarebbe stata prevista in maniera corretta. Il valore soglia utilizzato per considerare una partita nella zona arancio piuttosto che in quella rossa è -30%. Tale soglia è stata scelta in funzione della variabilità in merito alla differenza di probabilità registrata tra i risultati riferiti a queste partite. Avendo deciso tale soglia, 42 partite sono state classificate nella zona rossa e 88 in quella arancio per un totale di 130 partite classificate in maniera errata. Tramite questi dati possiamo confermare il tasso di errata classificazione ottenuto dal modello.

Osservando il grafico possiamo notare che, se il modello prevede una modalità con una probabilità decisamente elevata, quasi sicuramente sarà questo il risultato che si riscontrerà nella realtà. Quando, invece, le probabilità riferite alle modalità assumono valori molto simili fra loro è rischioso decidere quale, fra le tre, sarà quella che effettivamente si concretizzerà. Proprio per questo, al diminuire della probabilità associata al risultato effettivo si trovano le partite caratterizzate dai colori arancio e rosso.

Allo stesso modo abbiamo ritenuto interessante mostrare un grafico che metta in relazione la probabilità assegnata al risultato effettivo con la differenza reti calcolata come differenza tra le reti segnate dalla squadra di casa e quelle segnate dalla squadra ospite. Naturalmente ci aspettiamo che al crescere, in valore assoluto, della differenza reti, aumenti anche la probabilità assegnata dal modello al risultato effettivamente ottenuto.

Osservando il grafico ottenuto abbiamo notato che nel database vi era la presenza di un outlier, in quanto per una differenza reti pari a -7 la probabilità associata al risultato effettivo era di poco superiore al 41%. Questo dato era riferito alla partita Palermo-Udinese. Tale partita ha visto la supremazia netta della squadra bianconera sulla squadra di casa. Tuttavia è da notare che il modello, pur avendo previsto l'andamento della partita in modo corretto, non ha assegnato alla modalità "sconfitta" una probabilità molto elevata, tale da confermare la vittoria schiacciante dell'Udinese. Le considerazioni appena effettuate ci hanno portati a considerare tale dato come outlier e quindi ad eliminarlo dalle rappresentazioni sottostanti.

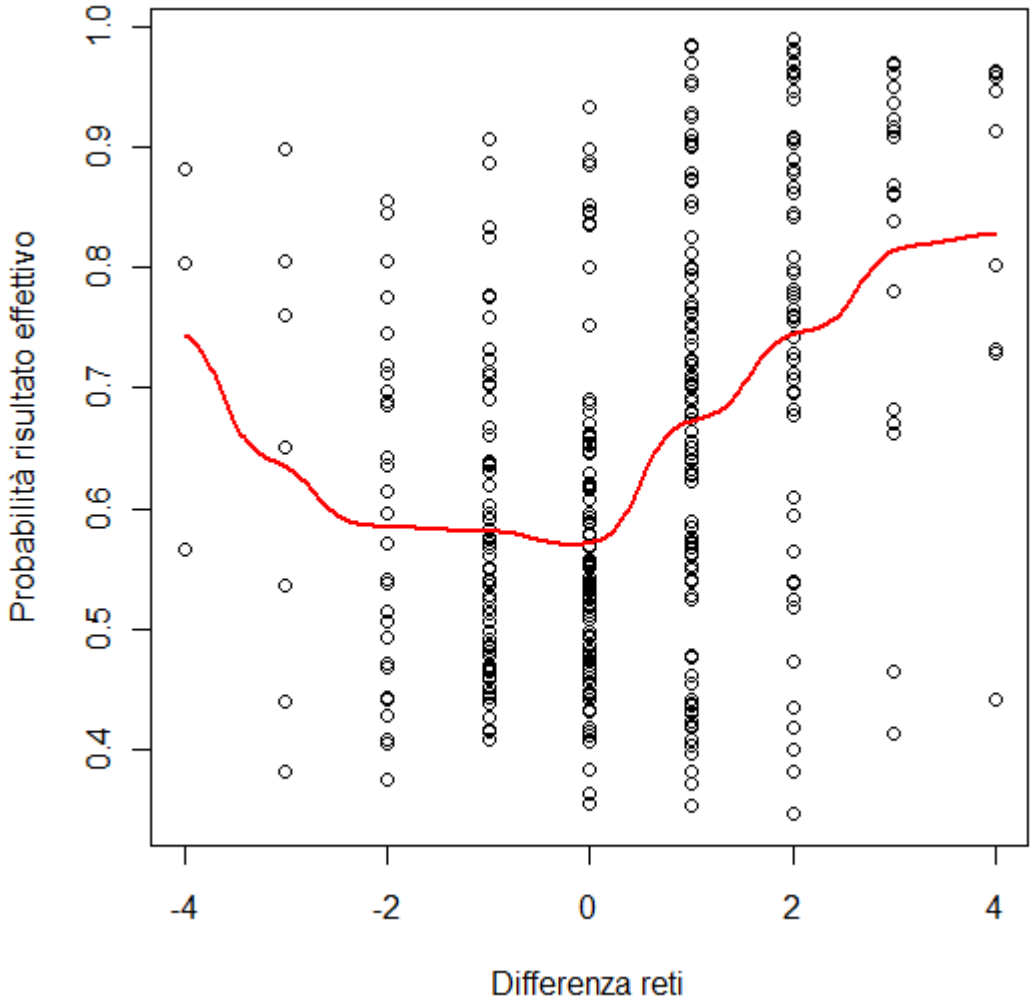


Figura 37: andamento della probabilità associata al risultato effettivo a seconda della differenza reti registrata.

Analizzando il grafico appena riportato troviamo sull'asse delle ascisse la differenza reti fra la squadra di casa e la squadra ospite. Più precisamente, nel lato sinistro del grafico, quindi quello caratterizzato da valori negativi per quanto riguarda l'asse in questione, troviamo le partite vinte dalla squadra ospite; in corrispondenza dell'origine troviamo tutti i match conclusi con un pareggio ed infine nella parte destra dello stesso troviamo le vittorie della

squadra di casa. Associato ad ogni dato osserviamo la corrispondente probabilità assegnata dal modello alla modalità che poi effettivamente si è rilevata nella realtà (asse delle ordinate). La linea rossa che attraversa il grafico la possiamo definire come una sorta di linea di tendenza che ci permette di analizzare come varia la probabilità assegnata dal modello a seconda del risultato effettivo dell'incontro. Osservando proprio tale linea notiamo che a parità di differenza reti, in termini di valore assoluto, il modello tende ad assegnare una percentuale più elevata alla vittoria della squadra di casa.

Per verificare se tale intuizione è corretta abbiamo deciso di costruire il medesimo grafico con l'ausilio dei Box Plot, i quali ci permettono di individuare con maggior dettaglio come il modello si comporta per ciascun valore *differenza reti* registrato nel Campionato di Serie A 2010/2011.

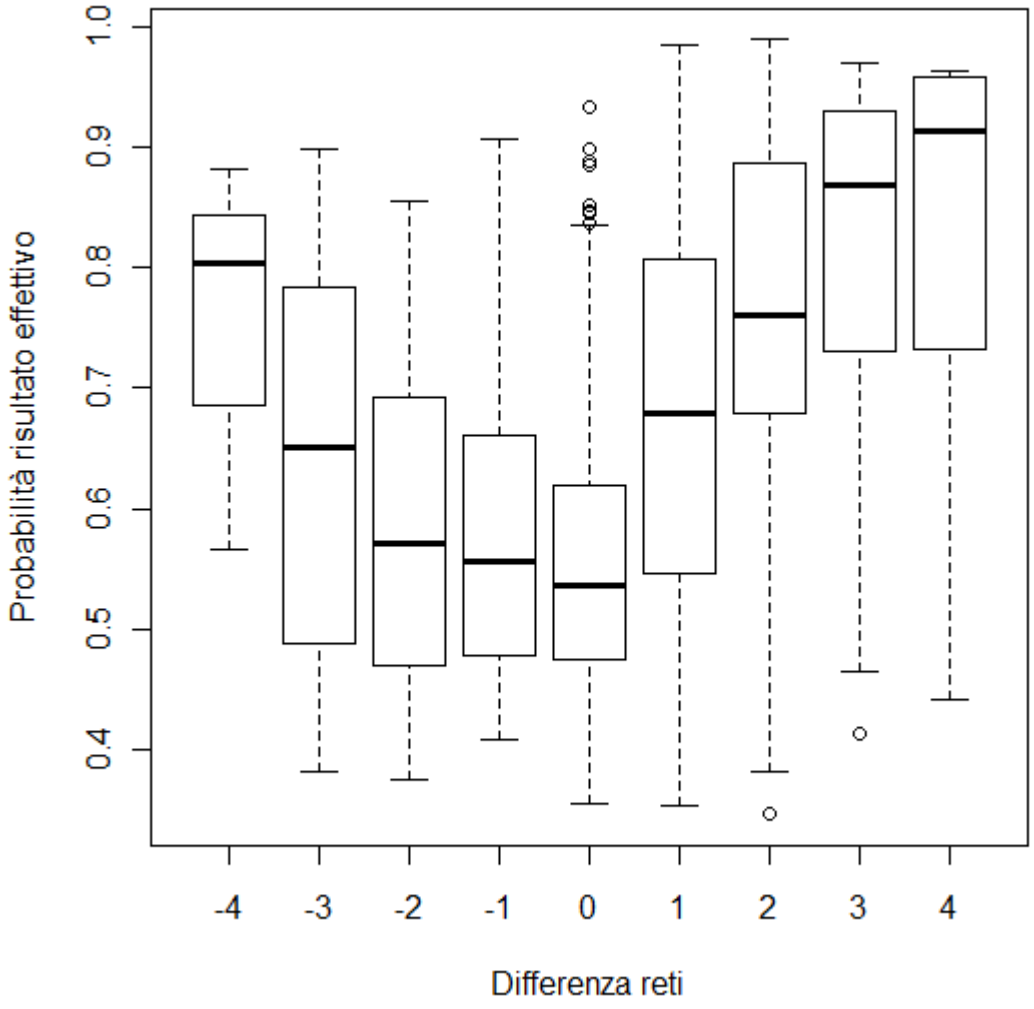


Figura 38: Box Plot che evidenzia la variabilità in merito alla probabilità associata alle varie partite concluse con una determinata differenza reti.

La considerazione prima effettuata sulla capacità di previsione del modello è confermata anche in questo grafico. Infatti, a parità di differenza reti in valore assoluto, registriamo una probabilità decisamente superiore per le partite concluse con la vittoria della squadra di casa. Questo lo si può facilmente percepire osservando attentamente la Figura 38, nella quale la probabilità media assegnata a ciascuna possibile modalità assunta dalla variabile *differenza reti* è evidenziata dalla linea orizzontale in grassetto racchiusa all'interno di ciascun rettangolo.

La variabilità in merito alla probabilità assegnata ai vari incontri, che si sono conclusi con un determinato scarto reti, la si può osservare nei cosiddetti "baffi" che caratterizzano i grafici Box Plot. Questi sono rappresentati dalle linee tratteggiate che fuoriescono da ciascuna "scatola" e si concludono con una piccola lineetta orizzontale. La lunghezza di questi "baffi" identifica la maggiore o minore variabilità nella probabilità associata agli incontri che si sono conclusi con quella particolare differenza reti.

Nel grafico vengono inoltre visualizzati alcuni puntini; questi rappresentano dati anomali (outliers) per i quali il modello ha previsto una probabilità molto diversa rispetto a quella media prevista per i match conclusi con quella particolare differenza punti.

Il codice R necessario per l'ottenimento dei grafici appena analizzati è disponibile in Appendice E.

Partendo dalle considerazioni fatte fino a questo punto sulla classificazione delle partite, abbiamo ritenuto interessante analizzare alcuni dei match classificati in maniera non corretta dal modello, per indagare sulle motivazioni per le quali il metodo della Random Forest era maggiormente propenso ad assegnare un risultato diverso all'incontro.

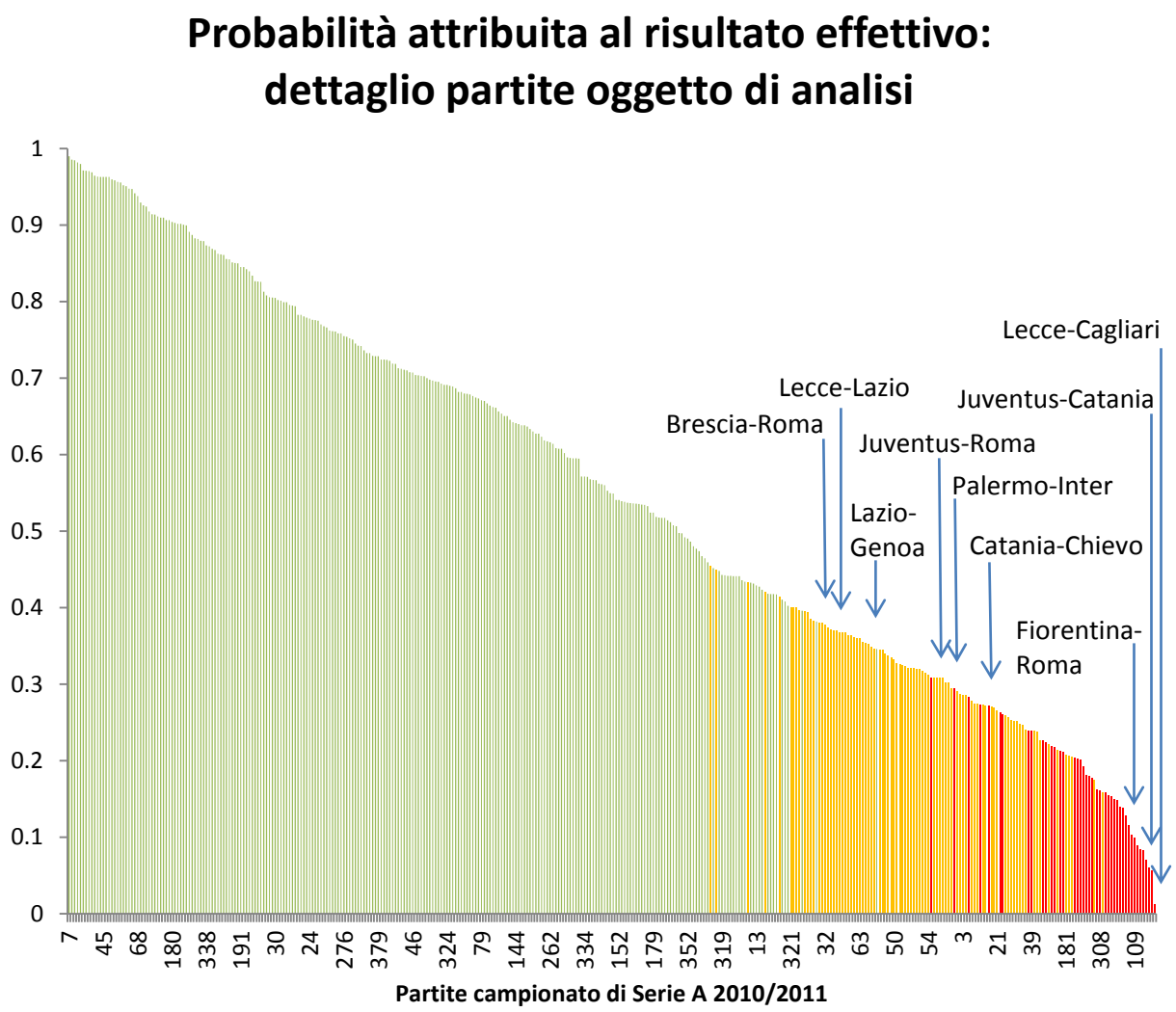


Figura 39: posizionamento degli incontri oggetto di analisi.

PALERMO – INTER 1 – 2

La modalità vittoria assegnata al Palermo era stata prevista con più del 57% di probabilità e la differenza tra la probabilità assegnata al risultato effettivo e quella assegnata al risultato previsto (57%) risulta pari a -28%. Pertanto, tale partita risulta essere nella zona arancio del grafico a barre rappresentato nelle Figure 36 e 39. Questa partita tuttavia si trova a confine tra le partite contrassegnate dal colore arancio e quelle contrassegnate con il colore rosso, in quanto è opportuno ricordare che il valore soglia scelto per classificare una partita prevista in maniera completamente errata è -30%.

Ma andiamo ad analizzare più nel dettaglio le motivazioni di questo grande divario tra realtà e previsione.

Il presidente del Palermo Zamparini era infuriato dopo l'incontro Palermo – Inter e dichiarò che l'arbitro Romeo aveva negato quattro rigori alla sua squadra. Questa vicenda si protrasse per numerosi giorni e risultarono evidenti errori arbitrali a scapito del Palermo.

Questa tesi è avvalorata anche dal nostro modello, il quale prevede che la partita si sarebbe dovuta concludere con la vittoria del Palermo, in quanto, questa squadra a livello di gioco meritava di vincere la sfida.

Può essere che gli errori arbitrali di Romeo abbiano compromesso fino a questo punto il risultato della partita?

BRESCIA – ROMA 2 - 1

Anche questo match, come il precedente, è stato ricordato a lungo per le varie polemiche sull'arbitraggio di Russo. Le maggiori critiche in merito a questa partita sono state che il calcio di rigore, a favore del Brescia era assolutamente inesistente perché fuori dall'area e perché l'intervento del giocatore della Roma, Mexes, era sulla palla e non nei confronti del giocatore della squadra avversaria. La Roma oltre a recriminare l'espulsione di Mexes ha denunciato di non aver avuto due rigori a favore.

Osservando le previsioni del nostro modello, notiamo che anche in questa occasione vi è una discrepanza tra risultato reale e risultato previsto. Infatti per la Random Forest questa partita si sarebbe dovuta concludere con un pareggio. Tuttavia i tifosi Romanisti non possono affermare che la Roma meritasse la vittoria, in quanto a livello di gioco le due squadre sono risultate equilibrate e pertanto un pareggio sarebbe stato il risultato migliore. Questa tesi è avvalorata anche dalle probabilità assegnate dal modello alle singole modalità, infatti, alla modalità pareggio era stata associata una probabilità pari al 42%, mentre a quella relativa alla sconfitta della squadra di casa solo del 20%.

JUVENTUS – ROMA 1 – 1

La partita si è conclusa con un pareggio, grazie al gol di Iaquinta (per la Juventus) e di Totti su rigore per la Roma. Il rigore della Roma è arrivato proprio allo scadere del primo tempo. Le tensioni sul campo erano a dir poco palpabili, si pensi che addirittura, alla fine del match, il capitano giallorosso ha insultato il portiere della Juventus, Storari, in quanto lo ha sempre ritenuto responsabile di aver fatto perdere l'anno precedente lo scudetto alla Roma. In realtà, il portiere, che fino all'anno prima giocava nella Sampdoria è stato colpevole solamente di

aver giocato una delle partite più spettacolari della sua carriera proprio durante la sfida con la Roma l'ultima giornata di campionato.

Da queste considerazioni si può facilmente capire che questo incontro ha scatenato notevoli polemiche sia a livello morale sia a livello tecnico. Per quanto riguarda quest'ultimo punto l'arbitro è stato duramente contestato sia perché ha concesso un calcio di rigore quando ormai il primo tempo era volto al termine, sia perché sembra, che per un'azione del genere non era assolutamente necessario concedere rigore.

La Juventus, tuttavia, ha giocato nettamente meglio rispetto alla sua avversaria, creando numerosissime occasioni da gol e per questo meritava la vittoria. Anche il modello sottolinea questa teoria, assegnando a tale modalità una probabilità pari al 52%.

JUVENTUS – CATANIA 2 – 2

Questo incontro ricade in una delle ultime linee rosse del grafico rappresentato nella Figura 36. Per tale incontro, la vittoria della squadra di casa era stata assicurata con una probabilità dell'84%, da qui potremmo dire che ci saremmo aspettati una vittoria schiacciante della vecchia signora sul Catania, ed invece non è andata in questo modo.

Possiamo dire che questa partita è stata veramente bizzarra, infatti fino a dieci minuti dalla fine, la Juventus conduceva la partita 2 a 0, poi all'ottantesimo il Catania segna il 2 – 1. Per poi pareggiare al 93°. Questo pareggio ha sollevato numerose polemiche, infatti l'arbitro è stato amaramente contestato per aver fatto calciare una punizione al limite dell'area a tempo scaduto, sulla quale il Catania è riuscito a realizzare il gol del pareggio.

È innegabile che il Catania sia stato molto bravo a non darsi per vinto e a tentare di recuperare il risultato fino all'ultimo minuto, tuttavia è anche vero che la Juventus ha dominato la partita sia in termini di risultati, sia in termini di gioco, almeno fino a dieci minuti prima della fine, pertanto, avrebbe meritato la vittoria, proprio come dimostra il nostro modello che, non tenendo in considerazione il fattore fortuna e gli errori di arbitraggio, aveva previsto la predominanza della Juventus sul Catania.

PARTITE SOSPETTE NELL'INCHIESTA SUL CALCIO SCOMMESSE.

È notizia di pochi giorni fa che altre partite del campionato di Serie A 2010/2011 sono ritenute sospette dagli inquirenti che indagano nell'ambito del calcio scommesse. Tutta la

vicenda ancora oggi è avvolta nell'incertezza, ma grazie alle svolte degli ultimi giorni sembra che gli inquirenti si stiano avvicinando sempre più alla verità.

Dato tutto questo fermento in merito all'argomento, vari siti internet, giornali e programmi televisivi continuano a pubblicare elenchi più o meno attendibili degli incontri che sarebbero oggetto d'inchiesta. Tali elenchi, purtroppo non sempre corrispondono fra loro, per cui è difficile apprendere precisamente quali tra questi siano veramente attendibili.

Alcuni elenchi vengono riportati per chiarezza espositiva in Appendice F, con il dettaglio di come tali incontri vengano classificati dal modello della Random Forest.

Dato che per il momento non abbiamo nulla di realmente ufficiale abbiamo deciso di analizzare come si comporta il modello con alcune di queste partite. La scelta è ricaduta sugli incontri che, al momento, presentano le maggiori perplessità in quanto non abbiamo nessun altro criterio, maggiormente oggettivo, per identificare le partite realmente coinvolte nella trama del Calcio Scommesse.

- CATANIA – CHIEVO 1 – 1
- FIORENTINA – ROMA 2 – 2
- LECCE – CAGLIARI 3 – 3
- LAZIO – GENOA 4 – 2
- LECCE – LAZIO 2 – 4

Le prime tre partite di questo elenco, le possiamo trovare nella parte contrassegnata dal colore rosso della Figura 39. Tali incontri, pertanto, vengono previsti in modo totalmente diverso dal nostro modello; più precisamente la modalità pareggio era stata prevista rispettivamente nelle tre partite al 26%, 15% ed al 6%. L'ultimo di questi tre incontri è quello che riscontra maggiori criticità data la bassissima probabilità assegnata dalla Random Forest a questa modalità. Cercando in rete i commenti in merito a questo incontro siamo arrivati a comprendere la motivazione per la quale questa partita è oggetto di indagine e forse, proprio queste implicazioni, potrebbero giustificare il diverso andamento del modello rispetto alla realtà. Questo match vedeva competere un Lecce in piena lotta retrocessione ed un Cagliari già salvo.

Una settimana prima dell'incontro c'era chi già parlava di combine tra sardi privi di motivazione e salentini obbligati a vincere, ma durante l'incontro, molti si erano ricreduti, visto che il Cagliari aveva creato numerosissime occasioni sin dall'inizio del match

conducendo fin da subito la partita. Al 48' il Cagliari poteva chiudere l'incontro, ma in quest'azione succede l'imprevedibile, Cossu, giocatore del Cagliari, parte in contropiede, il difensore del Lecce Tomovic scivola goffamente alla "Fantozzi", il pallone supera Rosati (portiere del Lecce) fino ad arrivare a Ragatzu, il quale a porta vuota incespica dimenticandosi anche di tirare e permettendo così alla difesa di rientrare. La partita in ultimo si conclude con un pareggio e con il sospetto che forse questi errori non siano stati del tutto casuali.

Passando alla partita successiva (Lazio – Genoa) la situazione previsionale del modello è realmente critica, infatti prevede il manifestarsi delle tre modalità con la medesima probabilità. Questo significa che per il modello ciascuna delle tre modalità è pressoché identica e la probabilità che questo preveda il risultato corretto è $\frac{1}{3}$.

L'ultima partita oggetto di analisi, Lecce - Lazio, è quella che per il momento è maggiormente sotto i riflettori in quanto, il portiere della squadra di casa è stato accusato di aver subito dei gol assurdi e pertanto gli inquirenti stanno indagando in questa direzione. Benassi, portiere del Lecce, negli ultimi giorni, ha dichiarato di essere del tutto estraneo ai fatti e molto addolorato per il suo coinvolgimento in questa inchiesta. Dato che per il momento non è stato rilasciato nessun verdetto ufficiale, anche in questo caso ci limiteremo ad osservare in maniera oggettiva come si caratterizza l'incontro con il metodo Random Forest. Nel nostro istogramma, questa partita, si trova nella zona caratterizzata dal colore arancio, la quale identifica le partite che sono state prevista in maniera errata, ma la cui differenza con il risultato previsto non supera la soglia -30%. In particolare, la probabilità assegnata dal modello alla vittoria del Lecce era pari al 51%, mentre quella della Lazio solo del 37%.

3.6.3 INDICI.

Arrivati a questo punto, abbiamo deciso di ricorrere al database contenente i fattori maggiormente interpretabili per ottenere una media di performance di ciascun indice per squadra.

Per maggiore chiarezza interpretativa della tabella sottostante riporto la descrizione corrispondente alle sigle identificative degli otto indici oggetto di questa analisi:

1) PericSqC1: giocate aeree effettuate dalla squadra di casa;
2) Peric_SqC2: giocate pericolose a palla bassa eseguite dalla squadra di casa;
3) DifSqC1: indice di difesa (o fase difensiva) della squadra di casa;
4) DifSqO1: fase difensiva protezione della porta della squadra ospite;

5) DifSqO2: barriera difensiva a protezione dell'area della squadra ospite;
6) PericSqO1: passaggi lunghi oltre il centrocampo attuati dalla squadra ospite;
7) PericSqO2: indice di pericolosità d'attacco della squadra ospite;
8) PericSqO3: tentativi di contropiede della squadra ospite.

I valori contrassegnati dai colori verde e rosso stanno ad indicare i migliori e i peggiori risultati ottenuti da ciascun indice oggetto di analisi.

Squadre	PericSqC1	Peric_SqC2	DifSqC1	DifSqO1	DifSqO2	PericSqO1	PericSqO2	PericSqO3
Bari	-0.304	-0.6	**-0.46**	**0.477**	-0.383	0.015	-0.262	**-0.584**
Bologna	-0.165	-0.54	-0.004	0.005	-0.202	0.324	**-0.528**	-0.164
Brescia	0.213	-0.35	-0.103	0.062	-0.271	-0.049	-0.369	-0.062
Cagliari	0.01	**-0.66**	0.44	0.189	-0.287	-0.197	-0.164	0.271
Catania	-0.105	-0.069	0.259	-0.064	0.351	0.153	-0.209	-0.383
Cesena	-0.301	-0.498	-0.337	0.242	**-0.637**	0.317	-0.479	-0.343
Chievo	0.216	-0.493	0.352	0.441	0.326	-0.434	-0.117	0.43
Fiorentina	0.551	0.182	-0.064	0.04	-0.262	0.038	0.04	-0.417
Genoa	0.009	0.204	-0.117	0.041	**0.419**	-0.004	-0.265	-0.094
Inter	-0.433	1.105	-0.55	**-0.359**	0.166	**0.732**	0.609	0.18
Juventus	0.498	0.102	0.448	0.027	0.24	-0.135	0.202	0.033
Lazio	-0.079	0.257	0.081	-0.227	0.098	-0.135	0.264	-0.138
Lecce	-0.073	-0.535	0.016	0.046	-0.076	0.229	-0.09	-0.514
Milan	-0.544	**1.134**	-0.044	-0.323	0.314	-0.151	0.497	**0.881**
Napoli	0.49	0.154	**0.497**	-0.18	0.249	-0.241	0.103	0.362
Palermo	**-0.561**	0.697	-0.324	0.224	0.075	-0.093	0.444	-0.3
Parma	**0.636**	-0.4	-0.207	-0.325	-0.117	0.233	-0.172	0.247
Roma	-0.215	0.573	0.232	-0.321	-0.111	-0.024	0.321	0.805
Sampdoria	0.445	-0.471	-0.17	-0.009	-0.242	**-0.458**	-0.456	-0.222
Udinese	-0.286	0.21	0.052	0.014	0.35	-0.118	**0.631**	0.014

In questo modo possiamo osservare le performance medie di ciascuna squadra sia quando questa gioca in casa, sia quando gioca come ospite.

Non abbiamo potuto fare a meno di constatare che, anche le squadre leader di questo campionato, sono state contrassegnate da valori negativi per quanto riguarda alcuni di questi indici. Si pensi ad esempio all'Inter; la quale è stata classificata come peggiore squadra del campionato per quanto riguarda *l'indice di fase difensiva a protezione della porta*, per le partite giocate come squadra ospite. Probabilmente questo indice non incide in maniera marcata sulla probabilità di vittoria della squadra, in quanto, a conti fatti sappiamo che l'Inter a fine campionato si è classificata seconda. In alcuni casi, tale risultato potrebbe essere in parte imputato ai cosiddetti *fenomeni*, spesso presenti all'interno delle squadre più facoltose.

Molto spesso, infatti, questi giocatori portano la propria squadra alla vittoria, seppur le performance di gioco di quest'ultima non siano da considerarsi pienamente soddisfacenti.

Naturalmente, ciascun fattore influisce in maniera differente sulla probabilità di vittoria o sconfitta, ma per riuscire a capire più nel dettaglio come effettivamente questi indici incidano su tali probabilità abbiamo deciso di analizzarli singolarmente e di osservare quanto ciascuno di questi è realmente importante. Per farlo, utilizzeremo il software statistico R, il cui codice è disponibile in Appendice G e il database contenente i valori di questi indici per ciascuna partita disputata nel campionato.

Per avere risultati maggiormente interessanti ai fini delle nostre analisi abbiamo deciso di depurare ciascun fattore dall'influenza dei restanti, in modo da osservare come effettivamente ciascuno di essi influisca sulle probabilità di vittoria o sconfitta della squadra di casa.

PRIMO FATTORE: GIOCATE AEREE EFFETTUATE DALLA SQUADRA DI CASA.

Dal grafico sottostante notiamo che all'aumentare del valore attribuito a questo fattore, la probabilità di vittoria della squadra di casa diminuisce progressivamente. Tuttavia, intorno al valore $+1$, sembra che questa diminuzione si attenui, diventando quasi una linea orizzontale.

Questo risultato è molto interessante, infatti il modello in questo modo ci comunica che le squadre non dovrebbero puntare troppo sulle cosiddette *giocate aeree*, caratterizzate da cross e tiri al volo in quanto il grafico ci comunica che queste incidono in maniera negativa sulla probabilità di vittoria. Forse abbiamo ottenuto questo risultato in quanto, a livello tecnico il tipo di gioco oggetto di analisi è decisamente spettacolare, ma a livello pratico, si concretizza in un risultato effettivo solo in poche occasioni.

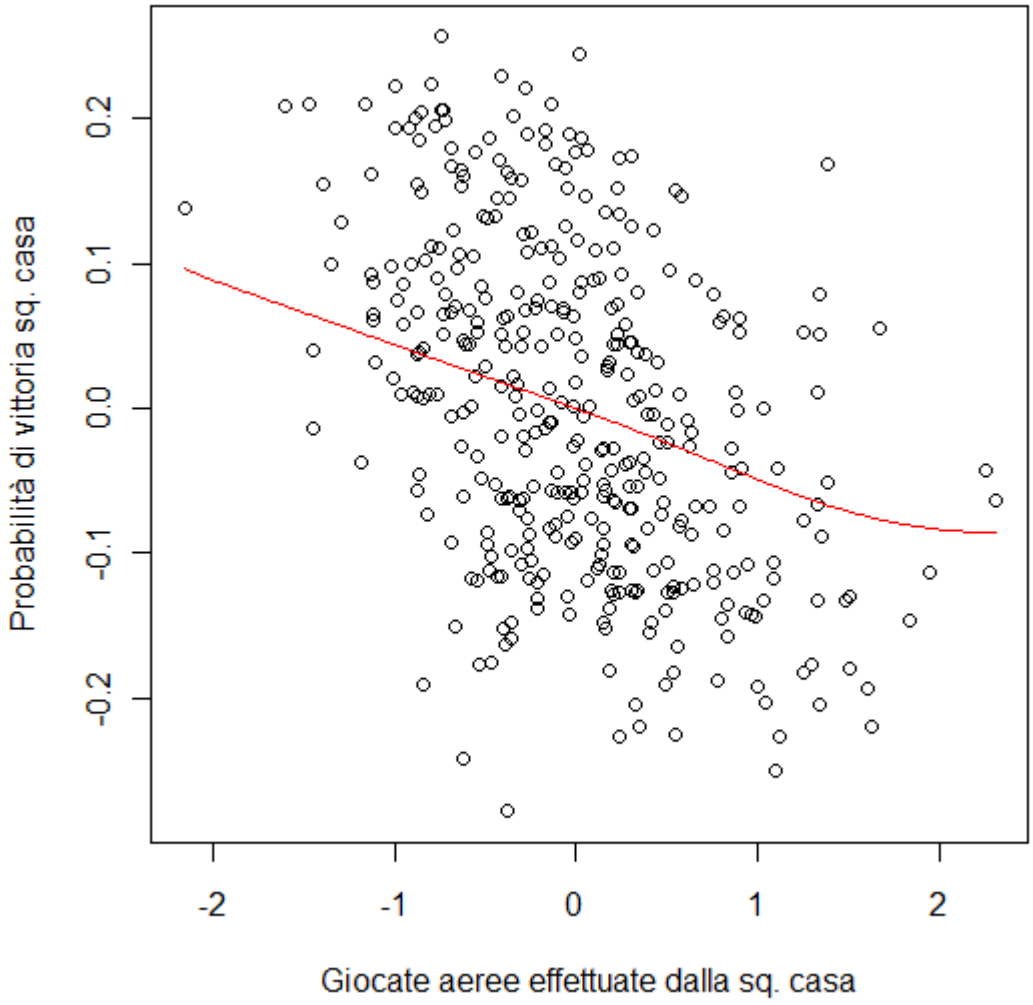


Figura 40: incidenza del primo fattore sulla modalità vittoria della squadra di casa.

SECONDO FATTORE: GIOCATE PERICOLOSE A PALLA BASSA ESEGUITE DALLA SQUADRA DI CASA.

Osservando l'andamento della Figura 41 non possiamo far altro che constatare che, questo particolare tipo di gioco, ha una marcata incidenza positiva sulla modalità vittoria riferita alla squadra di casa; infatti all'aumentare del valore assegnato all'indice *giocate pericolose a palla bassa*, si regista un corrispondente aumento della modalità vittoria.

Prendendo in considerazione anche quanto detto per il precedente fattore, possiamo affermare che per una squadra è maggiormente importante ai fini della vittoria di un incontro cercare di costruire un'azione attraverso *passaggi a palla bassa* (sotto l'altezza delle spalle) piuttosto che tentare di lanciare in avanti il pallone attraverso cross o passaggi lunghi, con la speranza che un compagno di squadra riesca ad intercettare il tiro e a dirigerlo verso la conclusione della porta. È necessario puntare meno alla spettacolarizzazione del gol e più alla concretezza di risultato.

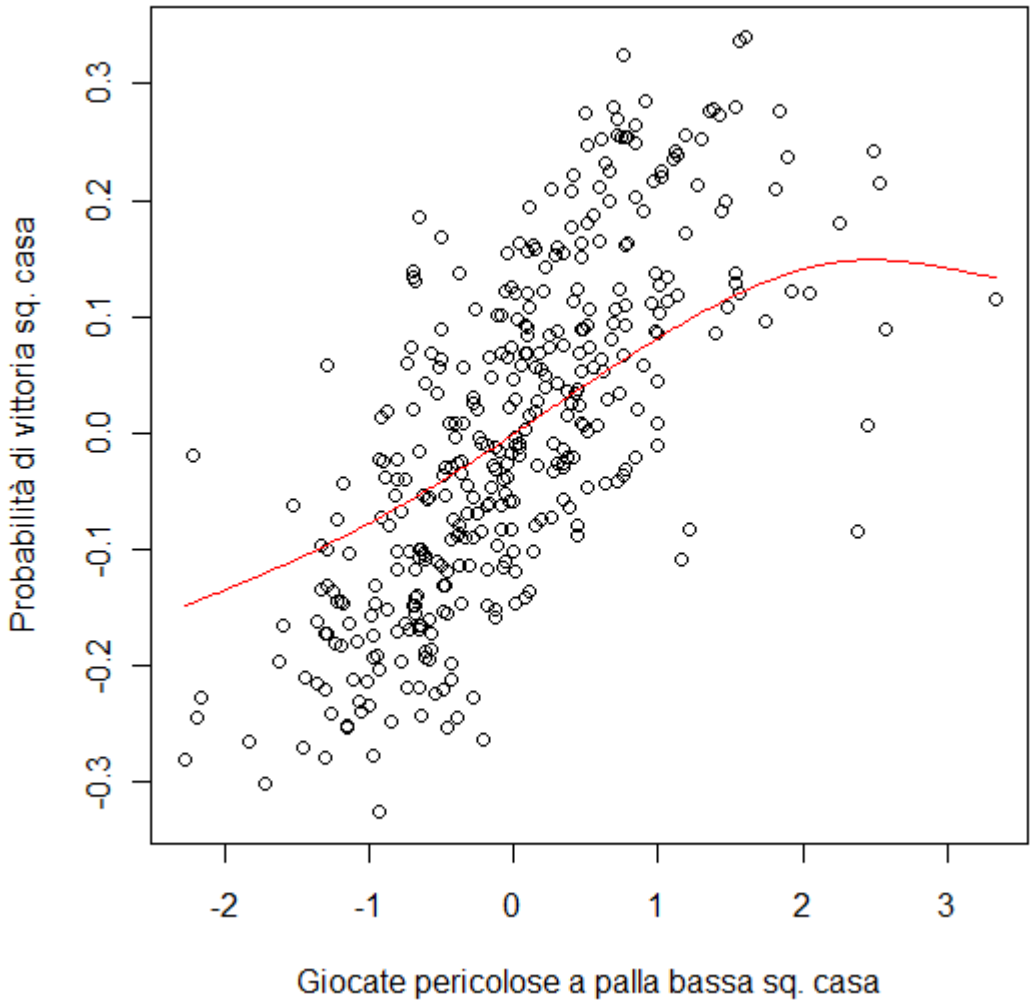


Figura 41: incidenza del secondo fattore sulla modalità vittoria della squadra di casa.

TERZO FATTORE: INDICE DI DIFESA (O FASE DIFENSIVA) DELLA SQUADRA DI CASA.

Osservando l'andamento della linea di tendenza notiamo che questo fattore non incide pesantemente sulla probabilità di vittoria della squadra di casa, infatti, tale linea, tende ad essere quasi orizzontale.

Naturalmente avere una buona difesa aiuta, ma non lo possiamo comunque considerare un fattore decisivo per la determinazione del risultato.

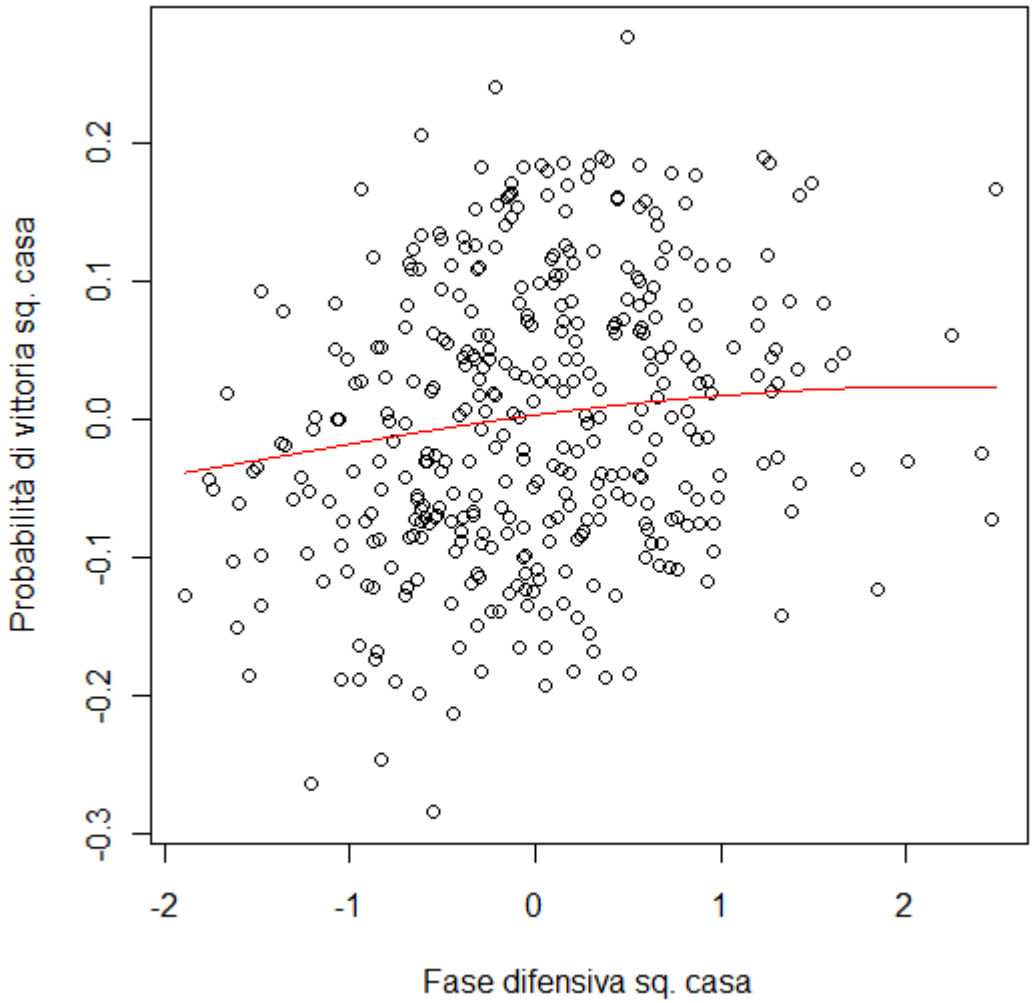


Figura 42: incidenza del terzo fattore sulla modalità vittoria della squadra di casa.

QUARTO FATTORE: FASE DIFENSIVA PROTEZIONE DELLA PORTA DELLA SQUADRA OSPITE.

Dato che questo indice è riferito alla squadra ospite, osserviamo la sua incidenza non più sulla probabilità di vittoria della squadra di casa, ma bensì sulla sua probabilità di sconfitta.

Dalla figura sotto riportata possiamo notare che questo fattore incide sulla probabilità di sconfitta della squadra di casa, seppur in maniera non troppo marcata.

Tutto questo ha senso in quanto all'aumentare dell'azione di protezione della porta da parte della squadra ospite diminuisce anche la probabilità di vittoria della squadra di casa, con conseguente aumento della sua probabilità di sconfitta. Essendo una difesa specifica a protezione della porta è normale registrare un tale andamento, mentre per quanto riguarda l'indice precedente, la fase difensiva era intesa più a livello di centro campo e di area e pertanto non così determinante per l'ostruzionismo degli avversari.

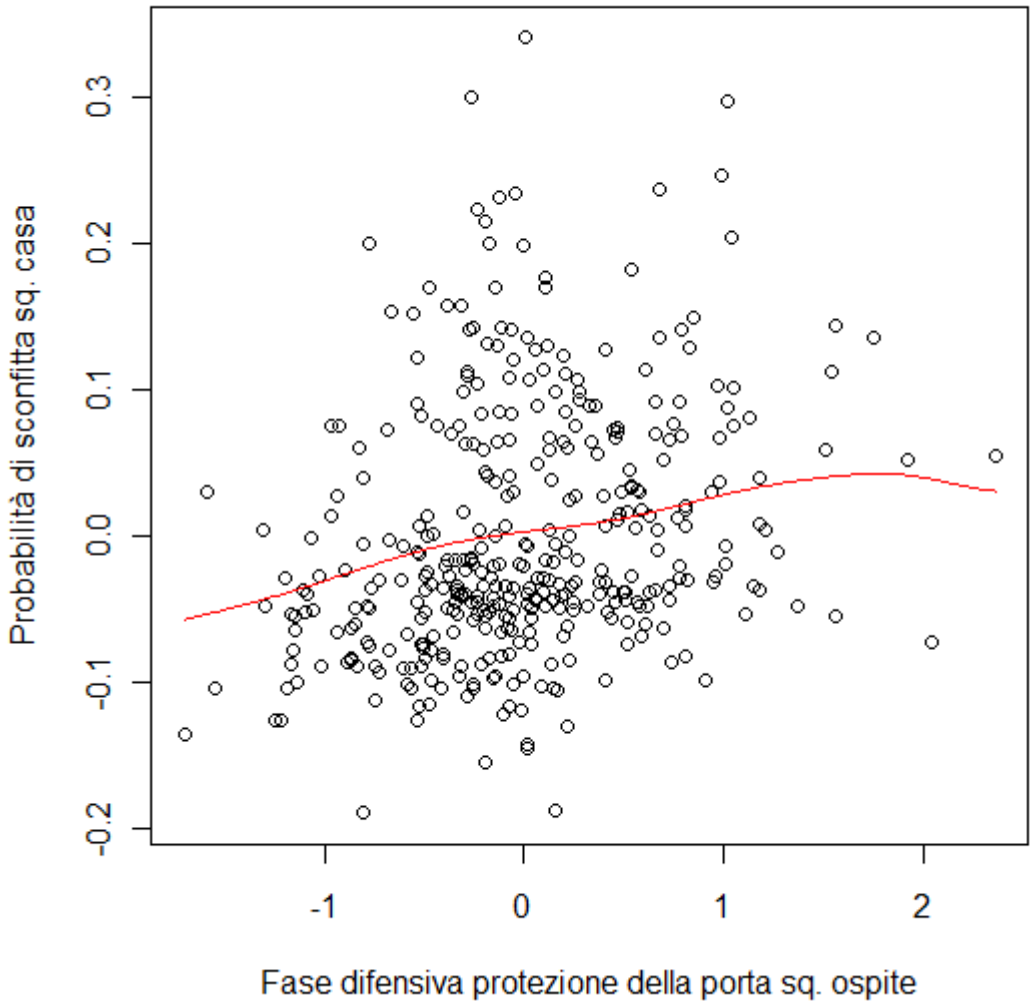


Figura 43: incidenza del quarto fattore sulla modalità sconfitta della squadra di casa.

QUINTO FATTORE: BARRIERA DIFENSIVA A PROTEZIONE DELL'AREA DELLA SQUADRA OSPITE.

Il fattore *barriera difensiva a protezione dell'area* della squadra ospite ha un andamento tendenzialmente orizzontale e pertanto non ha una forte incidenza sulla probabilità di sconfitta della squadra di casa. Tale andamento segue a grandi linee quello relativo al terzo fattore, seppur con qualche leggero picco all'aumentare dei valori associati a questo indice.

Come per il caso precedente, anche qui è stato deciso di osservare l'andamento del fattore rispetto alla probabilità di sconfitta della squadra di casa e non più in base a quella di vittoria, in quanto, il fattore era espresso in riferimento alla squadra ospite. Questa impostazione verrà utilizzata anche per i tre restanti fattori.

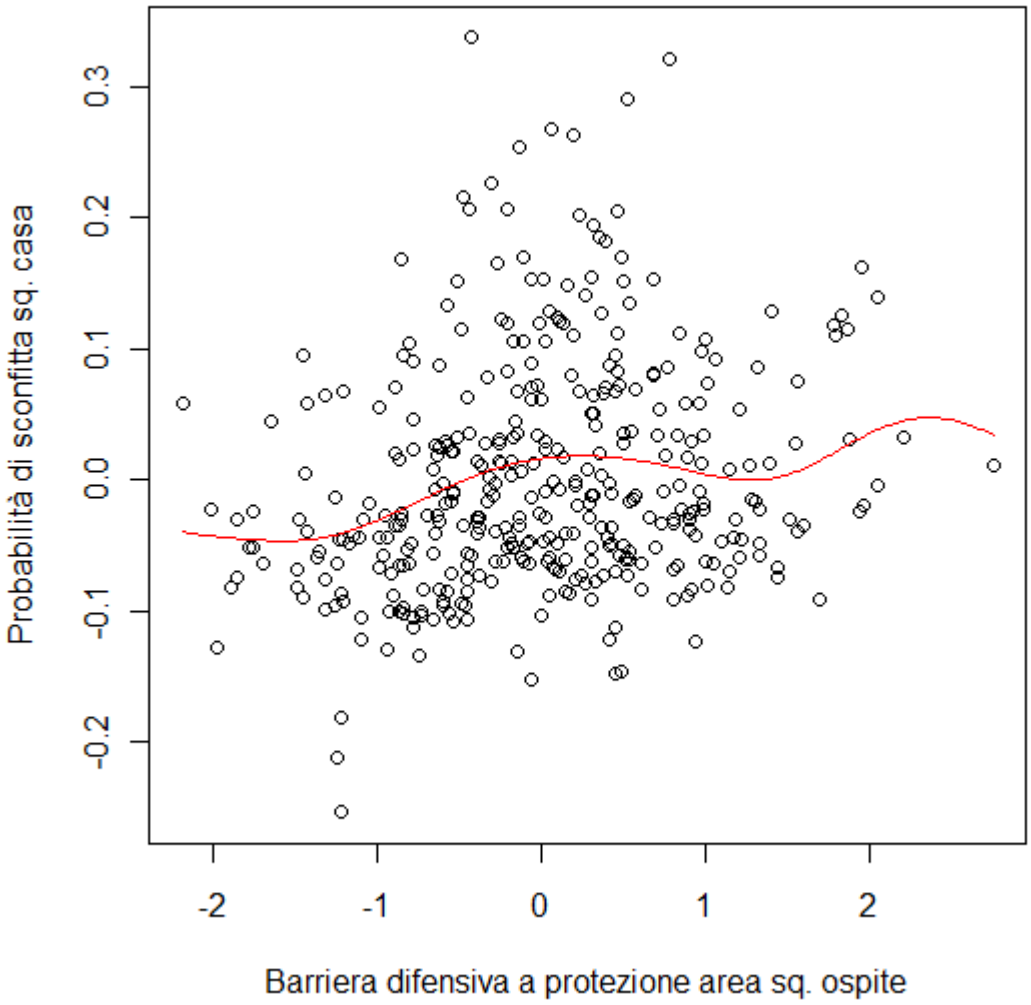


Figura 44: incidenza del quinto fattore sulla modalità sconfitta della squadra di casa.

SESTO FATTORE: PASSAGGI LUNGHI OLTRE IL CENTROCAMPO ATTUATI DALLA SQUADRA OSPITE.

Dalla figura notiamo un andamento decrescente della probabilità di sconfitta della squadra di casa all'aumentare dei *passi lunghi oltre il centrocampo* effettuati dalla squadra ospite.

Questo fatto sta ad indicare che, se la squadra ospite deve ricorrere pesantemente a questo tipo di passaggi, significa che la squadra di casa ha una difesa alta difficile da penetrare e pertanto questo fattore può favorire la probabilità di vittoria della squadra di casa.

È da sottolineare che la linea di tendenza decresce dolcemente all'aumentare del sesto fattore, questo naturalmente sottolinea un'incidenza negativa dello stesso sulla probabilità di sconfitta.

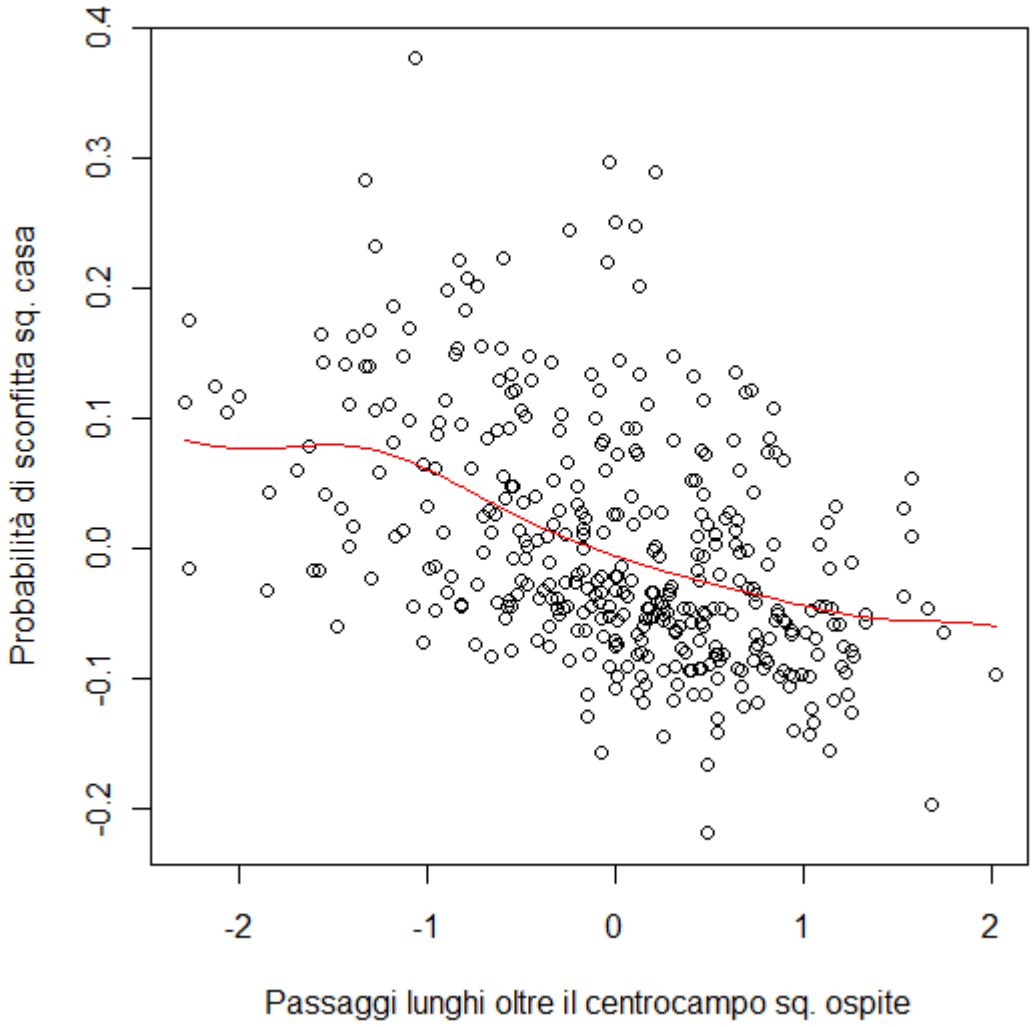


Figura 45: incidenza del sesto fattore sulla probabilità di sconfitta della squadra di casa.

SETTIMO FATTORE: INDICE DI PERICOLOSITA' D'ATTACCO DELLA SQUADRA OSPITE.

Solo osservando il grafico riusciamo ad intuire che questo fattore incide in maniera molto più marcata sulla probabilità di sconfitta della squadra di casa rispetto a quanto registrato per i precedenti indici analizzati.

La linea di tendenza riferita alla probabilità di sconfitta della squadra di casa aumenta all'aumentare della pericolosità di attacco della squadra ospite; pertanto più quest'ultima è pericolosa nella fase di attacco (riuscendo a creare numerose occasioni e tentando spesso di

tirare nello specchio della porta avversaria) e più la probabilità di sconfitta della squadra di casa aumenta.

Le squadre devono puntare a migliorare questo aspetto nella costruzione del loro gioco, al fine di migliorare i propri risultati all'interno del campionato.

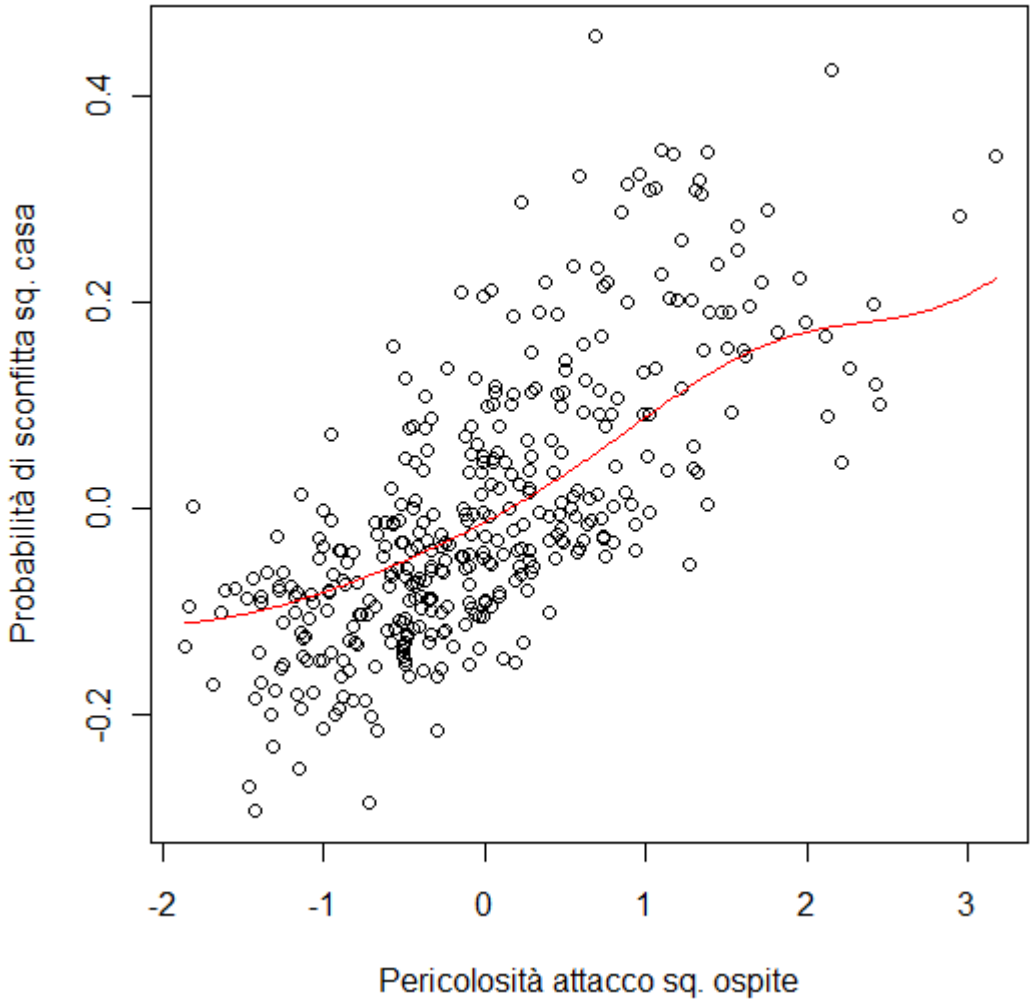


Figura 46: incidenza del settimo fattore sulla probabilità di sconfitta della squadra di casa.

OTTAVO FATTORE: TENTATIVI DI CONTROPIEDE DELLA SQUADRA OSPITE.

Tale fattore è quasi indifferente per la determinazione dei risultati, infatti la linea di tendenza è quasi completamente orizzontale, anche se leggermente inclinata verso il basso.

La probabilità di sconfitta della squadra di casa diminuisce, seppur di poco, all'aumentare dei *tentativi di contropiede* della squadra ospite. Se una squadra effettua numerosi *tentativi di contropiede*, significa che gli avversari hanno una difesa molto alta che costringe la squadra

ospite a rimanere chiusa all'interno della propria area. Questo comporta che per poter proseguire con la fase di attacco, gli ospiti sono costretti a cercare il contropiede non riuscendo per l'appunto a costruire un'azione senza ricorrere al pressing e ai passaggi alti.

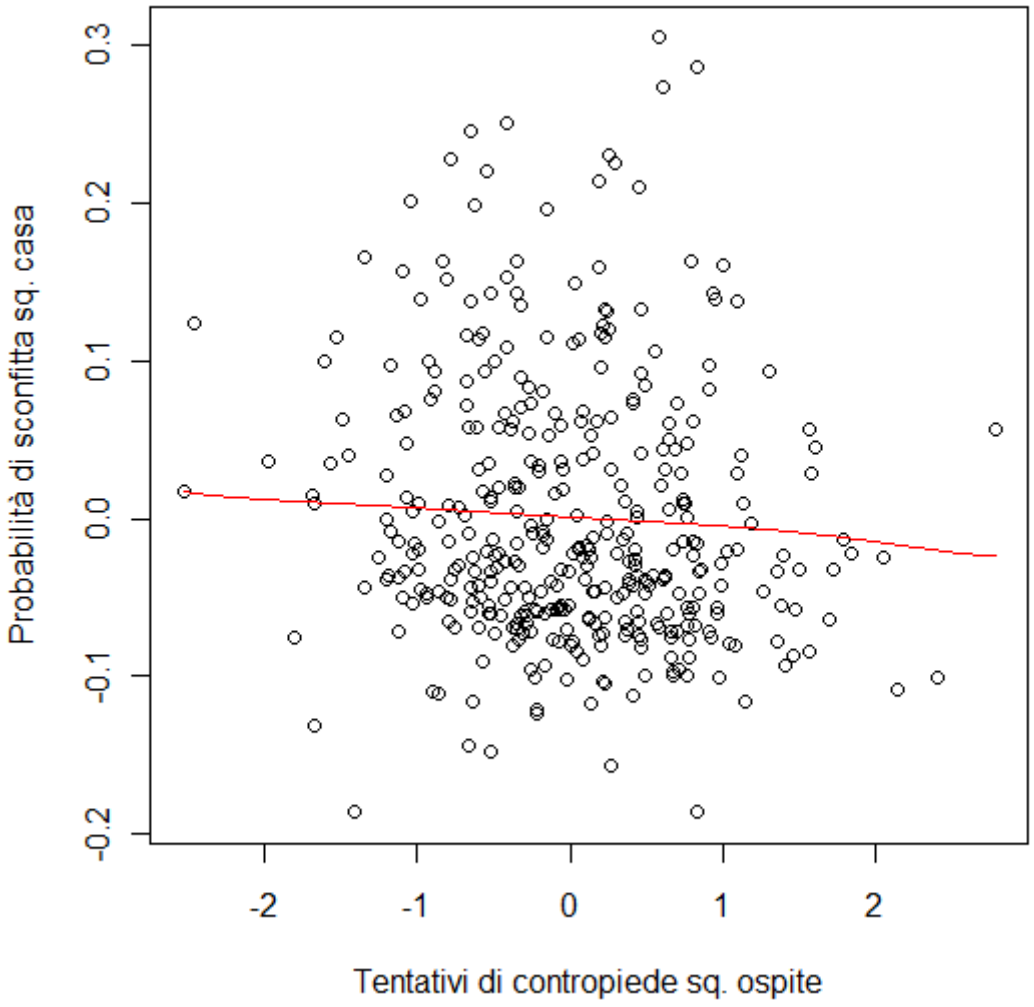


Figura 47: incidenza dell'ottavo fattore sulla probabilità di sconfitta della squadra di casa.

Avendo effettuato l'analisi di tutti gli otto fattori che compongono il nostro database abbiamo deciso di osservare se queste nostre interpretazioni si riflettono poi sui risultati reali.

Dalle analisi è emerso che i fattori più determinati per il raggiungimento della vittoria sono il secondo e il settimo, rispettivamente le *giocate pericolose a palla bassa* effettuate quando le squadre giocano in casa e l'indice di *pericolosità d'attacco* quando le squadre giocano come ospiti. Per verificare la correttezza dell'analisi abbiamo deciso di osservare i valori medi in merito a questi indici per quanto riguarda le capoliste della scorsa stagione (2010/2011): Milan ed Inter.

Tali valori medi sono stati riportati nella tabella all'inizio del paragrafo 3.6.3.

Osservando il fattore relativo alle *giocate pericolose a palla bassa*, effettuate quando le squadre giocano in casa, possiamo notare immediatamente che il Milan è la squadra che ottiene il valore medio migliore, seguita a ruota dall'Inter.

Passando all'indice relativo alla *pericolosità d'attacco* quando le squadre giocano come ospiti, notiamo che la squadra che ottiene il valore medio più elevato è l'Udinese. Tuttavia, questo non ci stupisce particolarmente in quanto questa squadra si è classificata al quarto posto nella classifica reale e pertanto è del tutto possibile che possa essere considerata la migliore in base ad alcuni fattori. A seguire i valori più elevati vengono assunti da Inter e Milan.

Da questi brevissimi apprezzamenti ottenuti raffrontando le linee di tendenza conseguite analizzando singolarmente i fattori e i valori medi ottenuti per ciascun indice per squadra, possiamo affermare che esiste un'effettiva corrispondenza tra l'andamento di ciascun fattore e il posizionamento delle squadre all'interno del campionato.

Alla luce di queste considerazioni gli allenatori potrebbero trarre beneficio da questo tipo di analisi, osservando come la propria squadra si posiziona per ciascuno di questi fattori identificando quelli che sarebbe maggiormente opportuno migliorare al fine di ottimizzare le prestazioni.

3.6.4 CLUSTER ANALYSIS.

Con il termine Cluster Analysis (analisi dei gruppi) si intende un metodo tipicamente esplorativo che consiste nella ricerca, all'interno delle n osservazioni p-dimensionali, di gruppi di unità tra loro simili, non sapendo a priori se tali gruppi omogenei esistono effettivamente nel data set. L'obiettivo, pertanto, è quello di classificare tali unità in gruppi con caratteristiche di coesione interna (le unità assegnate ad un medesimo gruppo devono essere tra loro simili) e di separazione esterna (i gruppi devono essere il più possibile distinti).

I metodi di formazione dei gruppi vengono distinti in gerarchici e non gerarchici.

I metodi gerarchici consentono di ottenere una famiglia di partizioni, con un numero di gruppi da n a 1, partendo da quella banale in cui tutte le unità sono distinte, per giungere a quella, pure banale, in cui tutti gli elementi sono riuniti in un unico gruppo.

I metodi non gerarchici forniscono un'unica partizione delle *n* unità in *g* gruppi, con *g* fissato a priori. Per questo metodo il numero ottimo di gruppi può essere determinato per tentativi, ripetendo più volte la procedura con diversi valori di *g*.

Dopo aver ricavato la famiglia di partizioni, occorre valutare la classificazione ottenuta, per vedere se essa soddisfa le condizioni di coesione interna e di separazione esterna. Possiamo dire che una partizione è soddisfacente quando la variabilità all'interno dei gruppi individuati è piccola (le unità di ogni gruppo presentano modeste differenze tra loro) ed inoltre i gruppi sono ben distinti l'uno dall'altro. È necessario trovare un compromesso tra numero di gruppi e omogeneità all'interno degli stessi in quanto riducendo il numero di gruppi si ottiene una classificazione più sintetica, e quindi generalmente più utile ai fini operativi, ma si deve pagare un prezzo in termini di maggiore variabilità nei gruppi poiché si aggregano unità maggiormente diverse tra loro (Sergio Zani e Andrea Cerioli, 2007).

Prendendo spunto da questi brevi cenni teorici, applichiamo questo metodo alla tabella riportata nel paragrafo precedente, che evidenzia per ogni squadra il valore medio associato a ciascuno degli otto indici individuati nelle precedenti analisi. Tale tabella prima di essere implementata attraverso il metodo Cluster Analysis deve essere standardizzata.

Il grafico sottostante evidenzia il numero ottimale di cluster per raggruppare le venti squadre oggetto della nostra analisi. Il codice R necessario per l'ottenimento di tale risultato è riportato all'interno dell'Appendice H.

Osservando il grafico possiamo notare quanto la varianza spiegata aumenta, all'aumentare del numero di cluster utilizzati. Più precisamente possiamo analizzare che passando da 2 a 3 gruppi si ha un notevole aumento nella varianza spiegata, pari addirittura al 54,75%. Incrementando ulteriormente i cluster notiamo che il miglioramento in varianza spiegata diminuisce in maniera marcata (infatti con 4 cluster si verifica un incremento del 18,99% mentre con 5 solo del 13,9%). Queste considerazioni ci portano ad optare per 3 raggruppamenti, in quanto rappresentano il massimo picco per il modello in miglioramento di varianza spiegata. Dopo il terzo cluster l'incremento in varianza diminuisce sempre più, fino a perdere completamente di significato.

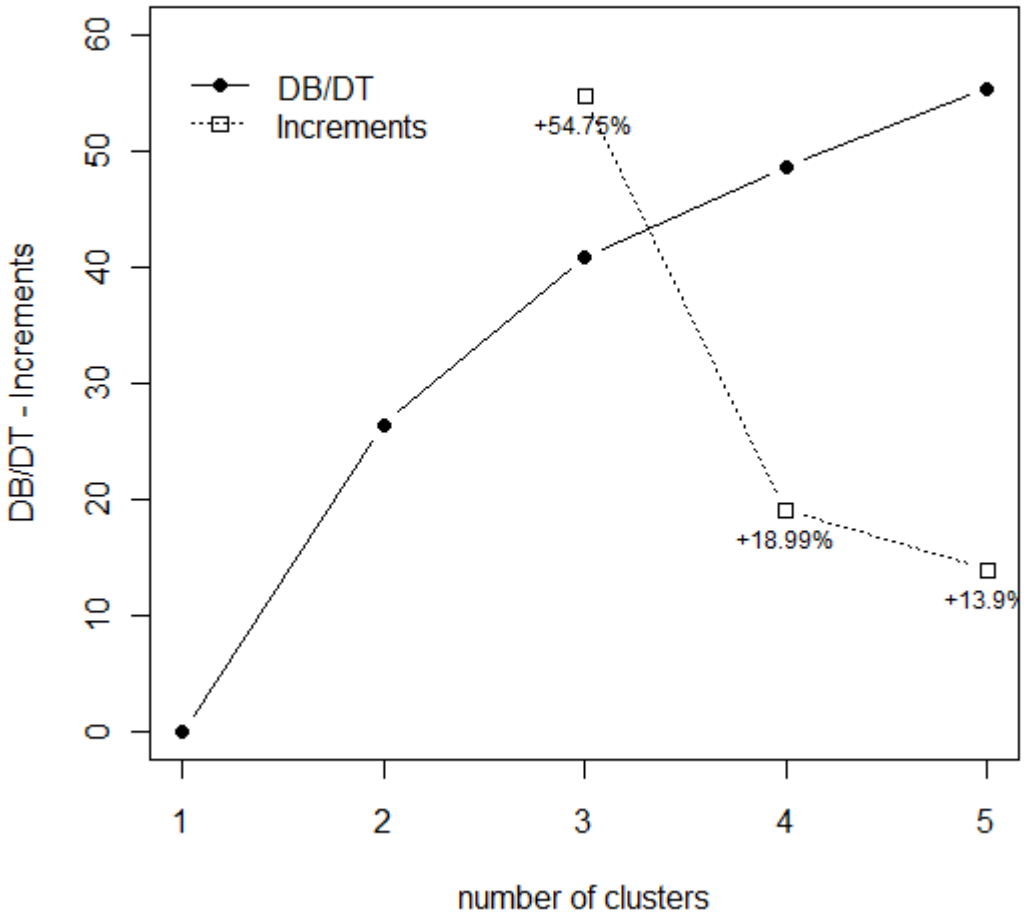


Figura 48: grafico che ci permette di osservare qual è il numero ottimale di cluster da considerare.

Il risultato ottenuto con il codice R corrisponde a quanto ottenuto attraverso il software statistico Systat, pertanto abbiamo deciso di utilizzare anche i grafici e le tabelle conseguiti da questo, per riuscire ad analizzare più nel dettaglio le classificazioni ottenute.

Cluster 1 of 3 Contains 6 Cases

Members		Statistics				
Case	**Distance**	**Variable**	**Minimum**	**Mean**	**Maximum**	**Standard Deviation**
Cagliari	0.781	PERICSQC1	-0.289	0.512	1.369	0.715
Catania	0.716	PERIC_SQC2	-1.214	-0.234	0.375	0.669
Chievo	0.782	DIFSQC1	-0.397	1.065	1.690	0.772
Genoa	0.711	DIFSQO1	-0.765	0.322	1.875	0.921
Juventus	0.475	DIFSQO2	-0.985	0.743	1.439	0.877
Napoli	0.648	PERICSQO1	-1.611	-0.531	0.567	0.750
		PERICSQO2	-0.742	-0.210	0.565	0.519
		PERICSQO3	-0.965	0.260	1.083	0.783

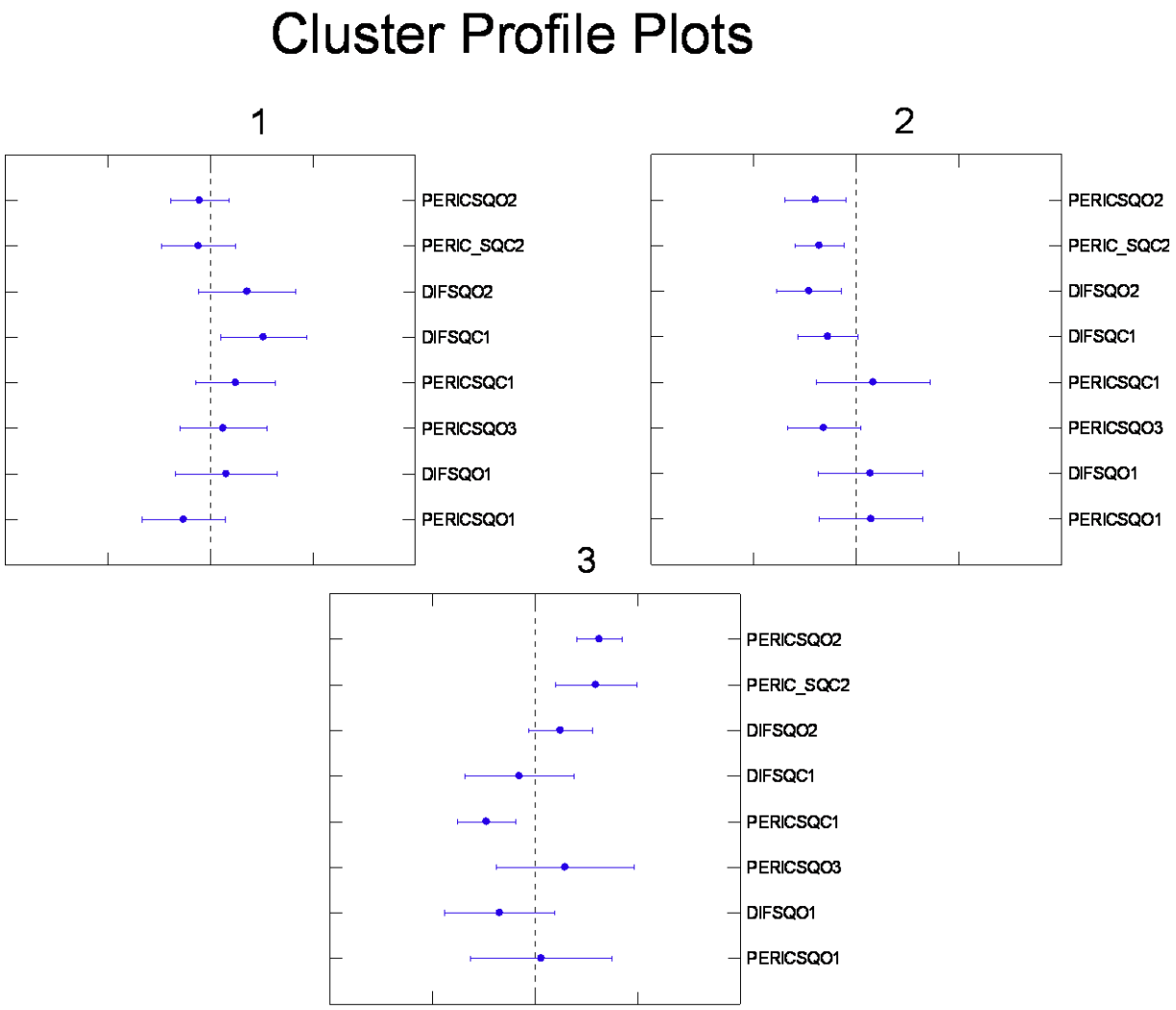
Cluster 2 of 3 Contains 8 Cases

Members		Statistics				
Case	**Distance**	**Variable**	**Minimum**	**Mean**	**Maximum**	**Standard Deviation**
Bari	0.897	PERICSQC1	-0.836	0.344	1.748	1.061
Bologna	0.553	PERIC_SQC2	-1.104	-0.739	0.335	0.458
Brescia	0.283	DIFSQC1	-1.563	-0.564	0.055	0.562
Cesena	0.786	DIFSQO1	-1.382	0.286	2.028	0.968
Fiorentina	0.680	DIFSQO2	-2.187	-0.940	-0.261	0.600
Lecce	0.529	PERICSQO1	-1.700	0.301	1.202	0.965
Parma	0.942	PERICSQO2	-1.478	-0.810	0.112	0.570
Sampdoria	0.801	PERICSQO3	-1.472	-0.649	0.622	0.679

Cluster 3 of 3 Contains 6 Cases

Members		Statistics				
Case	**Distance**	**Variable**	**Minimum**	**Mean**	**Maximum**	**Standard Deviation**
Inter	1.156	PERICSQC1	-1.543	-0.971	-0.217	0.529
Lazio	0.631	PERIC_SQC2	0.386	1.218	2.085	0.734
Milan	0.784	DIFSQC1	-1.869	-0.313	0.789	0.987
Palermo	0.854	DIFSQO1	-1.526	-0.703	0.952	0.995
Roma	0.778	DIFSQO2	-0.381	0.510	1.202	0.582
Udinese	0.602	PERICSQO1	-0.561	0.130	2.715	1.277
		PERICSQO2	0.739	1.291	1.766	0.417
		PERICSQO3	-0.756	0.605	2.220	1.244

Da queste tre tabelle possiamo individuare come le varie squadre che compongono il campionato possono essere raggruppate attraverso il metodo della Cluster Analysis. Inoltre in ogni gruppo, per ciascun fattore, viene evidenziato il valore minimo e massimo, la media e la deviazione standard registrata per le squadre contenute al proprio interno. Analizziamo ora per quali fattori si caratterizza ciascun cluster.



Il primo cluster, composto dalle squadre Cagliari, Catania, Chievo, Genoa, Juventus e Napoli si caratterizza in particolar modo per valori al di sopra della media per l'indice di *fase difensiva* quando queste giocano come squadra di casa. Il secondo cluster, composto dalle squadre Bari, Bologna, Brescia, Cesena, Fiorentina, Lecce, Parma e Sampdoria si caratterizza invece per valori al di sotto della media per i fattori: *giocate pericolose a palla bassa* quando queste squadre giocano in casa e *barriera difensiva a protezione dell'area* ed *indice di pericolosità d'attacco* quando giocano come squadre ospiti. Infine per quanto riguarda il terzo cluster composto dalle squadre Inter, Lazio, Milan, Palermo, Roma e Udinese i fattori maggiormente significativi, con valori al di sopra della media, sono le *giocate pericolose a palla bassa* quando le sqadre giocano in casa e la *pericolosità d'attacco* quando giocano con squadra ospite; inoltre si contraddistinguono anche per valori inferiori alla media relativamente al fattore *giocate aeree* quando le squadre giocano in casa.

Raffrontando questi risultati con l'analisi di ogni singolo fattore troviamo una buona corrispondenza tra il posizionamento delle squadre all'interno della classifica e i fattori che le

contraddistinguono maggiormente. Ma andiamo ad analizzare più nel dettaglio questa relazione.

Il primo cluster, come abbiamo accennato poco sopra, è composto dalle squadre Cagliari, Catania, Chievo, Genoa, Juventus e Napoli. Queste squadre, a parte il Napoli, hanno un posizionamento centrale all'interno della classifica e si contraddistinguono per valori al di sopra della media per quanto riguarda l'*indice di difesa o (fase difensiva)* quando si trovano a competere come squadra di casa. Tale fattore può essere considerato proprio come il carattere distintivo che contraddistingue questo cluster dai restanti. Dall'analisi dei singoli fattori era emerso che questo indice incideva positivamente, seppur in maniera non troppo marcata, sulla probabilità di vittoria della squadra di casa. Pertanto le squadre che si caratterizzano per elevati valori, solo in merito a questo fattore, sicuramente non saranno le peggiori all'interno della classifica, in quanto non comprendono indici per i quali è stata registrata una flessione della probabilità di vittoria. Tuttavia le squadre non comprendono nemmeno i due fattori per i quali l'incidenza positiva sulla probabilità di vittoria è molto marcata.

Per quanto riguarda il secondo cluster composto dalle squadre Bari, Bologna, Brescia, Cesena, Fiorentina, Lecce, Parma e Sampdoria non possiamo far altro che costatare il loro posizionamento nella parte finale della classifica, fatta eccezione per Fiorentina e Parma. Per quanto riguarda il carattere distintivo associato a queste squadre, emerso attraverso la Cluster Analisis, possiamo affermare che si contraddistinguono per valori al di sotto della media per quanto riguarda i fattori *indice di pericolosità d'attacco* e *barriera difensiva a protezione della porta*, quando le squadre si trovavano a competere come squadra ospite e *giocate aeree,* quando invece giocano come squadra di casa. Avere valori al di sotto della media per quanto riguarda il fattore riferito alle *giocate aeree* è un fatto positivo, in quanto, avevamo osservato attraverso l'analisi sull'influenza dei fattori che, all'aumentare dell'indice appena menzionato si registrava una flessione nella probabilità di vittoria in merito alla squadra di casa. Pertanto essendo un indice medio di performance delle squadre quando queste giocano in casa, avere valori al di sotto della media può essere considerato un fatto positivo. Tutt'altra considerazione deve essere fatta per i restanti due fattori. Questi, innanzitutto, si riferiscono ai valori medi associati alle squadre quando si trovano a gareggiare come squadre ospiti. Per questi indici l'analisi sull'influenza dei fattori era stata effettuata mettendo in luce come gli stessi influissero sulla probabilità di sconfitta della squadra di casa. Per entrambi, ma in particolar modo per l*'indice di pericolosità d'attacco*, era stata registrata una forte influenza positiva sulla probabilità di sconfitta della squadra di casa, questo significa che, all'aumentare

del valore attribuito a questi fattori la probabilità di vittoria assegnata alla squadra ospite subiva un netto miglioramento. Siccome stiamo considerando i fattori medi attribuiti alle squadre facenti parte del cluster (quando queste giocano come squadra ospite), non possiamo considerare in maniera positiva il fatto che queste assumano valori al di sotto della media per quanto riguarda questi indici. Questo le penalizza molto rispetto al posizionamento in classifica. In particolare, questo gruppo dovrebbe puntare maggiormente ad essere pericoloso in attacco quando si tratta di disputare un incontro come squadra ospite, in modo da veder migliorare anche il proprio posizionamento in graduatoria.

Passiamo ora ad analizzare i fattori che contraddistinguono il terzo ed ultimo cluster composto dalle squadre Inter, Lazio, Milan, Palermo, Roma e Udinese, posizionate nella parte alta della graduatoria. Come già osservato in precedenza queste squadre si differenziano dalle altre per valori al di sopra della media per quanto riguarda l'*indice di pericolosità d'attacco*, quando giocano come ospiti, e le *giocate pericolose a palla bassa*, quando giocano in casa. Inoltre le squadre assumono valori al di sotto della media per quanto riguarda l'*indice di difesa*, sempre riferito alle partite disputate in casa. Già da questi caratteri distintivi riusciamo a comprendere la motivazione per la quale queste squadre si trovano ai vertici della classifica. I fattori *indice di pericolosità d'attacco* e *giocate pericolose a palla bassa* sono quelli che maggiormente influenzano i risultati di una partita. Entrambi vengono caratterizzati da valori al di sopra della media e pertanto permettono una marcata influenza positiva per quanto riguarda la modalità vittoria. L'ultimo fattore che contraddistingue i cluster con valori al di sotto della media non influenza pesantemente il risultato della partita; tuttavia abbiamo osservato che, per valori elevati, rileva una lieve influenza positiva sulla vittoria quando le squadre giocano in casa. Di conseguenza l'unico appunto che può essere mosso a queste squadre è che, se migliorassero questo fattore, congiuntamente ai due riferiti alla difesa, attuata quando la quadra gioca come ospite, potrebbero ottenere risultati ancora migliori rispetto agli attuali.

3.6.5 PREVISIONI CAMPIONATO 2011/2012.

Arrivati a questo punto la curiosità ci ha spinti ad affrontare il discorso in merito all'attuale campionato italiano di Serie A. Mossi dall'ambizione di riuscire a prevedere l'andamento degli incontri di quest'anno, pur non avendo nessun dato a nostra disposizione, abbiamo deciso di addestrare la Random Forest con le partite dello scorso campionato, utilizzando sempre il database contenente gli otto fattori maggiormente interpretabili ottenuti attraverso

l'Analisi delle Componenti Principali. Il codice R relativo a questo tipo di analisi è disponibile in Appendice I.

Una volta addestrata la Random Forest abbiamo così ottenuto le previsioni riferite alle partite di quest'anno. I risultati non sono stati sfavillanti. Confrontando le previsioni con i risultati reali abbiamo constatato che il modello prevede il corretto andamento degli incontri nel 39,53% dei casi. Tale risultato, tuttavia, ci permettere di apprendere che le previsioni del modello sono leggermente più accurate rispetto alla scelta casuale della modalità corretta (probabilità pari a circa 33%).

Andiamo ad analizzare più nel dettaglio le motivazioni per le quali il modello non prevede al meglio le partite dell'attuale campionato:

- innanzi tutto alla fine di ogni campionato quasi tutte le squadre vengono stravolte dai cosiddetti calcio mercati, nei quali le grandi società cercano di accaparrarsi i migliori giocatori spendendo cifre da capogiro, mentre le piccole società cercano di capitalizzarsi vendendo i propri giocatori migliori e cercando di trovarne altrettanti bravi ma a basso costo. Data questa elevata variabilità nei componenti delle squadre è difficile che, passando da un anno all'altro, le performance riferite alle stesse rimangano invariate. Naturalmente se avessimo avuto uno storico abbastanza ampio tale varianza si sarebbe attenuata. Pertanto sarebbe stato molto interessante valutare il miglioramento nelle previsioni all'aumentare del numero di database disponibili.
- Oltre ai giocatori, molto spesso avviene allo stesso modo il cambio degli allenatori. Anche questo fatto influisce in maniera significativa sulle performance delle squadre, in quanto cambiano gli schemi di gioco e le formazioni considerate vincenti da ciascuno di essi.
- Terza motivazione, che avevamo già riscontrato per le previsioni relative al campionato 2010/2011 è che, il modello depura il gioco da tutti quegli elementi accidentali che possono essere legati alla fortuna. In questo ambito rientrano anche gli errori arbitrali, in quanto assumiamo che, quando si manifestano, non sono rivolti ad avvantaggiare una squadra rispetto ad un'altra, ma che avvengono in maniera del tutto casuale e che la probabilità che una squadra venga avvantaggiata oppure svantaggiata da un errore di questo tipo sia uguale per ogni squadra facente parte del campionato.
- Un altro fattore da tenere in considerazione è che, con i dati a nostra disposizione, non è stato possibile prevedere tutte le partite all'interno di ciascuna giornata, in quanto non abbiamo nessun dato in merito alle nuove tre squadre introdotte in Serie A. Inoltre

i dati relativi alle tre squadre retrocesse sono del tutto inutili ai fini delle previsioni del campionato in corso.

Seppur queste previsioni non siano state buonissime abbiamo provato a constatare l'accuratezza del modello per quelle partite nelle quali una delle tre possibili modalità (vittoria, sconfitta e pareggio) sia stata prevista con una probabilità maggiore del 50%. Da qui abbiamo scoperto che se si considerassero solo le partite in cui una modalità è stata prevista con una probabilità superiore alla soglia prima indicata, l'accuratezza del modello si amplierebbe al 54,55%, che tutto sommato, è un buonissimo risultato a fronte di tutte le considerazioni fatte in precedenza.

CONSIDERAZIONI FINALI E SVILUPPI FUTURI.

Arrivati alla conclusione di questa tesi, ripercorriamo i passi salienti che ci hanno condotti all'ottenimento dei vari risultati, illustrati nei capitoli precedenti.

Per procedere attraverso l'implementazione dei vari metodi statistici è stato necessario ridimensionare il database a nostra disposizione, in quanto, affetto da variabili ridondanti e di poco conto ai fini della determinazione delle variabili maggiormente significative per la definizione dei risultati delle varie partite di campionato; obiettivo principale di tutto questo lavoro.

Applicando la metodologia Random Forest al database contenente le variabili espresse come differenza di valori, tra quanto riferito alla squadra di casa e quanto riferito alla ospite, abbiamo ottenuto risultati particolarmente incoraggianti sotto il profilo delle previsioni, ma abbastanza difficili da interpretare. Per questo motivo sono stati utilizzati per redigere una nuova classifica del campionato, depurata dall'effetto aleatorietà che spesso influenza i risultati delle partite. Tale classifica quindi, tiene in considerazione solo gli elementi oggettivi, rilevabili durante un incontro e proprio sulla base di questi, prevede il risultato finale, attribuendo una probabilità a ciascuna delle tre possibili modalità con cui si può concludere una partita (vittoria, sconfitta, pareggio) e prendendo in considerazione quella con la probabilità più elevata. In questo modo il modello, valuta le performance di gioco effettive, riferite a ciascun incontro, senza considerare particolari effetti dovuti a fortuna, errori arbitrali o interferenze dovute, ad esempio, al recente scandalo Calcio Scommesse. Tale classifica pur non essendo identica a quanto riscontrato nella realtà presenta un buon indice di cograduazione pari a circa il 77% del massimo teorico. Attraverso questa analisi, è stato interessante valutare come il modello ha giudicato ogni singolo incontro e scoprire che le squadre che sarebbero dovute retrocedere in Serie B, congiuntamente alla squadra campione d'Italia, si differenziano dai risultati reali.

Data la scarsa interpretabilità dei fattori ottenuti attraverso questa prima analisi, abbiamo deciso di riapplicare il metodo relativo alla Random Forest al database depurato da tutte le *variabili differenza*. Tale analisi, ci ha permesso di ottenere variabili maggiormente interpretabili, a fronte di un lieve calo nella precisione delle previsioni. Grazie a questi risultati abbiamo applicato l'Analisi delle Componenti Principali alle sole variabili più significative, in modo da ottenere dei fattori, o indici di performance, relativi a ciascun

incontro; utili per valutare l'influenza di ognuno di essi sul risultato finale. Attraverso questa analisi sull'impatto dei fattori, si sono ottenuti risultati molto interessanti in merito alle strategie di gioco maggiormente premianti in termini di risultato, le quali potrebbero aiutare i commissari tecnici nella definizione degli schemi di gioco sui quali risulta più opportuno puntare. Oltre a tutto ciò, sono stati creati dei valori medi di performance per ciascuna squadra, sulla base dei valori attribuiti a tutte le partite disputate durante il campionato. Questi, congiuntamente alla Cluster Analysis, ci hanno permesso di apprendere quali sono i punti di forza e di debolezza relativi a ciascuna squadra e su quali di questi ogni allenatore dovrebbe puntare maggiormente per migliorare i propri risultati.

Proprio come accennato nella fase introduttiva, questo elaborato potrebbe essere utile ai vari allenatori e dirigenti sportivi, per valutare le performance di gioco effettive relative ad ogni incontro, captando in questo modo i principali errori commessi e cercando di migliorarsi in tale direzione.

Sempre nella fase introduttiva, si erano menzionate anche le società di scommesse, come soggetti interessati ai risultati di questa tesi. Tuttavia, per il momento, questi risultati non aiuterebbero in maniera significativa tali società, in quanto abbiamo notato che è presente una forte componente di variabilità legata al passaggio da un campionato al successivo. Per ovviare a questo problema, sarebbe necessario acquistare database riferiti a più stagioni e ripetere tutte le analisi compiute fino a questo momento, in modo da livellare la variabilità. Al fine di ottenere risultati ancora più precisi, sarebbe interessante valutare le performance relative a ciascun giocatore facente parte del campionato, e calcolare quindi le prestazioni delle squadre sulla base delle performance calcolate per ciascun giocatore che la compone. In questo modo, alla fine di ogni campionato, si potrebbero riunire, a livello numerico, i nuovi giocatori acquistati da ciascuna squadra e valutare le nuove prestazioni che questa presenterà nel nuovo campionato.

A fronte di tutto il lavoro appena presentato, sono convinta che il mondo del calcio si evolverà in questa direzione, in quanto la statistica può effettivamente aiutare, giocatori ed allenatori, a capire su quali variabili sia opportuno puntare per migliorare i propri risultati. Proprio come nel film *Money Ball*, sono sicura che una squadra possa conquistare risultati importanti non solo sulla base del budget di cui dispone per l'acquisto dei giocatori, ma anche e soprattutto sull'analisi approfondita ed accurata delle proprie performance alla fine di ogni match. Naturalmente non ho la presunzione di aver trovato l'analisi *perfetta* per l'ottenimento

di queste informazioni; tuttavia penso di aver dimostrato che con investimenti decisamente ridotti (un computer e un software statistico open source) si possano ottenere risultati molto interessanti, i quali potrebbero avere un elevato valore potenziale se divulgati sul mercato.

Pertanto mi auguro che, anche nel mondo del calcio, si faccia strada questa possibilità, la quale non deve essere vista solo come un'informazione accessoria fine a se stessa, ma come una grande opportunità; una sorta di secondo allenatore che scruta in maniera critica l'andamento di ogni incontro e che può fornire informazioni non banali, che difficilmente si potrebbero ottenere senza l'ausilio di metodi statistici.

APPENDICI.

APPENDICE A.

Qui di seguito verrà riportato l'elenco e la descrizione di ciascuna variabile contenuta all'interno del database. Per quanto riguarda il nome identificativo di ciascuna di esse, indicato a seguito della descrizione, è importante sottolineare che, ad esclusione della macro categoria *dati relativi alla giornata*, basterà aggiungere al nome della variabile il suffisso "_C" se si vuole considerare la variabile in merito alla squadra di casa, il suffisso "_O" se si considera quella riferita alla squadra ospite, ed infine, il prefisso "DIF_" se si vuole considerare la variabile differenza tra il valore attribuito alla squadra di casa e quello assunto dalla ospite.

Macro categorie e variabili contenute nel database:

1) DATI RELATIVI ALLA GIORNATA
   - *Nome della squadra che gioca in casa*: (NOME_C).
   - *Nome della squadra ospite*: (NOME_O).
   - *Giornata*: indica la giornata di campionato (GIORNATA).
   - *Risultato della partita*: indicato come nella tradizionale schedina del Totocalcio con i simboli 1, 2 ed X (RIS).
   - *Risultato positivo*: considera come risultato positivo la sola vittoria della squadra di casa, in caso di pareggio o sconfitta la variabile assume valore pari a zero (RIS_C1).
   - *Risultato positivo per almeno una delle squadre*: esclude dall'analisi i casi di pareggio, la variabile assume valore pari ad 1 nel caso di vittoria della squadra di casa e 0 nel caso di vittoria della squadra ospite. (RIS_C2)
   - *Meteo*: indica le condizioni meteorologiche registratesi durante le partite, questa variabile dovrà essere modificata in fase di preparazione dei dati in modo da trasformare le varie condizioni climatiche in condizioni favorevoli o sfavorevoli per la partita. Questa variabile è stata inserita per verificare se il tempo sfavorevole può arrivare addirittura ad essere una tra le variabili più significative per la determinazione del risultato di una partita. Naturalmente ci aspettiamo che il tempo sfavorevole incida per lo meno sulla precisione dei passaggi. (METEO)
2) DATI GENERALI

- *Punti*: si riferisce al classico punteggio che viene attribuito alle squadre alla fine di un match, 3 punti se la squadra ha vinto, 1 se ha pareggiato e 0 se ha perso. (G_PUNTI1)
- *Punti media inglese*: si ottiene partendo dal presupposto che per un piazzamento ad alti livelli in un campionato di calcio, le partite in casa si debbano vincere, mentre in trasferta si abbia il compito di perseguire almeno un pareggio; da qui la semplificazione su base statistica che una vittoria in casa ed un pareggio in trasferta sarebbero la condizione ideale per poter condurre ai vertici un campionato. Questa variabile quando è espressa in merito al risultato della partita per la squadra di casa, assume i seguenti valori: 0 punti in caso di vittoria, -2 in caso di pareggio e -3 in caso di sconfitta; mentre quando si riferisce al risultato della squadra ospite assume valori diversi, e più precisamente: +2 in caso di vittoria, 0 in caso di pareggio e -1 in caso di sconfitta. (G_PUNTI2)
- *Minuti di gioco*: misura l'effettiva durata dell'incontro in minuti, tenendo anche conto degli eventuali minuti di recupero assegnati dall'arbitro. (G_MINUTI)
- *Palle giocate*: numero di volte che i giocatori di una stessa squadra toccano la palla. In altre parole, indica per quanto tempo una squadra è riuscita a fraseggiare il possesso del pallone. (G_PALLE_GIOC)
- *Possesso palla*: viene espresso in secondi. (G_POS_PAL)
- *Supremazia territoriale*: indica il lasso di tempo in cui una squadra riesce a mantenere il pallone all'interno della metà campo avversaria. (G_SUP_TER)
- *Velocità di gioco*: viene rappresentata in termini di tempo di possesso palla medio di ciascun giocatore per effettuare una giocata. (G_VEL_GIOC)
- *Passaggi rapidi*: numero di passaggi rapidi, o meglio chiamati passaggi in prima. (G_PAS_RAP)
- *Passaggi tentati*: numero totale di passaggi tentati da una squadra, comprende sia quelli andati a buon fine, che quelli sbagliati, si considerano sbagliati quelli usciti dal campo o intercettati da un giocatore avversario. (G_PAS_TEN)
- *Percentuale passaggi riusciti*: è la percentuale di passaggi andati a buon fine rispetto al numero totale di passaggi effettuati. (G_%PAS_RIU)
- *Passaggi riusciti*: è l'insieme delle volte che la palla viene passata correttamente da un compagno ad un altro, in qualsiasi direzione o zona del campo. Questa variabile può essere ottenuta dalla somma delle tre variabili successive (passaggi

riusciti in difesa, passaggi riusciti a centrocampo e passaggi riusciti conteggiati in attacco). (G_PASRIU2)

- *Passaggi riusciti in difesa*: numero di passaggi effettuati correttamente in difesa. (G_PASRIU3)
- *Passaggi riusciti a centrocampo*: numero di passaggi effettuati correttamente a centrocampo. (G_PASRIU4)
- *Passaggi riusciti conteggiati in attacco*: numero passaggi effettuati correttamente in attacco. (G_PASRIU5)
- *Passaggi ricevuti*: numero totale di passaggi ricevuti; è formato dalla somma delle tre variabili seguenti (passaggi ricevuti in difesa, passaggi ricevuti a centrocampo e passaggi ricevuti in attacco). (G_PASRIC1)
- *Passaggi ricevuti in difesa*: numero di passaggi ricevuti correttamente nella zona di difesa. (G_PASRIC2)
- *Passaggi ricevuti a centrocampo*: numero di passaggi ricevuti a centrocampo. (G_PASRIC3)
- *Passaggi ricevuti in attacco*: numero di passaggi ricevuti nella zona di attacco. (G_PASRIC4)
- *Passaggi corti ricevuti*: numero totale di passaggi tra due giocatori della stessa squadra che si trovano a poca distanza l'uno dall'altro. (G_PASCORTI1)
- *Passaggi corti utili ricevuti*: il termine utile viene utilizzato quando l'azione considerata elimina dalla fase difensiva un avversario. Quindi questa variabile sta ad indicare il numero di passaggi corti ricevuti che hanno permesso di scavalcare un avversario che si trovava in fase difensiva, permettendo così al giocatore di proseguire con l'attacco. (G_PASCORTI2)
- *Passaggi di testa ricevuti*: numero di passaggi di testa effettuati correttamente ad un giocatore della propria squadra. (G_PASTESTA)
- *Passaggi lunghi ricevuti*: numero di passaggi ricevuti tra due giocatori della stessa squadra che si trovano a notevole distanza tra loro. (G_PASLUNGHI)
- *Lungo linea ricevuti*: numero di passaggi ricevuti in prossimità del lato lungo del campo. (G_LUNGO)
- *Passaggi filtranti ricevuti*: si verifica quando un giocatore mette in condizione favorevole per una conclusione un compagno, eludendo la difesa avversaria. (G_PASFIL)

- *Pallonetti ricevuti*: tiro, con traiettoria parabolica effettuato per scavalcare la linea avversaria, ricevuto da un giocatore della propria squadra. (G_PALLO)
- *Rinvii ricevuti*: numero totale di rinvii andati a buon fine. (G_RINVII)
- *Passaggi ricevuti su cross su azione*: il cross viene detto anche "traversone", ed è un passaggio costituito da un tiro lungo che alza il pallone da terra. In genere il cross mette in condizione l'attaccante di deviare la palla in gol con un colpo di testa o con un tiro al volo. È chiamato così proprio perché il passaggio incrocia il movimento dell'attaccante. Questa variabile rappresenta il numero di passaggi ricevuti su cross in zona di attacco. (G_PASRIC5)
- *Passaggi ricevuti su cross piazzato*: rappresenta il numero totale di passaggi ricevuti su cross piazzato, quindi ricevuti da un cross calciato nel momento in cui il pallone si trovava fermo sul campo (punizione o corner). (G_PASRIC6)
- *Passaggi alti ricevuti*: si considera passaggio alto quando supera l'altezza del collo. Questa variabile indica il numero di passaggi effettuati in questa modalità che sono stati ricevuti da un compagno della propria squadra. (G_PASALTI1)
- *Passaggi bassi ricevuti*: i passaggi bassi invece, sono tutti quelli che non superano l'altezza del collo, e come per la variabile precedente, la variabile in questione indica il numero di passaggi ricevuti correttamente da un compagno della propria squadra effettuati con questa modalità. (G_PASBASSI1)
- *Passaggi corti nella metà campo avversaria*: passaggi effettuati tra due giocatori della stessa squadra che si trovano a poca distanza tra di loro e che nel momento in cui viene effettuato il passaggio si trovano nella metà campo avversaria. (G_PASCORTI3)
- *Passaggi corti utili nella metà campo avversaria*: questa variabile rappresenta una frazione della variabile precedente. Infatti considera solo i passaggi effettuati in quella particolare zona del campo e tra giocatori che si trovano a poca distanza tra di loro. Questi passaggi sono risultati utili alla squadra, in quanto permettono di mantenere il possesso della palla. (G_PASCORTI4)
- *Passaggi corti utili*: passaggi effettuati tra due giocatori della stessa squadra che si trovano a breve distanza tra di loro e che tramite questo tipo di passaggio sono riusciti a mantenere il possesso della palla. (G_PASCORTI5)
- *Passaggi alti*: passaggi effettuati mandando la palla sopra l'altezza del collo. (G_PASALTI2)

- *Passaggi bassi*: passaggi effettuati mantenendo la palla sotto l'altezza del collo. (G_PASBASSI2)
- *Passaggi bassi nella metà campo avversaria*: passaggi effettuati dai giocatori di una squadra nella metà campo avversaria mantenendo la palla sotto l'altezza del collo. (G_PASBASSI3)
- *Passaggi bassi utili nella metà campo avversaria*: questa variabile rappresenta un sottogruppo della variabile precedente, in quanto di questa vengono considerati solo i passaggi che sono risultati utili in quanto hanno permesso di mantenere il possesso della palla. (G_PASBASSI4)
- *Giocate utili*: è formata dalla somma delle giocate utili in difesa, a centrocampo e in attacco. Sta ad indicare il numero di giocate in verticale che hanno permesso di eliminare la pressione avversaria. (G_GIOC_UT)
- *Percentuale giocate utili*: è il rapporto tra le giocate utili e il totale di palle giocate. (G_%GIOC_UT)
- *Giocate utili in difesa*: numero di giocate in verticale che hanno permesso di eliminare un giocatore avversario, effettuate nella zona di difesa. (G_GIOC_UT1)
- *Giocate utili a centrocampo*: esprime il numero di giocate, effettuate a centrocampo che hanno permesso di scavalcare un giocatore avversario. (G_GIOC_UT2)
- *Giocate utili in attacco*: numero di giocate nella zona di attacco che hanno permesso di scavalcare un giocatore avversario. (G_GIOC_UT3)
- *Colpi di testa*: con il termine colpo di testa si possono intendere due situazioni: l'intercettazione di testa, di un pallone calciato da un componente della squadra avversaria, normalmente si effettua in fase difensiva soprattutto in caso di cross avversari; o il passaggio effettuato sempre di testa tra compagni della medesima squadra, normalmente per smarcare un compagno o per tentare una realizzazione su cross. (G_COLPTESTA)
- *Colpi di testa in area avversaria*: ci troviamo nella fase offensiva dove normalmente il pallone viene colpito di testa per far si che un compagno riesca a smarcarsi o per tentare una realizzazione su cross. (G_COLPTESTA1)
- *Palle giocate sulla fascia destra*[1]: numero di palle giocate nella zona a destra del campo. (G_PALLE1)

[1] Destra e sinistra vengono sempre considerate in merito alla squadra che si è presa in considerazione.

- *Percentuale palle giocate sulla fascia destra*: palle giocate sulla fascia destra / palle giocate. (G_%PALLE)
- *Palle giocate sulla fascia destra in difesa*: palle giocate nella zona della propria porta sulla fascia destra. (G_PALLE2)
- *Palle giocate sulla fascia destra a centrocampo*: numero di palle giocate sulla fascia destra della metà campo. (G_PALLE3)
- *Palle giocate sulla fascia destra in attacco*: palle giocate sulla fascia destra della zona di attacco la quale si trova in prossimità della porta avversaria. (G_PALLE4)
- *Palle giocate sulla fascia centrale*: la fascia centrale è quella che si trova nella parte centrale del campo in senso verticale. La variabile rappresenta il numero di palle giocate in questa zona del campo. Può essere calcolata anche dalla somma di palle giocate sulla fascia centrale in difesa, palle giocate sulla fascia centrale a centrocampo e palle giocate sulla fascia centrale in attacco. (G_PALLE5)
- *Percentuale palle giocate sulla fascia centrale*: palle giocate sulla fascia centrale / palle giocate. (G_%PALLE2)
- *Palle giocate sulla fascia centrale in difesa*: numero di palle giocate nella zona centrale di difesa. (G_PALLE6)
- *Palle giocate sulla fascia centrale a centrocampo*: numero di palle giocate nella zona centrale di centrocampo. (G_PALLE7)
- *Palle giocate sulla fascia centrale in attacco*: numero di palle giocate nella parte centrale di attacco. (G_PALLE8)
- *Palle giocate sulla fascia sinistra*: numero di palle giocate nella zona a sinistra del campo. (G_PALLE9)
- *Percentuale palle giocate sulla fascia sinistra*: palle giocate sulla fascia sinistra / palle giocate (totali). (G_%PALLE3)
- *Palle giocate sulla fascia sinistra in difesa*: palle giocate nella zona difensiva della propria metà campo sulla fascia di sinistra. (G_PALLE10)
- *Palle giocate sulla fascia sinistra a centrocampo*: (G_PALLE11)
- *Palle giocate sulla fascia sinistra in attacco*: palle giocate sulla fascia sinistra della zona di attacco la quale si trova all'interno della metà campo avversaria. (G_PALLE12)
- *Palle giocate in difesa*: numero di volte che i giocatori di una squadra si sono passati palla in difesa. (G_PALLE13)

- *Percentuale palle giocate in difesa*: viene calcolata mettendo in rapporto il numero di palle giocate in difesa e il numero di palle giocate. (G_%PALLE4)
- *Palle giocate a centrocampo*: rappresenta il numero di volte in cui i giocatori di una squadra si sono passati palla nella zona di centrocampo. (G_PALLE14)
- *Percentuale palle giocate a centrocampo*: è ottenuta mediante il rapporto tra palle giocate a centrocampo e il numero totale di palle giocate. (G_%PALLE5)
- *Palle giocate in attacco*: numero di volte in cui i giocatori di una squadra si sono passati palla nella zona di attacco. (G_PALLE15)
- *Percentuale di palle giocate in attacco*: viene ricavata dal rapporto tra le palle giocate in attacco e il numero totale di palle giocate. (G_%PALLE6)
- *Misura della squadra*: è la distanza tra il difensore più arretrato e l'attaccante più avanzato. (G_MISSQU)
- *Larghezza della squadra*: rappresenta la distanza tra il giocatore più a destra e quello più a sinistra della stessa squadra. (G_GSQU)
- *Baricentro in metri*: è il punto medio in verticale in cui i giocatori di una squadra toccano la palla: 0 rappresenta la propria porta, mentre 100 rappresenta la porta avversaria. (G_BARIC)
- *Pressing in metri*: rappresenta il punto medio in verticale in cui i giocatori di una squadra recuperano la palla: 0 rappresenta la propria porta, mentre 100 rappresenta la porta avversaria. Il pressing è una strategia per recuperare la palla attraverso tutto il movimento della squadra. Questa tecnica mette in difficoltà la squadra avversaria che in quel momento detiene il possesso della palla. (G_PRES)
- *Giocate spettacolari*: sono tiri particolari, come la rovesciata, il tiro al volo o il colpo di tacco. (G_GIOC_SP)
- *Errori eclatanti*: sono errori clamorosi, davanti ai quali si rimane basiti. (G_ERRORI)
- *Pali*: numero di volte che la palla tocca i pali della porta avversaria; non viene considerato palo nel caso in cui la palla dopo aver colpito il palo finisce in rete. (G_PALI)
- *Reti*: somma totale di gol segnati da entrambe le squadre durante la partita. (G_RETI)
- *Autoreti*: rappresenta la somma delle reti causate dalle autoreti di entrambe le squadre durante la partita. Il termine autorete viene attribuito al giocatore che

devia la palla nella propria rete senza che precedentemente nessun avversario abbia effettuato un tiro. (G_AUTORETI)

- *Sostituito*: la sostituzione avviene quando un giocatore lascia il terreno di gioco per far posto ad un compagno. La variabile indica il numero totale di sostituzioni effettuate da una squadra nel corso di una partita. (G_SOSTITUTO)
- *Ammonizioni*: questa variabile sta ad indicare il numero totale di cartellini gialli ricevuti da una squadra a causa di infrazioni al regolamento del calcio[2]. (G_AMMON)
- *Espulsioni*: numero di giocatori fatti uscire dal campo a causa dei troppi falli commessi (alla seconda ammonizione) o a causa di un fallo molto grave nei confronti di un giocatore avversario. La squadra in cui il giocatore viene espulso rimane in inferiorità numerica rispetto all'avversaria. Il numero di espulsioni viene conteggiato a seconda di quanti cartellini rossi sono stati attribuiti ad una squadra. (G_ESPULS)
- *Ostruzionismo*: sono i così detti falli tattici. Si registrano quando un giocatore provoca fallo per bloccare un'azione pericolosa e per permettere anche al resto della squadra di tornare in difesa. (G_OSTRUZ)
- *Verticalizzazione*: passaggio da un lato all'altro del campo, principalmente per sviluppare una fase d'attacco. (G_VERTICA)
- *Intraprendenza*: avviene quando un giocatore prende l'iniziativa, magari saltando dei giocatori avversari. (G_INTRAPRE)

3) DATI SUI PORTIERI

- *Reti subite*: numero di tiri entrati nella propria porta. Determina il numero di goal segnati dalla squadra avversaria. (P_RETI_SUB)
- *Reti subite direttamente su calcio piazzato*: numero di reti subite da palla ferma a causa di una punizione o di una rimessa in gioco da angolo. (P_RETI_SUB1)
- *Reti subite indirettamente su calcio piazzato*: sta ad indicare il numero di gol subiti indirettamente da calcio piazzato, questo significa che la palla indirizzata tramite calcio piazzato verso l'area viene deviata in rete da un compagno. (P_RETI_SUB2)
- *Autoreti subite*: rappresenta il numero di gol segnati da una squadra nella propria porta. (P_AUTORETI)

[2] Tra i più noti casi di infrazione al regolamento del calcio possiamo trovare i falli, le proteste dei giocatori nei confronti dell'arbitro, le simulazioni, le quali rientrano nei comportamenti antisportivi e le esultanze eccessive come ad esempio il fatto di togliersi la maglietta dopo un gol.

- *Rigori subiti*: un rigore o massima punizione è un calcio piazzato assegnato alla squadra avversaria a causa di una propria infrazione commessa all'interno dell'area di rigore. (P_RIG_SUB)
- *Reti subite su rigore*: numero di gol subiti a causa di rigori assegnati alla squadra avversaria. (P_RETI_SUB3)
- *Reti totali subite su calcio piazzato*: numero di gol subiti a causa di corner, punizioni e rigori. (P_RETI_SUB4)
- *Tiri subiti*: numero totale di tiri subiti in direzione della propria porta. (P_TIRI_SUB)
- *Tiri dentro subiti*: numero di tiri subiti in direzione dello specchio della propria porta. (P_TIRI_SUB1)
- *Tiri subiti ribattuti o deviati*: numero di tiri deviati dal portiere o ribattuti da quest'ultimo e che quindi non hanno portato ad una occasione di gol per la squadra avversaria. (P_TIRI_SUB2)
- *Parate*: azione propria dei portieri; viene attribuita ogni volta che il portiere blocca o respinge un tiro degli avversari. (P_PARATE)
- *Parate su tiri da dentro area*: numero di parate effettuate dal portiere su tiri partiti all'interno della propria area. (P_PARATE1)
- *Parate su tiri da fuori area*: numero di parate effettuate dal portiere su tiri calciati al di fuori della propria area. (P_PARATE2)
- *Parate su occasione*: numero di parate effettuate su occasione degli avversari. (P_PARATE3)
- *Parate su rigore*: numero di volte in cui il portiere è stato in grado di parare i rigori assegnati alla squadra avversaria. (P_PARATE4)
- *Percentuale di reti / tiri dentro subiti*: rappresenta il rapporto tra reti subite e tiri dentro subiti, cioè i tiri effettuati nello specchio della porta avversaria. (P_TIRI_SUB5)
- *Uscite*: numero di volte in cui il portiere è uscito dalla propria porta. (P_USCITE)
- *Uscite su calcio piazzato*: numero di volte in cui il portiere è uscito dalla propria porta per recuperare un pallone proveniente da calcio piazzato. Durante la punizione da calcio piazzato si crea affollamento in area, quindi se il portiere esce dalla propria porta e recupera il pallone, provoca un vantaggio ai difensori della propria squadra. (P_USCITE1)

- *Uscite su azione*: rappresenta il numero di volte in cui il portiere è uscito dalla propria porta per bloccare un'azione della squadra avversaria, con palla in movimento. (P_USCITE2)
- *Uscite su cross su azione*: numero di volte in cui il portiere è uscito dalla propria porta per andare a bloccare un cross su azione effettuato dalla squadra avversaria. Su azione significa che non è stato effettuato da calcio piazzato. (P_USCITE3)
- *Uscite alte*: modalità di recupero della palla propria del portiere. L'azione viene attribuita quando il portiere recupera la palla con le mani (dopo uno spostamento verso la palla stessa, indipendentemente dalla posizione degli avversari) sopra l'altezza del collo. Se un portiere tocca la palla con le mani dopo un retropassaggio (che non può essere di piede) di un compagno non si considera uscita a meno che vi sia un avversario nelle vicinanze. I recuperi di testa fuori dall'area sono da considerarsi uscite alte (di testa). Se il portiere salta per prendere la palla è sempre uscita alta. (P_USCITE4)
- *Uscite alte su passaggio lungo*: numero di uscite effettuate dal portiere per recuperare palla proveniente da un passaggio lungo sopra l'altezza delle spalle. (P_USCITE5)
- *Uscite alte su pallonetto*: numero di uscite alte effettuate per recuperare palla che proviene da un passaggio con parabola la cui componente verticale è preminente e passa sopra l'altezza delle spalle dei giocatori. Questo passaggio per essere considerato pallonetto non deve avvenire a più di 30 metri di distanza. (P_USCITE6)
- *Uscite alte su cross*: numero di uscite alte effettuate dal portiere per recuperare palla che proviene da un cross della squadra avversaria. (P_USCITE7)
- *Uscite alte su cross su azione*: numero di uscite alte effettuate dal portiere per recuperare palla che proviene da un cross su azione (quindi non proveniente da un calcio piazzato) della squadra avversaria. (P_USCITE8)
- *Uscite alte su cross dal fondo su azione*: numero di uscite alte per recuperare una palla che proviene da un cross effettuato nella parte finale del campo, ma non su punizione. (P_USCITE9)
- *Uscite alte su cross da calcio piazzato*: numero di palle recuperate dal portiere sopra l'altezza del collo che provengono da un cross effettuato da calcio piazzato (punizione o corner), pertanto calciato nel momento in cui la palla si trova ferma. (P_USCITE10)

- *Uscite basse*: si tratta sempre di una modalità di recupero della palla propria del portiere. L'azione viene attribuita quando il portiere recupera palla con le mani (dopo uno spostamento verso la palla stessa, indipendentemente dalla posizione degli avversari) sotto l'altezza del collo. Se un portiere recupera la palla all'interno dell'area con i piedi non si considera uscita. Se il portiere recupera palla con le mani dopo un retropassaggio (che non può essere di piede) di un compagno non si considera uscita, anche se fuori dall'area, a meno che vi sia un avversario nelle vicinanze. I recuperi di piede fuori dall'area sono da considerarsi uscite basse. (P_USCITE11)
- *Uscite basse su passaggio lungo*: numero di palle recuperate dal portiere sotto l'altezza del collo che provenivano da un passaggio lungo. (P_USCITE12)
- *Uscite basse su pallonetto*: numero di palle recuperate dal portiere sotto l'altezza del collo e provenienti da un pallonetto, passaggio con parabola la cui componente verticale è preminente e passa sopra l'altezza delle spalle dei giocatori; questo passaggio per essere considerato pallonetto non deve avvenire a più di 30 metri di distanza. (P_USCITE13)
- *Uscite basse su dribbling*: numero di recuperi effettuati dal portiere sotto l'altezza del collo e provenienti da un tiro effettuato da un giocatore della squadra avversaria, dopo che quest'ultimo ha superato uno o più giocatori della propria squadra mediante finte. (P_USCITE14)
- *Uscite basse su cross*: numero di recuperi effettuati dal portiere al di sotto dell'altezza del collo su cross della squadra avversaria. (P_USCITE15)
- *Uscite basse su cross su azione*: numero di recuperi effettuati dal portiere sotto l'altezza del collo per recuperare il pallone che proviene da un cross effettuato non da calcio piazzato. (P_USCITE16)
- *Uscite basse su cross su azione dal fondo*: numero di uscite basse effettuate dal portiere per recuperare una palla che proviene da un cross effettuato nella parte finale del campo, ma non su calcio piazzato. (P_USCITE17)
- *Uscite basse su cross da calcio piazzato*: numero di palle recuperate dal portiere sotto l'altezza del collo che provengono da un cross effettuato da calcio piazzato (punizione, rigore o corner), quindi con palla calciata da ferma. (P_USCITE18)
- *Uscite basse su passaggio filtrante*: numero di recuperi effettuati dal portiere al di sotto dell'altezza del collo intercettando un passaggio filtrante avversario. (P_USCITE19)

- *Rinvii del portiere*: numero totale di rinvii effettuati dal portiere. Viene ottenuto sommando i rinvii di piede con quelli di mano. (P_RINVII)
- *Rinvii utili del portiere*: numero di rinvii che arrivano ad un giocatore della propria squadra. (P_RINVII1)
- *Rinvii di piede del portiere*: quando il portiere calcia al volo o di drop. Con il termine drop si intende quando il portiere con la palla in mano la lascia cadere verso il basso e la colpisce al volo con il piede. (P_RINVII2)
- *Rinvii di mano del portiere*: si usa solo quando il portiere rilancia la palla con le mani. (P_RINVII3)
- *Rilanci lunghi utili*: quando il pallone, rinviato lungo, arriva ad un giocatore della propria squadra. (P_RILANCI)
- *Rilanci lunghi*: numero di rilanci effettuati dal portiere verso i giocatori della propria squadra che non si trovano nelle immediate vicinanze. (P_RILANCI1)

4) FASE DIFENSIVA
   - *Percentuale di protezione dell'area*: viene calcolata facendo il rapporto tra il numero di palle recuperate dalla squadra in difesa e il numero totale di palle che gravitano in area. (D_%PROT)
   - *Rinvii*: numero di rinvii effettuati da tutta la squadra oggetto di analisi, ad esclusione dei rigori. Questa variabile non tiene conto del fatto che la palla sia arrivata ad un giocatore della propria squadra oppure no. (D_RINVII)
   - *Rinvii utili*: numero di rinvii effettuati da un qualsiasi giocatore di una squadra che hanno permesso di scavalcare almeno un giocatore avversario. (D_RINVII1)
   - *Rinvii di piede non del portiere*: possono essere le punizioni effettuate da tutti i componenti di una squadra ad esclusione del portiere. (D_RINVII2)
   - *Rinvii utili non del portiere*: sono i rinvii che hanno permesso di scavalcare almeno un giocatore della squadra avversaria che si trovava in fase difensiva, effettuati da un qualsiasi giocatore della squadra ad esclusione del portiere. (D_RINVII4)
   - *Rinvii di piede*: avvengono o quando il portiere calcia al volo o quando un difensore in difficoltà calcia la palla verso la metà campo avversaria senza volontà di direzionarla. (D_RINVII5)
   - *Rinvii di mano*: numero totale di rinvii effettuati di mano da tutti i componenti di una squadra; possono essere effettuati o dal portiere o da un qualsiasi giocatore che rinvia la palla da fallo laterale. (D_RINVII6)

- *Rinvii da zona arretrata*: rinvii effettuati in prossimità dei 30 metri della propria porta. Vi possono essere due particolari situazioni nelle quali questo tiro viene effettuato con maggior frequenza: o quando un difensore rinvia lungo in modo da allontanare il pericolo e far si che gli avversari prima di tornare all'attacco debbano andare a recuperare il pallone; oppure quando una squadra fa fatica a superare i giocatori della centrocampo avversaria ed allora opta per un passaggio lungo che arrivi direttamente agli attaccanti. La prerogativa del rinvio è di allontanare il pericolo. (D_RINVII7)
- *Rinvii utili da zona arretrata*: è costituita da un sottoinsieme di valori della variabile precedente, in quanto a differenza di quella, questa considera solo i rinvii che sono risultati utili, e che quindi hanno permesso di scavalcare almeno un giocatore avversario. (D_RINVII8)
- *Palle recuperate*: azione attribuita ogni volta che un giocatore recupera palla dopo un giocatore della squadra avversaria. (D_PALLE)
- *Palle recuperate e perse con rinvio non del portiere*: numero di palle ricevute da un rinvio effettuato da un giocatore diverso dal portiere ed immediatamente perse a vantaggio della squadra avversaria. (D_PALLE2)
- *Palle recuperate e perse con rinvio non del portiere in attacco*: numero di palle ricevute da un rinvio effettuato da un giocatore diverso dal portiere che si trova in zona di attacco, immediatamente, tuttavia, la palla viene persa a vantaggio della squadra avversaria. (D_PALLE3)
- *Recuperi in difesa*: la palla viene recuperata nella zona di difesa, all'interno della propria metà campo. (D_RECUPERI)
- *Recuperi a centrocampo*: numero di recuperi palle effettuati a centrocampo. (D_RECUPERI1)
- *Recuperi in attacco*: numero di recuperi palle effettuati in zona di attacco. (D_RECUPERI2)
- *Palle recuperate effettive*: numero di palle recuperate dalla squadra e rigiocate. (D_PALLE4)
- *Palle recuperate e perse subito*: numero totale di palle recuperate da una squadra ed immediatamente dopo riperse a vantaggio degli avversari. (D_PALLE5)
- *Palle recuperate e perse subito in attacco*: quando un giocatore in fase di attacco riesce a recuperare la palla dagli avversari ma immediatamente la riperde. Il

termine attacco si riferisce alla zona del campo in cui avviene l'azione appena considerata. (D_PALLE6)

- *Recuperi temporanei in attacco*: recuperi, che rimangono tali per un breve lasso di tempo e vengono effettuati in zona avanzata. (D_RECUPERI3)
- *Palle recuperate e perse subito in difesa*: quando un giocatore in fase di difensiva riesce a recuperare la palla dagli avversari ma immediatamente la riperde. (D_PALLE7)
- *Palle recuperate e rinviate non dal portiere in difesa*: si registra quando in zona difensiva il giocatore che è riuscito a recuperare il pallone lo rinvia lungo cercando di farlo arrivare ad un compagno che si trova in zona avanzata. (D_PALLE8)
- *Palle riconquistate*: numero di palle che arrivano ad un giocatore della propria squadra su errore degli avversari. (D_PALLE9)
- *Recuperi per fine azione avversaria*: numero di recuperi effettuati grazie al fatto che la squadra avversaria ha provocato la fuoriuscita dal campo della palla. (D_RECUPERI4)
- *Percentuale per fine azione avversaria*: rapporto fra recuperi per fine azione avversaria e palle riconquistate. (D_FINAZ)
- *Recuperi effettivi*: numero di recuperi che hanno fatto si che la palla rimanesse in possesso della squadra per impostare l'azione d'attacco. (D_RECUPERI5)
- *Percentuale di recuperi effettivi*: è ottenuta dal rapporto tra i recuperi effettivi e le palle riconquistate. (D_%RECUPEFF)
- *Recuperi temporanei*: numero totale di recuperi avvenuti per un breve lasso di tempo. (D_RECUPERI6)
- *Percentuale di recuperi temporanei*: ottenuta calcolando il rapporto fra i recuperi temporanei e le palle riconquistate. (D_%RECUPTEMP)
- *Recuperi in zona area*: recuperi effettuati nell'area difensiva. (D_RECUPERI7)
- *Recuperi effettivi in attacco*: numero di palle recuperate e poi mantenute tali in zona di attacco. (D_RECUPERI8)
- *Intercettazioni*: modalità di recupero dei palloni non compresa tra gli anticipi e i contrasti. (D_INTERCET)
- *Intercettazioni effettive*: modalità di recupero del pallone attraverso la quale quest'ultimo viene mantenuto. (D_INTERCET1)
- *Anticipi*: modalità di recupero del pallone. Il giocatore che anticipa è colui che si pone sulla riga di passaggio prima dell'avversario a cui era indirizzata la palla. È

anticipo anche quando il difensore, pur non interponendosi tra la linea di passaggio e l'attaccante avversario, giunge per primo sul pallone (in genere a seguito di un passaggio lungo avversario e di uno scatto comune con l'attaccante). Se invece l'attaccante rinuncia a rincorrere il pallone o è comunque distante dal difensore, l'azione difensiva si tramuta in un'intercettazione. (D_ANTICIPI)

- *Recuperi con anticipo effettivi*: tipologia di recupero della palla in cui la stessa viene mantenuta. (D_RECUPERI9)
- *Anticipi di tesa*: modalità di recupero del pallone in cui il giocatore che anticipa si pone sulla riga di passaggio prima dell'avversario a cui era indirizzata la palla e la devia o la passa ad un compagno con un colpo di testa. (D_ANTICIPI1)
- *Contrasti*: modalità di recupero del pallone. Avviene quando nell'uno contro uno il difensore toglie il possesso di palla all'avversario. (D_CONTRASTI)
- *Recuperi con contrasto effettivi*: numero di recuperi effettuati con la modalità del contrasto. (D_RECUPERI10)
- *Recuperi in elevazione*: numero di recuperi, non solo di testa, avvenuti mentre i giocatori si trovavano con i piedi sollevati da terra. (D_RECUPERI11)
- *Recuperi in elevazione effettivi*: modalità di recupero del pallone che avviene da parte di un giocatore in fase di elevazione colpendo la palla di testa. Viene considerato effettivo quando questo recupero viene mantenuto dalla squadra. (D_RECUPERI12)
- *Recuperi aerei*: recuperi di testa. (D_RECUPERI13)
- *Recuperi aerei in area*: numero di recuperi aerei avvenuti nell'aerea della porta avversaria. (D_RECUPERI14)
- *Duelli aerei*: numero di duelli che avvengono tra due giocatori avversari nel momento in cui entrambi sono in elevazione e quindi si trovano con i piedi sollevati da terra. È formato dalla somma dei duelli aerei vinti e di quelli persi. (D_DUELLI)
- *Duelli aerei vinti*: numero di duelli aerei vinti dalla squadra che si sta analizzando. (D_DUELLI1)
- *Percentuale di duelli aerei vinti*: è data dal rapporto tra duelli aerei vinti e duelli aerei (totali). (D_%DUELLI1)
- *Duelli aerei persi*: numero di duelli aerei persi dalla squadra che si sta analizzando. (D_DUELLI2)

- *Percentuale di duelli aerei persi*: è dato dal rapporto tra duelli aerei persi e duelli aerei (totali). (D_%DUELLI2)
- *Raddoppi*: l'azione viene attribuita quando conquista la palla il giocatore che è andato ad aiutare un compagno, il quale si trovava già impegnato a contrastare un avversario. (D_RADDOPPI)
- *Raddoppi effettivi*: rappresenta un sottoinsieme della variabile raddoppi e prende in considerazione solo il numero di questi che hanno fatto si che la palla fosse mantenuta dalla squadra. (D_RADDOPPI1)
- *Contropressing*: avviene quando il difensore recupera la palla dopo una respinta della difesa successiva ad un passaggio lungo (o rinvio) di un compagno. (D_CONTROPRES)
- *Palle perse*: numero di azioni attribuite ai giocatori di una squadra i quali perdono palla a favore degli avversari o determinano una interruzione del gioco, che determina la rimessa in gioco da parte della squadra avversaria. Può essere calcolata sommando: palle perse in difesa, palle perse a centrocampo e palle perse in attacco. (D_PALLE10)
- *Percentuale di palle perse*: palle perse / palle giocate (questa variabile è contenuta nella macrocategoria dati generali) (D_%PALLE_PERS)
- *Palle perse in difesa*: numero di palle perse in zona di difesa (all'interno della propria metà campo). (D_PALLE11)
- *Palle perse a centrocampo*: numero di palle perse a metà campo. (D_PALLE12)
- *Palle perse in attacco*: numero di palle perse nella zona di attacco (della metà campo avversaria). (D_PALLE13)
- *Palle perse effettive*: palle perse e non più riconquistate. (D_PALLE14)
- *Palle perse effettive in difesa*: palle perse e non più riconquistate in zona di difesa. (D_PALLE15)
- *Palle perse effettive a centrocampo*: palle perse e non più riconquistate a centrocampo. (D_PALLE16)
- *Palle perse effettive in attacco*: palle perse e non più riconquistate in zona d'attacco. (D_PALLE17)
- *Falli commessi*: avvengono quando un giocatore ferma un avversario in modo scorretto per sottrargli il pallone, oppure quando un giocatore in possesso di palla si libera di un avversario sempre in modo scorretto. (D_FALLI)
- *Falli commessi in difesa*: falli commessi nella zona di difesa. (D_FALLI1)

- *Falli commessi a centrocampo*: falli commessi nella zona di metà campo. (D_FALLI2)
- *Falli commessi in attacco*: falli commessi nella zona di attacco e quindi nei pressi dell'area avversaria. (D_FALLI3)
- *Falli commessi nei pressi della propria area*: numero totale di falli commessi nella zona di difesa della squadra presa in considerazione. (D_FALLI4)
- *Falli di mano*: vengono registrati quando un giocatore tocca il pallone con la mano (volontariamente o no) e l'arbitro fischia. (D_FALLI5)
- *Fuorigioco avversari*: numero totale di fuorigioco commessi dalla squadra avversaria. (D_FUORIGIOC)

5) FASE OFFENSIVA
   - *Fuorigioco*: avviene quando un giocatore, al momento del passaggio in avanti di un compagno, ha tra sé e la linea di fondo campo, un solo giocatore avversario (di solito il portiere) o nemmeno quello. (O_FUORIGIOC)
   - *Palle giocate da dietro in fase d'impostazione*: palle giocate nei propri 30 metri difensivi (O_PALLE1)
   - *Palle a scavalcare il centrocampo*: numero di volte in cui la squadra inizia l'azione con una palla lunga. (O_PALLE2)
   - *Palle a scavalcare il centrocampo utili*: numero di volte in cui la squadra inizia l'azione con una palla lunga, la quale permette di scavalcare almeno un giocatore avversario. (O_PALLE3)
   - *Percentuale palle a scavalcare il centrocampo*: è calcolato dal rapporto tra palle a scavalcare il centrocampo e palle giocate da dietro in fase d'impostazione. (O_%PALLE)
   - *Azioni manovrate da dietro*: l'azione parte dai difensori, si tratta di una sorta di fraseggio continuo tra i giocatori che si trovano in zona arretrata e quelli che si trovano più avanti. Grazie a questi continui passaggi la squadra avanza verso la porta avversaria. (O_AZ_MAN)
   - *Percentuale di azioni manovrate da dietro*: rapporto tra azioni manovrate da dietro e palle giocate da dietro in fase d'impostazione. (O_%AZ_MAN)
   - *Palle giocate in zona avanzata*: palle giocate nei 30 metri vicini alla porta avversaria. È ottenuta sommando le palle giocate in zona avanzata centralmente e le palle giocate in zona avanzata sulle fasce. (O_PALLE4)

- *Giocate utili in zona avanzata*: numero di giocate in verticale che eliminano un avversario dalla fase difensiva; queste giocate si sono registrate in zona avanzata e quindi a 30 metri dalla porta avversaria. (O_GIOC_UT)
- *Percentuale palle giocate in zona avanzata*: ottenuta dal rapporto tra giocate utili in zona avanzata e palle giocate in zona avanzata. (O_%PALLE1)
- *Palle giocate in zona avanzata centralmente*: numero di palle giocate a meno di 30 metri dalla porta avversaria in zona centrale. (O_PALLE5)
- *Palle giocate in zona avanzata sulle fasce*: numero di palle giocate sulle fasce a meno di 30 metri dalla porta avversaria. (O_PALLE6)
- *Recuperi in zona avanzata*: numero di palle recuperate a meno di trenta metri dalla porta avversaria. (O_RECUPERI)
- *Recuperi in zona avanzata centralmente*: numero di recuperi effettuati a meno di trenta metri dalla porta avversaria nella zona centrale del campo, la quale si trova esattamente di fronte alla porta. (O_RECUPERI1)
- *Recuperi in zona avanzata sulle fasce*: numero di recuperi effettuati sulle fasce a meno di trenta metri dalla porta avversaria. (O_RECUPERI2)
- *Giocate zona avanzata*: numero di giocate in verticale effettuate nei 30 metri antecedenti alla porta avversaria. (O_GIOC1)
- *Giocate in zona avanzata centralmente*: numero di giocate in zona avanzata in linea retta con la porta avversaria. (O_GIOC2)
- *Percentuale giocate in zona avanzata centralmente*: viene ottenuta dal rapporto tra giocate in zona avanzata centralmente e giocate in zona avanzata. (O_%GIOC1)
- *Giocate in zona avanzata sulle fasce*: numero di giocate non nella fascia di campo in linea retta con la porta ma nelle fasce di sinistra e destra, sempre considerando i trenta metri in prossimità della porta avversaria. (O_GIOC3)
- *Percentuale giocate in zona avanzata sulle fasce*: calcolata attraverso il rapporto tra giocate in zona avanzata sulle fasce e giocate in zona avanzata. (O_%GIOC2)
- *Cambi di gioco*: numero di volte che il pallone passa da una fascia all'altra. (O_CAMBI)
- *Ripartenze*: o più spesso chiamato "contropiede". La squadra che si trova in difesa riesce a recuperare il pallone e lo invia velocemente ai propri attaccanti che cercano di infilare la difesa per andare a fare gol. L'azione viene attribuita se c'è un'azione utile nelle due palle giocate successive ad un recupero palla; indica se

una squadra verticalizza in fretta quando entra in possesso palla. (O_RIPARTENZE)

- *Passaggi di testa*: numero totale di passaggi effettuati tramite un colpo di testa. (O_PASTEST)
- *Passaggi di testa utili*: numero di passaggi effettuati con un colpo di testa che hanno permesso di scavalcare almeno un giocatore avversario. (O_PASTEST1)
- *Passaggi lunghi*: passaggio di oltre 30 metri con parabola la cui componente orizzontale è preminente. Viene considerato passaggio lungo anche il rinvio a seguito di una rimessa dal fondo. (O_PAS1)
- *Passaggi lunghi utili*: numero di passaggi di oltre 30 metri che hanno permesso di scavalcare almeno un giocatore avversario. (O_PAS2)
- *Percentuale passaggi lunghi utili*: rapporto tra il numero di passaggi lunghi utili e i passaggi lunghi totali. (O_%PAS1)
- *Passaggi lunghi da dietro*: numero di passaggi lunghi effettuati dalla zona difensiva. (O_PAS3)
- *Passaggi lunghi utili da dietro*: passaggi lunghi effettuati dalla zona difensiva che hanno permesso di scavalcare almeno un avversario che di trovava in fase difensiva. (O_PAS4)
- *Passaggi lunghi senza cambio di gioco*: il termine cambio di gioco viene attribuito quando la palla passa da una fascia all'altra. Ha funzione di aggettivo con la variabile passaggio lungo; e questo avviene quando il cambio di gioco ha una parabola aerea la quale può essere anche in avanti. La variabile che stiamo analizzando in questo momento rappresenta il numero totale di passaggi lunghi senza cambio di gioco e pertanto non viene conteggiato il passaggio della palla da una fascia all'altra. (O_PAS5)
- *Passaggi lunghi utili senza cambio di gioco*: numero totale di passaggi lunghi effettuati sulla stessa fascia che permettono di superare un avversario che si trova in fase difensiva. (O_PAS6)
- *Pallonetti*: passaggio con parabola, che passa sopra l'altezza delle spalle dei giocatori, la cui componente verticale è preminente. È pallonetto qualsiasi lancio in avanti effettuato da tre quarti di campo in su nella zona centrale che raggiunge un compagno a non più di 30 metri. (O_PALLON)
- *Pallonetti utili*: quelli che permettono di superare almeno un avversario. (O_PALLON1)

- *Percentuale di pallonetti utili*: pallonetti utili / pallonetti totali (O_%PALLON)
- *Pallonetti nella metà campo avversaria*: numero di passaggi avvenuti nella metà la cui componente verticale è preminente e raggiungono un compagno a meno di trenta metri,. (O_PALLON2)
- *Pallonetti utili nella metà campo avversaria*: numero di pallonetti avvenuti nella metà campo avversaria che hanno permesso di scavalcare almeno un avversario. (O_PALLON3)
- *Passaggi filtranti*: passaggio rasoterra (o comunque ad altezza inferiore al livello delle spalle) che passa in mezzo a due giocatori avversari prima di essere ricevuto da un compagno o venire intercettato da un avversario. (O_PAS7)
- *Passaggi filtranti utili*: passaggio filtrante che permette di scavalcare almeno un avversario. (O_PAS8)
- *Percentuale di passaggi filtranti utili*: passaggi filtranti utili / passaggi filtranti. (O_%PAS2)
- *Passaggi filtranti nella metà campo avversaria*: numero di passaggi effettuati nella metà campo avversaria ad altezza inferiore al livello delle spalle che passano solitamente tra due giocatori avversari prima di essere intercettati dagli stessi o ricevuti da un proprio compagno di squadra. (O_PAS9)
- *Passaggi filtranti utili nella metà campo avversaria*: numero di passaggi filtranti ricevuti da un giocatore della propria squadra nella metà campo avversaria. (O_PAS10)
- *Sovrapposizioni*: azione attribuita ad un giocatore che riceve palla dopo un inserimento (interno o esterno) da dietro il portatore di palla. Normalmente si verificano sulle ali, ovvero la parte laterale del campo ed avviene quando il terzino, che solitamente sta dietro, si sovrappone all'ala. Il terzino passa davanti all'ala in modo da mettere in crisi il difensore che non sa più quale dei due giocatori difendere. Il terzino, generalmente, va a sovrapporsi alla zona del giocatore in ala per creare superiorità numerica. (O_SOVRAPP)
- *Cross*: azione eseguita nella tre quarti d'attacco dalla fascia laterale verso l'area di rigore; il cross può essere indistintamente basso o alto e può essere effettuato anche da calcio piazzato. (O_CROSS)
- *Cross utili*: cross che permettono di scavalcare almeno un giocatore della squadra avversaria. (O_CROSS1)
- *Percentuale di cross utili*: cross utili / cross. (O_%CROSS)

- *Cross a rientrare*: si registra quando la parabola del cross è convessa (ad esempio il tipo cross con l'interno sinistro da destra). (O_CROSS2)
- *Cross su azione da destra*: cross effettuato con palla in movimento, quindi non da calcio piazzato, che parte dalla fascia destra del campo. (O_CROSS3)
- *Percentuale di cross su azione da destra*: cross su azione da destra / (cross su azione da destra + cross su azione da sinistra). (O_%CROSS1)
- *Cross su azione da sinistra*: cross effettuato con palla in movimento, calciato dalla fascia sinistra del campo. La palla in movimento esclude tutti i possibili cross avvenuti su calcio piazzato. (O_CROSS4)
- *Percentuale cross su azione da sinistra*: cross su azione da sinistra / (cross su azione da sinistra + cross su azione da destra). (O_%CROSS2)
- *Cross ad uscire*: si registra quando la parabola del cross è concava (ad esempio il tipico cross con l'interno destro da destra). (O_CROSS5)
- *Cross a rientrare su calcio piazzato*: cross con parabola convessa su calcio piazzato (corner e punizioni). (O_CROSS6)
- *Cross ad uscire su calcio piazzato*: cross con parabola concava su calcio piazzato (corner e punizioni). (O_CROSS7)
- *Cross a rientrare su azione*: numero di cross con parabola convessa effettuati in fase d'azione; quindi con palla in movimento. (O_CROSS8)
- *Cross ad uscire su azione*: numero di cross con parabola concava effettuati in fase d'azione. (O_CROSS9)
- *Cross a rientrare su azione dal fondo*: quando la parabola del cross è convessa ed è effettuato nel settore compreso tra la linea di fondo campo e il prolungamento dell'area di rigore verso l'esterno. (O_CROSS10)
- *Cross ad uscire su azione dal fondo*: quando la parabola del cross è concava ed è effettuato nel settore compreso tra la linea di fondo campo e il prolungamento dell'area di rigore verso l'esterno. (O_CROSS11)
- *Cross su azione*: numero di cross effettuati con palla in movimento. (O_CROSS12)
- *Cross su azione utili*: numero di cross su azione che hanno permesso di superare almeno un giocatore avversario. (O_CROSS13)
- *Cross su azione dal fondo*: cross effettuato dal settore compreso tra la linea di fondo campo e il prolungamento dell'area di rigore verso l'esterno. (O_CROSS14)

- *Cross su azione dal fondo utili*: cross effettuati su azione nella zona avanzata del campo, i quali hanno permesso di scavalcare almeno un giocatore avversario che si trovava in fase difensiva. (O_CROSS15)
- *Cross su calcio piazzato*: numero totale di cross (senza distinzione di tipo) che avvengono su calcio piazzato (punizioni e corner). (O_CROSS16)
- *Cross di mano su fallo laterale*: azione attribuita al giocatore che effettua una rimessa lunga su fallo laterale verso l'area di rigore avversaria, questa azione è possibile solo se ci si trova in zona d'attacco. (O_CROSS17)
- *Passaggi lungo linea*: azione attribuita quando la palla è giocata parallelamente alla linea laterale. Può essere effettuato in tutte le zone esterne del campo purché non ci siano avversari tra la linea di passaggio e la linea laterale. (O_PAS11)
- *Passaggi lungo linea utili*: passaggi lunghi giocati parallelamente alla linea laterale che hanno permesso di scavalcare almeno un giocatore avversario. (O_PAS12)
- *Percentuale di passaggi lungo linea utili*: passaggi lungo linea utili / passaggi lungo linea. (O_%PAS3)
- *Lungo linea nella metà campo avversaria*: numero di passaggi effettuati parallelamente alla linea laterale nella metà campo avversaria. (O_LUNGO)
- *Lungo linea utili nella metà campo avversaria*: passaggi lungo linea effettuati nella metà campo avversaria che hanno permesso di scavalcare almeno un giocatore della squadra avversaria. (O_LUNGO1)
- *Dai e vai*: avviene quando un giocatore passa la palla ad un compagno e questo gliela torna, dopo che il primo ha effettuato un movimento in avanti superando il compagno. (O_DAIVAI)
- *Accelerazioni*: azione attribuita quando un giocatore, in possesso palla, aumenta la propria velocità in uno spazio vuoto; con l'avversario al fianco si può avere comunque un'accelerazione quando si punta ad esempio il fondo del campo. Questa tattica è utile quando si supera in velocità almeno un avversario o comunque si giunge al tiro tramite cross (non ribattuti immediatamente). (O_ACCELLER)
- *Colpi di testa offensivi*: quando un giocatore stacca di testa per colpire la palla e fare gol. (O_COLPITEST)
- *Sponde*: l'azione viene attribuita quando un giocatore riceve palla dal dietro e la rigioca indietro rispetto all'asse del proprio corpo. Se il passaggio di ritorno viene effettuato "di prima" è sempre sponda, ma è sempre considerata sponda anche

quando i tocchi sono due, se in tempi brevi, e se la palla ritorna al giocatore da cui è provenuta oppure ad un altro giocatore situato nella stessa zona di campo. (O_SPONDE)

- *Sponde di piede*: azione attribuita quando un giocatore respinge la palla di piede verso la direzione da cui è arrivata, oppure la gira in direzione della porta avversaria. (O_SPONDE1)
- *Sponde di piede riuscite*: sponde di piede andate a buon fine e quindi arrivate ad un giocatore della propria squadra. (O_SPONDE2)
- *Sponde riuscite*: sponde andate a buon fine che hanno permesso che il possesso della palla fosse mantenuto dalla propria squadra. (O_SPONDE4)
- *Sponde di testa*: azione attribuita quando un giocatore respinge la palla di testa verso la direzione da cui è arrivata; oppure la gira volontariamente in direzione della porta avversaria. (O_SPONDE6)
- *Sponde di testa riuscite*: sponde di testa arrivate ad un giocatore della stessa squadra. (O_SPONDE7)
- *Sponde di testa utili*: se la palla giungendo ad un compagno elimina un avversario dalla fase difensiva. (O_SPONDE8)
- *Sponde di testa perse*: numero di sponde di testa che non sono andate a buon fine e quindi, invece di arrivare ad un giocatore della propria squadra, sono arrivate ad un avversario. (O_SPONDE9)
- *Spizzichi di testa*: viene attribuito quando un giocatore devia la palla di testa facendola proseguire nella direzione opposta rispetto a quella di arrivo della stessa cambiandogli leggermente la traiettoria. (O_SPIZZ)
- *Spizzichi di testa utili*: spizzichi di testa che hanno permesso di superare almeno un giocatore avversario. (O_SPIZZ1)
- *Veli*: azione attribuita al giocatore che è sulla palla e la lascia sfilare per eludere la marcatura di un avversario, non è catalogata come palla giocata perché l'esecutore non ne entra in possesso. Il velo viene considerato utile quando elimina un difensore dalla fase difensiva (tranne il caso in cui il compagno vada al tiro). (O_VELI)
- *Tagli dentro*: consiste in un movimento di smarcamento dell'attaccante. Si tratta di una corsa in diagonale dall'esterno verso il centro e può avvenire in due circostanze: con palla al piede o prima di ricevere la palla, attraverso un movimento che interseca la linea che congiunge palla e porta avversaria. Ad

esempio quando i centrocampisti si infilano passando la difesa e questo provoca sorpresa da parte della stessa che si trova a dover gestire un maggior numero di giocatori offensivi. (O_TAGLI)

- *Tagli fuori*: si tratta sempre di un movimento di smarcamento dell'attaccante che corre in diagonale dal centro verso l'esterno. Può avvenire nelle medesime due circostanze sopra menzionate per i tagli dentro. (O_TAGLI1)
- *Allargamenti*: movimento di smarcamento che effettua l'attaccante che si allarga per andare a ricevere la palla verso la linea laterale, con un movimento dalla parte opposta rispetto alla posizione del compagno che lo lancia. La squadra allarga i propri giocatori verso l'esterno, andando, in questo modo, ad occupare la maggior parte di campo possibile. Allargando il gioco si mettono in difficoltà i difensori. (O_ALLARG)
- *Dribbling*: avviene quando un giocatore in possesso di palla tenta di superare uno o più avversari con le sue finte. (O_DRIBBLING)
- *Dribbling utili*: il dribbling viene considerato utile quando si riesce a superare l'avversario avvicinandosi alla porta avversaria, o comunque si giunge al tiro o al cross (non ribattuti immediatamente). (O_DRIBBLING1)
- *Percentuale di dribbling utili*: percentuale dei dribbling che possono essere ritenuti utili sul totale dei dribbling. (O_%DRIBBLING)
- *Dribbling di spalle a sinistra*: quando l'attaccante riceve palla con le spalle rivolte verso la porta e tenta di liberarsi del marcatore girandosi alla sua sinistra. (O_DRIBBLING2)
- *Dribbling di spalle a destra*: quando l'attaccante riceve palla con le spalle rivolte verso la porta e tenta di liberarsi del marcatore girandosi alla sua destra. (O_DRIBBLING3)
- *Dribbling tunnel*: consiste nel fare dribbling facendo passare la palla in mezzo alle gambe dell'avversario. È utile quando l'attaccante riprende la palla dopo aver superato il difensore avversario. (O_DRIBBLING4)
- *Dribbling sterzata*: è un particolare tipo di dribbling nel quale il giocatore che sta correndo con la palla blocca improvvisamente la corsa e scavalca il giocatore avversario passando dalla parte opposta (sterzata) mantenendo sempre il possesso palla. (O_DRIBBLING5)
- *Simulazioni*: le così dette "finte". È un infrazione commessa da un giocatore in possesso palla, che simula di aver subito un fallo. Avvengono principalmente in

due occasioni: o nella zona dell'area quando un giocatore fa in modo di essere toccato per cercare rigore o quando un giocatore fa finta di essere stato colpito e si butta a terra in modo da far fermare il gioco. (O_SIMULA)

- *Colpi di tacco*: numero di volte in cui il pallone viene colpito con la parte posteriore del piede. Possono avvenire in ogni parte del campo. (O_COLPITAC)
- *Azioni in rovesciata*: la rovesciata è una tecnica spettacolare per calciare il pallone al volo. Essa viene effettuata lanciandosi in aria in una rotazione all'indietro, per poi colpire il pallone violentemente con il collo del piede. Il tiro parte così oltre la testa del giocatore, nel verso opposto a quello verso cui egli era rivolto prima di spiccare il salto. Questa variabile indica il numero di volte in cui è avvenuta un'azione con rovesciata per ciascuna delle squadre oggetto di analisi. (O_AZ_ROV)
- *Falli subiti*: avviene quando un giocatore in possesso palla subisce fallo da un avversario, oppure quando tenta di recuperare palla e l'avversario glielo impedisce in modo scorretto. È ottenuto dalla somma delle tre variabili successive (falli subiti in difesa, a centrocampo e in attacco). (O_FALLI)
- *Falli subiti in difesa*: numero di falli subiti da un giocatore che si trova in zona di difesa. (O_FALLI1)
- *Falli subiti a centrocampo*: numero di falli subiti dai giocatori a centro campo. (O_FALLI2)
- *Falli subiti in attacco*: numero di falli subiti dai giocatori nella zona di attacco. (O_FALLI3)
- *Assist*: viene attribuito in automatico al giocatore che passa la palla al compagno per il tiro. Pertanto la variabile rappresenta il numero totale di passaggi di questo tipo. Può essere ottenuto sommando gli assist su sviluppo di azione e quelli su sviluppo di calcio piazzato. (O_ASSIST)
- *Assist su sviluppo di azione*: numero di volte che il passaggio decisivo al giocatore che tira verso la porta è arrivato da un'azione in movimento. (O_ASSIST1)
- *Assist su sviluppo di calcio piazzato*: numero di volte che il passaggio decisivo verso il giocatore che tira verso la porta è arrivato da calcio piazzato (punizioni e corner). (O_ASSIST2)
- *Assist vincenti*: viene attribuito in automatico al giocatore che passa la palla al compagno che segna una rete. Questa variabile rappresenta il numero totale di assist che hanno permesso la realizzazione di un gol. (O_ASSIST3)

- *Assist vincenti su sviluppo di azione*: numero di assist provenienti da un'azione in movimento che hanno permesso di segnare una rete. (O_ASSIST4)
- *Assist vincenti su sviluppo di calcio piazzato*: numero di assist provenienti da calcio piazzato che hanno permesso di segnare una rete. (O_ASSIST5)
- *Percentuale di attacco alla porta*: è data dal rapporto tra numero di palle giocate dalla squadra in area avversaria e numero totale di palle che gravitano in area avversaria. (O_%ATT)
- *Percentuale attacco aereo alla porta*: è data dal rapporto tra palle giocate di testa dalla squadra in area avversaria e numero totale di palle che vengono colpite di testa in area avversaria. (O_%ATT1)
- *Occasioni*: indica chiare occasioni da rete (non importa se l'azione è coronata o meno da una finalizzazione). Vi possono essere più occasioni all'interno della stessa azione solo se vi sono più finalizzazioni pericolose. Nel caso di rigore, palo o rete si ha sempre un'occasione e questa viene inserita in automatico in questa variabile. (O_OCCAS)
- *Capacità realizzativa*: è data dal rapporto tra numero di reti e numero di tiri. (O_CAPREAL)
- *Capacità realizzativa (new)*: è data dal rapporto tra reti e numero di occasioni da rete. (O_CAPREAL1)
- *Tiri*: numero totale di tiri tentati da una squadra. (O_TIRI)
- *Precisione nel tiro*: è il rapporto fra i tiri nello specchio della porta ed il totale dei tiri. (O_PREC_TIRO)
- *Tiri dentro*: conclusioni direzionate nello specchio della porta e quindi può portare a diverse situazioni: rete, parata, palo o salvataggio di un difensore nel caso in cui il portiere sia stato battuto. (O_TIRI1)
- *Tiri dentro di destro*: numero di tiri direzionati dalla squadra con l'arto destro, nello specchio della porta avversaria. (O_TIRI2)
- *Tiri dentro di sinistro*: numero di tiri direzionati da una squadra con l'arto sinistro nello specchio della porta avversaria. (O_TIRI3)
- *Tiri dentro di testa*: numero totale di tiri direzionati nello specchio della porta colpendo la palla di testa. (O_TIRI4)
- *Tiri fuori*: tiri che vanno al di fuori dello specchio della porta o intercettati da un qualsiasi giocatore. (O_TIRI5)

- *Tiri di destro*: numero di tiri effettuati da tutti i giocatori della squadra con l'arto destro. (O_TIRI6)
- *Tiri di sinistro*: numero di tiri effettuati da tutti i giocatori di una squadra con l'arto sinistro. (O_TIRI7)
- *Tiri di testa*: specifica il tipo di tiro che può essere dovuto ad una sponda od a una intercettazione e si riferisce sempre al numero totale di tiri effettuati tramite questa modalità. (O_TIRI8)
- *Tiri in rovesciata*: numero di tiri effettuati lanciandosi in aria in una rotazione all'indietro e colpendo il pallone violentemente con il collo del piede, nel verso opposto a quello verso cui era rivolto il giocatore prima di spiccare il salto. (O_TIRI9)
- *Tiri al volo*: numero di tiri effettuati, verso la porta avversaria, senza far rimbalzare la palla in terra dopo un passaggio o una rimessa. (O_TIRI10)
- *Tiri in acrobazia*: numero di tiri effettuati in direzione della porta facendo un'acrobazia (quasi sempre "in aria"). (O_TIRI11)
- *Tiri di piede con palla bassa*: numero di tiri effettuati di piede mantenendo sempre la palla bassa. (O_TIRI12)
- *Tiri da fuori area*: numero di tiri effettuati da una squadra al di fuori dell'area avversaria. (O_TIRI13)
- *Tiri da dentro area di rigore*: numero di tiri effettuati all'interno dell'area di rigore. (O_TIRI14)
- *Tiri da dentro area di porta*: numero di tiri effettuati all'interno dell'area di porta, la quale si trova all'interno dell'area di rigore. (O_TIRI15)
- *Tiri ribattuti o deviati*: numero di tiri respinti o deviati in una diversa direzione. (O_TIRI16)
- *Tiri fuori dallo specchio*: numero di tiri effettuati e che vanno al di fuori dello specchio della porta. (O_TIRI17)
- *Tiri su azione*: tiri verso la porta non provenienti da calcio piazzato; ma provenienti da un passaggio da parte di un compagno in un'azione d'attacco. (O_TIRI18)
- *Tiri di testa su azione*: tiri verso la porta effettuati tramite un colpo di testa. (O_TIRI19)
- *Tiri in rovesciata su azione*: tiri effettuati verso la porta con un salto all'indietro. (O_TIRI20)

- *Tiri al volo su azione*: tiri effettuati verso la porta colpendo il pallone proveniente da un passaggio senza farlo rimbalzare in terra. (O_TIRI21)
- *Tiri in acrobazia su azione*: numero di tiri verso la porta colpendo il pallone in modo particolare. È ottenuto dalla somma dei tiri in rovesciata su azione e di quelli al volo su azione. (O_TIRI22)
- *Tiri di piede su palla bassa da azione*: quando il passaggio proveniente da un compagno arriva ad altezza inferiore al collo al giocatore che poi effettuerà il tiro in porta. (O_TIRI23)
- *Tiri da piazzati in attacco*: numero di tiri effettuati da calcio piazzato, quindi con pallone fermo, che avvengono in zona avanzata. (O_TIRI24)
- *Tiri totali da calcio piazzato*: numero totale di tiri effettuati da calcio piazzato e quindi da punizioni, corner e rigori. (O_TIRI25)
- *Tiri diretti da calcio piazzato*: o meglio chiamati "tiri di prima", in quanto il giocatore può decidere di calciare direttamente in porta. Rappresenta il numero di tiri effettuati da calcio piazzato ed indirizzati direttamente verso la porta avversaria, senza cercare l'appoggio dei compagni di squadra. (O_TIRI26)
- *Tiri indiretti da calcio piazzato*: o meglio chiamati "tiri di seconda", possono riguardare solo le punizioni, in quanto alcune di esse presuppongono che la palla sia toccata da almeno un giocatore (oltre a quello che calcia la punizione) prima di indirizzarlo verso la porta. (O_TIRI27)
- *Tiri su rigore*: numero di rigori assegnati ad una squadra durante una partita. (O_TIRI28)
- *Tiri da punizione da destra*: numero di tiri effettuati su punizione nella parte destra della zona d'attacco riferito in merito alla squadra che viene considera. (O_TIRI29)
- *Tiri da punizione centrale*: numero di tiri effettuati su punizione nella zona centrale di attacco, quella che si trova in linea d'aria di fronte alla porta. (O_TIRI30)
- *Tiri da punizione da sinistra*: numero di tiri effettuati su punizione nella parte sinistra della zona d'attacco. (O_TIRI31)
- *Tiri da angolo da destra*: numero di corner calciati dall'angolo a destra della porta avversaria. (O_TIRI32)
- *Tiri da angolo da sinistra*: numero di corner calciati dall'angolo a sinistra della porta avversaria. (O_TIRI33)

- *Tiri da laterale da destra*: tiri effettuati nella zona leggermente posteriore alla zona di attacco sul lato destro. La zona di attacco è leggermente più grande dell'area. (O_TIRI34)
- *Tiri da laterale da sinistra*: tiri effettuati prima della zona di attacco sul lato sinistro del campo. (O_TIRI35)
- *Tiri di testa da un calcio piazzato*: numero di tiri effettuati di testa verso la porta, ma provenienti da un passaggio effettuato da calcio piazzato. (O_TIRI36)
- *Tiri in rovesciata da un calcio piazzato*: numero di tiri effettuati in rovesciata verso la porta e provenienti da un passaggio da calcio piazzato. (O_TIRI37)
- *Tiri al volo da un calcio piazzato*: rappresenta il numero di volte in cui i giocatori di una squadra hanno ricevuto una palla proveniente da un calcio piazzato e l'hanno tirata direttamente in porta senza farla rimbalzare a terra. (O_TIRI38)
- *Tiri in acrobazia su calcio piazzato*: numero di tiri verso la porta effettuati con un'acrobazia e provenienti da un passaggio effettuato da calcio piazzato. (O_TIRI39)
- *Tiri di piede su palla bassa da calcio piazzato*: numero di tiri effettuati di piede verso la porta e provenienti da un passaggio su calcio piazzato. (O_TIRI40)
- *Palle giocate in zona area*: numero di volte in cui i giocatori di una squadra hanno toccato palla in zona area. (O_PALLE7)
- *Rapidità nell'arrivare al tiro misurata in numero di passaggi*: numero medio di passaggi che una squadra effettua prima di arrivare al tiro verso la porta avversaria. (O_RAPID1)
- *Reti da fuori area*: numero di reti segnate, dalla squadra presa in considerazione, su tiri effettuati al di fuori dell'area avversaria. (O_RETI)
- *Reti da dentro area di porta*: numero di reti effettuate calciando il pallone all'interno dell'area della porta avversaria. (O_RETI1)
- *Reti di destro*: numero di reti segnate da una squadra, i cui giocatori hanno tirato il pallone in porta con l'arto destro. (O_RETI2)
- *Reti di sinistro*: numero di reti segnate dai giocatori di una squadra tirando la palla con l'arto sinistro. (O_RETI3)
- *Reti di testa*: numero di reti segnate grazie ai colpi di testa dei giocatori della squadra in questione. (O_RETI4)
- *Reti in rovesciata*: numero di reti segnate dai giocatori di una squadra lanciandosi in aria con una rotazione all'indietro e colpendo il pallone violentemente con il

collo del piede, nella direzione opposta rispetto a quella verso cui era rivolto il giocatore prima del salto. (O_RETI5)

- *Reti al volo*: numero di tiri andati a segno effettuati senza far rimbalzare il pallone a terra. (O_RETI6)
- *Reti totali su azione*: qualsiasi rete effettuata non da calcio piazzato. (O_RETI7)
- *Reti totali su calcio piazzato*: numero totale di reti segnate su rigori, punizioni e corner. (O_RETI8)
- *Reti dirette su calcio piazzato*: numero di tiri "di prima" andati a buon fine, o meglio numero di rigori realizzati. (O_RETI9)
- *Reti indirette su calcio piazzato*: numero di tiri "di seconda" che hanno portato la squadra alla realizzazione di un gol. (O_RETI10)

6) CALCI PIAZZATI BATTUTI
   - *Rimesse dal fondo*: azione di squadra attribuita ogni volta che la palla toccata da un avversario esce dalla propria linea di fondo campo. (CP_RIMESSE)
   - *Piazzati in attacco*: numero di calci piazzati battuti in attacco, quindi nella metà campo avversaria. (CP_ATT)
   - *Falli laterali*: si tratta di un calcio piazzato eseguito con le mani dietro la linea laterale nella zona dove la palla era precedentemente uscita. (CP_FALLI)
   - *Falli laterali in difesa*: numero di volte in cui il pallone uscito lateralmente dal campo viene rimesso in gioco nella zona di difesa. (CP_FALLI1)
   - *Falli laterali in difesa da destra*: quando il pallone esce lateralmente dal campo e viene rimesso in gioco nella zona di difesa; da destra si intende sempre considerando la squadra presa in considerazione con la propria porta alle sue spalle. (CP_FALLI2)
   - *Falli laterali in difesa da sinistra*: numero di volte in cui il pallone uscito lateralmente dal campo viene rimesso in gioco nella zona di difesa sul lato sinistro del campo. (CP_FALLI3)
   - *Falli laterali a centrocampo*: numero di volte in cui il pallone uscito lateralmente dal campo viene rimesso in gioco nella zona centrale di quest'ultimo. (CP_FALLI4)
   - *Falli laterali a centrocampo da destra*: numero di volte in cui il pallone uscito lateralmente dal campo viene rimesso in gioco sul lato destro del centrocampo. (CP_FALLI5)

- *Falli laterali a centrocampo da sinistra*: numero di volte in cui il pallone uscito lateralmente dal campo viene rimesso in gioco sul lato sinistro del centrocampo. (CP_FALLI6)
- *Falli laterali in attacco*: numero di rimesse laterali avvenute nella zona di attacco della metà campo avversaria, a causa della fuoriuscita del pallone dalle linee laterali che delimitano il campo di gioco. (CP_FALLI7)
- *Falli laterali in attacco da destra*: numero di rimesse laterali avvenute nella zona di attacco della metà campo avversaria alla destra del campo, a causa della fuoriuscita del pallone dalla linea laterale destra che delimita il campo di gioco. (CP_FALLI8)
- *Falli laterali in attacco da destra con tiro*: numero di rimesse laterali in attacco sul lato destro del campo che arrivano nei pressi dell'area e permettono ad un giocatore della propria squadra di crosssare e tentare il tiro verso la porta. (CP_FALLI9)
- *Falli laterali in attacco da destra senza tiro*: numero di rimesse laterali in attacco sul lato destro del campo che non permettono ad un giocatore della propria squadra di effettuare un tiro verso la porta avversaria. (CP_FALLI10)
- *Falli laterali in attacco da sinistra*: numero di rimesse laterali avvenute nella zona di attacco della metà campo avversaria alla sinistra del campo, a causa della fuoriuscita del pallone dalla linea laterale sinistra che delimita il campo di gioco. (CP_FALLI11)
- *Falli laterali in attacco da sinistra con tiro*: numero di rimesse laterali in attacco sul lato sinistro del campo che arrivano nei pressi dell'area e permettono ad un giocatore della propria squadra di tentare il tiro verso la porta. (CP_FALLI12)
- *Falli laterali in attacco da sinistra senza tiro*: numero di rimesse laterali in attacco sul lato sinistro del campo che non permettono ad un giocatore della propria squadra di effettuare un tiro verso la porta avversaria. (CP_FALLI13)
- *Falli laterali da destra*: numero totale di rimesse laterali effettuate da una squadra sul lato destro del campo di gioco. (CP_FALLI14)
- *Falli laterali da sinistra*: numero totale di rimesse laterali effettuate da una squadra sul lato sinistro del campo di gioco. (CP_FALLI15)
- *Angoli*: numero totale di corner calciati da una squadra. (CP_ANGOLI)
- *Angoli da destra*: numero di angoli calciati da una squadra sul lato destro della porta avversaria. (CP_ANGOLI1)

- *Angoli da destra con tiro*: numero di angoli calciati sul lato destro, rispetto alla porta avversaria, che hanno portato ad un tiro verso la porta. (CP_ANGOLI2)
- *Angoli da destra senza tiro*: numero di corner calciati sul lato destro della porta avversaria che non hanno portato ad un tiro verso la stessa. (CP_ANGOLI3)
- *Angoli da sinistra*: numero di angoli calciati da una squadra sul lato sinistro della porta avversaria. (CP_ANGOLI4)
- *Angoli da sinistra con tiro*: numero di angoli calciati sul lato sinistro che hanno portato ad un tiro verso la porta. (CP_ANGOLI5)
- *Angoli da sinistra senza tiro*: numero di corner calciati sul lato sinistro della porta avversaria che non hanno portato ad un tiro verso la stessa. (CP_ANGOLI6)
- *Punizione*: calcio piazzato eseguito dopo un fallo o un fuorigioco. (CP_PUNIZ)
- *Punizioni in difesa*: il difensore calcia una punizione dalla zona di difesa nella propria metà campo. (CP_PUNIZ1)
- *Punizioni in difesa da destra*: numero di punizioni calciate da una squadra sulla fascia destra della zona di difesa. (CP_PUNIZ2)
- *Punizioni in difesa centrali*: numero di punizione calciate da una squadra nella fascia centrale della zona di difesa. (CP_PUNIZ3)
- *Punizioni in difesa da sinistra*: numero di punizioni calciate da una squadra sulla fascia sinistra della zona di difesa. (CP_PUNIZ4)
- *Punizioni a centrocampo*: numero di punizioni calciate da una squadra nella zona di centrocampo. (CP_PUNIZ5)
- *Punizioni a centrocampo da destra*: numero di punizioni calciate da una squadra nella fascia destra di centrocampo. (CP_PUNIZ6)
- *Punizioni a centrocampo centrali*: numero di punizioni calciate da una squadra nella zona centrale di centrocampo. (CP_PUNIZ7)
- *Punizioni a centrocampo da sinistra*: numero di punizioni calciate da una squadra nella fascia sinistra di centrocampo. (CP_PUNIZ8)
- *Punizioni in attacco*: l'attaccante calcia una punizione dalla zona di attacco nella metà campo avversaria, più precisamente dalla tre quarti di campo in poi. (CP_PUNIZ9)
- *Punizioni in attacco da destra*: numero di punizioni calciate dalla tre quarti avversaria sulla fascia destra. (CP_PUNIZ10)
- *Punizioni in attacco da destra con tiro*: numero di punizioni calciate verso la porta dalla tre quarti avversaria sulla fascia destra. (CP_PUNIZ11)

- *Punizioni in attacco da destra senza tiro*: numero di punizioni calciate verso un compagno dalla tre quarti avversaria sulla fascia destra. (CP_PUNIZ12)
- *Punizioni in attacco centrali*: numero di punizioni calciate da una squadra dalla tre quarti avversaria nella zona centrale rispetto alla porta. (CP_PUNIZ13)
- *Punizioni in attacco centrali con tiro*: numero di punizioni calciate verso la porta dalla tre quarti avversaria nella zona centrale. (CP_PUNIZ14)
- *Punizioni in attacco centrali senza tiro*: numero di punizioni calciate verso un compagno dalla tre quarti avversaria per un'azione d'attacco. (CP_PUNIZ15)
- *Punizioni in attacco da sinistra*: numero di punizioni calciate dalla tre quarti avversaria sulla fascia sinistra. (CP_PUNIZ16)
- *Punizioni in attacco da sinistra con tiro*: numero di punizioni calciate verso la porta dalla tre quarti avversaria sulla fascia sinistra. (CP_PUNIZ17)
- *Punizioni in attacco da sinistra senza tiro*: numero di punizioni calciate verso un compagno dalla tre quarti avversaria sulla fascia sinistra. (CP_PUNIZ18)
- *Rigori*: numero totale di rigori assegnati ad una squadra. (CP_RIGORI)
- *Rigori sbagliati*: numero di rigori che non hanno prodotto una rete per la squadra che li ha calciati. (CP_RIGORI1)
- *Rigori trasformati*: numero di rigori andati a segno per la squadra presa in considerazione. (CP_RIGORI2)

APPENDICE B.

Vengo di seguito riportate le varie fasi di riduzione del database principale e le corrispondenti variabili eliminate per ciascuna di esse.

1° FASE DI ELIMINAZIONE: queste variabili sono state eliminate in quanto risultavano essere funzioni di altre variabili e pertanto dati ridondati che non avrebbero apportato alcun miglioramento al nostro modello. Le variabili sotto elencate sono state eliminate in riferimento alla squadra di casa (suffisso _C), alla squadra ospite (suffisso _O), e come valore differenza tra quanto associato alla squadra di casa rispetto all'avversaria (prefisso DIF_).

- Parate (P_PARATE)
- Tiri dentro subiti (P_TIRI_SUB1)
- Reti / tiri dentro subiti (%) (P_RETI_SUB5)
- Percentuale di duelli aerei vinti (D_%DUELLI1)
- Percentuale di duelli aerei persi (D_%DUELLI2)
- Fuorigioco avversari (D_FUORIGIOC)
- Percentuale di palle a scavalcare il centrocampo (O_%PALLE)
- Percentuale sulle palle giocate in zona avanzata (O_%PALLE1)
- Palle perse (D_PALLE10)
- Percentuale di palle perse (D_%PALLE_PERS)
- Palle perse effettive (D_PALLE14)
- Palle giocate sulla fascia destra (G_PALLE1)
- Percentuale di palle giocate sulla fascia destra (G_%PALLE)
- Palle giocate sulla fascia centrale (G_PALLE5)
- Percentuale palle giocate sulla fascia centrale (G_%PALLE2)
- Palle giocate sulla fascia sinistra (G_PALLE9)
- Percentuale di palle giocate sulla fascia sinistra (G_%PALLE3)
- Percentuale di palle giocate in difesa (G_%PALLE4)
- Percentuale di palle giocate a centrocampo (G_%PALLE5)
- Percentuale di palle giocate in attacco (G_%PALLE6)
- Recuperi in zona avanzata (O_RECUPERI)
- Percentuale per fine azione avversaria (D_%FINAZ)
- Percentuale di recuperi effettivi (D_%RECUPEFF)
- Percentuale di recuperi temporanei (D_%RECUPTEMP)
- Percentuale giocate in zona avanzata centralmente (O_%GIOC1)

- Giocate in zona avanzata sulle fasce (O_GIOC3)
- Giocate utili (G_GIOC_UT)
- Percentuale di giocate utili (G_%GIOC_UT)
- Percentuale di pallonetti utili (O_%PALLON)
- Percentuale di passaggi filtranti utili (O_%PAS2)
- Percentuale di passaggi lungo linea utili (O_%PAS3)
- Percentuale di passaggi lunghi utili (O_%PAS1)
- Percentuale di passaggi riusciti (G_%PAS_RIU)
- Passaggi riusciti (G_PASRIU2)
- Passaggi ricevuti (G_PASRIC1)
- Percentuale di cross utili (O_%CROSS)
- Percentuale di cross su azione da destra (O_%CROSS1)
- Percentuale di cross su azione da sinistra (O_%CROSS2)
- Sponde riuscite (O_SPONDE4)
- Percentuale di dribbling utili (O_%DRIBBLING)
- Percentuale di azioni manovrate da dietro (O_%AZ_MAN)
- Assist (O_ASSIST)
- Assist vincenti (O_ASSIST3)
- Tiri (O_TIRI)
- Precisione nel tiro (O_PREC_TIRO)
- Tiri su azione (O_TIRI18)
- Tiri totali da calcio piazzato (O_TIRI25)
- Tiri su rigore (O_TIRI28)
- Reti totali subite su calcio piazzato (P_RETI_SUB4)
- Falli subiti (O_FALLI)
- Falli laterali in difesa (CP_FALLI1)
- Falli laterali a centrocampo (CP_FALLI4)
- Falli laterali in attacco (CP_FALLI7)
- Falli laterali in attacco da destra senza tiro (CP_FALLI10)
- Falli laterali in attacco da sinistra senza tiro (CP_FALLI13)
- Falli commessi (D_FALLI)
- Angoli (CP_ANGOLI)
- Angoli da destra senza tiro (CP_ANGOLI3)
- Angoli da sinistra senza tiro (CP_ANGOLI6)

- Punizioni in difesa (CP_PUNIZ1)
- Punizioni a centrocampo (CP_PUNIZ5)
- Punizioni in attacco (CP_PUNIZ9)
- Punizioni in attacco da destra senza tiro (CP_PUNIZ12)
- Punizioni in attacco centrali senza tiro (CP_PUNIZ15)
- Punizioni in attacco da sinistra senza tiro (CP_PUNIZ18)

2° FASE DI ELIMINAZIONE: le variabili sotto riportate sono state eliminate dal database in quanto "troppo significative" per la determinazione del risultato. Se le avessimo lasciate, il modello ci avrebbe restituito solo queste come le più significative ai fini della determinazione del risultato delle partite. Come per la prima fase di eliminazione, ogni variabile citata corrisponde a 3 variabili eliminate all'interno del database, quelle relative alla squadra di casa e alla squadra ospite e quella relativa alla differenza di valori tra queste due variabili.

- Punti (G_PUNTI1)
- Punti media inglese (G_PUNTI2)
- Assist vincenti su sviluppo di azione (O_ASSIST4)
- Assist vincenti su sviluppo di calcio piazzato (O_ASSIST5)
- Capacità realizzativa (O_CAPREAL)
- Capacità realizzativa (new) (O_CAPREAL1)
- Reti da fuori area (O_RETI)
- Reti da dentro area di porta (O_RETI1)
- Reti di destro (O_RETI2)
- Reti di sinistro (O_RETI3)
- Reti di testa (O_RETI4)
- Reti in rovesciata (O_RETI5)
- Reti al volo (O_RETI6)
- Reti totali su azione (O_RETI7)
- Reti totali su calcio piazzato (O_RETI8)
- Reti dirette su calcio piazzato (O_RETI9)
- Reti subite su rigore (P_RETI_SUB3)
- Reti subite (P_RETI_SUB)
- Reti subite direttamente su calcio piazzato (P_RETI_SUB1)
- Reti subite indirettamente su calcio piazzato (P_RETI_SUB2)
- Autoreti subite (P_AUTORETI)

- Reti (G_RETI)
- Autoreti (G_AUTORETI)
- Reti indirette su calcio piazzato (O_RETI10)
- Rigori sbagliati (CP_RIGORI1)
- Rigori trasformati (CP_RIGORI2)
- Rigori subiti (P_RIG_SUB)

3° FASE DI ELIMINAZIONE: in questa fase abbiamo individuato le variabili che sarebbero risultate poco significative ai fini della nostra analisi. Sono state eliminate principalmente per due motivi: in primis, avrebbero aumentato il "disturbo" ai fini dell'individuazione delle variabili più importanti, ed inoltre era necessario ridurre la dimensione del database in quanto eccessivamente rettangolare (380 osservazioni per 1204 variabili). Anche in questa fase, come nelle precedenti, per ogni variabile sotto menzionata corrispondono tre variabili eliminate.

- Colpi di testa (G_COLPTESTA)
- Baricentro (mt) (G_BARIC)
- Pali (G_PALI)
- Ostruzionismo (G_OSTRUZ)
- Uscite su calcio piazzato (P_USCITE1)
- Uscite su azione (P_USCITE2)
- Uscite su cross su azione (P_USCITE3)
- Uscite alte (P_USCITE4)
- Uscite alte su passaggio lungo (P_USCITE5)
- Uscite alte su pallonetto (P_USCITE6)
- Uscite alte su cross (P_USCITE7)
- Uscite alte su cross su azione (P_USCITE8)
- Uscite alte su cross dal fondo su azione (P_USCITE9)
- Uscite alte su cross da calcio piazzato (P_USCITE10)
- Uscite basse (P_USCITE11)
- Uscite basse su passaggio lungo (P_USCITE12)
- Uscite basse su pallonetto (P_USCITE13)
- Uscite basse su dribbling (P_USCITE14)
- Uscite basse su cross (P_USCITE15)
- Uscite basse su cross su azione (P_USCITE16)
- Uscite basse su cross su azione dal fondo (P_USCITE17)

- Uscite basse su cross da calcio piazzato (P_USCITE18)
- Uscite basse su passaggio filtrante (P_USCITE19)
- Rilanci lunghi (P_RILANCI1)
- Rinvii ricevuti (G_RINVII)
- Rinvii (D_RINVII)
- Rinvii di piede non del portiere (D_RINVII2)
- Rinvii di piede (D_RINVII5)
- Rinvii di piede del portiere (P_RINVII2)
- Rinvii di mano del portiere (P_RINVII3)
- Raddoppi (D_RADDOPPI)
- Recuperi in zona area (D_RECUPERI7)
- Recuperi temporanei in attacco (D_RECUPERI3)
- Pallonetti nella metà campo avversaria (O_PALLON2)
- Passaggi filtranti nella metà campo avversaria (O_PAS9)
- Passaggi lungo linea (O_PAS11)
- Passaggi di testa (O_PASTEST)
- Passaggi lunghi (O_PAS1)
- Passaggi lunghi da dietro (O_PAS3)
- Passaggi lunghi senza cambio di gioco (O_PAS5)
- Lungo linea ricevuti (G_LUNGO)
- Passaggi corti ricevuti (G_PASCORTI1)
- Passaggi corti nella metà campo avversaria (G_PASCORTI3)
- Passaggi corti utili (G_PASCORTI5)
- Passaggi alti (G_PASALTI2)
- Passaggi bassi (G_PASBASSI2)
- Passaggi bassi nella metà campo avversaria (G_PASBASSI3)
- Passaggi filtranti ricevuti (G_PASFIL)
- Cross a rientrare su calcio piazzato (O_CROSS6)
- Cross ad uscire su calcio piazzato (O_CROSS7)
- Cross ad uscire su azione (O_CROSS9)
- Cross a rientrare su azione dal fondo (O_CROSS10)
- Cross ad uscire su azione dal fondo (O_CROSS11)
- Lungo linea nella metà campo avversaria (O_LUNGO)
- Sponde di testa perse (O_SPONDE9)

- Spizzichi di testa (O_SPIZZ)
- Tiri dentro di destro (O_TIRI2)
- Tiri dentro di sinistro (O_TIRI3)
- Rimesse dal fondo (CP_RIMESSE)
- Falli laterali (CP_FALLI)
- Falli laterali a centrocampo da destra (CP_FALLI5)
- Falli laterali a centrocampo da sinistra (CP_FALLI6)
- Falli laterali da destra (CP_FALLI14)
- Falli di mano (D_FALLI5)
- Falli laterali da sinistra (CP_FALLI15)
- Punizioni a centrocampo da destra (CP_PUNIZ6)
- Punizioni a centrocampo centrali (CP_PUNIZ7)
- Punizioni a centrocampo da sinistra (CP_PUNIZ8)

4° FASE DI ELIMINAZIONE: in questa fase sono state eliminate le variabili riferite alla squadra ospite e alla differenza di valori tra le due squadre per quanto riguarda la variabile minuti di gioco (G_MINUTI); in quanto i minuti giocati in una partita sono i medesimi per entrambe le squadre che si stanno scontrando e per cui sarebbe risultato inutile mantenere questa ripetizione nel database.

APPENDICE C.

Codice R relativo al funzionamento della Random Forest.

Tale codice può essere utilizzato per diversi database, in quanto è sufficiente modificare il nome assegnato al file nel quale sono stati memorizzati i dati da caricare.

```
rm(list=ls())
memory.size(max=TRUE)

# Directory di lavoro, nella quale verranno estrapolati i dati e salvati i risultati.
setwd("C:/Users/Roberta/Desktop/Tesi/Panini2/")

# Caricamento libreria necessaria per l'implementazione del metodo Random Forest.
library(randomForest)

#########################
#       Caricamento dati        #
#########################

# Definizione del nome assegnato al file contenente il database iniziale oggetto di analisi.
dati <- read.table("DATABASEfase3tiri1.txt", na.string=".", header=TRUE)

# I dati iniziali vengono inseriti all'interno di una matrice.
# Successivamente viene visualizzata la dimensione relativa alla matrice dei dati.
dimdati <- dim(dati)
print(dimdati)

# Definizione delle tre variabili obiettivo.
c("y1","y2","y3") -> names(dati)[1:3]
dati$y1 <- factor(dati$y1,labels=c("sconfitta o pareggio", "vittoria"))
dati$y2 <- factor(dati$y2,labels=c("sconfitta", "vittoria"))
dati$y3 <- factor(dati$y3,labels=c("vittoria", "sconfitta", "pareggio"))

################################
#    Scrematura per la variabile y2       #
################################

datinomiss <- na.omit(dati)
dimdatinomiss <- dim(datinomiss)
print(dimdatinomiss)

# Le prime sette colonne non vengono considerate come variabili da utilizzare nella RF.
# Infatti si riferiscono alle tre variabili obiettivo sopra definite e ad ulteriori informazioni
# aggiuntive utili per riuscire ad interpretare al meglio i risultati.
col <- 7                              # Colonna di inizio delle vars ind

varind <- names(dati)[col:dimdati[2]]
```

```
# Numero variabili indipendenti
ind <- as.matrix(varind)
nindrc <- dim(ind)
nind <- nindrc[1]

print(nind)

################################################
#     RF considerando come variabile dipendente Y1       #
################################################

y=dati$y1
x=dati[,col:dimdati[2]]

nt <- 8000

rf.estimate <- randomForest(x, y, importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=FALSE)

print(rf.estimate)

imp <- round(importance(rf.estimate), 2)

#######################################
#          Variable Importance Measures          #
#   Colonna 3: Mean Decrease in Accuracy    #
#             Colonna 4: Gini impurity             #
#######################################

# E' sufficiente modificare il valore assegnato alla variabile colonna per ottenere i grafici
# relativi alla seconda misura di Variable Importance.
colonna <- 3

idx <- order(imp[,colonna])
imp.ord <- imp[idx,]
write.table(imp.ord, file = "imp7.txt", row.names = TRUE, col.names = TRUE)

jpeg(file="imp7_bar.jpg", width = 800, height = 800, quality = 100)
par(las=2, cex.axis=0.8)

# Definizione dell'intervallo di variabili che si intende visualizzare all'interno del grafico.
barplot(imp.ord[600:720,colonna], horiz=TRUE)

# Grafico di dispersione con definizione delle soglie.
plot(imp[,3],imp[,4])
lines(c(-1,10),c(1.05,1.05))
lines(c(0.90,0.90),c(-1,10))

selvar <- which((imp.ord[,3]>0.90 & imp.ord[,4]>1.05))
impsel <- imp.ord[selvar,3:4]
```

```
selv <- c()
for(i in 1:length(selvar)){
selv <- rbind(selv,which(names(dati)==names(selvar)[i]))
}

dev.off()

################################################
#       RF considerando come variabile dipendente Y2       #
################################################

y=datinomiss$y2
x=datinomiss[,col:dimdatinomiss[2]]

nt <- 8000

rf.estimate <- randomForest(x, y, importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=FALSE)

# Previsioni relative alla RF.
print(rf.estimate)

imp <- round(importance(rf.estimate), 2)

#######################################
#          Variable Importance Measures          #
#   Colonna 3: Mean Decrease in Accuracy    #
#             Colonna 4: Gini impurity             #
#######################################

colonna <- 3

idx <- order(imp[,colonna])
imp.ord <- imp[idx,]
write.table(imp.ord, file = "impy21.txt", row.names = TRUE, col.names = TRUE)

jpeg(file="impy21_bar.jpg", width = 800, height = 800, quality = 100)
par(las=2, cex.axis=0.8)
barplot(imp.ord[701:720,colonna], horiz=TRUE)
dev.off()

################################################
#       RF considerando come variabile dipendente Y3       #
################################################

y=dati$y3
x=dati[,col:dimdati[2]]

nt <- 8000

rf.estimate <- randomForest(x, y, importance=TRUE, proximity=FALSE,
```

```
ntree=nt, replace=TRUE, keep.forest=FALSE)

print(rf.estimate)

imp <- round(importance(rf.estimate), 2)

#####################################
#          Variable Importance Measures          #
#   Colonna 4: Mean Decrease in Accuracy    #
#              Colonna 5: Gini impurity              #
#####################################

colonna <- 5

idx <- order(imp[,colonna])
imp.ord <- imp[idx,]
write.table(imp.ord, file = "impy37.txt", row.names = TRUE, col.names = TRUE)

jpeg(file="impy37_bar.jpg", width = 800, height = 800, quality = 100)
par(las=2, cex.axis=0.8)

barplot(imp.ord[600:720,colonna], horiz=TRUE)

plot(imp[,4],imp[,5])
lines(c(-1,10),c(1.05,1.05))
lines(c(0.9,0.9),c(-1,10))

selvar <- which((imp.ord[,4]>0.9 & imp.ord[,5]>1.05))
impsel <- imp.ord[selvar,3:4]

selv <- c()
for(i in 1:length(selvar)){
selv <- rbind(selv,which(names(dati)==names(selvar)[i]))
}

dev.off()
```

APPENDICE D.

Vengono di seguito riportate tutte le partite disputate durante il campionato italiano di Serie A 2010/2011. Più precisamente vengono illustrate: le squadre, i risultati previsti dal modello, i risultati reali (con annessi gol per ciascuna squadra), e l'indicatore che identifica le partite concluse con il medesimo risultato, sia in termini di previsione che in termini reali.

Giorn	Nome sq. Di Casa	Nome sq. Ospite	Risultato previsto dal modello	Risultato reale (schedina)	Risultato effettivo	Corrispondenza fra previsione e realtà
1	BARI	JUVENTUS	1	1	1-0	OK
1	BOLOGNA	INTER	2	X	0-0	
1	CHIEVO	CATANIA	X	1	2-1	
1	FIORENTINA	NAPOLI	X	X	1-1	OK
1	MILAN	LECCE	1	1	4-0	OK
1	PALERMO	CAGLIARI	X	X	0-0	OK
1	PARMA	BRESCIA	1	1	2-0	OK
1	ROMA	CESENA	1	X	0-0	
1	SAMPDORIA	LAZIO	1	1	2-0	OK
1	UDINESE	GENOA	X	2	0-1	
2	BRESCIA	PALERMO	1	1	3-2	OK
2	CAGLIARI	ROMA	1	1	5-1	OK
2	CATANIA	PARMA	1	1	2-1	OK
2	CESENA	MILAN	1	1	2-0	OK
2	GENOA	CHIEVO	2	2	1-3	OK
2	INTER	UDINESE	1	1	2-1	OK
2	JUVENTUS	SAMPDORIA	1	X	3-3	
2	LAZIO	BOLOGNA	X	1	3-1	
2	LECCE	FIORENTINA	1	1	1-0	OK
2	NAPOLI	BARI	2	X	2-2	
3	BARI	CAGLIARI	1	X	0-0	
3	CESENA	LECCE	1	1	1-0	OK
3	CHIEVO	BRESCIA	2	2	0-1	OK
3	FIORENTINA	LAZIO	2	2	1-2	OK
3	MILAN	CATANIA	2	X	1-1	
3	PALERMO	INTER	1	2	1-2	
3	PARMA	GENOA	X	X	1-1	OK
3	ROMA	BOLOGNA	1	X	2-2	
3	SAMPDORIA	NAPOLI	1	2	1-2	
3	UDINESE	JUVENTUS	2	2	0-4	OK
4	BOLOGNA	UDINESE	2	1	2-1	
4	BRESCIA	ROMA	X	1	2-1	
4	CAGLIARI	SAMPDORIA	X	X	0-0	OK
4	CATANIA	CESENA	1	1	2-0	OK

4	GENOA	FIORENTINA	1	X	1-1	
4	INTER	BARI	1	1	4-0	OK
4	JUVENTUS	PALERMO	2	2	1-3	OK
4	LAZIO	MILAN	1	X	1-1	
4	LECCE	PARMA	1	X	1-1	
4	NAPOLI	CHIEVO	2	2	1-3	OK
5	BARI	BRESCIA	1	1	2-1	OK
5	CATANIA	BOLOGNA	1	X	1-1	
5	CESENA	NAPOLI	2	2	1-4	OK
5	CHIEVO	LAZIO	2	2	0-1	OK
5	FIORENTINA	PARMA	1	1	2-0	OK
5	JUVENTUS	CAGLIARI	1	1	4-2	OK
5	MILAN	GENOA	1	1	1-0	OK
5	PALERMO	LECCE	1	X	2-2	
5	ROMA	INTER	2	1	1-0	
5	SAMPDORIA	UDINESE	2	X	0-0	
6	BOLOGNA	SAMPDORIA	X	X	1-1	OK
6	CHIEVO	CAGLIARI	2	X	0-0	
6	FIORENTINA	PALERMO	2	2	1-2	OK
6	GENOA	BARI	X	1	2-1	
6	INTER	JUVENTUS	2	X	0-0	
6	LAZIO	BRESCIA	1	1	1-0	OK
6	LECCE	CATANIA	1	1	1-0	OK
6	NAPOLI	ROMA	1	1	2-0	OK
6	PARMA	MILAN	2	2	0-1	OK
6	UDINESE	CESENA	1	1	1-0	OK
7	BARI	LAZIO	X	2	0-2	
7	BRESCIA	UDINESE	X	2	0-1	
7	CAGLIARI	INTER	1	2	0-1	
7	CATANIA	NAPOLI	X	X	1-1	OK
7	CESENA	PARMA	X	X	1-1	OK
7	JUVENTUS	LECCE	1	1	4-0	OK
7	MILAN	CHIEVO	1	1	3-1	OK
7	PALERMO	BOLOGNA	1	1	4-1	OK
7	ROMA	GENOA	1	1	2-1	OK
7	SAMPDORIA	FIORENTINA	1	1	2-1	OK
8	BOLOGNA	JUVENTUS	1	X	0-0	
8	CHIEVO	CESENA	1	1	2-1	OK
8	FIORENTINA	BARI	1	1	2-1	OK
8	GENOA	CATANIA	1	1	1-0	OK
8	INTER	SAMPDORIA	1	X	1-1	
8	LAZIO	CAGLIARI	1	1	2-1	OK
8	LECCE	BRESCIA	1	1	2-1	OK
8	NAPOLI	MILAN	2	2	1-2	OK
8	PARMA	ROMA	X	X	0-0	OK

8	UDINESE	PALERMO	1	1	2-1	OK
9	BARI	UDINESE	2	2	0-2	OK
9	BRESCIA	NAPOLI	2	2	0-1	OK
9	CAGLIARI	BOLOGNA	1	1	2-0	OK
9	CATANIA	FIORENTINA	1	X	0-0	
9	CESENA	SAMPDORIA	2	2	0-1	OK
9	GENOA	INTER	2	2	0-1	OK
9	MILAN	JUVENTUS	X	2	1-2	
9	PALERMO	LAZIO	X	2	0-1	
9	PARMA	CHIEVO	1	X	0-0	
9	ROMA	LECCE	1	1	2-0	OK
10	BARI	MILAN	2	2	2-3	OK
10	BOLOGNA	LECCE	X	1	2-0	
10	FIORENTINA	CHIEVO	1	1	1-0	OK
10	INTER	BRESCIA	1	X	1-1	
10	JUVENTUS	CESENA	1	1	3-1	OK
10	LAZIO	ROMA	2	2	0-2	OK
10	NAPOLI	PARMA	X	1	2-0	
10	PALERMO	GENOA	1	1	1-0	OK
10	SAMPDORIA	CATANIA	X	X	0-0	OK
10	UDINESE	CAGLIARI	1	X	1-1	
11	BRESCIA	JUVENTUS	1	X	1-1	
11	CAGLIARI	NAPOLI	1	2	0-1	
11	CATANIA	UDINESE	X	1	1-0	
11	CESENA	LAZIO	1	1	1-0	OK
11	CHIEVO	BARI	X	X	0-0	OK
11	GENOA	BOLOGNA	1	1	1-0	OK
11	LECCE	INTER	2	X	1-1	
11	MILAN	PALERMO	1	1	3-1	OK
11	PARMA	SAMPDORIA	2	1	1-0	
11	ROMA	FIORENTINA	1	1	3-2	OK
12	BARI	PARMA	X	2	0-1	
12	BOLOGNA	BRESCIA	2	1	1-0	
12	CAGLIARI	GENOA	X	2	0-1	
12	FIORENTINA	CESENA	1	1	1-0	OK
12	INTER	MILAN	X	2	0-1	
12	JUVENTUS	ROMA	1	X	1-1	
12	LAZIO	NAPOLI	1	1	2-0	OK
12	PALERMO	CATANIA	1	1	3-1	OK
12	SAMPDORIA	CHIEVO	X	X	0-0	OK
12	UDINESE	LECCE	1	1	4-0	OK
13	BRESCIA	CAGLIARI	2	2	1-2	OK
13	CATANIA	BARI	1	1	1-0	OK
13	CESENA	PALERMO	2	2	1-2	OK
13	CHIEVO	INTER	2	1	2-1	

13	GENOA	JUVENTUS	X	2	0-2	
13	LECCE	SAMPDORIA	X	2	2-3	
13	MILAN	FIORENTINA	1	1	1-0	OK
13	NAPOLI	BOLOGNA	1	1	4-1	OK
13	PARMA	LAZIO	1	X	1-1	
13	ROMA	UDINESE	1	1	4-0	OK
14	BARI	CESENA	1	X	1-1	
14	BOLOGNA	CHIEVO	1	1	2-1	OK
14	BRESCIA	GENOA	X	X	0-0	OK
14	CAGLIARI	LECCE	1	1	3-2	OK
14	INTER	PARMA	1	1	5-2	OK
14	JUVENTUS	FIORENTINA	1	X	1-1	
14	LAZIO	CATANIA	X	X	1-1	OK
14	PALERMO	ROMA	1	1	3-1	OK
14	SAMPDORIA	MILAN	1	X	1-1	
14	UDINESE	NAPOLI	1	1	3-1	OK
15	CATANIA	JUVENTUS	2	2	1-3	OK
15	CESENA	BOLOGNA	X	2	0-2	
15	CHIEVO	ROMA	2	X	2-2	
15	FIORENTINA	CAGLIARI	1	1	1-0	OK
15	LAZIO	INTER	1	1	3-1	OK
15	LECCE	GENOA	2	2	1-3	OK
15	MILAN	BRESCIA	1	1	3-0	OK
15	NAPOLI	PALERMO	1	1	1-0	OK
15	PARMA	UDINESE	1	1	2-1	OK
15	SAMPDORIA	BARI	1	1	3-0	OK
16	BOLOGNA	MILAN	2	2	0-3	OK
16	BRESCIA	SAMPDORIA	1	1	1-0	OK
16	CAGLIARI	CATANIA	1	1	3-0	OK
16	GENOA	NAPOLI	X	2	0-1	
16	INTER	CESENA	1	1	3-2	OK
16	JUVENTUS	LAZIO	1	1	2-1	OK
16	LECCE	CHIEVO	1	1	3-2	OK
16	PALERMO	PARMA	1	1	3-1	OK
16	ROMA	BARI	1	1	1-0	OK
16	UDINESE	FIORENTINA	1	1	2-1	OK
17	BARI	PALERMO	1	X	1-1	
17	CATANIA	BRESCIA	1	1	1-0	OK
17	CESENA	CAGLIARI	1	1	1-0	OK
17	CHIEVO	JUVENTUS	2	X	1-1	
17	FIORENTINA	INTER	2	2	1-2	OK
17	LAZIO	UDINESE	1	1	3-2	OK
17	MILAN	ROMA	1	2	0-1	
17	NAPOLI	LECCE	1	1	1-0	OK
17	PARMA	BOLOGNA	1	X	0-0	

17	SAMPDORIA	GENOA	2	2	0-1	OK
18	BOLOGNA	FIORENTINA	2	X	1-1	
18	BRESCIA	CESENA	X	2	1-2	
18	CAGLIARI	MILAN	X	2	0-1	
18	GENOA	LAZIO	X	X	0-0	OK
18	INTER	NAPOLI	1	1	3-1	OK
18	JUVENTUS	PARMA	2	2	1-4	OK
18	LECCE	BARI	2	2	0-1	OK
18	PALERMO	SAMPDORIA	1	1	3-0	OK
18	ROMA	CATANIA	1	1	4-2	OK
18	UDINESE	CHIEVO	1	1	2-0	OK
19	BARI	BOLOGNA	1	2	0-2	
19	CATANIA	INTER	1	2	1-2	
19	CESENA	GENOA	X	X	0-0	OK
19	CHIEVO	PALERMO	X	X	0-0	OK
19	FIORENTINA	BRESCIA	X	1	3-2	
19	LAZIO	LECCE	2	2	1-2	OK
19	MILAN	UDINESE	2	X	4-4	
19	NAPOLI	JUVENTUS	1	1	3-0	OK
19	PARMA	CAGLIARI	2	2	1-2	OK
19	SAMPDORIA	ROMA	1	1	2-1	OK
20	BRESCIA	PARMA	1	1	2-0	OK
20	CAGLIARI	PALERMO	1	1	3-1	OK
20	CATANIA	CHIEVO	1	X	1-1	
20	CESENA	ROMA	2	2	0-1	OK
20	GENOA	UDINESE	1	2	2-4	
20	INTER	BOLOGNA	1	1	4-1	OK
20	JUVENTUS	BARI	1	1	2-1	OK
20	LAZIO	SAMPDORIA	1	1	1-0	OK
20	LECCE	MILAN	2	X	1-1	
20	NAPOLI	FIORENTINA	X	X	0-0	OK
21	BARI	NAPOLI	X	2	0-2	
21	BOLOGNA	LAZIO	1	1	3-1	OK
21	CHIEVO	GENOA	2	X	0-0	
21	FIORENTINA	LECCE	X	X	1-1	OK
21	MILAN	CESENA	1	1	2-0	OK
21	PALERMO	BRESCIA	1	1	1-0	OK
21	PARMA	CATANIA	X	1	2-0	
21	ROMA	CAGLIARI	1	1	3-0	OK
21	SAMPDORIA	JUVENTUS	2	X	0-0	
21	UDINESE	INTER	1	1	3-1	OK
22	BOLOGNA	ROMA	X	2	0-1	
22	BRESCIA	CHIEVO	2	2	0-3	OK
22	CAGLIARI	BARI	1	1	2-1	OK
22	CATANIA	MILAN	2	2	0-2	OK

22	GENOA	PARMA	1	1	3-1	OK
22	INTER	PALERMO	1	1	3-2	OK
22	JUVENTUS	UDINESE	2	2	1-2	OK
22	LAZIO	FIORENTINA	1	1	2-0	OK
22	LECCE	CESENA	1	X	1-1	
22	NAPOLI	SAMPDORIA	1	1	4-0	OK
23	BARI	INTER	2	2	0-3	OK
23	CESENA	CATANIA	1	X	1-1	
23	CHIEVO	NAPOLI	1	1	2-0	OK
23	FIORENTINA	GENOA	2	1	1-0	
23	MILAN	LAZIO	2	X	0-0	
23	PALERMO	JUVENTUS	1	1	2-1	OK
23	PARMA	LECCE	1	2	0-1	
23	ROMA	BRESCIA	X	X	1-1	OK
23	SAMPDORIA	CAGLIARI	X	2	0-1	
23	UDINESE	BOLOGNA	2	X	1-1	
24	BOLOGNA	CATANIA	1	1	1-0	OK
24	BRESCIA	BARI	1	1	2-0	OK
24	CAGLIARI	JUVENTUS	2	2	1-3	OK
24	GENOA	MILAN	1	X	1-1	
24	INTER	ROMA	1	1	5-3	OK
24	LAZIO	CHIEVO	1	X	1-1	
24	LECCE	PALERMO	2	2	2-4	OK
24	NAPOLI	CESENA	1	1	2-0	OK
24	PARMA	FIORENTINA	2	X	1-1	
24	UDINESE	SAMPDORIA	1	1	2-0	OK
25	BARI	GENOA	X	X	0-0	OK
25	BRESCIA	LAZIO	1	2	0-2	
25	CAGLIARI	CHIEVO	1	1	4-1	OK
25	CATANIA	LECCE	X	1	3-2	
25	CESENA	UDINESE	2	2	0-3	OK
25	JUVENTUS	INTER	2	1	1-0	
25	MILAN	PARMA	1	1	4-0	OK
25	PALERMO	FIORENTINA	2	2	2-4	OK
25	ROMA	NAPOLI	X	2	0-2	
25	SAMPDORIA	BOLOGNA	1	1	3-1	OK
26	BOLOGNA	PALERMO	2	1	1-0	
26	CHIEVO	MILAN	2	2	1-2	OK
26	FIORENTINA	SAMPDORIA	X	X	0-0	OK
26	GENOA	ROMA	1	1	4-3	OK
26	INTER	CAGLIARI	1	1	1-0	OK
26	LAZIO	BARI	1	1	1-0	OK
26	LECCE	JUVENTUS	1	1	2-0	OK
26	NAPOLI	CATANIA	1	1	1-0	OK
26	PARMA	CESENA	X	X	2-2	OK

26	UDINESE	BRESCIA	X	X	0-0	OK
27	BARI	FIORENTINA	2	X	1-1	
27	BRESCIA	LECCE	X	X	2-2	OK
27	CAGLIARI	LAZIO	1	1	1-0	OK
27	CATANIA	GENOA	1	1	2-1	OK
27	CESENA	CHIEVO	1	1	1-0	OK
27	JUVENTUS	BOLOGNA	2	2	0-2	OK
27	MILAN	NAPOLI	1	1	3-0	OK
27	PALERMO	UDINESE	2	2	0-7	OK
27	ROMA	PARMA	1	X	2-2	
27	SAMPDORIA	INTER	2	2	0-2	OK
28	BOLOGNA	CAGLIARI	2	X	2-2	
28	CHIEVO	PARMA	X	X	0-0	OK
28	FIORENTINA	CATANIA	1	1	3-0	OK
28	INTER	GENOA	1	1	5-2	OK
28	JUVENTUS	MILAN	X	2	0-1	
28	LAZIO	PALERMO	1	1	2-0	OK
28	LECCE	ROMA	2	2	1-2	OK
28	NAPOLI	BRESCIA	2	X	0-0	
28	SAMPDORIA	CESENA	X	2	2-3	
28	UDINESE	BARI	1	1	1-0	OK
29	BRESCIA	INTER	2	X	1-1	
29	CAGLIARI	UDINESE	2	2	0-4	OK
29	CATANIA	SAMPDORIA	1	1	1-0	OK
29	CESENA	JUVENTUS	1	X	2-2	
29	CHIEVO	FIORENTINA	2	2	0-1	OK
29	GENOA	PALERMO	1	1	1-0	OK
29	LECCE	BOLOGNA	2	2	0-1	OK
29	MILAN	BARI	1	X	1-1	
29	PARMA	NAPOLI	2	2	1-3	OK
29	ROMA	LAZIO	1	1	2-0	OK
30	BARI	CHIEVO	2	2	1-2	OK
30	BOLOGNA	GENOA	2	X	1-1	
30	FIORENTINA	ROMA	2	X	2-2	
30	INTER	LECCE	1	1	1-0	OK
30	JUVENTUS	BRESCIA	1	1	2-1	OK
30	LAZIO	CESENA	1	1	1-0	OK
30	NAPOLI	CAGLIARI	1	1	2-1	OK
30	PALERMO	MILAN	1	1	1-0	OK
30	SAMPDORIA	PARMA	X	2	0-1	
30	UDINESE	CATANIA	1	1	2-0	OK
31	BRESCIA	BOLOGNA	1	1	3-1	OK
31	CATANIA	PALERMO	1	1	4-0	OK
31	CESENA	FIORENTINA	2	X	2-2	
31	CHIEVO	SAMPDORIA	X	X	0-0	OK

31	GENOA	CAGLIARI	2	2	0-1	OK
31	LECCE	UDINESE	1	1	2-0	OK
31	MILAN	INTER	1	1	3-0	OK
31	NAPOLI	LAZIO	2	1	4-3	
31	PARMA	BARI	2	2	1-2	OK
31	ROMA	JUVENTUS	2	2	0-2	OK
32	BARI	CATANIA	X	X	1-1	OK
32	BOLOGNA	NAPOLI	2	2	0-2	OK
32	CAGLIARI	BRESCIA	X	X	1-1	OK
32	FIORENTINA	MILAN	2	2	1-2	OK
32	INTER	CHIEVO	1	1	2-0	OK
32	JUVENTUS	GENOA	1	1	3-2	OK
32	LAZIO	PARMA	1	1	2-0	OK
32	PALERMO	CESENA	1	X	2-2	
32	SAMPDORIA	LECCE	2	2	1-2	OK
32	UDINESE	ROMA	X	2	1-2	
33	CATANIA	LAZIO	1	2	1-4	
33	CESENA	BARI	1	1	1-0	OK
33	CHIEVO	BOLOGNA	1	1	2-0	OK
33	FIORENTINA	JUVENTUS	X	X	0-0	OK
33	GENOA	BRESCIA	2	1	3-0	
33	LECCE	CAGLIARI	2	X	3-3	
33	MILAN	SAMPDORIA	1	1	3-0	OK
33	NAPOLI	UDINESE	X	2	1-2	
33	PARMA	INTER	X	1	2-0	
33	ROMA	PALERMO	X	2	2-3	
34	BARI	SAMPDORIA	2	2	0-1	OK
34	BOLOGNA	CESENA	2	2	0-2	OK
34	BRESCIA	MILAN	2	2	0-1	OK
34	CAGLIARI	FIORENTINA	2	2	1-2	OK
34	GENOA	LECCE	1	1	4-2	OK
34	INTER	LAZIO	1	1	2-1	OK
34	JUVENTUS	CATANIA	1	X	2-2	
34	PALERMO	NAPOLI	1	1	2-1	OK
34	ROMA	CHIEVO	1	1	1-0	OK
34	UDINESE	PARMA	2	2	0-2	OK
35	BARI	ROMA	1	2	2-3	
35	CATANIA	CAGLIARI	1	1	2-0	OK
35	CESENA	INTER	2	2	1-2	OK
35	CHIEVO	LECCE	2	1	1-0	
35	FIORENTINA	UDINESE	1	1	5-2	OK
35	LAZIO	JUVENTUS	X	2	0-1	
35	MILAN	BOLOGNA	1	1	1-0	OK
35	NAPOLI	GENOA	X	1	1-0	
35	PARMA	PALERMO	1	1	3-1	OK

35	SAMPDORIA	BRESCIA	X	X	3-3	OK
36	BOLOGNA	PARMA	1	X	0-0	
36	BRESCIA	CATANIA	2	2	1-2	OK
36	CAGLIARI	CESENA	2	2	0-2	OK
36	GENOA	SAMPDORIA	X	1	2-1	
36	INTER	FIORENTINA	1	1	3-1	OK
36	JUVENTUS	CHIEVO	2	X	2-2	
36	LECCE	NAPOLI	1	1	2-1	OK
36	PALERMO	BARI	2	1	2-1	
36	ROMA	MILAN	X	X	0-0	OK
36	UDINESE	LAZIO	1	1	2-1	OK
37	BARI	LECCE	2	2	0-2	OK
37	CATANIA	ROMA	1	1	2-1	OK
37	CESENA	BRESCIA	1	1	1-0	OK
37	CHIEVO	UDINESE	2	2	0-2	OK
37	FIORENTINA	BOLOGNA	1	X	1-1	
37	LAZIO	GENOA	2	1	4-2	
37	MILAN	CAGLIARI	1	1	4-1	OK
37	NAPOLI	INTER	2	X	1-1	
37	PARMA	JUVENTUS	2	1	1-0	
37	SAMPDORIA	PALERMO	1	2	1-2	
38	BOLOGNA	BARI	X	2	0-4	
38	BRESCIA	FIORENTINA	1	X	2-2	
38	CAGLIARI	PARMA	X	X	1-1	OK
38	GENOA	CESENA	1	1	3-2	OK
38	INTER	CATANIA	1	1	3-1	OK
38	JUVENTUS	NAPOLI	1	X	2-2	
38	LECCE	LAZIO	1	2	2-4	
38	PALERMO	CHIEVO	2	2	1-3	OK
38	ROMA	SAMPDORIA	1	1	3-1	OK
38	UDINESE	MILAN	1	X	0-0	

APPENDICE E.

Codice R relativo alla costruzione dei grafici, necessari per l'analisi dell'andamento della probabilità, assegnata dal modello Random Forest, al variare della differenza reti riscontrata in ciascun incontro.

```
# Pulizia dell'area di lavoro
rm(list=ls())
memory.size(max=TRUE)

# Directory di lavoro
setwd("C:/Users/Roberta/Desktop/Tesi/")

######################
#      caricamento dati      #
######################
dati <- read.table("BoxPlotDiffReti.txt", na.string=".", header=TRUE)

dimdati <- dim(dati)
print(dimdati)

# Grafico con linea di tendenza per verificare l'andamento della probabilità
par(las=2, cex.axis=0.8)
par (mar=c(5,4,2,2))
plot(dati[,1],dati[,3], xlab='Differenza reti', ylab='Probabilità risultato effettivo' )
lines(ksmooth(dati[,1],dati[,3],"normal",1),col='red',lwd=2)
dev.off()

# Box Plot
par(las=2, cex.axis=0.8)
par (mar=c(5,4,2,2))
plot(as.factor(dati[,1]),dati[,3], xlab='Differenza reti', ylab='Probabilità risultato effettivo')
dev.off()
```

APPENDICE F.

Qui potrete trovare alcuni degli elenchi disponibili sulle partite "sospette" del campionato italiano di Serie A 2010/2011. Tali elenchi non sempre corrispondono. Tuttavia, come più volte ribadito, non essendoci ancora nulla di certo non possiamo che considerare tutte queste varianti, in modo da non correre il rischio di ignorare alcuni aspetti che in un secondo momento potrebbero rilevarsi importanti.

Gli elenchi, verranno contrassegnati con i colori verde, arancio e rosso, in modo da identificare come il modello prevede queste partite. Tali colori riprendono quelli identificati per la creazione della Figura 36.

Elenco pubblicato sul giornale Repubblica il giorno 03/02/12.

1. **Napoli – Sampdoria** 4 – 0 (30/01/11)
2. **Brescia – Chievo** 0 – 3 (31/01/11)
3. **Brescia – Bari** 2 – 0 (06/02/11)
4. **Genoa – Roma** 4 – 3 (26/02/11)
5. **Bari – Chievo** 1 – 2 (20/03/11)
6. **Brescia – Bologna** 3 – 1 (02/04/11)
7. **Parma – Bari** 1 – 2 (03/04/11)
8. **Chievo – Sampdoria** 0 – 0 (03/04/11)
9. **Inter – Lecce** 1 – 0 (12/04/11)
10. **Bari – Sampdoria** 0 – 1 (23/04/11)
11. **Bari – Roma** 2 – 3 (01/05/11)
12. **Bari – Palermo** 1 – 2 (07/05/11)
13. **Lazio – Genoa** 4 – 2 (14/05/11)
14. **Lecce – Lazio** 2 – 4 (22/05/11)

Elenco trovato su vari siti internet come: http://it.eurosport.yahoo.com e www.epicfootball.org. L'elenco in questione, era già stato reso noto a partire dal 21/12/11 ed ancora oggi i telegiornali e i vari notiziari parlano di queste 22 partite. Gli inquirenti stanno ancora indagando in tale direzione. Pare, che quest'ultimo, si allungherà di molto grazie alle svolte dell'ultimo periodo. Per il momento, tuttavia, non sono ancora trapelate altre indiscrezioni.

1. **Catania – Chievo** 1 – 1 (11/01/11)
2. **Napoli – Sampdoria** 4 – 0 (30/01/11)
3. **Brescia – Chievo** 0 – 3 (31/01/11)
4. **Brescia – Bari** 2 – 0 (06/02/11)
5. **Genoa – Roma** 4 – 3 (26/02/11)
6. **Brescia – Lecce** 2 – 2 (27/02/11)
7. **Fiorentina – Roma** 2 – 2 (20/03/11)
8. **Genoa – Lecce** 4 – 2 (20/03/11)
9. **Bari – Chievo** 1 – 2 (20/03/11)
10. **Brescia – Bologna** 3 – 1 (02/04/11)
11. **Parma – Bari** 1 – 2 (03/04/11)
12. **Chievo – Sampdoria** 0 – 0 (03/04/11)
13. **Bologna – Napoli** 0 – 2 (10/04/11)
14. **Chievo – Bologna** 2 – 0 (17/04/11)
15. **Lecce – Cagliari** 3 – 3 (17/04/11)
16. **Catania – Cagliari** 2 -0 (01/05/11)
17. **Lazio – Genoa** 4 – 2 (14/05/11)
18. **Chievo – Udinese** 0 – 2 (15/05/11)
19. **Catania – Roma** 2 – 1 (15/05/11)
20. **Genoa – Cesena** 3 – 2 (22/05/11)
21. **Lecce – Lazio** 2 – 4 (22/05/11)
22. **Bologna – Bari** 0 – 4 (22/05/11)

APPENDICE G.

Codice R necessario per valutare l'influenza dei fattori sulle probabilità di vittoria e sconfitta della squadra di casa.

```
#Pulizia dell'area di lavoro
rm(list=ls())
memory.size(max=TRUE)

#Directory di lavoro
setwd("C:/Users/Roberta/Desktop/Tesi/Systat/")

library(randomForest)

#################################
#              caricamento dati             #
#################################

dati <- read.table("DBSSD.txt", na.string=".", header=TRUE)

dimdati <- dim(dati)
print(dimdati)

c("y1","y2","y3") -> names(dati)[1:3]
dati$y1 <- factor(dati$y1,labels=c("sconfitta o pareggio", "vittoria"))
dati$y2 <- factor(dati$y2,labels=c("sconfitta", "vittoria"))
dati$y3 <- factor(dati$y3,labels=c("vittoria", "sconfitta", "pareggio"))

col <- 7
varind <- names(dati)[col:dimdati[2]]

# numero variabili indipendenti
ind <- as.matrix(varind)
nindrc <- dim(ind)
nind <- nindrc[1]

print(nind)

#########################################################
# RF considerando come variabile dipendente Y3                    #
#########################################################

y=dati$y3
x=dati[,col:dimdati[2]]

nt <- 8000

rf.estimate <- randomForest(x, y, importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)
```

```
print(rf.estimate)
predict(rf.estimate)
imp <- round(importance(rf.estimate), 2)
voti <-rf.estimate$votes

#####################################################
#   Osserviamo come ciascun fattore influenza le probabilità    #
#    depurandolo dall'effetto prodotto dai restanti fattori          #
#####################################################

###################################
#                Primo fattore                  #
###################################

rf.estimate1 <- randomForest(x[,-1], voti[,1], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate1)
yt <- predict(rf.estimate1)
resy <- voti[,1]-yt

rf.estimate2 <- randomForest(x[,-1], x[,1], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate2)
xt <- predict(rf.estimate2)
resx <- x[,1]-xt

par (mar=c(5,4,2,2))
plot(resx, resy, type='p',xlab='Giocate aeree effettuate dalla sq. casa',ylab='Probabilità di
vittoria sq. casa')
 lines(ksmooth(resx, resy, "normal",1.5),col='red')

###################################
#                Secondo fattore                #
###################################

rf.estimate1 <- randomForest(x[,-2], voti[,1], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate1)
yt <- predict(rf.estimate1)
resy <- voti[,1]-yt

rf.estimate2 <- randomForest(x[,-2], x[,2], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate2)
xt <- predict(rf.estimate2)
resx <- x[,2]-xt
```

```
par (mar=c(5,4,2,2))
plot(resx, resy, type='p',xlab='Giocate pericolose a palla bassa sq. casa',ylab='Probabilità di
vittoria sq. casa')
 lines(ksmooth(resx, resy, "normal",2),col='red')

###################################
#              Terzo fattore              #
###################################

rf.estimate1 <- randomForest(x[,-3], voti[,1], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate1)
yt <- predict(rf.estimate1)
resy <- voti[,1]-yt

rf.estimate2 <- randomForest(x[,-3], x[,3], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate2)
xt <- predict(rf.estimate2)
resx <- x[,3]-xt

par (mar=c(5,4,2,2))
plot(resx, resy, type='p',xlab='Fase difensiva sq. casa',ylab='Probabilità di vittoria sq. casa')
 lines(ksmooth(resx, resy, "normal",2),col='red')

###################################
#              Quarto fattore              #
###################################

rf.estimate1 <- randomForest(x[,-4], voti[,2], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate1)
yt <- predict(rf.estimate1)
resy <- voti[,2]-yt

rf.estimate2 <- randomForest(x[,-4], x[,4], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate2)
xt <- predict(rf.estimate2)
resx <- x[,4]-xt

par (mar=c(5,4,2,2))
plot(resx, resy, type='p',xlab='Fase difensiva protezione della porta sq.
ospite',ylab='Probabilità di sconfitta sq. casa')
 lines(ksmooth(resx, resy, "normal",1),col='red')
```

```
###################################
#               Quinto fattore                #
###################################

rf.estimate1 <- randomForest(x[,-5], voti[,2], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate1)
yt <- predict(rf.estimate1)
resy <- voti[,2]-yt

rf.estimate2 <- randomForest(x[,-5], x[,5], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate2)
xt <- predict(rf.estimate2)
resx <- x[,5]-xt

par (mar=c(5,4,2,2))
plot(resx, resy, type='p',xlab='Barriera difensiva a protezione area sq. ospite',ylab='Probabilità
di sconfitta sq. casa')
 lines(ksmooth(resx, resy, "normal",1),col='red')

###################################
#               Sesto fattore                 #
###################################

rf.estimate1 <- randomForest(x[,-6], voti[,2], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate1)
yt <- predict(rf.estimate1)
resy <- voti[,2]-yt

rf.estimate2 <- randomForest(x[,-6], x[,6], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)
print(rf.estimate2)
xt <- predict(rf.estimate2)
resx <- x[,6]-xt

par (mar=c(5,4,2,2))
plot(resx, resy, type='p',xlab='Passaggi lunghi oltre il centrocampo sq.
ospite',ylab='Probabilità di sconfitta sq. casa')
 lines(ksmooth(resx, resy, "normal",1),col='red')

###################################
#               Settimo fattore               #
###################################

rf.estimate1 <- randomForest(x[,-7], voti[,2], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)
```

```
print(rf.estimate1)
yt <- predict(rf.estimate1)
resy <- voti[,2]-yt

rf.estimate2 <- randomForest(x[,-7], x[,7], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate2)
xt <- predict(rf.estimate2)
resx <- x[,7]-xt

par (mar=c(5,4,2,2))
plot(resx, resy, type='p',xlab='Pericolosità attacco sq. ospite',ylab='Probabilità di sconfitta sq.
casa')
 lines(ksmooth(resx, resy, "normal",1.5),col='red')

###################################
#                Ottavo fattore                #
###################################

rf.estimate1<- randomForest(x[,-8], voti[,2], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate1)
yt <- predict(rf.estimate1)
resy <- voti[,2]-yt

rf.estimate2 <- randomForest(x[,-8], x[,8], importance=TRUE, proximity=FALSE,
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate2)
xt <- predict(rf.estimate2)
resx <- x[,8]-xt

par (mar=c(5,4,2,2))
plot(resx, resy, type='p',xlab='Fase di protezione area sq. ospite',ylab='Probabilità di sconfitta
sq. casa')
 lines(ksmooth(resx, resy, "normal",2),col='red')
```

APPENDICE H.

Di seguito viene riportato il codice R utilizzato per applicare il modello della Cluster Analysis ai valori medi riferiti agli indici standardizzati.

```
#Pulizia dell'area di lavoro
rm(list=ls())
memory.size(max=TRUE)

#Directory di lavoro
setwd("C:/Users/Roberta/Desktop/Tesi/Panini2/")

#Caricamento dei dati
config <- read.table("IndiciMediStandardizzati.txt", na.string=".", header=TRUE)

#Numero di cluster massimo e numero di iterazioni
 nclumax <- 5
 runs <- 5

sapply(as.data.frame(config),var)

row.names(config) <- config[,1]
 config <- config[,-1]

nvars <- dim(config)[1]
 n <- dim(config)[1]
dim <- dim(config)[2]

varfra.nclu <- array(0,nclumax)
 clu.k <- matrix(1,n,nclumax)
 clumean <- matrix(0,nclumax,dim*nclumax)

for(nclu in 2:nclumax){
 varfra.tot.runs <- array(0,runs)
 clumean.r <- matrix(0,nclu,runs*dim)
 clu.r <- array(0,n*runs)

 for(r in 1:runs){
  ca.k <- kmeans(config, nclu, algorithm = "Hartigan-Wong")
  clu.r[((r-1)*n+1):(r*n)] <- ca.k$cluster
  clumean.r[,((r-1)*dim+1):(r*dim)] <- ca.k$centers

  varfra <- matrix(0,dim,1)
  varnei <- matrix(0,dim,1)
    for(jj in 1:dim){
     varfra[jj] <- sum((ca.k$centers[,jj]-mean(config[,jj]))^2*table(ca.k$cluster))/n
     varnei[jj] <- sum(tapply(config[,jj],ca.k$cluster,var)*table(ca.k$cluster))/n
     }

   varfra.tot.runs[r] <- sum(varfra)/sum(varfra+varnei)
```

```
}

maxvarfra <- max(varfra.tot.runs)
varfra.nclu[nclu] <- maxvarfra

r <- which(varfra.tot.runs==maxvarfra)[1]

clu.k[,nclu] <- clu.r[((r-1)*n+1):(r*n)]
clumean[1:nclu,((nclu-1)*dim+1):(nclu*dim)] <- clumean.r[,((r-1)*dim+1):(r*dim)]
}

clu.k <- as.data.frame(clu.k)
row.names(clu.k) <- row.names(config)

v <- round((varfra.nclu[3:nclumax]/varfra.nclu[2:(nclumax-1)]-1)*100,2)
etch <- paste("+",v,"%",sep="")

plot(c(1:nclumax),varfra.nclu*100,
    xlab="number of clusters",ylab="DB/DT - Increments",type='b',pch=19,
    lty="solid",xaxp=c(1,nclumax,nclumax-1),ylim=c(0,60))
lines(c(3:nclumax),v,type='b',lty="dotted",pch=0)
text(c(3:nclumax),v,etch,cex=0.75,pos=1)
legend(1,59,c("DB/DT","Increments"),lty=c("solid","dotted"),pch=c(19,0,8),bty='n')

#Identificazione del file nel quale verranno salvati i risultati ottenuti
write.table(clu.k,file="clusters.txt")
```

APPENDICE I.

Codice R necessario per prevedere i risultati delle partite del campionato 2011/2012.

```
# Pulizia area di lavoro.
rm(list=ls())
memory.size(max=TRUE)

# Directory di lavoro.
setwd("C:/Users/Roberta/Desktop/Tesi/Systat/")

# Libreria necessaria per l'utilizzo del metodo.
library(randomForest)

######################
#      caricamento dati      #
######################

dati <- read.table("DBSSD.txt", na.string=".", header=TRUE)

dimdati <- dim(dati)
print(dimdati)

c("y1","y2","y3") -> names(dati)[1:3]
dati$y1 <- factor(dati$y1,labels=c("sconfitta o pareggio", "vittoria"))
dati$y2 <- factor(dati$y2,labels=c("sconfitta", "vittoria"))
dati$y3 <- factor(dati$y3,labels=c("vittoria", "sconfitta", "pareggio"))

col <- 7

varind <- names(dati)[col:dimdati[2]]

# numero variabili indipendenti
ind <- as.matrix(varind)
nindrc <- dim(ind)
nind <- nindrc[1]

print(nind)

##########################################
#    RF su tutti i dati con grafici e salvataggio dati   #
#    considerando come variabile dipendente Y3     #
##########################################

y=dati$y3
x=dati[,col:dimdati[2]]

nt <- 8000

rf.estimate <- randomForest(x, y, importance=TRUE, proximity=FALSE,
```

```
ntree=nt, replace=TRUE, keep.forest=TRUE)

print(rf.estimate)

predict(rf.estimate)
imp <- round(importance(rf.estimate), 2)
rf.estimate$votes

voti <-rf.estimate$votes
write.table(voti, file = "voti1.txt", row.names = TRUE, col.names = TRUE)

############################
#       Dati campionato 2012        #
############################

datic <- read.table("DB2012.txt", na.string=".", header=TRUE)

dimdatic <- dim(datic)
print(dimdatic)

xnew <- datic[,-c(1:3)]
prevs <- predict(rf.estimate,xnew)
prevs.p <- predict(rf.estimate,xnew,type='prob')

# Salvataggio delle previsioni.
write.table(prevs.p, file = "previsioni2012.txt", row.names = TRUE, col.names = TRUE)
```

BIBLIOGRAFIA

EUGENIO BRENTARI, LIVIA DANCELLI e PAOLA ZUCCOLOTTO, *Analisi dei fattori di maggior impatto sulla formazione del prezzo del vino: un esempio di utilizzo delle Random Forest*, paper.

LEO BREIMAN, *Random Forest*, Robert E. Schapire, Berkely

MARCO SANDRI e PAOLA ZUCCOLOTTO, *Analysis of a bias effect in a tree-based variable importance measure*, paper.

MARCO SANDRI e PAOLA ZUCCOLOTTO, *Classification with Random Forests: the theoretical frame work*, paper.

MAURIZIO CARPITA, *Modelli per il Data Mining*, lucidi lezioni in aula.

PAOLA ZUCCOLOTTO, *Reti neurali per lo studio del mercato*, lucidi lezioni in aula.

SERGIO ZANI e ANDREA CERIOLI, *Analisi dei dati e data mining per le decisioni aziendali*, Giuffrè editore, Milano.

SITOGRAFIA

http://msdn.microsoft.com/it-it/library/ms174949.aspx

http://it.wikipedia.org/wiki/Serie_A

http://it.wikipedia.org/wiki/Calcio_(sport)

http://segnalazionit.org/2011/08/la-statistica-nel-pallone/

http://www.ilmeteo.it/portale/archivio-meteo/

http://www.legaseriea.it/it/sala-stampa/archivio-match-report/

http://soccerstatistically.blogspot.com/

http://www.paninidigital.com/site/IT/IT/azienda

http://it.wikipedia.org/wiki/

http://www.sport.it/

http://www.gazzetta.it/Calcio/

http://www.aforismario.it/aforismi-calcio.htm

http://it.eurosport.yahoo.com

www.epicfootball.org

RINGRAZIAMENTI.

Arrivati al termine di questo percorso di studi, non posso far altro che esprimere la mia più sincera gratitudine a quanti mi sono stati vicini.

In particolar modo, vorrei ringraziare la professoressa Paola Zuccolotto per la grande disponibilità prestata durante tutto il periodo relativo alla stesura della tesi e per la passione con la quale affronta il suo lavoro. Il professor Maurizio Carpita il quale ha contribuito con varie idee all'ampliamento di questo elaborato e il professor Eugenio Brentari finanziatore dei dati necessari per l'attuazione di questo lavoro.

Inoltre, vorrei ringraziare la Panini Digital per aver fornito in tempi celeri tutti i dati di cui avevamo bisogno. In particolar modo il mio grazie va al Dott. Aldo Maccagni e al Sig. Andrea Pompili.

Un sincero ringraziamento va ai miei genitori Marina e Giuseppe, a mia sorella Monica, al mio fidanzato Corrado e ai miei nonni Caterina, Giorgio e Luigia, poiché sono stati il mio punto di riferimento, trasmettendomi importanti valori che mi hanno permesso di crescere. Vorrei esprimere un'ulteriore ringraziamento a mio padre per il suo sostanziale aiuto durante lo svolgimento di questa tesi. Senza di lui avrei avuto moltissime difficoltà, data la mia scarsa conoscenza in merito agli aspetti tecnici legati al mondo del calcio. Pertanto grazie di cuore papà, parte della buona riuscita di questa tesi è anche tua.

Ringrazio calorosamente tutte le persone con cui ho condiviso la mia esperienza universitaria, dallo studio per la preparazione degli esami, fino ai momenti di maggiore divertimento, in particolare vorrei ringraziare i miei compagni di corso Alessandro, Emanuele, Mara, Michele e Simone. Non da meno, sento il bisogno di ringraziare anche tutti i miei amici sui quali so di poter sempre fare affidamento, in particolare Catia, Claudio, Eva, Federica, Mirko, Paolo, Patrizia B., Patrizia V., tutti i soci ed ex soci facenti parte del Rotaract Lovere-Iseo-Breno e tutti i componenti della Gelateria Mej.

Infine, l'ultimo ringraziamento va a due persone che, purtroppo, non potranno leggere queste righe ma che sono state molto importanti nella mia vita: mio nonno Giovanni che mi ha sempre spinta a ragionare con la mia testa e a non darmi per vinta anche nelle situazioni di maggiore difficoltà e Cristina, carissima amica, la quale riusciva sempre ad affrontare ogni situazione con il sorriso stampato sulle labbra.