

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE STATISTICHE

CORSO DI LAUREA TRIENNALE IN  
STATISTICA PER LE TECNOLOGIE E LE SCIENZE

**Clustering dei calciatori della  
Serie A 2022-2023 per stile di gioco**

*Relatore:*  
PROF. STEFANO MAZZUCO

*Laureando:*  
EDOARDO LOVATO  
2001319

Anno Accademico 2023/2024



*A mio papà Simone e a mia mamma Sandra,  
alle mie sorelle Clara e Camilla,  
ai miei nonni.*



## **Abstract**

Le tradizionali modalità di classificazione dei giocatori in base alle loro posizioni specifiche in campo (es. difensore centrale, terzino sinistro, centrocampista centrale, ala sinistra, ecc.) stanno perdendo di rilevanza nel delineare efficacemente il loro ruolo. Si sta assistendo, invece, a un crescente interesse verso l'effettiva funzione dei giocatori durante la partita.

Questo studio rappresenta un esempio di modifica del metodo di raggruppamento dei giocatori, basandosi sulle azioni durante il gioco per determinare la loro effettiva funzione all'interno della tattica di squadra. Attraverso i dati statistici della stagione 2022/2023, è stata condotta un'analisi di clustering (cluster analysis) per raggruppare i giocatori in categorie, basate sulla loro funzione in campo. L'analisi è stata condotta tramite model-based clustering, che fornisce a ciascun giocatore una probabilità di appartenenza a ciascuno dei cluster identificati. Questo metodo consente anche di individuare i giocatori ibridi, che possono mostrare appartenenze multiple a diversi cluster.

I risultati ottenuti possono trovare applicazione nello scouting e offrono un nuovo approccio basato sui dati, che fornisce uno strumento immediato per identificare le caratteristiche dei giocatori. Di conseguenza, può essere utilizzato in concomitanza con i metodi tradizionali per identificare il tipo di giocatore necessario all'interno della rosa di una squadra di calcio.



# Indice

<b>Introduzione</b>	<b>7</b>
<b>1 Preparazione dei Dati</b>	<b>11</b>
1.1 Costruzione del Dataset . . . . .	11
1.2 Descrizione delle Variabili . . . . .	12
1.3 Normalizzazione dei Dati . . . . .	15
1.4 Standardizzazione dei Dati . . . . .	17
1.5 Riduzione della Dimensionalità . . . . .	18
1.5.1 UMAP . . . . .	18
1.5.2 PCA . . . . .	19
1.5.3 Scelta di UMAP per la Riduzione della Dimensionalità	20
<b>2 Concetti e Metodi della Cluster Analysis</b>	<b>23</b>
2.1 Apprendimento Supervisionato . . . . .	23
2.2 Apprendimento Non Supervisionato . . . . .	23
2.3 Classificazione e Raggruppamento . . . . .	24
2.4 Model-based Clustering . . . . .	25
2.4.1 Modelli Mistura . . . . .	25
2.4.2 Decomposizione VSO . . . . .	26
2.4.3 Stima del Modello tramite Massima Verosimiglianza .	27
2.4.4 Scelta del Numero di Cluster e del Modello di Clustering	29
2.5 Confronto con il K-means . . . . .	30
<b>3 Clustering</b>	<b>33</b>
3.1 Clustering dei Giocatori . . . . .	33
3.1.1 Cluster 2 - Regista Arretrato . . . . .	37
3.1.2 Cluster 8 - Marcatore . . . . .	38
3.1.3 Cluster 9 - Laterale Creatore . . . . .	38
3.1.4 Cluster 1 - Regista . . . . .	39
3.1.5 Cluster 7 - Ruba Palloni . . . . .	40
3.1.6 Cluster 6 - Equilibratore . . . . .	40
3.1.7 Cluster 3 - Dribblatore . . . . .	41
3.1.8 Cluster 4 - Creatore di Occasioni . . . . .	41
3.1.9 Cluster 5 - Finalizzatore . . . . .	42
3.2 Giocatori Ibridi . . . . .	43
3.3 Composizione delle Squadre . . . . .	44
<b>4 Conclusioni</b>	<b>47</b>
<b>Bibliografia e Sitografia</b>	<b>49</b>



## Introduzione

L’evoluzione del calcio negli ultimi vent’anni ha seguito un ritmo straordinario. In ambito scouting, si è assistito a un passaggio di grande rilievo: da un’osservazione diretta sul campo a un’era in cui dati e statistiche sono diventati indispensabili. Questo trend, proveniente dagli Stati Uniti e dagli sport americani, ha influenzato anche il modo di lavorare nel calcio professionistico, trasformando gradualmente l’approccio tradizionale al lavoro.

Nel contesto del calcio moderno, l’analisi dei calciatori ha assunto un ruolo sempre più cruciale nell’ottimizzazione delle prestazioni delle squadre e nella formulazione di strategie di gioco vincenti. Mentre in passato la categorizzazione dei giocatori si basava principalmente sulle loro posizioni in campo, come terzino destro o attaccante centrale, oggi l’attenzione si sta spostando verso l’effettiva funzione dei giocatori durante il gioco. I moduli sono diventati più fluidi, con squadre che difendono con una tattica e attaccano con un’altra. Le interpretazioni tattiche nelle due fasi di gioco, insieme alle diverse caratteristiche tecniche e atletiche dei giocatori, determinano lo stile di gioco di una squadra. Questo cambiamento ha portato a un’evoluzione, con calciatori sempre più versatili e adattabili, in grado di esprimere al meglio le loro qualità in base ai compiti assegnati. Quando si valuta di acquistare sul mercato un nuovo giocatore, non si dovrebbe più individuare quale posizione ricopre (come terzino o centrocampista centrale), ma capire quali sono le sue funzioni principali. È fondamentale chiedersi se il giocatore sia un costruttore di gioco o sia più bravo ad incidere nella fase offensiva della squadra attraverso i suoi inserimenti. Le abilità del giocatore dovrebbero esaltarsi rispetto ai compiti e alle funzioni della squadra. I numeri e le statistiche svolgono un ruolo cruciale nell’identificare le caratteristiche tecniche e atletiche di ciascun giocatore, permettendo una valutazione più approfondita e accurata.

L’adozione della tecnologia e dell’analisi dei dati da parte dei top club calcistici sta rapidamente accelerando in molti settori chiave, compreso quello dello scouting. Professionisti come i match e data analyst, infatti, sono diventati figure indispensabili nell’ambito della costruzione di squadre competitive, fornendo un supporto prezioso nella valutazione e interpretazione dei dati. Questo cambio di prospettiva, alimentato anche dalla crescente importanza della statistica nello sport, offre agli allenatori e agli osservatori uno strumento per valutare le prestazioni dei calciatori e prendere decisioni più consapevoli durante il mercato. L’osservazione diretta e la conoscenza approfondita del gioco sono ancora essenziali per valutare i calciatori. L’utilizzo dei dati ha avuto un impatto importante sulle strategie e sui metodi di lavoro nel calcio professionistico, rendendo indispensabile il supporto numerico e

statistico. Questo approccio permette un'analisi dettagliata delle prestazioni dei giocatori e una migliore comprensione delle dinamiche di gioco. L'obiettivo principale di questa ricerca è quello di identificare i diversi stili di gioco dei calciatori della Serie A 2022/23 attraverso l'applicazione di tecniche di clustering avanzate.

Il clustering rappresenta un'importante tecnica analitica nel mondo dello sport, in particolare nel contesto del calcio. Questo approccio consente di identificare pattern e strutture nascoste nei dati relativi alle prestazioni dei giocatori, ai modelli di gioco delle squadre e alle dinamiche di squadra. Nel calcio moderno, caratterizzato da un'ampia varietà di ruoli e stili di gioco, il clustering offre un metodo efficace per categorizzare i giocatori in base alle loro caratteristiche e funzioni specifiche. Questo può essere particolarmente utile per gli allenatori e gli osservatori nel processo decisionale, aiutandoli a comprendere meglio le esigenze della squadra e a identificare i giocatori che meglio si adattano al sistema di gioco. In questo contesto, l'analisi dei cluster diventa uno strumento fondamentale per ottimizzare le prestazioni delle squadre e massimizzare il loro potenziale sul campo. Questo approccio permette di andare oltre la semplice classificazione basata sulle posizioni in campo e di individuare le caratteristiche distintive dei diversi giocatori in base al loro contributo effettivo durante le partite. In questo modo, non solo si fornisce agli allenatori e agli osservatori uno strumento più sofisticato per valutare le prestazioni dei giocatori, ma si contribuisce anche alla comprensione più approfondita del calcio moderno e delle sue dinamiche.

La metodologia utilizzata in questa ricerca prevede l'applicazione del model-based clustering ai dati delle prestazioni dei calciatori della Serie A 2022/23. Si considera un dataset contenente un'ampia gamma di variabili, che comprendono statistiche rappresentative delle diverse fasi di gioco presenti in una partita, spaziando dalla fase difensiva a quella offensiva (come contrasti, passaggi, tiri, dribbling, etc). Attraverso un'analisi dettagliata di questi dati, si sarà in grado di identificare i diversi stili di gioco dei calciatori raggruppandoli in cluster in base a ciò che effettivamente producono in campo. È altrettanto utile considerare la categoria degli "ibridi". Questi sono giocatori che non si limitano ad appartenere ad un solo gruppo, ma sono atleti versatili in grado di svolgere più compiti in campo e che, di conseguenza, appartengono a diversi cluster anche a livello statistico.

La ricerca mira, dunque, a presentare un approccio moderno per raggruppare i calciatori in base al loro stile di gioco. I risultati che emergono possono fungere da supporto nelle decisioni tattiche e nello scouting delle squadre.

Nel corso della ricerca, il libro [1] *The Clustering Project* di A. Gagliardi e Soccerment, ha rappresentato un valido spunto per esplorare la possibilità di applicare il clustering ai calciatori, fornendo un supporto nell’orientare l’approccio metodologico.

- **Nella prima sezione**, si presenterà tutta la preparazione dei dati necessaria per prepararli al clustering, descrivendo le tecniche di pre-processing, come la pulizia dei dati e la gestione dei dati mancanti. Verranno trattate, inoltre, la normalizzazione, la standardizzazione e la riduzione della dimensionalità.
- **Nella seconda sezione**, si introdurranno i concetti teorici e i metodi della *cluster analysis*, con particolare attenzione al *model-based clustering*, alla selezione del modello e al numero di cluster.
- **Nella terza sezione**, si affronterà l’adattamento dei dati al *model-based clustering*, descrivendo il processo di applicazione del modello e i cluster risultanti dall’analisi. Inoltre, verranno approfonditi il profilo del giocatore ibrido e il tema della composizione delle squadre.
- **Nella conclusione**, si discuteranno gli esiti della ricerca, suggerendo possibili direzioni future per migliorare o ampliare i risultati ottenuti.



# 1 Preparazione dei Dati

La fase di pre-processing (preparazione e pulizia dei dati) è essenziale in preparazione alla cluster analysis (come per tutte le altre tipologie di analisi dei dati). Si procede attraverso diverse fasi per garantire che i dati siano affidabili, coerenti e adatti all'analisi successiva. Questo processo comprende la costruzione del dataset, che include la pulizia e la pre-elaborazione dei dati per garantire la coerenza e l'integrità delle informazioni. Successivamente, si procede con la normalizzazione e la standardizzazione dei dati, per uniformare le scale delle diverse variabili e renderle confrontabili tra loro. Infine, la riduzione della dimensionalità è un passo cruciale per semplificare la complessità del dataset, in prospettiva del passo successivo del processo. In questa sezione, si esploreranno ciascuna di queste fasi, illustrando l'importanza di una preparazione accurata dei dati in vista dell'analisi di clustering.

## 1.1 Costruzione del Dataset

Nell'ambito del presente studio, l'obiettivo primario consiste nella realizzazione di una caratterizzazione stilistica dei giocatori basata sulle loro azioni in campo. È stata effettuata una selezione di 24 variabili, ritenute idonee a evidenziare il contributo individuale dei giocatori nelle varie fasi di gioco, quali difesa, costruzione, rifinitura e finalizzazione (si veda Sezione 1.2). Nel processo di costruzione del dataset, le 24 variabili pertinenti sono state acquisite da fonti online, utilizzando i siti web [2]FBref e [3]WhoScored.

La volontà dell'analisi è quella di mettere in evidenza le diverse tipologie di stili e funzioni di gioco, senza dare importanza al livello di performance individuale. Pur essendo consapevoli che le statistiche selezionate possano fungere da indicatori di performance, si è scelto di utilizzare i conteggi totali delle azioni, anziché i loro risultati o le percentuali di successo. La scelta di non utilizzare misure di performance è stata fatta per evitare di distinguere i giocatori in base alla qualità e concentrarsi, invece, sullo stile di gioco.

Il dataset comprende le statistiche per tutti i giocatori con almeno 450 minuti giocati nella stagione 2022/2023 nel campionato di Serie A, per un totale di 406 giocatori. Questa analisi non contempla i portieri, data la funzione specifica del loro ruolo in campo. Nonostante le molteplici interpretazioni possibili del ruolo, il numero limitato di giocatori non consente di condurre un'analisi di clustering con la stessa metodologia applicata ai giocatori di movimento. Inoltre, nel dataset sono presenti giocatori che risultano appartenere a due squadre diverse a causa di trasferimenti avvenuti nel mercato di gennaio. Tali giocatori sono considerati nell'analisi come due entità distinte. Questo approccio consente di valutare, al termine del clustering, se

entrambe le versioni del giocatore si trovano nello stesso cluster. In alternativa, è possibile determinare se il cambiamento dello stile di gioco, dovuto alla diversa strategia della nuova squadra e del nuovo allenatore, ha influenzato il loro stile, portandoli a essere raggruppati in due cluster differenti.

## 1.2 Descrizione delle Variabili

Per comprendere appieno le abilità di un calciatore, è essenziale utilizzare una gamma di variabili che coprano diversi aspetti del gioco. Le variabili analizzate sono state selezionate per rappresentare in modo chiaro e completo le principali fasi di gioco: difesa, costruzione, rifinitura e finalizzazione.

Ogni statistica riflette un'abilità specifica o una dimensione del gioco che, combinata alle altre, aiuta a delineare le caratteristiche individuali e il ruolo del giocatore. Per esempio, le statistiche difensive rivelano l'efficacia nell'ostacolare le azioni avversarie, mentre le metriche di costruzione e rifinitura mostrano il contributo nella gestione del possesso palla e nella creazione di opportunità offensive. La fase di finalizzazione si concentra sulle capacità del calciatore nel concludere le azioni offensive, fornendo dati su come e dove cerca di effettuare i tiri, se da dentro o fuori area, e con quali modalità, come il tiro di piede o di testa. Queste statistiche rivelano lo stile di finalizzazione del giocatore e il suo livello di incisività nelle vicinanze della porta avversaria, aspetto cruciale per la valutazione del potenziale realizzativo di un attaccante o di un centrocampista avanzato. In questo senso, ciascuna variabile è fondamentale per delineare i tratti distintivi e il contributo complessivo di ogni giocatore alla squadra.

### Difesa

- **Intercetti:** numero di situazioni in cui il giocatore interrompe un passaggio avversario, anticipando l'azione. Misura l'abilità nel leggere il gioco e prevenire avanzamenti.
- **Contrasti:** numero di azioni in cui il giocatore ingaggia fisicamente un avversario per sottrargli il pallone. Indica la capacità difensiva e la propensione al contatto fisico.
- **Contrasti fuori dal terzo difensivo:** numero di contrasti eseguiti in zone più avanzate del campo, dimostrando una strategia difensiva più aggressiva o orientata alla pressione.
- **Tiri bloccati:** numero di tiri avversari fermati dal giocatore con il corpo o i piedi, evidenziando la capacità di proteggere la propria porta.

- **Spazzate:** numero di azioni in cui il giocatore libera l'area difensiva con un calcio lungo, solitamente in situazioni di emergenza.
- **Duelli aerei:** numero di volte in cui il giocatore compete con un avversario per il pallone in aria. Il duello aereo è fondamentale per vincere la palla su passaggi lunghi.
- **Falli commessi:** numero di irregolarità commesse dal giocatore. Un indicatore della sua intensità o aggressività difensiva.
- **Cartellini gialli:** numero di ammonizioni ricevute. Possono riflettere lo stile di gioco e la gestione del rischio nel difendere.

## Costruzione

- **Passaggi brevi:** numero di passaggi brevi tentati, tra 4 e 13 iarde<sup>1</sup> (circa 3.7-12 metri), che permettono di mantenere il possesso palla. Usati soprattutto per costruire il gioco in modo sicuro.
- **Scambi in ampiezza:** numero di passaggi lunghi, oltre 35 iarde (circa 32 metri), che spostano il gioco sull'ampiezza del campo. Utili per cambiare rapidamente il lato dell'attacco e disorientare la difesa avversaria.
- **Passaggi nel terzo offensivo:** numero di passaggi completati entrati nell'ultimo terzo di campo più vicino alla porta avversaria. Non sono conteggiate le azioni derivanti da calci piazzati, come calci d'angolo o punizioni, in cui la palla è ferma e la squadra può organizzare uno schema prestabilito.
- **Passaggi progressivi:** numero di passaggi effettuati da un giocatore che avanzano il pallone verso la porta avversaria di almeno 10 metri rispetto al punto più lontano in cui si trovava nei precedenti sei passaggi della squadra. Questo indicatore include qualsiasi passaggio che porta il pallone nell'area di rigore avversaria ed esclude i passaggi che finiscono nella metà campo della propria difesa. Misura la capacità di un giocatore di fare avanzare l'azione e avvicinare la squadra alla zona di tiro, spesso aumentando la pressione sull'avversario o preparando opportunità di attacco.
- **Palle lunghe:** numero di passaggi tentati superiori a 27 iarde (circa 25 metri), usati per guadagnare rapidamente campo o superare la difesa avversaria.

---

<sup>1</sup>1 iarda corrisponde a 0.9144 metri

- **Cross:** numero di passaggi tentati da ampie posizioni laterali che solitamente viaggiano a mezz'aria, mirati a servire i compagni nei pressi dell'area di rigore.
- **Conduzioni progressive:** distanza complessiva, espressa in iarde, percorsa da un giocatore mentre controlla il pallone con i piedi in direzione della porta avversaria. Questa variabile misura la capacità di avanzare nel campo mantenendo il possesso, portando la squadra in posizione più offensiva e spesso superando uno o più avversari nella progressione.

## Rifinitura / Creazione di Occasioni

- **Dribbling:** numero di azioni di dribbling tentate, che misura l'abilità di superare l'avversario uno contro uno.
- **Tocchi in area avversaria:** numero di tocchi di palla dentro l'area di rigore avversaria. Indica quanto il giocatore sia presente nelle zone pericolose.
- **Passaggi filtranti:** numero di passaggi completati che attraversano la difesa avversaria trovando un compagno in una posizione vantaggiosa.
- **Passaggi completati in area di rigore:** numero di passaggi effettuati e completati dentro l'area avversaria. Sono fondamentali per creare occasioni di tiro ravvicinato. Non sono conteggiate le azioni derivanti da calci piazzati, come calci d'angolo o punizioni, in cui la palla è ferma e la squadra può organizzare uno schema prestabilito.
- **Cross in area di rigore:** numero di cross eseguiti all'interno dell'area di rigore. Spesso sono usati per servire attaccanti già posizionati. Non sono conteggiate le azioni derivanti da calci piazzati, come calci d'angolo o punizioni, in cui la palla è ferma e la squadra può organizzare uno schema prestabilito.

## Finalizzazione

- **Tiri da fuori area:** numero di conclusioni tentate da fuori dell'area di rigore avversaria. Rappresenta un indicatore di abilità tecnica e di capacità di concludere anche da lontano.

- **Tiri da dentro area:** numero di conclusioni tentate da dentro l'area di rigore. Corrisponde a un indicatore della capacità di posizionarsi in zone pericolose e sfruttare le occasioni ravvicinate.
- **Tiri di piede:** numero di conclusioni eseguite calciando il pallone con il piede. È una misura della preferenza o abilità di concludere con precisione.
- **Tiri di testa:** numero di conclusioni tentate colpendo il pallone di testa. Riflette sia l'abilità nei duelli aerei sia la capacità di finalizzare i cross in area.

Considerare molteplici statistiche ci permette di cogliere la complessità del gioco moderno e di valutare le abilità dei calciatori in modo più accurato. Anche se alcune variabili possono sembrare simili, ognuna di esse porta un contributo unico che arricchisce la valutazione complessiva. Ad esempio, un difensore che si distingue per i contrasti mostra uno stile di gioco più fisico e diretto, mentre un difensore con alte capacità di intercetto è probabilmente abile nella lettura anticipata delle azioni avversarie. In modo simile, un centrocampista che eccelle nei passaggi progressivi aggiunge dinamismo al gioco, mentre uno che opta per passaggi laterali contribuisce al controllo del possesso. Combinare queste informazioni ci permette di identificare con maggiore precisione i ruoli, le strategie e le attitudini individuali, dipingendo un quadro esaustivo delle capacità di ogni atleta.

### 1.3 Normalizzazione dei Dati

I dati devono essere normalizzati per tenere conto della diversa quantità di minuti giocati dai calciatori. Come viene presentato nel libro [1] *The Clustering Project* di A. Gagliardi e Soccernet, il metodo più comune per questo scopo è la normalizzazione "per novanta minuti", che consiste nel dividere la statistica di interesse per "minuti giocati / 90". In questo modo, le statistiche vengono convertite in una scala uniforme, rappresentando il numero di volte che il giocatore compie una determinata azione ogni 90 minuti. Questo approccio presenta alcune criticità, tra cui l'equivalenza implicita tra minuti giocati e opportunità di compiere le varie azioni di gioco. Tale equivalenza non è necessariamente reale, poiché dipende dal grado di controllo del gioco esercitato dalla squadra. I giocatori appartenenti a squadre che tendono a dominare la partita effettueranno generalmente più azioni di costruzione, come passaggi, e meno azioni difensive, mentre il contrario avverrà per i giocatori di squadre che concedono il controllo del gioco agli avversari. Per tenere conto di questo aspetto, viene utilizzata una normalizzazione diversa

per le statistiche difensive e per quelle più semplici relative ai passaggi, concentrandosi non sui minuti giocati, ma sul totale dei tocchi di palla. Il valore della statistica viene diviso per "tocchi / 100", indicando così quante volte il giocatore compie una determinata azione ogni cento tocchi di palla. Nella Tabella 1 (Table 1), si può osservare il tipo di normalizzazione scelto per ogni variabile del dataset. Questa normalizzazione facilita l'interpretazione dei dati poiché mette sullo stesso piano tutti i calciatori, permettendo un confronto equo tra loro. La normalizzazione per novanta minuti (P90) e quella per tocchi di palla (P100) assicurano che le statistiche siano comparabili indipendentemente dal tempo di gioco o dal numero di tocchi. Senza normalizzazione, infatti, le distorsioni nei dati potrebbero emergere a causa delle differenze nelle scale delle variabili. Ad esempio, le statistiche di giocatori con più minuti di gioco potrebbero risultare molto più alte rispetto a quelle dei giocatori con meno minuti, anche se entrambi hanno prestazioni simili. Questo potrebbe portare a una percezione inaccurata delle capacità dei calciatori e delle loro contribuzioni al gioco. Inoltre, senza normalizzazione, le variabili con unità di misura diverse potrebbero influenzare in modo sproporzionato l'analisi, enfatizzando alcuni aspetti rispetto ad altri. La normalizzazione aiuta quindi a garantire che le variabili siano confrontabili su una scala comune (in questo caso per 90 minuti o per 100 tocchi), riducendo al minimo il rischio di distorsioni nei risultati dell'analisi.

Table 1: Tipo di normalizzazione per ognuna delle 24 variabili.

P90: normalizzazione per 90 minuti. P100: normalizzazione per 100 tocchi.

Fase di gioco	Statistica	Normalizzazione
Difesa	Intercetti	P100
	Contrasti	P100
	Contrasti fuori dal terzo difensivo	P100
	Tiri bloccati	P100
	Spazzate	P100
	Dueli aerei	P100
	Falli commessi	P100
	Cartellini gialli	P90
	Passaggi brevi	P90
Costruzione	Scambi ampiezza	P90
	Passaggi nel terzo offensivo	P100
	Passaggi progressivi	P100
	Palle lunghe	P90
	Cross	P90
Rifinitura / Creazione di Occasioni	Conduzioni progressive	P90
	Dribbling	P90
	Tocchi in area avversaria	P90
	Passaggi filtranti	P90
	Passaggi completati in area di rigore	P90
Finalizzazione	Cross in area di rigore	P90
	Tiri da fuori area	P90
	Tiri da dentro area	P90
	Tiri di piede	P90
	Tiri di testa	P90

## 1.4 Standardizzazione dei Dati

Il dataset in esame richiede un’adeguata preparazione prima di poter applicare un algoritmo di clustering. Un ulteriore passo consiste nella standardizzazione dei dati, per uniformare ulteriormente la scala delle varie statistiche, che potrebbero presentare valori numerici di gran lunga diversi a causa delle disparità tra le diverse misurazioni [4](Scikit-Learn) [5](James et al., 2013) [6](Towards Data Science, 2021). Ad esempio, il valore P90 dei passaggi brevi potrebbe essere considerevolmente superiore al valore P90 dei passaggi filtranti. In questa analisi è stato impiegato il metodo dello standard scaling, che standardizza i dati affinché abbiano una media di zero e una deviazione standard di uno. La standardizzazione dei dati prepara il

dataset alla successiva trasformazione, ovvero la riduzione dimensionale attraverso la metodologia UMAP. Senza questa standardizzazione, la riduzione delle dimensioni potrebbe essere distorta, poiché UMAP potrebbe considerare alcune variabili come più influenti di altre semplicemente a causa dei loro valori numerici più elevati. Standardizzando i dati, tutte le variabili vengono portate su una scala comune, garantendo che nessuna variabile domini il processo di riduzione dimensionale.

## 1.5 Riduzione della Dimensionalità

Per ottenere risultati di rilievo e più facilmente interpretabili dagli algoritmi di clustering, è spesso necessario ridurre la dimensionalità del dataset. Questi algoritmi, infatti, si basano sulla distanza tra i punti. Un numero elevato di variabili implica che queste distanze vengano calcolate in uno spazio ad alta dimensionalità (nel caso specifico, 24 dimensioni), rendendo meno probabile che due punti risultino vicini. Questo fenomeno, che è uno degli aspetti principali del "curse of dimensionality", può compromettere l'efficacia di molti algoritmi di analisi. La riduzione della dimensionalità è un'importante tecnica di analisi dei dati, utilizzata per semplificare i dataset complessi mantenendo al contempo la struttura essenziale delle informazioni. È cruciale non solo per la visualizzazione dei dati, ma anche per migliorare l'efficienza degli algoritmi di machine learning. Tra le varie tecniche di riduzione della dimensionalità, due degli approcci più noti sono l'analisi delle componenti principali (PCA) e l'Uniform Manifold Approximation and Projection (UMAP).

### 1.5.1 UMAP

UMAP (Uniform Manifold Approximation and Projection) è una metodologia innovativa per la riduzione della dimensionalità, sviluppata da [7] McInnes et al. (2020), che si basa su un framework teorico radicato nella geometria riemanniana e nella topologia algebrica. L'obiettivo principale di UMAP è preservare la struttura globale e locale dei dati ad alta dimensione durante la proiezione in uno spazio di dimensione ridotta [8] (Hastie et al. 2009) [9] (UMAP Documentation).

Funziona in due fasi principali:

1. **Costruzione del Grafo di Vicinato:** UMAP costruisce un grafo di vicinato utilizzando una misura di similarità basata sulla distanza tra i punti. Le distanze vengono calcolate attraverso una funzione di densità, come il kernel gaussiano. I punti vengono quindi collegati in base alla loro similarità, creando un grafo che rappresenta le relazioni locali nel dataset originale.

2. **Proiezione in Spazio Ridotto:** La metodologia cerca di trovare un'incorporazione in uno spazio a dimensione ridotta che minimizza la differenza tra le distribuzioni delle distanze nel grafo originale e quelle nell'incorporamento ridotto. Questo obiettivo di ottimizzazione viene formalizzato come una funzione di perdita, che quantifica la differenza tra le distribuzioni delle distanze nel grafo originale e quelle nell'incorporamento ridotto.

UMAP si dimostra particolarmente efficace per la visualizzazione dei dati, poiché mantiene le relazioni rilevanti tra i punti, facilitando l'identificazione di cluster e strutture all'interno dei dati.

### 1.5.2 PCA

L'analisi delle componenti principali (PCA) è una delle tecniche più tradizionali per la riduzione della dimensionalità. L'obiettivo principale della PCA è trasformare un dataset originale in un nuovo sistema di coordinate, chiamato componenti principali, in cui ogni componente è una combinazione lineare delle variabili originali [10](Jolliffe, 2002) [11](Scikit-Learn PCA Documentation) [12](Bishop, 2006).

La PCA opera attraverso i seguenti passaggi:

1. **Standardizzazione dei Dati:** Prima di applicare la PCA, è importante standardizzare i dati, in modo che ogni variabile contribuisca equamente all'analisi. Questo è particolarmente rilevante quando le variabili hanno scale diverse.
2. **Calcolo della Matrice di Covarianza:** La matrice di covarianza viene calcolata per determinare le relazioni tra le variabili del dataset.
3. **Calcolo degli Autovalori e degli Autovettori:** Gli autovalori e gli autovettori della matrice di covarianza vengono calcolati per identificare le direzioni di massima varianza.
4. **Selezione delle Componenti Principali:** Gli autovettori corrispondenti agli autovalori più elevati vengono selezionati per formare le nuove componenti principali. Queste componenti sono ordinate in base alla quantità di varianza che rappresentano.

A differenza di UMAP, la PCA è una tecnica lineare e, di conseguenza, può non essere in grado di catturare relazioni non lineari nei dati. Tuttavia, è spesso più veloce e computazionalmente meno costosa, rendendola una scelta popolare per dataset di grandi dimensioni.

### 1.5.3 Scelta di UMAP per la Riduzione della Dimensionalità

In sintesi, sia UMAP che PCA sono tecniche preziose per la riduzione della dimensionalità, ognuna con i propri punti di forza e limitazioni. La scelta tra i due metodi dipende spesso dalle specifiche esigenze del dataset e dall'obiettivo dell'analisi.

In questo studio, è stata scelta la metodologia UMAP poiché conserva meglio la struttura locale e globale dei dati, fondamentale per ottenere risultati rilevanti nel clustering. A differenza della PCA, che è una tecnica lineare e punta a massimizzare la varianza spiegata nelle componenti principali, UMAP è progettato per preservare le relazioni locali tra i punti, mantenendo in modo accurato la vicinanza tra elementi simili. Inoltre, UMAP permette di scegliere la metrica di distanza e bilanciare la conservazione della struttura locale e globale tramite parametri regolabili, rendendolo più adatto per analisi in cui la struttura dei dati gioca un ruolo cruciale. UMAP opera creando una rappresentazione grafica dei dati, mantenendo la vicinanza tra i punti più simili e rappresentando accuratamente le distanze relative tra di essi.

Questa riduzione della dimensionalità è essenziale per preparare i dati al processo di clustering, poiché facilita la visualizzazione e l'interpretazione dei gruppi formati. È stata esaminata visivamente la rappresentazione bidimensionale dei dati attraverso il diagramma di dispersione delle coordinate ridotte UMAP (Figura 1), per verificare se i giocatori erano disposti nello spazio in modo coerente con il contesto calcistico. Questa valutazione è stata effettuata basandosi sull'esperienza e sulla conoscenza del dominio. Pertanto, l'uso di UMAP non solo migliora la comprensione dei dati, ma anche l'efficacia degli algoritmi di clustering applicati successivamente.

L'intuizione iniziale per la scelta del numero di cluster può essere suggerita dall'osservazione visiva del diagramma di dispersione, il quale mostra una prima rappresentazione della distribuzione dei dati nello spazio ridotto. Tuttavia, il corretto numero di cluster, il quale servirà nell'analisi di clustering, verrà determinato con maggiore precisione utilizzando il metodo descritto nella Sezione 2.4.4.

## Diagramma di Dispersione

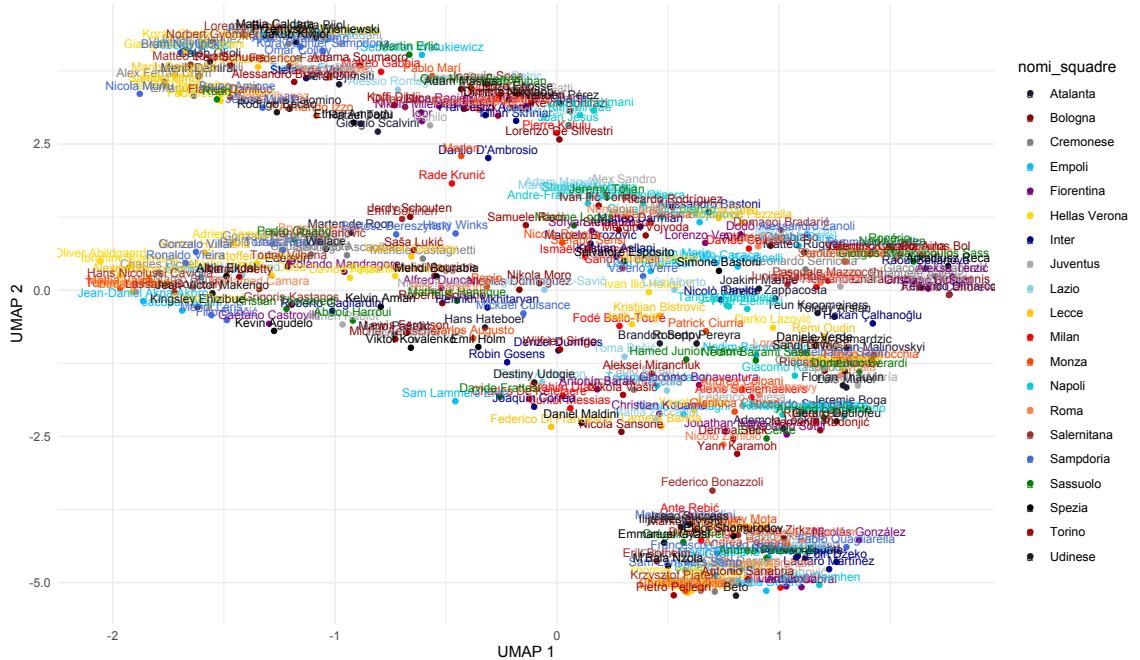


Figure 1: Diagramma di dispersione delle coordinate ridotte in due dimensioni tramite il metodo UMAP, con le etichette dei giocatori colorate in base alla squadra di appartenenza.



## 2 Concetti e Metodi della Cluster Analysis

L’obiettivo della Cluster Analysis è trovare gruppi rilevanti nei dati. Tipicamente, questi gruppi saranno internamente coerenti e separati l’uno dall’altro. Lo scopo è identificare gruppi i cui membri hanno qualcosa in comune che non condividono con i membri di altri gruppi.

In questa sezione, si discuteranno le metodologie di clustering e gli approcci statistici per fornire una base teorica al model-based clustering. Verranno esplorati i concetti di apprendimento supervisionato e non supervisionato, con un focus sulle differenze tra classificazione e raggruppamento. Inoltre, verrà presentato il model-based clustering, seguito da un confronto con il K-means per mettere in evidenza i vantaggi del model-based clustering. Quest’ultimo approccio di clustering sarà poi applicato ai dati dei giocatori della Serie A 2022/2023, dopo la fase di preparazione dei dati descritta nella sezione precedente.

### 2.1 Apprendimento Supervisionato

Nell’ambito dell’apprendimento automatico, uno dei principali paradigmi è l’apprendimento supervisionato [13](Murphy, 2012) [14](Scikit-learn Documentation). Questo approccio si basa su un insieme di dati in cui ogni esempio è associato a un’etichetta o a un valore obiettivo. L’obiettivo dell’algoritmo è apprendere una funzione o un modello che, dato un nuovo input, sia in grado di predire correttamente l’etichetta associata. I modelli supervisionati sono comunemente utilizzati in contesti come la classificazione e la regressione. Nel caso della classificazione, l’algoritmo assegna ogni istanza a una classe discreta (ad esempio, “spam” o “non spam” nel caso di filtri di posta elettronica). Nella regressione, invece, il modello predice un valore continuo (ad esempio, il prezzo di una casa basato sulle sue caratteristiche). La principale forza dell’apprendimento supervisionato risiede nella possibilità di valutare le performance del modello tramite metriche standardizzate, come l’accuratezza o l’errore quadratrico medio, utilizzando un insieme di dati di test etichettato. Un esempio comune di algoritmo supervisionato è la regressione logistica, spesso utilizzata per problemi di classificazione binaria. Altri algoritmi supervisionati includono le reti neurali, le macchine a vettori di supporto (SVM) e gli alberi decisionali.

### 2.2 Apprendimento Non Supervisionato

Al contrario, l’apprendimento non supervisionato si occupa di dati che non hanno etichette o valori obiettivo. In questo contesto, l’obiettivo dell’algoritmo

è individuare pattern nascosti nei dati o suddividere gli esempi in gruppi basati su caratteristiche comuni. Tra i principali esempi di apprendimento non supervisionato vi è il clustering, che mira a suddividere i dati in insiemi o "cluster", in modo tale che gli esempi all'interno di un cluster siano più simili tra loro rispetto agli esempi appartenenti ad altri cluster. Un esempio classico di apprendimento non supervisionato è il K-means clustering, un algoritmo che suddivide i dati in  $K$  gruppi, basandosi sulla minimizzazione della distanza tra gli esempi e i centroidi di ciascun gruppo. Questo approccio permette di individuare strutture nei dati senza la necessità di una supervisione esterna. Altri algoritmi non supervisionati includono le tecniche di riduzione dimensionale, come l'Analisi delle Componenti Principali (PCA), che riducono il numero di variabili mantenendo al contempo la variabilità più rilevante dei dati.

## 2.3 Classificazione e Raggruppamento

La distinzione tra classificazione e raggruppamento risiede nel fatto che la classificazione appartiene all'apprendimento supervisionato, mentre il raggruppamento è un processo di apprendimento non supervisionato.

La classificazione assegna ogni esempio di un dataset a una delle categorie predefinite. In un problema di classificazione binaria, ad esempio, l'algoritmo è addestrato su dati etichettati e tenta di predire quale tra le due classi assegnare a un nuovo esempio in base alla funzione appresa. La qualità del modello di classificazione può essere valutata tramite misure come l'accuratezza, la precisione, il richiamo e l'F1 score. La classificazione è utilizzata in molte applicazioni reali, come il riconoscimento di immagini, il rilevamento di spam e i sistemi di raccomandazione.

Il raggruppamento, o clustering, è un processo nel quale i dati vengono suddivisi in gruppi basati sulla loro somiglianza senza la necessità di etichette predefinite. La mancanza di supervisione implica che l'algoritmo deve identificare da solo i gruppi rilevanti nei dati. Gli esempi all'interno dello stesso cluster sono considerati simili tra loro in base a una misura di distanza o di somiglianza, mentre quelli appartenenti a cluster diversi sono considerati diversi. L'obiettivo del clustering è massimizzare la similarità interna ai cluster e minimizzare la similarità tra cluster differenti. Un esempio molto comune è l'uso del clustering nei dati di mercato, dove i clienti vengono raggruppati in base ai loro comportamenti di acquisto. Questa segmentazione consente alle aziende di sviluppare strategie di marketing mirate per ciascun gruppo. Nel contesto dell'analisi dello stile di gioco dei calciatori della Serie A 2022/2023, l'uso del clustering permette di superare la tradizionale classificazione dei giocatori in base alle loro posizioni fisse (difensore, centrocampista, attac-

cante) e di raggrupparli in funzione delle loro azioni specifiche sul campo, come il numero di intercetti, passaggi filtranti, dribbling e tiri.

In sintesi, mentre la classificazione si basa su dati etichettati e mira ad assegnare un’etichetta specifica a ciascun esempio, il clustering identifica strutture nascoste all’interno dei dati senza alcun tipo di supervisione. Pertanto, la classificazione richiede un set di dati etichettato, mentre il clustering può operare su dati privi di etichette, rendendolo una tecnica estremamente versatile quando non si dispone di informazioni pregresse su come i dati dovrebbero essere strutturati.

## 2.4 Model-based Clustering

Nella presente sezione si esaminano le idee fondamentali del model-based clustering. Questo è un approccio strutturato alla cluster analysis, basato su un modello probabilistico e che utilizza i metodi standard dell’inferenza statistica. Il modello probabilistico su cui si basa è una combinazione finita di distribuzioni multivariate, descritto nella Sezione 2.4.1, con particolare enfasi sulla famiglia di modelli più utilizzata, ovvero le combinazioni di distribuzioni normali multivariate.

Nella Sezione 2.4.3 si descrive la stima della massima verosimiglianza per questo modello, concentrandosi sull’algoritmo Expectation-Maximization (EM). Questo algoritmo è garantito per convergere verso un ottimo locale della funzione di verosimiglianza, ma non necessariamente verso un massimo globale; di conseguenza, la scelta del punto di partenza può essere cruciale.

Due domande ricorrenti nell’uso pratico dell’analisi dei cluster sono: quanti cluster ci sono? e quale metodo di clustering dovrebbe essere usato? Si scopre che determinare il numero di cluster può essere considerato un problema di scelta del modello, e la scelta del metodo di clustering è spesso collegata, almeno approssimativamente, alla scelta del modello probabilistico. Queste due domande trovano quindi risposta simultaneamente nel contesto del model-based clustering attraverso la scelta di un modello probabilistico appropriato.

La teoria riportata in questo capitolo si basa su [15] *Model-Based Clustering and Classification for Data Science* di Bouveyron et al. (2019). Per ulteriori approfondimenti dei contenuti trattati, si invita alla consultazione di questo testo.

### 2.4.1 Modelli Mistura

Si consideri un dataset con  $n$  osservazioni  $y_1, \dots, y_n$  in  $d$  variabili, tali che  $y_i = (y_{i,1}, \dots, y_{i,d})$  per  $i = 1, \dots, n$ . Nei modelli di mistura, la funzione

di densità o la funzione di probabilità di un'osservazione  $y_i$  è definita dalla media pesata di  $G$  funzioni di densità, chiamate componenti della mistura. Nel modello di mistura gaussiano, in particolare, tali funzioni di densità provengono tutte da distribuzioni normali:

$$p(y_i) = \sum_{g=1}^G \tau_g \phi_g(y_i | \theta_g), \quad (2.1)$$

dove  $\tau_g$  indica la probabilità che l'osservazione sia generata dalla  $g$ -esima componente, con  $\tau_g \geq 0$  per  $g = 1, \dots, G$  e  $\sum_{g=1}^G \tau_g = 1$ , mentre  $\phi_g(\cdot | \theta_g)$  indica la funzione di densità della  $g$ -esima componente, dati i suoi parametri  $\theta_g$ . Poiché le componenti della mistura hanno distribuzione normale, si ha  $\theta_g = (\mu_g, \Sigma_g)$ , dove  $\mu_g$  indica il vettore  $d$ -dimensionale delle medie della  $g$ -esima componente e  $\Sigma_g$  la corrispondente matrice  $d \times d$  di varianza-covarianza.

#### 2.4.2 Decomposizione VSO

Un problema che emerge dall'utilizzo del modello di mistura normale è tuttavia rappresentato dal numero di parametri da utilizzare. Si hanno infatti  $G - 1$  parametri per gli elementi  $\tau_g$ ,  $Gd$  parametri per i vettori delle medie e, infine,  $\frac{Gd(d+1)}{2}$  parametri per le matrici di varianza-covarianza. Al crescere del numero di componenti della mistura e del numero di variabili, il modello diventa dunque poco parsimonioso, creando possibili problemi di stima.

Una soluzione a questa problematica è l'utilizzo della decomposizione VSO (Volume-Shape-Orientation) delle matrici di varianza-covarianza delle componenti della mistura:

$$\Sigma_g = \lambda_g D_g A_g D_g^T.$$

- $D_g$  è la matrice degli autovettori di  $\Sigma_g$ . Essa determina l'orientamento della  $g$ -esima componente.
- $A_g$  è la matrice diagonale con valori proporzionali agli autovalori di  $\Sigma_g$  in ordine decrescente. Essa determina la forma della  $g$ -esima componente.
- $\lambda_g$  è la costante di proporzionalità associata. Essa determina il volume della  $g$ -esima componente.

Il vantaggio dell'utilizzo di questa decomposizione sta nel fatto che, imponendo alcuni vincoli sugli elementi di quest'ultima, nel caso multivariato è possibile definire 14 tipi di modelli più o meno parsimoniosi. Ciascuno di

questi modelli è indicato con un identificativo di tre lettere: la prima rappresenta il volume dei cluster, la seconda la forma e la terza l'orientamento. Nello specifico:

- Se la prima lettera è  $E$ , allora  $\lambda_g = \lambda \forall g = 1, \dots, G$ . Il volume è dunque costante in tutte le componenti. In caso contrario si utilizza la lettera  $V$ .
- Se la seconda lettera è  $E$ , allora  $A_g = A \forall g = 1, \dots, G$ . La forma è dunque la stessa per tutte le componenti. Se, oltre a ciò,  $A = I$ , con  $I$  intesa come matrice di identità, la forma delle componenti risulta sferica e si utilizza la lettera  $I$ . Negli altri casi si utilizza la lettera  $V$ .
- Se la terza lettera è  $E$ , allora  $D_g = D \forall g = 1, \dots, G$ . L'orientamento è dunque lo stesso per tutte le componenti. Se, oltre a ciò,  $D = I$ , con  $I$  intesa come matrice di identità, si utilizza la lettera  $I$ . Negli altri casi si utilizza la lettera  $V$ .

Con tali modelli, il numero di parametri per le matrici di varianza-covarianza va da un minimo di  $d$  (nel caso  $EII$ , dove  $\Sigma_g = \lambda I \forall g = 1, \dots, G$ ) a un massimo di  $\frac{Gd(d+1)}{2}$  (nel caso  $VVV$ , dove  $\Sigma_g$  è differente per ciascun cluster).

#### 2.4.3 Stima del Modello tramite Massima Verosimiglianza

I modelli descritti in precedenza possono essere stimati tramite massima verosimiglianza, usando in particolare l'algoritmo Expectation-Maximization, anche detto EM ([16]Dempster Laird e Rubin, 1977; [17]McLachlan e Krishnan, 1997).

Assumiamo che i dati consistano di  $n$  osservazioni  $(y_i, z_i)$  per  $i = 1, \dots, n$ , dove sono osservati  $y_i$ , ma non  $z_i$ . Siano inoltre  $(y_i, z_i)$  indipendenti e identicamente distribuite secondo una distribuzione  $f$  con parametri  $\theta$ . Allora la verosimiglianza dei dati completi risulta essere:

$$L_C(y, z|\theta) = \prod_{i=1}^n f(y_i, z_i|\theta), \quad (2.2)$$

dove  $y = (y_1, \dots, y_n)$  e  $z = (z_1, \dots, z_n)$ . La verosimiglianza dei dati osservati si ottiene, invece, integrando (2.2) rispetto a  $z$ :

$$L_O(y|\theta) = \int L_C(y, z|\theta) dz.$$

Nel caso gaussiano, essa può anche essere riscritta come:

$$L_O(y|\theta) = \sum_{g=1}^G \sum_{i=1}^n \tau_g \phi_g(y_i|\mu_g, \Sigma_g), \quad (2.3)$$

Nello specifico, lo stimatore di massima verosimiglianza di  $\theta$  basato sui dati osservati massimizza  $L_O(y|\theta)$  rispetto a  $\theta$ .

Come detto in precedenza, per stimare i parametri del modello verrà utilizzato l'algoritmo EM. Tale algoritmo prevede l'iterazione di due passaggi: l'E-step e l'M-step. Il primo fornisce una stima dei dati non osservati alla luce di quelli osservati e dei parametri stimati. Il secondo, invece, massimizza la log-verosimiglianza dei dati completi (basata sulla stima dei dati non osservati ottenuta al punto precedente) rispetto ai parametri da stimare. Questi due step vengono ripetuti fino a convergenza o, perlomeno, fino al raggiungimento di una determinata condizione.

Nel caso del modello di mistura normale, i dati completi sono dati da  $(y_i, z_i)$ , dove  $z_i = (z_{i,1}, \dots, z_{i,G})$  sono i dati non osservati, con:

$$z_{i,g} = \begin{cases} 1 & \text{se } y_i \text{ appartiene al gruppo } g \\ 0 & \text{altrimenti} \end{cases}$$

Assumiamo che gli  $z_i$  siano i.i.d. e provenienti da una distribuzione multinomiale dei  $G$  gruppi con probabilità  $\tau_1, \dots, \tau_G$ . Inoltre, assumiamo che la probabilità di osservare  $y_i$ , data  $z_i$ , sia  $\sum_{g=1}^G \phi_g(y_i|\theta_g) z_{i,g}$ . In questo caso, la log-verosimiglianza dei dati completi è:

$$l_C(\theta_g, \tau_g, z_{i,g}|y) = \sum_{i=1}^n \sum_{g=1}^G z_{i,g} \log[\tau_g \phi_g(y_i|\theta_g)], \quad (2.5)$$

All's-esima iterazione dell'algoritmo EM vengono, quindi, calcolate le seguenti quantità:

- **E-step:** viene fornita una stima di  $z_{i,g}$ , calcolata come:

$$\hat{z}_{i,g} = \frac{\hat{\tau}_g^{(s-1)} \phi_g(y_i|\theta_g^{(s-1)})}{\sum_{h=1}^G \hat{\tau}_h^{(s-1)} \phi_h(y_i|\theta_h^{(s-1)})}, \quad (2.6)$$

dove  $\hat{z}_{i,g}$  è il valore di  $\hat{\tau}$  che l'osservazione  $i$  appartenga al gruppo  $g$ , date le osservazioni  $y_i$  e i parametri  $\theta_g$ .

- **M-step:** viene massimizzata (2.5) rispetto a  $\tau_g$  e  $\theta_g$ , con  $z_{i,g}$  fissato dall'E-step precedente. Essendo nel caso gaussiano, le stime dei parametri risultano:

–  $\hat{\tau}_g^{(s)} = \frac{1}{n} \sum_{i=1}^n \hat{z}_{i,g}^{(s-1)}$  è la stima della probabilità condizionata.

$$\begin{aligned}
- \hat{\mu}_g^{(s)} &= \frac{1}{\hat{n}_g^{(s-1)}} \sum_{i=1}^n \hat{z}_{i,g}^{(s-1)} y_i. \\
- \hat{\Sigma}_g^{(s)} &= \frac{1}{\hat{n}_g^{(s-1)}} \sum_{i=1}^n \hat{z}_{i,g}^{(s-1)} (y_i - \hat{\mu}^{(s)}) (y_i - \hat{\mu}^{(s)})^T.
\end{aligned}$$

L'algoritmo viene reiterato, nel caso specifico, fino a quando la differenza tra la log-verosimiglianza al passo  $s$  e al passo  $s - 1$  non è più piccola di una certa soglia (tipicamente  $10^{-5}$ ).

La verosimiglianza per i modelli di mistura non è tuttavia generalmente convessa, perciò, è possibile ci siano dei massimi locali [18](Biernacki, Celeus e Govaert, 2003). A conseguenza di ciò, la stima ottenuta dall'algoritmo EM può dipendere dai valori iniziali che sono stati scelti per i parametri. Una soluzione computazionalmente efficiente, che può essere usata per l'inizializzazione dei parametri dell'algoritmo EM, è quella della classificazione gerarchica basata sui modelli [19](Banoeld e Raftery, 1993). Analogamente ai classici metodi gerarchici agglomerativi, questa tecnica identifica inizialmente ciascuna osservazione in un proprio cluster e, successivamente, ad ogni passo unisce due gruppi in base a uno specifico criterio. Tale criterio è, in questo caso, la verosimiglianza di classificazione, definita come:

$$L_{CL}(\theta, z|y) = \sum_{i=1}^n f_{z_i}(y_i|\theta_{z_i}). \quad (2.7)$$

Tale verosimiglianza viene massimizzata stimando  $\theta$  e  $z$  (quindi parametri e gruppo di appartenenza) contemporaneamente. Si noti che tali stime in generale non sono asintoticamente consistenti [20](Mariott, 1975), ma, essendo computazionalmente efficienti, risultano utili come stime iniziali per l'algoritmo EM.

#### 2.4.4 Scelta del Numero di Cluster e del Modello di Clustering

Poiché di fatto non si conosce il numero di gruppi presenti nei dati, la selezione del modello da utilizzare prevede la scelta di due fattori: il modello di clustering (tra i 14 possibili) e il numero di cluster. Nella decisione da prendere ci sarà da valutare il compromesso (tradeoff) tra un modello più semplice (quindi con maggiori restrizioni per quanto riguarda volume, forma e orientamento dei cluster), ma con un maggior numero di gruppi, o un modello più complesso (dunque con maggiore elasticità per quanto riguarda le varie matrici di varianza-covarianza) con meno gruppi.

Un approccio possibile per tale scelta è la selezione del modello utilizzando la probabilità a posteriori dei modelli [21](Kass e Raftery, 1995). Immaginiamo di avere  $K$  possibili modelli  $M_1, \dots, M_K$  con probabilità a priori  $p(M_k)$ ,

$k = 1, \dots, K$  (solitamente poste uguali per ogni modello). Siano  $D$  i dati a nostra disposizione, allora per il teorema di Bayes si ha:

$$p(M_k|D) \propto p(D|M_k)p(M_k) \quad (2.8)$$

La scelta ricade poi sul modello con la maggiore probabilità a posteriori e, poiché  $p(M_k)$  è costante per ogni  $k$ , allora si guarda la probabilità a priori più grande. Quest'ultima viene ottenuta tramite il teorema della probabilità totale:

$$p(D|M_k) = \int p(D|\theta_{M_k}, M_k)p(\theta_{M_k}|M_k) d\theta_{M_k} \quad (2.9)$$

dove  $p(\theta_{M_k}|M_k)$  è la distribuzione a priori di  $\theta_{M_k}$ , cioè dei parametri del modello  $M_k$ . L'integrale di cui sopra è tuttavia difficile da calcolare. Sotto condizioni di regolarità dei modelli però, la (2.9) può essere approssimata dal criterio di informazione di Bayes (BIC):

$$2 \log p(D|M_k) \approx 2 \log p(D|\hat{\theta}_{M_k}, M_k) - \nu_{M_k} \log(n) = BIC_{M_k} \quad (2.10)$$

dove  $\nu_{M_k}$  è il numero di parametri indipendenti da stimare nel modello  $M_k$  [22] (Haughton, 1988).

Sebbene i modelli di mistura non soddisfino le condizioni di regolarità richieste, l'uso del criterio è appropriato anche nel caso della classificazione basata sui modelli ([23]Leroux, 1992; [24]Keribin, 1998). La scelta ricade dunque sul modello con BIC più elevato.

## 2.5 Confronto con il K-means

Il metodo K-means è una delle procedure non gerarchiche più popolari [25] (Johnson and Wichern, 2007). Il termine K-means descrive un algoritmo che assegna ogni elemento al cluster il cui centroide (media) è più vicino. Nella sua versione più semplice, il processo si compone di questi tre passaggi:

1. Partizionare gli elementi in  $K$  cluster iniziali.
2. Procedere attraverso la lista degli elementi, assegnando un elemento al cluster il cui centroide (media) è più vicino. La distanza è solitamente calcolata utilizzando la distanza euclidea con osservazioni standardizzate o non standardizzate. Ricalcolare il centroide per il cluster che riceve il nuovo elemento e per il cluster che perde l'elemento.
3. Ripetere il Passaggio 2 fino a quando non ci sono più riassegnazioni.

Invece di partire con una partizione di tutti gli elementi in  $K$  gruppi preliminari nel Passaggio 1, si potrebbero specificare  $K$  centroidi iniziali (punti centrali) e poi procedere al Passaggio 2.

L'assegnazione finale degli elementi ai cluster dipenderà, in una certa misura, dalla partizione iniziale o dalla selezione iniziale dei punti centrali. L'esperienza suggerisce che la maggior parte delle principali modifiche nell'assegnazione si verifica nel primo passaggio di riallocazione.

K-means è un algoritmo rapido e intuitivo, ma presenta limiti importanti, soprattutto in termini di rigidità nella forma dei cluster, sensibilità agli outliers e difficoltà nella determinazione del numero ottimale di cluster. Questo metodo assume che i cluster abbiano una forma sferica e dimensioni simili, utilizzando la distanza euclidea per assegnare i punti ai centroidi. Di conseguenza, in situazioni dove i dati non seguono una struttura regolare, K-means può portare a risultati inaccurati, generando cluster che non riflettono la vera distribuzione dei dati.

In contrasto, Model-Based Clustering offre una soluzione a questi problemi grazie alla sua maggiore flessibilità. Questo approccio rappresenta ogni cluster come una distribuzione probabilistica, generalmente una gaussiana, permettendo di modellare cluster di forme e dimensioni differenti. Ciò consente una migliore rappresentazione di strutture complesse nei dati, in cui i cluster potrebbero essere ellittici o avere orientamenti diversi, cosa che K-means non è in grado di catturare.

Un ulteriore vantaggio del Model-Based Clustering è la sua capacità di determinare il numero ottimale di cluster in modo automatico, tramite criteri statistici come il BIC (Bayesian Information Criterion). Al contrario, K-means richiede che il numero di cluster  $K$  venga specificato in anticipo, il che può essere problematico in assenza di una chiara indicazione su quante suddivisioni siano necessarie nei dati.

Per quanto riguarda la gestione degli outliers, K-means è molto sensibile ai valori anomali, poiché questi possono influenzare negativamente i centroidi, distorcendo l'intera struttura dei cluster. Model-Based Clustering, invece, gestisce gli outliers in modo più naturale, trattandoli come punti con una bassa probabilità di appartenenza a qualsiasi cluster. Questo rende l'approccio più robusto, soprattutto in contesti dove la presenza di anomalie è comune.

Infine, mentre K-means assegna i punti ai cluster in maniera deterministica, senza considerare l'incertezza, Model-Based Clustering utilizza un'assegnazione probabilistica, dove ogni punto può appartenere a più cluster con diverse probabilità. Questa capacità di modellare l'incertezza nei dati è par-

ticolarmente vantaggiosa per gestire casi ambigui, in cui i dati potrebbero essere equamente vicini a più cluster.

In sintesi, Model-Based Clustering si distingue per la sua versatilità e capacità di adattarsi a dati complessi, in cui la variabilità interna e le dipendenze tra variabili giocano un ruolo importante. Questo approccio si rivela particolarmente efficace in applicazioni che richiedono una segmentazione accurata e dettagliata, superando i limiti strutturali del K-means.

## 3 Clustering

In questa sezione verra` applicato il *model-based clustering* ai dati a nostra disposizione. Si procedera` a selezionare il modello e il numero di gruppi tramite il criterio BIC descritto nella sezione 2.4.4. Seguira` quindi una descrizione delle caratteristiche di ciascun cluster, si analizzerà la situazione dei giocatori ibridi e, infine, la composizione delle squadre.

### 3.1 Clustering dei Giocatori

La fase di applicazione del *model-based clustering* ai dati può dunque iniziare. Infatti, come descritto nella Sezione 1, è stata eseguita la fase di pre-processing, normalizzazione, standardizzazione e riduzione della dimensionalità dei dati tramite il metodo UMAP. A questo punto, tramite il software R e il pacchetto `mclust()`, è stato possibile individuare il modello ottimale e selezionare il numero di gruppi corretto tramite il criterio BIC, descritto nella Sezione 2.4.4.

La Figura 2 mostra il grafico del BIC plot. I valori sono visualizzati per un massimo di  $G_{\max} = 9$  componenti della miscela e per i 14 modelli di covarianza stimati, vale a dire per  $9 \times 14 = 126$  diversi modelli concorrenti in tutto. Una versione ingrandita della Figura 2 è mostrata nella Figura 3. In base al valore di BIC più alto, è stato selezionato il miglior modello con nove componenti di miscela e la specifica di covarianza EEI. Questa configurazione specifica che tutti i cluster sono diagonali, hanno stessa forma, stesso volume e sono orientati parallelamente agli assi principali.

Dunque, si può affermare con certezza che il numero corretto di cluster è nove, mentre in precedenza, dal Diagramma di dispersione delle coordinate ridotte UMAP (Figura 1), era possibile solo intuire questa informazione.

Ad ogni unità statistica è assegnato un vettore di probabilità che indica l'affinità di ogni giocatore a ciascun gruppo. Il giocatore viene quindi inserito nel cluster per cui tale probabilità ha valore maggiore. La Figura 4 riporta il numero di giocatori appartenenti a ciascun cluster.

Nei paragrafi successivi viene fornita una descrizione dettagliata per ciascuno dei cluster individuati. Per individuare più chiaramente i nove cluster, si riporta il grafico di classificazione per i giocatori (Figura 5) usando il modello EEI con nove cluster selezionato dal BIC.

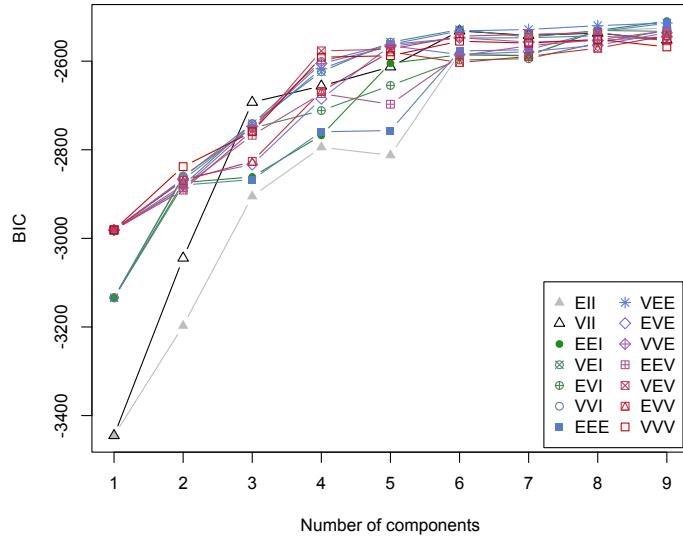


Figure 2: Selezione del modello tramite il BIC plot.

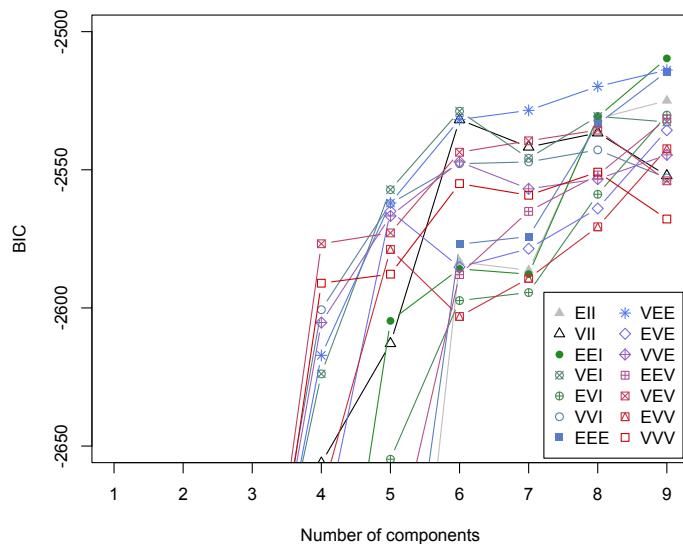


Figure 3: Versione ingrandita del BIC plot Figura 2.

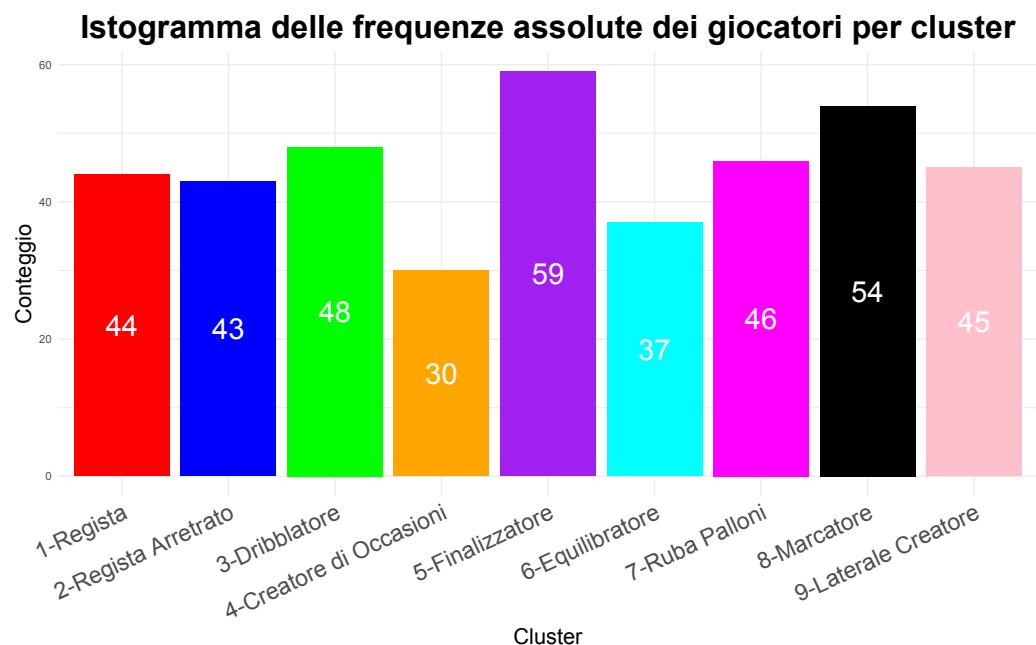


Figure 4: Istogramma delle frequenze assolute dei giocatori per cluster.

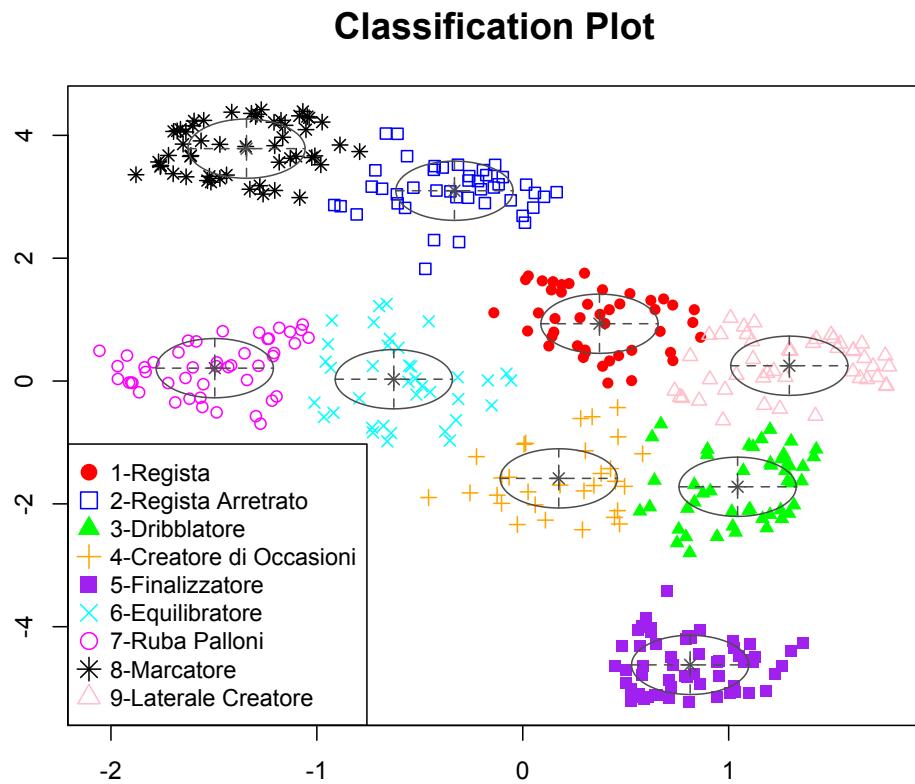


Figure 5: Grafico di classificazione per i giocatori usando il modello selezionato. Modello selezionato dal BIC con 9 gruppi, modello EEI

Si riporta di seguito l'heatmap delle statistiche standardizzate medie per cluster (Figura 6), da cui si può osservare in quali cluster ci sono valori alti o bassi di determinate variabili. L'heatmap utilizza una scala cromatica che varia dal blu al rosso: il blu indica valori più bassi, il bianco rappresenta valori centinati attorno a zero, mentre il rosso evidenzia valori più alti. È bene sottolineare che il modello è una semplificazione della realtà; perciò, il fatto che due giocatori siano presenti nello stesso cluster non significa necessariamente che siano altamente simili, soprattutto se hanno un profilo di ibridazione diverso in base alle probabilità di classificazione in ogni cluster. È possibile che due giocatori che differiscono in alcune statistiche appartengano allo stesso cluster, se hanno valori simili per la maggior parte delle altre statistiche.

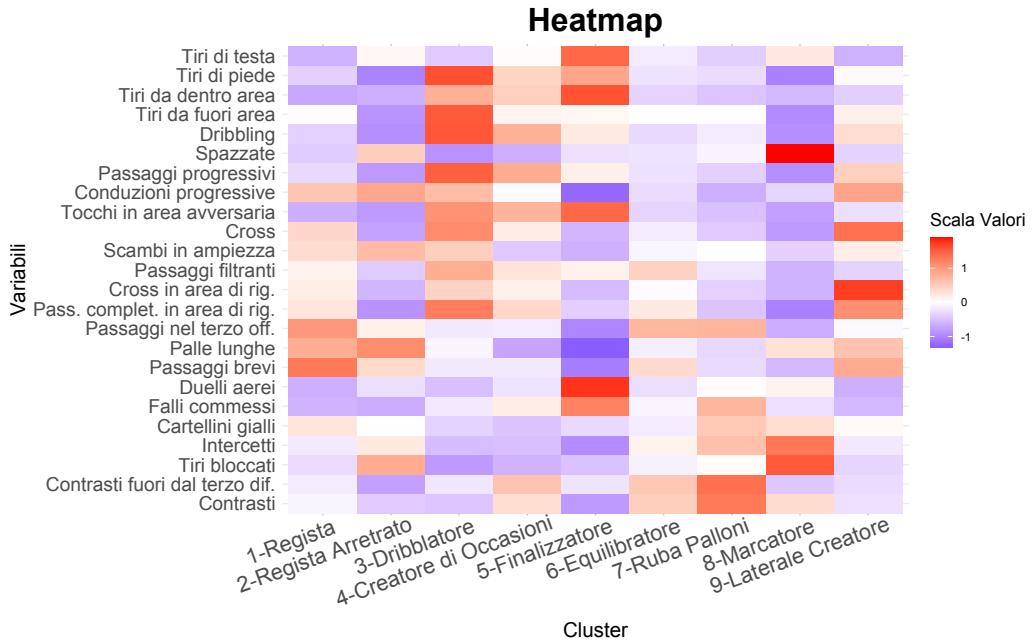


Figure 6: Heatmap delle statistiche standardizzate medie per cluster. Per la scala cromatica che va dal blu al rosso: blu indica valori più bassi, bianco rappresenta valori centrati attorno a zero, rosso evidenzia valori più alti.

### 3.1.1 Cluster 2 - Regista Arretrato

Sono difensori centrali che si distinguono per la loro attitudine a costruire il gioco dalla retroguardia. Presentano valori particolarmente elevati nei passaggi corti, negli scambi laterali e nella conduzione palla, il che li rende importanti per avviare le azioni e sviluppare il possesso palla della squadra. Tra le loro statistiche, spicca un valore elevato nei passaggi lunghi (9.15 ogni 90 min), suggerendo una propensione a variare il gioco e tentare di raggiungere direttamente i compagni distanti, un'abilità che aggiunge versatilità alla manovra offensiva. A livello difensivo, questi giocatori non sono caratterizzati da una marcata aggressività nei contrasti (2.33 ogni 100 tocchi), il che riduce il rischio di sanzioni come falli commessi (1.59 ogni 100 tocchi) e ammonizioni. Tuttavia, mantengono un buon livello nella capacità di bloccare tiri (1.01 ogni 100 tocchi) e di spazzare i palloni (3.88 ogni 100 tocchi). Riescono, dunque, ad evitare situazioni pericolose dimostrando una certa efficacia nel posizionamento difensivo e nella lettura delle azioni avversarie.

Fanno parte di questo cluster giocatori come Kim, Tomori, Danilo, Acerbi, Romagnoli e molti altri difensori appartenenti a squadre di alta classifica, il

cui obiettivo è dominare la partita. Tra i membri di questo gruppo si trovano anche difensori di squadre come la Fiorentina di Italiano, il Sassuolo di Dionisi e il Monza di Stroppa (prima parte di stagione) e Palladino (restante parte di stagione). Quest'ultime sono formazioni guidate da allenatori noti per il loro stile di gioco basato sul mantenimento del possesso e sulla costruzione dell'azione partendo dal basso.

### 3.1.2 Cluster 8 - Marcatore

Questi giocatori mostrano valori più alti in statistiche difensive di base, come intercetti (2.68 ogni 100 tocchi), contrasti (3.36 ogni 100 tocchi), tiri bloccati (1.39 ogni 100 tocchi), spazzate (7.11 ogni 100 tocchi) e duelli aerei (6.72 ogni 100 tocchi). Proprio a causa di questa predisposizione alla marcatura dell'avversario, tendono a commettere più falli (2.19 ogni 100 tocchi) e a ricevere qualche ammonizione in più. Eseguono meno azioni di costruzione rispetto al cluster “Regista Arretrato”; per caratteristiche tecniche o stile di gioco della squadra, risultano quindi meno coinvolti nella manovra offensiva.

Fanno parte di questo cluster giocatori come Bremer, Buongiorno, Hien, Djimsiti e Bijol, oltre a molti altri difensori appartenenti a squadre di medio-bassa classifica, che spesso subiscono la pressione dell'avversario e sono quindi costretti a eseguire molte azioni difensive. Rientrano in questo gruppo anche difensori di squadre, come l'Atalanta di Gasperini o il Torino di Juric, che adottano una fase difensiva definita "uomo su uomo", caratterizzata da numerosi duelli fisici.

### 3.1.3 Cluster 9 - Laterale Creatore

I giocatori appartenenti a questo gruppo occupano principalmente le zone laterali del campo e includono sia quinti di centrocampo sia terzini offensivi. In alcuni casi, vi rientrano anche centrocampisti che ricoprono ruoli di mezzala. Questi giocatori tendono quindi a posizionarsi in aree avanzate e svolgono un ruolo fondamentale nella creazione di occasioni da gol. Infatti, presentano il numero più alto di cross (4,34 ogni 90 minuti), cross in area (0,70 ogni 90 minuti) e passaggi completati in area di rigore (1,40 ogni 90 minuti). I valori delle conduzioni progressive (111,09 ogni 90 minuti) e dei passaggi progressivi (4,60 ogni 100 tocchi), oltre al numero di dribbling (2,31 ogni 90 minuti), confermano la loro tendenza ad avanzare e a spostare il gioco verso la porta avversaria. Questi giocatori partecipano alla costruzione del gioco attraverso passaggi corti (23,57 ogni 90 minuti) e lanci lunghi (7,55 ogni 90 minuti), ma contribuiscono anche in modo rilevante alla fase offensiva. Dimostrano

pericolosità tramite i cross, la loro esuberanza fisica e, talvolta, anche con conclusioni dalla distanza (0,58 tiri da fuori area ogni 90 minuti). Forniscono, comunque, anche un contributo difensivo, come evidenziano i valori di contrasti (2,57 ogni 100 tocchi), intercetti (1,30 ogni 100 tocchi) e spazzate (1,90 ogni 100 tocchi).

Fanno parte di questo cluster terzini di spinta come Hernández e quinti di centrocampo come Dimarco e Kostić, oltre a numerosi altri giocatori che ricoprono ruoli simili in moduli come il 3-4-3 o il 3-5-2. Anche centrocampisti offensivi come Koopmeiners, Barella, Çalhanoğlu e Zieliński rientrano in questo gruppo, dimostrando che partecipano alla fase offensiva con una costante pericolosità.

### 3.1.4 Cluster 1 - Regista

Questo cluster raggruppa giocatori che, pur occupando ruoli diversi in campo, condividono una funzione fondamentale per il gioco della squadra: la regia. Comprende infatti centrocampisti centrali, difensori impiegati come "braccetti" (cioè i difensori che, in una linea a tre, occupano una delle due posizioni laterali accanto al difensore centrale), e anche terzini che operano sulle fasce. Questi giocatori svolgono il ruolo di "fulcro del gioco" della propria formazione: quasi tutte le azioni di costruzione passano attraverso di loro. Dal punto di vista statistico, i dati mostrano che i giocatori di questo cluster eccellono nel gioco di costruzione. Sono il gruppo con i valori più elevati sia per passaggi brevi (26,32 per 90 minuti) sia per passaggi completati nel terzo offensivo (6,61 per 100 tocchi). Si distinguono anche per il numero di lanci lunghi (8,20 per 90 minuti) e di scambi in ampiezza (0,49 per 90 minuti), sottolineando la loro capacità di variare il gioco. La conduzione progressiva (99,43 per 90 minuti) indica inoltre una continua relazione con il pallone e ribadisce la capacità di essere coinvolti nel gioco. Non spiccano sul piano delle statistiche difensive e, infatti, presentano valori simili o poco superiori rispetto al cluster "Laterale Creatore". Analizzando questo gruppo emerge, quindi, che una squadra può disporre di registi non solo a centrocampo, ma anche in altre zone del campo, come le fasce laterali.

Appartengono a questo cluster terzini come Di Lorenzo, Dodô, Rodriguez e Alessandro Bastoni, che gioca come braccetto nella difesa a tre dell'Inter. Sono calciatori che effettuano meno cross rispetto al cluster "Laterale Creatore", ma contribuiscono maggiormente all'inizio del gioco da dietro. Avere un giocatore in grado di fare da regista anche sulle fasce laterali arretrate è un vantaggio per molte squadre, in quanto fornisce maggiore flessibilità nella

costruzione delle azioni. Questi calciatori devono possedere eccellenti abilità tecniche e una spiccata sicurezza nel controllo palla, caratteristiche apprezzate dagli allenatori che puntano su uno stile di gioco basato sul dominio del possesso.

Tra i centrocampisti si trovano Lobotka, Tonali e Brozovic, classici registi di centrocampo, ma anche Luis Alberto della Lazio, che, pur essendo più offensivo, è centrale nella manovra della squadra allenata da Sarri, grazie alle sue qualità di costruzione del gioco.

### 3.1.5 Cluster 7 - Ruba Palloni

Questo cluster include i centrocampisti con il maggior numero di azioni difensive. Si tratta dei giocatori con il più alto numero di contrasti (4,59 ogni 100 tocchi) e contrasti fuori dal terzo difensivo (2,65 ogni 100 tocchi). La loro funzione principale è schermare e supportare la difesa, con l'obiettivo di recuperare rapidamente il possesso palla. Questi calciatori sono coinvolti in numerosi duelli aerei (6,22 ogni 100 tocchi) e sono piuttosto fallosi (3,59 ogni 100 tocchi), risultando i più ammoniti tra i cluster (0,30 cartellini gialli ogni 90 minuti). Questo indica che spesso sono incaricati di interrompere le azioni avversarie, anche a costo di commettere falli. Questi giocatori sono principalmente focalizzati sulla fase difensiva e sulle azioni senza possesso palla, risultando poco coinvolti nella costruzione del gioco.

Fanno parte di questo gruppo calciatori come Éderson, Locatelli, Cristante e molti altri centrocampisti di squadre di bassa classifica, spesso costretti a svolgere un numero elevato di azioni difensive durante le partite.

### 3.1.6 Cluster 6 - Equilibratore

Questo cluster include centrocampisti dinamici, capaci di coprire ampie porzioni di campo e percorrere molti chilometri durante una partita. Non eccellono particolarmente in caratteristiche esclusivamente offensive o difensive, ma si distinguono per l'equilibrio tra le due fasi di gioco, riuscendo a svolgere bene sia la fase di costruzione che quella di interdizione. Contribuiscono alla fase di possesso e alla costruzione dell'azione, anche se in misura minore rispetto al cluster dei "Registi", come si nota dai valori di passaggi brevi (20,73 ogni 90 minuti) e passaggi completati nel terzo offensivo (6,05 ogni 100 tocchi). Risultano tuttavia più incisivi nei passaggi filtranti (0,16 ogni 90 minuti), il che dimostra la loro capacità di preparare situazioni pericolose nella fase di rifinitura. Rispetto al cluster "Ruba Palloni", il giocatore "Equilibratore" possiede una maggiore efficacia nella costruzione, pur mantenendo un buon

contributo anche in fase difensiva. I valori di contrasti (3,53 ogni 100 tocchi) e di contrasti fuori dal terzo difensivo (1,99 ogni 100 tocchi) confermano infatti la capacità di supportare la squadra nel recupero palla.

Appartengono a questo gruppo giocatori come Rabiot, Pašalić e McKennie.

### 3.1.7 Cluster 3 - Dribblatore

Questo cluster include attaccanti esterni, centrocampisti offensivi e seconde punte che eccellono nelle statistiche offensive. È il gruppo di giocatori con i valori più alti nei dribbling (4,16 ogni 90 minuti), nei tiri di piede (2,50 ogni 90 minuti), nei tiri da fuori area (1,24 ogni 90 minuti) e nei passaggi progressivi (7,26 ogni 100 tocchi). Grazie alla loro notevole abilità nel puntare l'avversario e superarlo in velocità, abbinata a una grande qualità tecnica, questi giocatori rappresentano una risorsa unica e preziosa per qualsiasi squadra, poiché riescono a rompere gli schemi difensivi avversari e creare superiorità numerica. Anche i tiri da fuori area costituiscono un'arma importante, soprattutto in situazioni in cui mancano altre soluzioni per arrivare al gol. Questi giocatori dimostrano, inoltre, una spiccata tendenza ad attaccare la porta, come si evince dai valori elevati di tocchi in area avversaria (3,82 ogni 90 minuti) e passaggi completati in area di rigore (1,49 ogni 90 minuti), a testimonianza della loro volontà di contribuire alla fase di rifinitura e finalizzazione. Altri indicatori come le conduzioni progressive (102,94 ogni 90 minuti), i cross (3,91 ogni 90 minuti), i cross in area di rigore (0,34 ogni 90 minuti) e i passaggi filtranti (0,20 ogni 90 minuti) sottolineano ulteriormente il loro ruolo chiave nella creazione di occasioni pericolose. Come prevedibile, non emergono per statistiche difensive rilevanti, poiché la loro funzione principale è offensiva.

Appartengono a questo cluster attaccanti esterni come Leão, Kvaratskhelia, Lookman e Berardi, oltre a giocatori che possono operare come seconde punte, come Dybala. Ne fanno parte anche centrocampisti offensivi come Samardzic, Bonaventura e Pellegrini.

### 3.1.8 Cluster 4 - Creatore di Occasioni

Questo cluster include centrocampisti offensivi di inserimento, esterni d'attacco di qualità e quinti di centrocampo con un ruolo offensivo. Pur avendo valori delle statistiche offensive inferiori rispetto al cluster “Dribblatore”, questi giocatori sono comunque in grado di rendersi molto pericolosi in zona d'attacco.

Ciò è evidenziato da un numero elevato di tocchi in area avversaria (3,28 ogni 90 minuti), dribbling (2,97 ogni 90 minuti) e passaggi progressivi (5,51 ogni 100 tocchi). Nonostante il loro orientamento portato all'attacco, questi giocatori si sacrificano molto anche in fase di non possesso, offrendo un notevole contributo in fase di interdizione, diversamente dal "Dribblatore". Questo è confermato dai valori nei contrasti (3,34 ogni 100 tocchi), contrasti fuori dal terzo difensivo (2,01 ogni 100 tocchi) e falli commessi (2,82 ogni 100 tocchi). I tiri effettuati sono prevalentemente di piede (1,47 ogni 90 minuti) e, in misura minore, di testa (0,24 ogni 90 minuti). Inoltre, calciano maggiormente da dentro l'area (1,16 ogni 90 minuti), piuttosto che tentare delle conclusioni da fuori area (0,57 ogni 90 minuti). Questo suggerisce che siano giocatori abili negli inserimenti e pronti a entrare in area per finalizzare. Inoltre, dimostrano una buona capacità di incidere nella fase di rifinitura, come evidenziato dai passaggi filtranti (0,13 ogni 90 minuti), cross (2,15 ogni 90 minuti) e passaggi in area di rigore (1,00 ogni 90 minuti).

Fanno parte di questo cluster trequartisti e centrocampisti offensivi come Brahim Díaz, Miranchuk e Frattesi; esterni d'attacco come Felipe Anderson e Zaccagni; nonché quinti di centrocampo offensivi come Dumfries, Udogie e Singo.

### 3.1.9 Cluster 5 - Finalizzatore

Questo gruppo include i giocatori che incarnano il prototipo del "numero 9" classico: fisicamente possenti e particolarmente efficaci nei duelli aerei. Infatti, questo cluster registra i valori più alti di duelli aerei (16,01 ogni 100 tocchi), tiri di testa (0,60 ogni 90 minuti) e falli commessi (4,28 ogni 100 tocchi). Al contrario, questi giocatori partecipano molto meno alla fase di creazione delle occasioni, e tutte le altre statistiche risultano poco rilevanti per comprendere il loro ruolo specifico. Si distinguono anche per l'elevato numero di tiri da dentro l'area (1,94 ogni 90 minuti) e tocchi in area avversaria (4,41 ogni 90 minuti), evidenziando la loro costante ricerca del gol. Questi attaccanti giocano gran parte della gara con l'obiettivo di segnare, offrendo anche un supporto fondamentale nelle situazioni di lancio lungo, sfruttando la propria altezza. Solitamente sono giocatori fisicamente strutturati, ben preparati a vincere duelli aerei e a mantenere il possesso, utilizzando la loro qualità tecnica per proteggere il pallone e subire falli.

A questo cluster appartengono giocatori come Osimhen, Lautaro Martínez, Vlahović e Giroud. Tutte le squadre cercano di avere almeno un attaccante con queste caratteristiche in rosa, poiché il ruolo del finalizzatore è determi-

nante per il successo della squadra.

### 3.2 Giocatori Ibridi

Uno dei requisiti fondamentali del model-based clustering era quello di estrarre non solo il cluster di appartenenza, ma anche le probabilità di classificazione relative a tutti i cluster per ogni singolo giocatore. La complessità del calcio, unita alle limitazioni dei dati disponibili, rappresenta una sfida importante per ogni tentativo di generalizzazione e classificazione. Pertanto, la possibilità di ricostruire un profilo di ibridazione completo per i giocatori nel modello di clustering consente di ottenere una caratterizzazione più dettagliata (sebbene non esclusiva) dei vari atleti all'interno di un sistema con un numero limitato di categorie.

In questo contesto, sono stati definiti come “ibridi” quei giocatori per i quali la probabilità massima di classificazione in uno dei cluster è inferiore a due terzi (ossia 66.7 %). La funzione di questi giocatori in campo rappresenta una combinazione delle caratteristiche distintive di ciascun cluster, il che consente di adottare due interpretazioni che potrebbero coesistere. Da un lato, un profilo ibrido può indicare che il giocatore ha ricoperto ruoli diversi nel corso della stagione, sia per inclinazioni personali sia per diverse strategie tattiche. Dall'altro lato, può anche suggerire che la funzione di un giocatore ibrido sia sufficientemente rara da non emergere come un cluster autonomo nel dataset.

Un esempio di giocatore con un elevato livello di ibridazione nel profilo di clustering è Sergej Milinković-Savić. Il nostro modello caratterizza Milinković-Savić come un ibrido tra il cluster "Equilibratore" (66.0%) e "Regista" (33.0%). In qualità di "Regista", è molto presente nel gioco di costruzione, evidenziato dai passaggi brevi (29,11 per 90 minuti) e dai passaggi completati nel terzo offensivo (5,72 per 100 tocchi). Risaltano anche il numero di lanci lunghi (6,43 per 90 minuti) e di scambi in ampiezza (0,65 per 90 minuti), che mettono in luce la sua abilità di variare il gioco. Ha la capacità di essere sempre coinvolto nel gioco. Inoltre, come "Equilibratore", si distingue come un centrocampista dinamico, in grado di svolgere bene anche la fase di interdizione. Infatti, supporta la squadra nel recupero della palla, come evidenziato dai contrasti (2,41 ogni 100 tocchi) e dai contrasti effettuati al di fuori del terzo difensivo (1,49 ogni 100 tocchi). Inoltre, Milinković-Savić, grazie ai suoi passaggi filtranti (0,45 ogni 90 minuti), dimostra una notevole abilità nella fase di rifinitura. Questo giocatore rappresenta un chiaro esempio di un profilo ibrido che combina diverse caratteristiche e funzioni in campo. È possibile che anche l'allenatore Maurizio Sarri abbia influenzato

il suo gioco, dato che alla Lazio propone un calcio offensivo, incentrato sul possesso del pallone.

### 3.3 Composizione delle Squadre

Un'applicazione interessante del modello di clustering consiste nello studio della composizione delle squadre. Squadre molto diverse stilisticamente presenteranno percentuali effettivamente distinte dei nove cluster. Una grande squadra nel calcio moderno ha una percentuale alta di giocatori con un elevato tasso tecnico, come "Regista Arretrato", "Laterale Creatore", "Regista", "Creatore di Occasioni" e "Dribblatore". In fase difensiva, le squadre di vertice schierano difensori a proprio agio con la palla al piede, dotati di visione di gioco e abilità nei passaggi necessarie per apportare un contributo fondamentale alla fase di costruzione. Sulle fasce, terzini e ali sono quasi sempre giocatori tecnici in grado di creare opportunità. Nel reparto avanzato, le grandi squadre tendono a utilizzare sempre meno attaccanti puri, preferendo una combinazione di attaccanti di movimento e giocatori forti nell'uno contro uno. All'estremo opposto, le squadre di livello inferiore si caratterizzano per una composizione più tradizionale, basata su centrali "Marcatori" e terzini bloccati in difesa, centrocampisti difensivi come i "Ruba Palloni" e come attaccanti i classici "Finalizzatori".

Nella Figura 7 si osservano le percentuali dei cluster nel Napoli, una squadra considerata di alto livello in quanto vincitrice del campionato nella stagione 2022/23. Invece, nella Figura 8 ci sono le percentuali dei cluster nella Sampdoria, squadra di basso livello poiché ultima nella stessa stagione.

Nel Napoli, le percentuali più alte si registrano per i cluster del "Regista" (25%), del "Regista Arretrato" (18,8%), del "Dribblatore" (25%) e del "Laterale Creatore" (25%), mentre la percentuale per il "Finalizzatore" è più bassa (6,2%), rappresentata unicamente dall'attaccante Osimhen. Non sono presenti, invece, giocatori appartenenti ai cluster di "Creatore di Occasioni," "Equilibratore," "Ruba Palloni," e "Marcatore". Questa analisi conferma che le squadre più competitive tendono ad avere una prevalenza di cluster tecnici, con maggiore qualità individuale. Inoltre, lo stile di gioco offensivo e orientato al possesso palla dell'allenatore Spalletti può aver influenzato i giocatori a soddisfare le richieste tattiche, portando a un incremento nelle statistiche di costruzione.

Per quanto riguarda la Sampdoria, si notano basse percentuali nei cluster di "Regista" (4,3%), "Dribblatore" (4,3%) e "Laterale Creatore" (8,7%), con una totale assenza di giocatori nei cluster "Regista Arretrato" e "Creatore di Occasioni". Al contrario, le percentuali risultano più alte per i cluster di "Equilibratore" (13%), "Ruba Palloni" (21,7%), "Marcatore" (30,4%) e "Final-

izzatore" (17,4%). Questi dati confermano che una squadra di basso livello tende a includere giocatori con ruoli difensivi e meno orientati alla costruzione o al gioco offensivo.

Pur non essendo esaustive, queste osservazioni offrono spunti riguardo ai profili di cui una squadra dovrebbe cercare di dotarsi se desidera scalare le classifiche giocando un calcio simile a quello delle grandi squadre contemporanee.

**Percentuali dei Cluster Napoli**

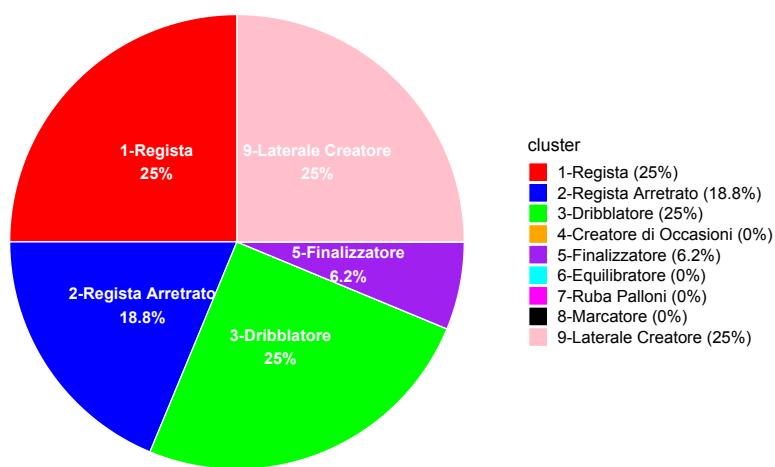


Figure 7: Grafico a torta che rappresenta le percentuali dei cluster nella rosa del Napoli, che è una squadra considerata di alto livello in quanto vincitrice del campionato nella stagione 2022/23.

**Percentuali dei Cluster Sampdoria**

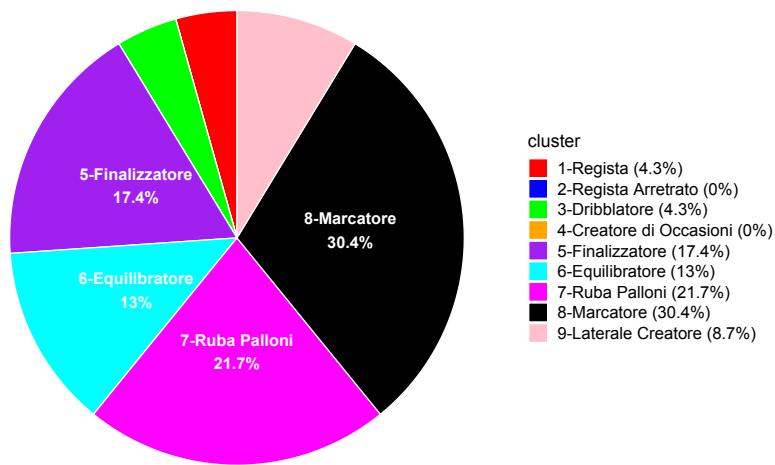


Figure 8: Grafico a torta che rappresenta le percentuali dei cluster nella rosa della Sampdoria, che è considerata una squadra di basso livello poiché si è classificata ultima nella stagione 2022/23.

## 4 Conclusioni

Da questo studio, sono emersi nove cluster che rappresentano altrettanti profili di giocatori, ognuno caratterizzato da uno stile di gioco e da abilità specifiche. Questo raggruppamento illustra come, nel calcio moderno, i ruoli non siano più rigidamente definiti: esistono, infatti, profili differenti anche per la stessa posizione, ciascuno con caratteristiche e funzioni uniche per le varie fasi della partita. Ogni cluster esprime un tipo di contributo tattico preciso, che può spaziare dalla costruzione della manovra alla finalizzazione, fino alla copertura difensiva, sottolineando come una squadra possa necessitare di una combinazione di questi profili per ottenere un equilibrio tattico.

Oltre ai profili definiti per ciascun cluster, è fondamentale considerare anche l'esistenza di giocatori ibridi, figure che non appartengono in modo chiaro e netto a un singolo cluster, ma presentano caratteristiche di più cluster contemporaneamente. Questi giocatori sono in grado di svolgere più ruoli e di interpretare differenti funzioni tattiche, rendendoli elementi estremamente preziosi per una squadra. I profili ibridi non solo ampliano la versatilità tattica, ma consentono anche agli allenatori di adattarsi a situazioni di gioco differenti senza necessità di sostituzioni, offrendo alternative dinamiche e aumentandone l'imprevedibilità in campo.

Inoltre, l'utilizzo di profili ibridi permette di colmare eventuali lacune in rosa, offrendo la possibilità di coprire più funzioni in una sola figura. Per esempio, un giocatore capace di combinare abilità di un "Regista" e un "Equilibratore" può adattarsi sia alla costruzione della manovra sia alla protezione della difesa in momenti cruciali della partita. Questa multifunzionalità rende i giocatori ibridi strumenti chiave nella pianificazione strategica di una squadra, poiché consentono di variare approccio tattico e di reagire in modo efficace alle sfide poste da ogni singolo avversario.

La creazione di una rosa bilanciata è quindi un aspetto cruciale per rispondere sia alle ambizioni di classifica che alle risorse economiche del club. Una squadra che mira a obiettivi di vertice, ad esempio, e che adotta un gioco offensivo, potrà beneficiare sia di profili fortemente orientati alla creazione del gioco, come i "Registi" e i "Dribblatori", sia di profili che contribuiscono alla solidità difensiva come i "Ruba Palloni" o gli "Equilibratori". Allo stesso modo, le squadre con risorse limitate possono sfruttare la versatilità dei profili ibridi per creare una rosa competitiva anche con un numero più contenuto di giocatori.

Infine, questa segmentazione di profili mette in evidenza il valore delle statistiche e dei dati per una costruzione della squadra sempre più mirata e scientifica. In un mercato altamente competitivo, l'analisi dei dati rappresenta uno strumento strategico che permette alle squadre di fare scelte più

consapevoli e di individuare, anche con risorse limitate, i profili adatti per il proprio sistema di gioco. Inoltre, un approccio data-driven permette di ottimizzare la gestione del talento, prevedendo il rendimento di un giocatore in contesti specifici o contro determinati avversari. Un approccio “data-driven” si basa sull’analisi di grandi volumi di dati per prendere decisioni informate, affidandosi a informazioni oggettive e dettagliate anziché su sole valutazioni soggettive. Questo tipo di approccio aiuta, infatti, a individuare i ruoli in cui ogni atleta può dare il massimo e a migliorare la pianificazione strategica e tattica della rosa, supportando la squadra nella costruzione di un profilo completo e anticipando le prestazioni future.

Questa ricerca dimostra, quindi, come un’analisi dei profili possa migliorare la composizione della rosa, rendendo più efficace e adattabile la strategia di squadra. Grazie a questi cluster, è possibile immaginare squadre capaci di massimizzare il rendimento, mantenendo sempre la possibilità di modificare lo stile di gioco per rispondere a qualsiasi tipo di sfida o avversario.

Infine, per allargare questo studio si potrebbero proporre dei possibili miglioramenti futuri. Estendere il dataset includendo altri campionati e, in particolare, sfruttando i dati posizionali del tracking permetterebbe di avvicinarsi ulteriormente alle reali funzioni di ciascun giocatore in campo. Il tracking, infatti, consentirebbe di integrare informazioni tecniche (come tiri e dribbling), atletico-fisiche (come il numero di sprint) e tattiche (come ricezioni fra le linee e inserimenti senza palla), creando un quadro più completo e dinamico delle prestazioni individuali. Inoltre, ampliando l’analisi a più stagioni, le fluttuazioni statistiche della singola annata sarebbero mitigate. Mediare i dati su più stagioni, applicando i metodi di normalizzazione descritti, renderebbe il dataset più solido e il clustering più affidabile.

## Bibliografia e Sitografia

- [1] Gagliardi, A., Soccernet. *The Clustering Project*. Soccernet.
- [2] FBref. Disponibile su: <https://www.fbref.com>
- [3] WhoScored. Disponibile su: <https://www.whoscored.com>
- [4] Scikit-Learn. *StandardScaler Documentation*. Disponibile su: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
- [5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.
- [6] Towards Data Science. (2021). "Data Preprocessing Techniques."
- [7] McInnes, L., Healy, J., & Melville, J. (2020). "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction."
- [8] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- [9] UMAP Documentation. Disponibile su: <https://umap-learn.readthedocs.io/>
- [10] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics.
- [11] Scikit-Learn PCA Documentation. Disponibile su: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>
- [12] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- [13] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
- [14] Scikit-learn Documentation. "Supervised Learning". Disponibile su: [https://scikit-learn.org/stable/supervised\\_learning.html](https://scikit-learn.org/stable/supervised_learning.html)
- [15] Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-Based Clustering and Classification for Data Science*. Cambridge University Press.

- [16] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1-38.
- [17] McLachlan, G. J., & Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley.
- [18] Biernacki, C., Celeux, G., & Govaert, G. (2003). "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(4), 583-586.
- [19] Banfield, J. D., & Raftery, A. E. (1993). "Model-Based Gaussian and Non-Gaussian Clustering." *Biometrics*, 49(3), 803-821.
- [20] Mariott, J. (1975). "Some Properties of the Likelihood Ratio Test for the Null Hypothesis of Homogeneity of Variance." *Biometrika*, 62(2), 491-497.
- [21] Kass, R. E., & Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90(430), 773–795.
- [22] Haughton, D. (1988). "On the Choice of a Model to Fit Data from an Exponential Family." *Annals of Statistics*, 16(1), 342–355.
- [23] Leroux, B. G. (1992). "Consistent Estimation of a Mixing Distribution." *Annals of Statistics*, 20(3), 1350–1360.
- [24] Keribin, C. (1998). "Consistent Estimation of the Order of Mixture Models." *Sankhya: The Indian Journal of Statistics*, 60(1), 49-66.
- [25] Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. 6th ed. Prentice Hall.