UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

DIPARTIMENTO DI ECONOMIA

Corso di Laurea
in Economia e Azienda Digitale

Relazione Finale

Pallacanestro e Statistica: uno studio della performance dei giocatori NBA 2022-2023

Relatore: Chiar.ma Prof.ssa Marica Manisera

Laureando:
Simone Spiritelli

Matricola n. 731891

Anno Accademico 2022/2023

INDICE

INTRODUZIONE

Questa relazione finale nasce dalla curiosità e dall'interesse sviluppato durante i corsi di "Statistica", "Statistica per l'Economia e l'Azienda Digitale" e "Data Analysis and Big Data Lab", che hanno portato a voler sperimentare e intraprendere un progetto legato all'analisi statistica nel contesto sportivo della pallacanestro, in particolare nell'NBA, campionato di basket più prestigioso al mondo e caratterizzato da numerosi dati da poter analizzare. L'obiettivo principale era quello di sviluppare un algoritmo in grado di estrarre dati appartenenti ai giocatori NBA con il fine di poter calcolare la probabilità che un giocatore superi una soglia specifica per una variabile d'interesse.

Questa relazione finale è composta da tre capitoli distinti. Nel primo capitolo si esplorerà l'importanza delle analisi statistiche e dell'utilizzo dei dati nel contesto sportivo, con particolare attenzione alla pallacanestro, dopo aver definito regole e funzionamento di questo sport. Si evidenzieranno l'importanza dei dati e le influenze che questi possono avere sulle prestazioni e sulle strategie delle squadre.

Nel secondo capitolo ci si concentrerà sull'analisi e la descrizione di due database creati per questa relazione finale e per il progetto che compone l'obiettivo principale della relazione. Grazie a questi database sarà possibile effettuare analisi statistiche di base come, ad esempio, il calcolo di indicatori statistici per alcuni giocatori NBA e l'analisi dell'influenza dei ruoli in campo dei giocatori sulle medie delle loro prestazioni.

Nel terzo capitolo l'attenzione si sposterà sull'obiettivo principale da raggiungere e da cui è nata questa relazione finale, ovvero la creazione dell'algoritmo in grado di calcolare la probabilità per un giocatore di superare una determinata soglia per quanto riguarda una variabile di performance particolare, come ad esempio il numero di rimbalzi. Per il raggiungimento di tale obiettivo, è stato necessario imparare ad utilizzare un determinato linguaggio di programmazione e realizzare uno script per l'estrazione e l'analisi dei dati, che sarà descritto in maniera dettagliata anche attraverso l'applicazione a dati reali.

All'interno delle pagine che seguiranno, si cercherà di evidenziare il potenziale che l'analisi dei dati offre per ottimizzare le strategie di gioco e migliorare in generale la comprensione e la percezione di questo sport ormai sempre più seguito.

CAPITOLO 1: L'IMPORTANZA DEI DATI NEL BASKET

Nel primo capitolo, si esplorerà l'importanza delle analisi statistiche e dell'utilizzo dei dati nel contesto sportivo, con un'attenzione speciale al basket. Verrà definito il concetto di Big Data, per poi analizzare il ruolo dell'analisi dati nello sport in generale, con un focus poi sul basket. Prima di approfondire l'impatto dei dati sul basket, verrà fornita una panoramica generale delle regole e di come funzioni questo sport, per poi mostrare come i dati influenzino le prestazioni e come questi possano cambiare le strategie e l'allenamento dei giocatori.

1.1. I BIG DATA

"Il mondo sta diventando un computer. L'informatica è sempre più connessa a tutto quello che facciamo, a ogni aspetto della nostra vita quotidiana". Questa affermazione è tratta dall'intervento di Satya Nadella, CEO di Microsoft, presso l'Università Bocconi a Milano in data 30 maggio 2019.

Nel quotidiano, infatti, ciascun individuo scambia dati con dispositivi e applicazioni fornendo a soggetti terzi che gestiscono tali dati grandi quantità di informazioni. Questi dati vengono elaborati, organizzati e condivisi con terze parti, diventando al giorno d'oggi fondamentali. Oggi il termine "Big Data" è onnipresente in articoli, blog e sui social media, ed è difficile non averne almeno sentito parlare, anche se non si è familiari con il concetto. Per comprenderlo, però, è importante partire dalla base e capire di cosa si sta parlando. Che cosa sono i big data?

In "Big Data. Come stanno cambiando il nostro mondo" di Marco Delmastro e Antonio Nicita, gli autori utilizzano questo termine, riportando la definizione formulata dall'Unione Europea: <<grandi quantità di tipi diversi di dati prodotti da varie fonti, fra cui persone, macchine e sensori>>. Inoltre, i due autori approfondiscono il significato di "big", citando e mettendo in evidenza il modello realizzato da Douglas Laney nel 2001 che si basa sulle "tre v" come variabili chiave per descrivere i big data: velocità, volume e varietà. Con il passare degli anni, sono state integrate al modello altre due v mirate a delineare come questi nuovi dati debbano essere sfruttati: veridicità e variabilità.

1.1.1. LE CINQUE V DEI BIG DATA

Nel 2001, Douglas Laney, all'epoca vicepresidente e Service Director dell'azienda Meta Group, presentò attraverso il report "il modello delle 3V", le caratteristiche fondamentali dei Big Data: Volume, Velocità e Varietà.

Nel corso del tempo questo modello si è evoluto includendo altre variabili, portando le caratteristiche chiave dei Big Data a diventare inizialmente 5, per poi aumentare ulteriormente grazie a studiosi ed esperti con l'obiettivo di catturare la complessità e diversità di questi dati.

In generale, si definiscono Big Data quei dati che abbiano almeno una delle seguenti caratteristiche:

- Volume: Si riferisce alla grandezza dei dati generati quotidianamente in varie attività mondane. Identificare una soglia precisa per definire il volume dei Big Data è difficile, ma solitamente si considera una massa di dati superiore a 50 Terabyte. Questa massa di informazioni è in costante aumento e supera la capacità delle tecnologie tradizionali.
- Velocità: Indica la rapidità con cui i dati vengono generati, trasferiti e accumulati, grazie anche alla crescente presenza di dispositivi e reti con sensori in grado di raccogliere dati in tempo reale. La parte complessa non è solo quella di raccogliere rapidamente questi dati, ma anche di analizzarli in tempo reale per prendere decisioni tempestive ed efficaci.
- Varietà: Fa riferimento alla diversità di tipologie di dati provenienti da una vasta gamma di fonti eterogenee, come sensori e social network. I dati possono essere strutturati (organizzati con una lunghezza e un formato definiti), semi strutturati (parzialmente organizzati, come i file di registro) o non strutturati (senza una struttura rigida, come testi, immagini e video).

Queste tre sono le caratteristiche fondamentali e iniziali presenti nel modello elaborato da Douglas Laney. Con il tempo sono state introdotte una quarta, quella di veridicità, e poi una quinta, quella di valore, per riflettere ulteriori sfaccettature dei Big Data:

- Veridicità: Indica il grado di affidabilità o inaffidabilità dei dati, sottolineando la necessità di dati precisi e affidabili per condurre analisi attendibili, in quanto data la grande quantità di fonti diverse, spesso è complicato organizzare i dati in maniera efficace senza causare disordine e incongruenze. La veridicità implica che i dati debbano essere accurati e rappresentare fedelmente la realtà, sottolineando il fatto che quelli di scarsa qualità possono essere più dannosi rispetto all'assenza di dati.

- Valore: Riguarda l'abilità di trasformare dati in informazioni utili, consentendo decisioni mirate e azioni ben orientate. Non conta solamente la quantità dei dati, che possono pure essere sbagliati, ma soprattutto la qualità di questi, ovvero il valore che possiamo trarne.

Con il passare del tempo ulteriori caratteristiche si sono aggiunte al modello[1], quali ad esempio la variabilità:

- Variabilità: Si riferisce alla diversità dei dati provenienti da varie fonti e formati. Questa diversità è importante poiché i dati possono avere significati differenti a seconda del contesto in cui vengono interpretati, contribuendo a distinguere tra dati utili e dati inconsistenti, fornendo informazioni e previsioni attendibili.

1.1.2. LA DIFFERENZA TRA BIG DATA E DATI TRADIZIONALI

Il mondo dei dati può essere suddiviso in due categorie principali: i dati tradizionali, conosciuti anche come "small data", e i big data. La principale distinzione tra questi risiede nella dimensione, nella complessità e negli obiettivi d'uso. I dati tradizionali, o small data, rappresentano informazioni strutturate, solitamente raccolte e organizzate in formati tabellari, database relazionali o fogli di calcolo, che consentono facili analisi e manipolazioni. Sono ideali per scopi dalla portata limitata e operazioni quotidiane, come ad esempio il monitoraggio delle vendite, la gestione delle relazioni con i clienti o la supervisione dei processi aziendali. D'altra parte, i big data sono enormi aggregati di dati, spesso non strutturati, che richiedono un approccio completamente diverso per essere analizzati in maniera efficace, che utilizzi strumenti avanzati di elaborazione e analisi. Questi dati, come già detto nel paragrafo 1.1.1, si caratterizzano per il loro volume elevato, la varietà di formati (compresi testi, immagini, audio e video) e la velocità con cui vengono generati e accumulati.

Per distinguere tra big data e dati tradizionali, si utilizzano diverse caratteristiche, tra cui le dimensioni dei dati, le modalità con cui sono organizzati, l'architettura necessaria per gestire i dati, le fonti da cui derivano i dati e i metodi utilizzati per analizzare i dati. I database di dati tradizionali tendono ad essere misurati in gigabyte e terabyte, e possono essere gestiti in modo centralizzato, inclusa la conservazione su un singolo server. Al contrario, i big data si distinguono per il loro volume, misurato

[1] https://www.bigdata4innovation.it/big-data/5-v-dei-big-data-cosa-sono-quale-ruolo-rivestono/

in petabyte (1000 terabyte), zettabyte (1000 petabyte) o exabyte (1000 zettabyte), per questo vengono richieste soluzioni di gestione più moderne, ad alta capacità e basate su cloud.

Per quanto riguarda l'organizzazione dei dati, quelli tradizionali si basano solitamente su relazioni semplici e sono generalmente strutturati e organizzati in record, file e tabelle. Invece, i big data, utilizzano uno schema dinamico e non sono solitamente strutturati.

L'architettura per gestire i dati tradizionali è in genere centralizzata, risultando più conveniente e sicura per dati strutturati di piccole dimensioni, mentre i big data richiedono un'architettura complessa e distribuita a causa della loro larga scala. I sistemi distribuiti collegano più server o computer su una rete, garantendo stabilità e funzionamento continuo anche in caso di guasti di singoli nodi.

Le fonti dei dati tradizionali provengono principalmente da software come, ad esempio, quelli relativi alla pianificazione delle risorse aziendali, relativi alla gestione delle relazioni con i clienti e transazioni online, mentre i big data derivano da fonti dinamiche ed in continua evoluzione come ad esempio social media, dati dispositivi e sensori, dati audiovisivi.

I big data consentono di scoprire relazioni complesse tra variabili e previsioni più accurate relative a possibili tendenze di mercato o al comportamento di consumatori in tempo reale, consentendo inoltre decisioni rapide e garantendo un importante vantaggio sulla concorrenza, in qualsiasi ambito.

In sintesi, mentre i dati tradizionali sono più agevoli da manipolare e interpretare, i big data offrono un potenziale di analisi più profondo e immediato. Comprendere ed integrare entrambi è essenziale per ottenere un quadro completo, non si tratta infatti di scegliere tra big data e dati tradizionali, ma di comprendere come utilizzare e gestire entrambi, garantendo un vantaggio competitivo in un mondo sempre più orientato ai dati[2].

[2] https://www.purestorage.com/it/knowledge/big-data/big-data-vs-traditional-data.html

1.1.3. LA STORIA DEI BIG DATA

L'evoluzione della gestione e analisi dei big data rappresenta un cambiamento importante nel modo in cui l'umanità ha affrontato e sfruttato le informazioni nel corso del tempo. Fin dall'antichità, infatti, l'uomo ha sempre dimostrato un profondo interesse nella conservazione e consultazione delle informazioni e della conoscenza, per poterla utilizzare nel futuro, come dimostra ad esempio la biblioteca di Alessandria nel III secolo a.C., considerata la più grande dell'antichità.

Tra il XIX e il XX secolo si osserva una notevole svolta nel modo in cui le informazioni venivano gestite. Nel 1880, infatti, Herman Hollerith, dipendente dell'ufficio di censimento degli Stati Uniti, ha rivoluzionato il processo di catalogazione dei dati censuari, riducendo drasticamente i tempi necessari da 10 anni a soli 3 mesi. Questo è stato per l'appunto il primo passo verso l'automazione dei processi di calcolo e gestione dati, aprendo la strada a una vera e propria rivoluzione nel settore.

Successivamente, nel 1958, il ricercatore e inventore tedesco Hans Peter Luhn, lavorando per IBM, ha coniato il termine "Business Intelligence", indicando come raccogliere ed esaminare dati potesse dare un vantaggio competitivo e prendere decisioni strategiche informate[3].

Nel 1965 si registra un altro passo importante: viene creato il primo data center, che è una struttura dedicata all'archiviazione, elaborazione, gestione e distribuzione di grandi quantità di dati e servizi informatici, negli Stati Uniti, segnando l'inizio della più efficiente gestione centralizzata dei dati.

Con l'avvento di Internet alla fine del XX secolo, particolarmente nel 1991, il panorama dei dati ha subito un'accelerazione importante. Questo evento ha segnato l'inizio di un'era in cui i dati sono diventati sempre più vasti e facili da raggiungere, modificando il modo in cui venivano raccolti, elaborati e utilizzati.

Nel 1999 è stato coniato il termine "Big Data" da John Mashey, Chief Scientist alla Silicon Graphics. L'introduzione dell'Internet of Things (IoT) in quell'anno ha mostrato come gli oggetti e i dispositivi potessero essere connessi a Internet, generando ancora più dati da elaborare e comprendere[4].

Al giorno d'oggi si vive in un'epoca in cui i big data continuano ad evolversi ed a portare all'adozione di nuove tecnologie, dove l'intelligenza artificiale (IA) e il machine learning hanno guadagnato un ruolo fondamentale, consentendo previsioni più accurate e gestione di sistemi complessi. Guardando al futuro, si prevedono grandi progressi nell'elaborazione dei dati, nella privacy, nella sicurezza e

[3] https://www.andreapacchiarotti.it/archivio/big-data.html
[4] https://www.themarketingfreaks.com/2019/11/big-data-cosa-sono-la-storia-le-caratteristiche-le-analisied-esempi/

nell'integrazione con l'intelligenza artificiale. I big data continueranno a guidare l'innovazione in settori come la sanità, l'energia, l'ambiente, lo sport, su cui ci si focalizzerà in seguito, cambiando il modo di prendere decisioni prediligendo uno basato su informazioni derivanti dall'analisi di questi dati.

1.2. L' ANALISI DEI DATI

L'analisi dei dati, anche chiamata data analytics, è diventata sempre più importante nell'era digitale odierna, tanto che i dati vengono considerati il "nuovo petrolio", ma per trarne il massimo beneficio è necessario impiegare in maniera accurata strumenti e tecniche specifiche. L'analisi dei dati si concentra su come trasformare dati grezzi in informazioni più preziose e utili a livello decisionale e funzionale, cercando di estrarre valore dai dati stessi. Per condurre un'analisi efficace dei dati, è necessario avere competenze specifiche nel campo dell'analisi dei dati: un professionista specializzato in questo settore, chiamato Data Analyst o Analista Dati, è responsabile di gestire e analizzare i dati in modo accurato ed efficiente.

Esistono tre principali tipi di data analytics, ognuno dei quali è caratterizzato da un obiettivo distinto e contribuisce in modo diverso all'analisi dei dati ed alle decisioni finali: analisi descrittiva, analisi predittiva e analisi prescrittiva. Nel primo capitolo di "Data analytics per tutti: imparare ad analizzare, visualizzare e raccontare i dati" di Andrea De Mauro è possibile trovare una accurata descrizione di ciascuna tra le tipologie di analisi esistenti[5].

- Analisi descrittiva (descriptive analytics): L'analisi descrittiva rappresenta il punto di partenza di ogni processo analitico, ovvero una serie di azioni mirate all'analisi e comprensione di dati o informazioni per trarre conclusioni, identificare modelli e prendere decisioni. Si concentra sulla comprensione dei dati passati per fornire una visione chiara e interpretare al meglio lo stato dell'attività considerata. Risponde alla domanda "Che cosa è successo?" attraverso l'uso di tabelle, grafici e statistiche riassuntive con indicatori di prestazione. La descriptive analytics si manifesta attraverso strumenti che permettono di accedere ai dati, filtrarli e visualizzarli, come report statistici, ovvero file in diversi formati come PDF o fogli Excel, oppure con dashboard interattive, che offrono un'interfaccia web per esplorare i dati attraverso una visione più approfondita degli aspetti più rilevanti. L'analisi descrittiva è fondamentale per comprendere il passato e trarre indicazioni per il futuro, rendendo i dati più comprensibili.

- Analisi predittiva (predictive analytics): L'analisi predittiva rappresenta un approccio più avanzato rispetto alla descrittiva e mira a rispondere a due domande fondamentali: "Perché è successo?" e "Che cosa succederà?". Sfrutta tecniche analitiche sofisticate, spesso basate sull'intelligenza artificiale, per creare modelli probabilistici attraverso i quali è possibile anticipare diversi scenari futuri, come le vendite, i prezzi, le dimensioni di mercato e i comportamenti dei clienti. La predictive analytics si manifesta attraverso strumenti probabilistici come previsioni, modelli di propensione e rilevanza di anomalie per ricercare relazioni ed essere in grado di prendere le decisioni migliori, cercando di anticipare i comportamenti di clienti e concorrenti.
- Analisi prescrittiva (prescriptive analytics): L'analisi prescrittiva va oltre alla descrizione del passato e la previsione del futuro, ma fornisce indicazioni dirette su quali azioni sia meglio intraprendere per raggiungere un obiettivo specifico, rispondendo alla domanda "Che cosa fare?". La prescriptive analytics utilizza algoritmi sofisticati che simulano scenari alternativi e consigliano la migliore azione per massimizzare i risultati derivanti dalle decisioni prese, ma anche algoritmi che agiscono autonomamente come sistemi di raccomandazione, prendendo decisioni in tempo reale in base alle informazioni ricevute.

In sintesi, il processo di analisi dati inizia con la descrizione del passato tramite l'analisi descrittiva successivamente, si avanza con l'analisi predittiva, che punta ad anticipare il futuro utilizzando modelli e tecniche avanzate, ed infine, si arriva all'analisi prescrittiva, che non solo prevede cosa accadrà, ma consiglia azioni specifiche per massimizzare gli obiettivi prefissati da raggiungere.

1.2.1. L'ANALISI DEI DATI NELLO SPORT

L'analisi dei dati e la statistica sono strumenti che stanno riscontrando sempre più successo e importanza in numerosi settori. Il loro utilizzo non si limita solamente al piano accademico o aziendale, ma riscontrano sempre più importanza anche nel mondo dello sport, diventandone un vero e proprio pilastro.

Per quanto riguarda il mondo sportivo, la raccolta dettagliata dei dati dei giocatori e delle partite, insieme all'utilizzo di algoritmi sofisticati, ha dato origine ad una vera e propria scienza chiamata "Sports Analytics", relativa alla gestione di database strutturati, l'impiego di modelli analitici descrittivi e predittivi e l'utilizzo di sistemi informativi per ottimizzare le decisioni al fine di ottenere

vantaggi competitivi e maggiori probabilità di successo. Per quanto riguarda l'analisi sportiva, questa si distingue in:

- Analisi dei dati sul campo, che monitora le informazioni relative alle prestazioni individuali e collettive attraverso le quali è possibile migliorare le strategie di gioco e la performance degli atleti, permettendo loro di adottare tattiche e strategie più precise e mirate, come ad esempio una personalizzazione di allenamenti finalizzata ad un miglioramento specifico di determinate caratteristiche carenti.
- Analisi dei dati fuori dal campo, che riguarda per esempio gli aspetti commerciali e finanziari delle organizzazioni sportive con l'obiettivo di aumentare la crescita e la redditività, come ad esempio l'ottimizzazione della vendita di biglietti, il merchandising e il coinvolgimento dei tifosi.

Ad oggi, sempre più squadre in qualsiasi sport investono in data scientist specializzati per analizzare in modo efficiente i dati, sia sul campo che fuori, con il fine di raggiungere tattiche e strategie vincenti più mirate, considerando anche analisi simili su avversari per cercare di evidenziare i loro punti di forza e di debolezza.

Il baseball è stato uno dei precursori in questo campo, con il libro "Percentage Baseball" pubblicato da Earnshaw Cook nel 1964, in cui vengono utilizzate leggi matematiche per analizzare il baseball. Tuttavia, il vero introduttore dell'analisi sportiva in questo sport è stato Billy Beane, che ha ricoperto la carica di direttore generale degli Oakland Athletics, squadra di baseball, dal 1997 al 2016. Prima del 2002 l'analisi dei dati nel baseball era spesso trascurata e completamente sconosciuta, con allenatori e giocatori che si affidavano principalmente ad istinto e bravura dei giocatori. Billy Beane in quell'anno ha rivoluzionato questo approccio, in quanto con risorse finanziarie limitate e di conseguenza acquisti limitati, Beane ha compreso che far avanzare i corridori in base era fondamentale per segnare punti e vincere partite; quindi, ha focalizzato la sua strategia sull'acquisizione di lanciatori sottovalutati a costi ridotti, ma che avessero alte percentuali in base. L'analisi sportiva è diventata però popolare solo nel 2011, con l'uscita del film "MoneyBall: l'arte di vincere" che riprendeva questa storia.

Come detto precedentemente, l'analisi dei dati sportivi e la statistica hanno assunto importanza esponenziale al giorno d'oggi; tuttavia, ogni sport utilizza approcci diversi per raccogliere ed elaborare dati al fine di ottimizzare prestazioni e decisioni.

Nel calcio, i club calcistici di tutto il mondo investono nella raccolta e nell'analisi di dati per ottenere informazioni relative a variabili molto importanti come il posizionamento dei giocatori, la distanza

percorsa e le statistiche individuali come dribbling e passaggi completati, con il fine di ottenere una visione approfondita sui giocatori, aiutando allenatori e preparatori atletici a sviluppare strategie mirate e di successo. Esempio importante di ciò è avvenuto ad esempio durante la supercoppa europea del 2021 contro il Villareal, quando Thomas Tuchel, ex allenatore del Chelsea, squadra londinese, decise di far entrare in campo il portiere Kepa Arrizabalaga al 119esimo minuto basandosi solamente su dati statistici relativi alla sua bravura nel parare i rigori, portando la sua squadra alla vittoria.

Il baseball, che, come detto prima, è stato uno dei precursori all'analisi sportiva, si basa su una vasta gamma di metriche come, ad esempio la media battuta, ovvero il rapporto tra il numero totale di battute valide e il numero totale di turni alla battuta, e la percentuale di arrivo in base, che misura la frequenza con cui un giocatore raggiunge una delle basi durante i suoi turni di battuta, per guidare le decisioni strategiche squadra per squadra. Inoltre, l'utilizzo di telecamere avanzate, sensori e tecnologie indossabili è stato fondamentale per il reclutamento dei giocatori, permettendo di raccogliere una vasta quantità di dati per evidenziare l'importanza di selezionare l'atleta giusto in base alle esigenze di ciascuna squadra.

Nella NFL (National Football League), la competizione più importante al mondo di football americano, le squadre hanno iniziato ad utilizzare etichette elettroniche a radiofrequenza per monitorare con precisione la rapidità, l'accelerazione, la decelerazione e la distanza percorsa dai giocatori durante le partite.

Nella pallacanestro, Daryl Morey, su cui ci si focalizzerà nel paragrafo successivo, è stato uno dei primi a portare l'innovazione di considerare indicatori statistici nella valutazione dei giocatori, per massimizzare le loro prestazioni. Fondamentali a livello strategico sono anche le informazioni relative alle telecamere di tracciamento che registrano ogni movimento di ciascun giocatore. In virtù di queste, nel passato sono state effettuate analisi sull'efficienza dei tiri da tre punti, con conseguente aumento nei tentativi per squadra nel corso degli anni. Anche se la percentuale di successo potrebbe essere leggermente inferiore, l'analisi dei dati ha dimostrato, ad esempio, che in media i tiri da tre punti hanno un rendimento superiore rispetto ai tiri da due punti in prossimità della linea dei tre punti, in quanto il maggior valore di ogni canestro da tre punti compensa questa differenza. Nella NBA (National Basketball Association), lega di pallacanestro più importante al mondo, la startup israeliana RSPCT con cui la NBA collabora, raccoglie dati avanzati attraverso videocamere 3D e sensori, per analizzare traiettoria di tiri e stato di salute degli atleti, per cercare di identificare potenziali rischi di

infortuni, considerando anche una analisi relativa all'intensità dell'attività e agli impatti derivanti dalle collisioni[6].

1.2.2. L'ANALISI DEI DATI NEL BASKET

L'analisi dei dati nel mondo del basket ha assunto un ruolo di fondamentale importanza, contribuendo a fornire alle squadre maggiori informazioni per cercare di ottenere vantaggi competitivi. Attraverso la raccolta e l'elaborazione di statistiche, le squadre possono identificare le abilità individuali e i punti di forza dei propri atleti, permettendo agli allenatori di prendere decisioni e attuare strategie più mirate.

Una delle figure più importanti ad aver introdotto l'utilizzo dell'analisi dati nel basket è Daryl Morey, direttore generale degli Houston Rockets dal 2007, con i quali nella stagione 2017-2018 nonostante gli infortuni, grazie all'approccio statistico, è riuscito a raggiungere le finali della Western Conference e stabilire il record di tiri da tre punti in una sola stagione.

Un'innovazione importante apportata nel basket dall'analisi dei dati è stata, come detto nel paragrafo 1.2.1, l'incremento dei tentativi di tiro da tre punti. Infatti, grazie ad analisi approfondite le squadre hanno adottato una nuova strategia basata sulla convenienza statistica dei tiri da tre punti. I dati hanno dimostrato che, pur essendo la percentuale di successo dei tentativi da tre punti inferiore, il valore superiore di ogni canestro rispetto ai tentativi da due punti in prossimità della linea da tre punti compensa la differenza in modo statisticamente vantaggioso. È stato analizzato inoltre che, in media, le squadre che facevano maggiore affidamento sui tiri da tre punti rispetto alle avversarie, avevano maggiori probabilità di vincere la partita, in quanto tendevano a segnare complessivamente più punti durante la partita. È possibile notare l'adozione di questa nuova strategia osservando la differenza di medie di tentativi nel tempo: nel 2012 le squadre presentavano una media di circa 18 tentativi da tre punti a partita, mentre nel 2017 questo numero è salito a 27 tentativi totali a partita.

Un altro aspetto fondamentale ottimizzato dall'utilizzo dei dati è la preparazione tattica della partita: analizzando le prestazioni passate delle squadre avversarie e le loro tattiche è possibile valutare i loro

[6] https://www.culturedigitali.org/analytics-e-sport-dati-prestazioni-sportive/

punti di forza e debolezza e per gli allenatori realizzare strategie mirate a sfruttare le vulnerabilità degli avversari, ottenendo un vantaggio tattico importante.

Ad esempio, è possibile analizzare in quali situazioni un difensore ha avuto più o meno successo contro varie tipologie di attacchi offensivi e capire quali difensori abbiano determinate caratteristiche per impedire con più probabilità un tentativo

avversario, oppure cercare di individuare situazioni in cui un giocatore della propria squadra eccella in un particolare aspetto del gioco contro un determinato difensore avversario meno efficace.

Oltre che per la preparazione tattica della partita, l'analisi dei dati viene utilizzata anche per la gestione generale della squadra, tenendo traccia di informazioni utili come i minuti giocati dai singoli atleti per ottimizzare la gestione, evitando il sovraccarico di lavoro per ciascun giocatore e garantendogli il necessario riposo. Un esempio di questo può essere mostrato con lo slogan "Strength in numbers" dei Golden State Warriors, squadra NBA nella Western Conference, che più che uno slogan è una vera e propria filosofia di gioco basata su questo concetto, la forza della squadra si basa su una strategia volta a far riposare i giocatori chiave per prevenire infortuni e fare in modo che tutti i componenti della squadra diano il proprio contributo per la vittoria.

L'analisi dei dati nel basket ha portato anche all'adozione di tecnologie avanzate per la raccolta e la visualizzazione dei dati, con la collaborazione tra l'NBA e la società STATS che ha portato all'installazione di sei telecamere in ogni arena di basket, monitorando i movimenti dei giocatori a 25 fotogrammi al secondo, in modo tale da tracciare i punti di forza e debolezza dei giocatori, offrendo ad esempio mappe dettagliate delle aree in cui un giocatore tira con maggiore precisione, monitorando anche l'andamento delle prestazioni individuali. Con queste tecnologie le azioni di gioco possono essere registrate e analizzate consentendo, come detto precedentemente, di personalizzare le proprie strategie difensive ed offensive in base all'avversario e valutare le aree in cui ciascun giocatore può migliorare.

Negli Stati Uniti, l'NBA è nota per essere all'avanguardia nell'utilizzo dell'analisi dei dati, con squadre che investono pesantemente in tecnologie avanzate e analisi statistiche per prendere decisioni tattiche e strategiche. L'Italia, e l'Europa in generale, per il momento non possono competere con quanto sta avvenendo in America; tuttavia, sta emergendo un sempre più crescente interesse in questo settore, per cercare di colmare il divario esistente con l'NBA. In una intervista[7] per il Corriere della

[7] https://www.corriere.it/tecnologia/22_giugno_04/gallinari-belinelli-dai-sensori-metaverso-vi-raccontiamo-nuovo-basket-dd27ce8a-e0f0-11ec-a138-4bfa3d154041.shtml

sera, documentata in data 4 Giugno 2022, Marco Belinelli, guardia della Virtus Bologna, ha affermato come nonostante la supremazia americana, anche l'Italia stia facendo importanti passi avanti, facendo riferimento alla sua attuale squadra. Belinelli ha evidenziato il fatto che, anche in Italia, alla Virtus Bologna, si sia sviluppato uno staff specializzato per l'analisi dei dati, composto da videomaker e data analyst, che hanno contribuito attraverso videoanalisi ad ottimizzare gli aspetti tecnici del gioco, come ad esempio miglioramenti per quanto riguarda gli angoli di tiro e movimenti senza palla. In Italia sono state introdotte importanti innovazioni tecnologiche anche per quanto riguarda la preparazione atletica. Matteo Panichi, preparatore fisico e responsabile delle performance sia della Virtus Bologna che della Nazionale italiana, nell'intervista per il Corriere della sera ha illustrato l'adozione di strumenti avanzati come dinamometri, accelerometri e cardiofrequenzimetri per valutare la forza, asimmetrie muscolari, la velocità e la qualità muscolare dei giocatori. Inoltre, ha affermato come l'integrazione di app e software collegati alle videocamere abbia reso più semplice scaricare e confrontare i dati, offrendo analisi più approfondite delle performance dei giocatori.

In breve, la tecnologia e l'analisi dei dati stanno diventando sempre più centrali nel modo in cui le squadre si preparano e competono nel mondo dello sport, offrendo una vasta gamma di informazioni utili per prendere decisioni più accurate, ottimizzare le prestazioni e garantire anche per i tifosi appassionati un'esperienza più avvincente. È utile analizzare nel paragrafo 1.3 questo sport, per ottenere una panoramica generale di come questo funzioni e delle competizioni principali che lo caratterizzano.[8]

1.3. LE REGOLE DEL BASKET

Le regole del basket sono numerose e complesse, perciò in questo paragrafo ci si limiterà a presentare in modo chiaro i principi fondamentali su cui si basa questo sport, per una comprensione generale. In breve, una partita di basket, o pallacanestro, consiste nel cercare di segnare punti lanciando la palla nel canestro avversario durante la fase offensiva e allo stesso tempo, in fase difensiva, impedire all'altra squadra di fare lo stesso.

Nel mondo del basket, possiamo dire che esistano tre elementi chiave, ovvero la palla, il tabellone e il canestro. Secondo il regolamento tecnico della Federazione Italiana Pallacanestro (FIP), la palla da

[8] https://towardsdatascience.com/nba-data-analytics-changing-the-game-a9ad59d1f116

basket deve avere caratteristiche fisse per tutti i campionati, ma bisogna differenziare tra campionati di diverse nazioni, in America ad esempio si applicano norme e regole diverse. In Italia il pallone di gioco, di colore arancione, ha forma sferica di circonferenza che varia a seconda della tipologia di gara, dai 72,4 - 73,7 cm nelle partite femminili ai 74,9 - 78 cm in quelle maschili e può essere realizzata in diversi materiali come cuoio, gomma o materiale sintetico. Per quanto riguarda il peso, questo può variare tra i 510-567 grammi per le donne e i 567-650 grammi per gli uomini. Ci sono anche indicazioni relative alla pressione che deve avere la palla, che viene lasciata cadere da un'altezza di 1,8 metri e ci si aspetta che rimbalzi a un'altezza compresa tra gli 1,2 e 1,4 metri dalla superficie del campo da gioco.

I tabelloni invece, altro elemento fondamentale in campo, sono installati su supporti posizionati al di fuori delle linee che delimitano l'area di gioco. A questi tabelloni sono ancorati i canestri, composti da un anello di ferro posizionato ad un'altezza da terra di 3,05 metri dalla superficie di gioco, con un diametro di 45 cm al quale è attaccata una rete di corda bianca. Oltre a questi elementi, un cronometro di gara, spesso collegato a un tabellone luminoso su cui è possibile vedere, oltre al tempo rimanente, anche il punteggio della partita.

Una partita ufficiale si svolge all'interno di un palazzetto dello sport, dove il soffitto deve trovarsi ad almeno 7 metri di altezza dal suolo, mentre il campo da gioco ha una forma rettangolare con una larghezza di 15 metri e una lunghezza di 28 metri, ma la Federazione Internazionale Pallacanestro (FIBA) può concedere deroghe. Il pavimento del campo può essere realizzato in vari materiali come legno, gomma o altri materiali sintetici, con linee spesso bianche di larghezza di 5 cm, ognuna con un ruolo e una funzione specifica, ad esempio la linea centrale divide il campo in due parti perfettamente uguali.

Una partita ha durata standard di 40 minuti, suddivisa in quattro periodi, anche detti quarti, ciascuno della durata di 10 minuti, mentre nell'NBA la durata totale si estende a 48 minuti, con ciascun quarto della durata di 12 min. Tra il secondo e il terzo periodo si tiene un intervallo più lungo, solitamente di 10-15 minuti, in cui le squadre cambiano metà campo, mentre alla fine di ogni altro periodo vi è un intervallo più breve di circa due minuti.

Per quanto riguarda la composizione delle squadre, oltre ai titolari in campo, che sono 5 per squadra, ognuna dispone di un numero variabile di riserve, solitamente tra 5 e 7 a seconda della tipologia di campionato a cui ci si riferisce, che possono entrare in campo senza alcuna restrizione, potendo entrare e uscire liberamente. Ci sono però alcune regole comportamentali che si riflettono sulle sostituzioni: accumulando 5 falli (o 6 nel caso dell'NBA), il giocatore viene sostituito e deve lasciare

il campo. Allo stesso modo, se un giocatore viene espulso, non può più partecipare alla partita. Se, a causa di falli, espulsioni o infortuni, una squadra si ritrova con un solo giocatore in campo, la partita termina automaticamente e la vittoria va all'altra squadra.

Per quanto riguarda le restrizioni temporali legate alle diverse fasi di gioco, ogni squadra ha un limite di tempo di 24 secondi, che inizia non appena prende possesso della palla, per completare un'azione offensiva, ovvero per effettuare un tiro a canestro, dopo questi è necessario effettuare un cambio di possesso palla, conseguentemente al quale si resettano i 24 secondi. Ci sono poi regole relative al possesso della palla, un giocatore non può trattenere la palla senza palleggiare per più di 5 secondi, inoltre è vietato restare nell'area della squadra avversaria per più di 3 secondi, per evitare situazioni di stallo accanto al canestro. Infine, una squadra non può trattenere la palla nella propria metà campo per più di 8 secondi, per garantire ritmi più elevati, propositivi e coinvolgenti per giocatori e spettatori.

Per quanto riguarda altre infrazioni, ovvero momenti in cui il gioco si interrompe momentaneamente, le principali sono relative all'uscita del pallone o del giocatore dal campo di gioco, oppure se il giocatore, concluso il palleggio, afferra la palla con due mani e ricomincia nuovamente a palleggiare (doppio palleggio). Inoltre, avviene infrazione se un giocatore nel palleggiare porta la palla a fermarsi momentaneamente, portando la mano al di sotto del pallone (palla accompagnata), oppure quando un giocatore muove il piede perno dopo che ha concluso un palleggio (passi), oppure salta con il possesso di palla e atterra prima che questa abbia lasciato la sua mano o effettua più di due passi senza effettuare palleggi.

Infine, ci sono le infrazioni riguardanti l'anello o il canestro in generale, che si verificano ad esempio quando, durante un tiro a canestro, un giocatore tocca la palla mentre questa è completamente al di sopra dell'altezza dell'anello e sta seguendo una traiettoria discendente oppure ha già toccato il tabellone. Se questa violazione è commessa dall'attaccante, l'azione si ferma e si concede una rimessa agli avversari, mentre se è un difensore a compiere l'interferenza, si assegna al tiro il punteggio che avrebbe ottenuto se il canestro fosse stato realizzato. È vietato inoltre toccare la retina del canestro o il tabellone durante un tiro verso il canestro, se un difensore commette questa infrazione, il canestro viene considerato valido.

Per quanto riguarda i falli, esistono falli su tiro, ovvero quando il difensore ostacola in maniera irregolare il tiro dell'attaccante, falli di sfondamento, quando l'attaccante colpisce un difensore fermo durante l'azione di attacco, oppure falli antisportivi, contatti fallosi di un giocatore che, a discrezione dell'arbitro, non possono essere considerati un legittimo tentativo di giocare la palla. Due falli antisportivi comportano l'espulsione, così come l'aver commesso cinque falli.

Il punteggio nel basket è assegnato in modo diverso a seconda del tipo di canestro: un tiro libero, simile al calcio di rigore nel calcio, posto a distanza di 3,6 metri dal canestro, vale un punto. Un canestro realizzato all'interno dell'area delimitata dalla linea dei 3 punti è valutato 2 punti, mentre se il tiro viene eseguito al di fuori di questa area ed entrambi i piedi del tiratore non toccano la linea, vale 3 punti. Alla fine del quarto periodo di gioco, la squadra con il punteggio più alto vince la partita. In caso di parità, si procede a un tempo supplementare della durata di cinque minuti, e se la situazione di parità persiste ci saranno ulteriori tempi supplementari[9].

1.3.1. LE COMPETIZIONI

Il basket è caratterizzato da numerose competizioni internazionali e nazionali che coinvolgono squadre appartenenti ad ogni parte del mondo. Tra le competizioni di basket più prestigiose e seguite al mondo, spicca innanzitutto l'NBA, la National Basketball Association, considerata la regina indiscussa del basket, campionato americano celebre per partite spettacolari e giocatori leggendari che hanno fatto la storia di questo sport.

A livello europeo, molto importante, anche se non paragonabile alla maestosità dell'NBA, rimane l'Eurolega, che sta guadagnando sempre più appassionati della pallacanestro in Europa, una competizione alla quale partecipano 18 club che fanno parte della FIBA Europe (Federazione Internazionale Pallacanestro). Di questi 18 club, 11 hanno un posto fisso nella competizione per prestigio, mentre gli altri 7 accedono annualmente in base ai risultati ottenuti.

Altra competizione amata è l'Eurocup, che prevede la partecipazione delle 20 squadre più forti provenienti dai vari campionati nazionali europei, divise in due gironi da 10 squadre. Le ultime due squadre per ogni girone vengono eliminate, mentre ottavi, quarti, semifinali e finale si giocano con gara secca.

In Europa inoltre è stata creata nel 2016 la Champions League di pallacanestro, in cui si sfidano le 32 squadre più forti che non partecipano all'Eurolega o all'Eurocup, 24 qualificate direttamente e 8 a seguito di turni preliminari. Ci sono quattro gironi composti da otto squadre, in cui dopo partite andata e ritorno le prime quattro accedono ai playoff, mentre la quinta e la sesta partecipano alla Eurocup.

[9] https://it.wikipedia.org/wiki/Pallacanestro

A livello nazionale, le migliori competizioni in cui vengono offerte partite di maggior livello sono rappresentate dalla Bundesliga (Germania), dalla Liga ACB (Spagna), dalla Turkish Basketball Super League (Turchia), dalla Greek League (Grecia) e dalla Lega Basket Serie A (Italia).

Per quanto riguarda il nostro paese, la pallacanestro italiana è cresciuta molto negli ultimi anni, con la Lega Basket Serie A (LBA) e la Lega Basket Serie A2[10].

1.3.2. FOCUS SULL'NBA

La NBA, acronimo di National Basketball Association, rappresenta la principale lega professionistica di pallacanestro negli Stati Uniti. Per ogni giocatore di basket, conquistare il titolo NBA o anche solo avere l'opportunità di giocare per una delle squadre della competizione rappresenta il traguardo più ambito. Le squadre partecipanti sono suddivise in due conference, la Western e la Eastern, ciascuna delle quali è ulteriormente frammentata in 3 divisioni, ognuna delle quali contiene 5 squadre.

Nella Western Conference si trovano:

- Nella Northwest Division: Utah Jazz, Portland Trail Blazers, Denver Nuggets, Oklahoma Thunder, Minnesota Timberwolves.
- Nella Southwest Division: San Antonio, Memphis Grizzlies, Dallas Mavericks, New Orleans Pelicans, Houston Rockets.
- Nella Pacific Division**:** Phoenix Suns, Los Angeles Lakers, Los Angeles Clippers, Golden State Warriors, Sacramento Kings.

Nella Eastern Conference si trovano:

- Nell'Atlantic Division: Philadelphia 76ers, Brooklyn Nets, Boston Celtics, New York Knicks, Toronto Raptors.
- Nella Central Division: Milwaukee Bucks, Indiana Pacers, Chicago Bulls, Cleveland Cavaliers, Detroit Pistons.

[10] https://it.wikipedia.org/wiki/Pallacanestro

- Nella Southeast Division: Miami Heat, Charlotte Hornets, Atlanta Hawks, Washington Wizards, Orlando Magic.

Per quanto riguarda la competizione, questa si articola in tre fasi chiave: la stagione regolare (regular season), i playoff e le fasi finali (Finals). Durante la regular season, che ha inizio a fine ottobre, ogni squadra disputa 82 partite, 4 partite contro ognuna delle altre 4 squadre della stessa division (per un totale di 16 partite), 3 o 4 partite contro ciascuna delle altre 10 squadre della stessa conference (per un totale di 36 partite) e 2 partite contro ogni squadra dell'altra conference (30 partite). Per quanto riguarda la classifica NBA nella Regular Season si adotta un sistema basato sulla percentuale di vittorie, anziché sull'assegnazione di punti per ogni risultato, come invece avviene nel caso della Lega Basket Serie A italiana. Questa percentuale, nota come "winning percentage" (PCT), deriva dal rapporto tra numero di vittorie e partite giocate. Al termine della stagione regolare, le prime otto squadre di ciascuna conference avanzano ai playoff, che si giocano al meglio delle sette gare e che culminano nelle Finals, in cui si sfidano i campioni di ciascuna conference, sempre al meglio delle sette gare[11].

1.3.3. LE DIFFERENZE TRA NBA E FIBA

Quando si parla di basket, per quanto riguarda gli ambiti NBA e FIBA, emerge una importante distinzione non solo nelle filosofie di gioco ma anche nei regolamenti, che pur condividono molti aspetti. Le principali differenze tra la NBA e la FIBA sono evidenti sia nelle regole sia nelle dinamiche di gioco.

Innanzitutto, come già detto, nella NBA ciascun quarto dura 12 minuti, con un tempo totale di 48 minuti, mentre in Europa sono di 10 minuti per 40 minuti totali di gioco. Per quanto riguarda i falli, un giocatore in NBA può commetterne sei prima di essere espulso, mentre nella FIBA il limite è di cinque falli.

Altre differenze riguardano invece le dimensioni del campo, con il campo NBA leggermente più grande (28,65m x 15,24m) rispetto a quelli Europei (28m x 15m), e la distanza della linea del tiro da

[11] https://it.wikipedia.org/wiki/NBA

tre punti, che mentre in FIBA è uniforme a 6,75 metri dal canestro in qualsiasi punto della linea, nella competizione americana di trova a 7,24 metri frontalmente e 6,70 metri lateralmente.

Per quanto riguarda i time-out, nella NBA ogni squadra ne ha a disposizione sette durante i 48 minuti di gioco, che possono essere richiamati anche dai giocatori durante l'azione e durano 75 secondi ciascuno, mentre in FIBA, solo l'allenatore può chiamare il time-out, due nel primo tempo e tre nel secondo.

Altra differenza chiave riguarda l'interferenza a canestro: nella NBA, una volta che la palla tocca il ferro e rimane sopra il canestro, nessun giocatore può toccarla, mentre secondo le regole FIBA, la palla può essere rigiocata dopo aver toccato il ferro, consentendo all'attaccante di segnare facilmente o al difensore di spazzarla via[12].

Nel corso di questo capitolo, è stata realizzata una panoramica generale sui big data e sull'importanza dell'analisi statistica, con un focus specifico sul contesto sportivo, in particolare la pallacanestro. Si è osservato che, nel mondo dello sport in generale, la raccolta e l'interpretazione dei dati sono diventati strumenti indispensabili per comprendere al meglio le dinamiche di gioco ed avere più opzioni, per ottenere una maggiore probabilità di successo. In particolare, è stata posta l'attenzione sul basket e sull'NBA, con una esplorazione delle regole fondamentali e dei meccanismi generali per garantire una panoramica generale ed una base per il prossimo capitolo. Nel capitolo successivo si procederà all'applicazione di metodologie statistiche ad un database creato appositamente per la presente relazione finale e contenente le statistiche dei giocatori NBA del 2022.

[12] https://it.wikipedia.org/wiki/NBA

CAPITOLO 2: ANALISI STATISTICA DEI GIOCATORI NBA

In questo capitolo si andrà ad analizzare un database creato appositamente per il progetto che si descriverà nel terzo capitolo. Dopo aver realizzato una panoramica generale sulla statistica descrittiva definito i principali indicatori riassuntivi, come ad esempio la moda, la mediana, i quantili e la media, si calcoleranno questi indici per alcuni giocatori NBA relativamente a quattro variabili fondamentali: minuti, punti, assist e rimbalzi.

In seguito, si andranno a calcolare altri indicatori statistici considerando altre variabili non presenti nel primo database. Si utilizzerà un secondo database creato scaricando i dati da Basketball Reference e modificato attraverso operazioni di pulizia e miglioramento dati.

2.1. DESCRIZIONE DEL PRIMO DATABASE

In questo capitolo ci si concentrerà sull'applicazione di statistiche di base e indici di sintesi ad un database appositamente creato, dopo aver illustrato il processo di creazione di questo database, che sarà fondamentale soprattutto per il progetto che verrà mostrato nel terzo capitolo. Il database è formato da un file riassuntivo generale e da 30 file Excel distinti, ognuno dedicato a una squadra partecipante alla NBA, come visibile nella Figura 1.

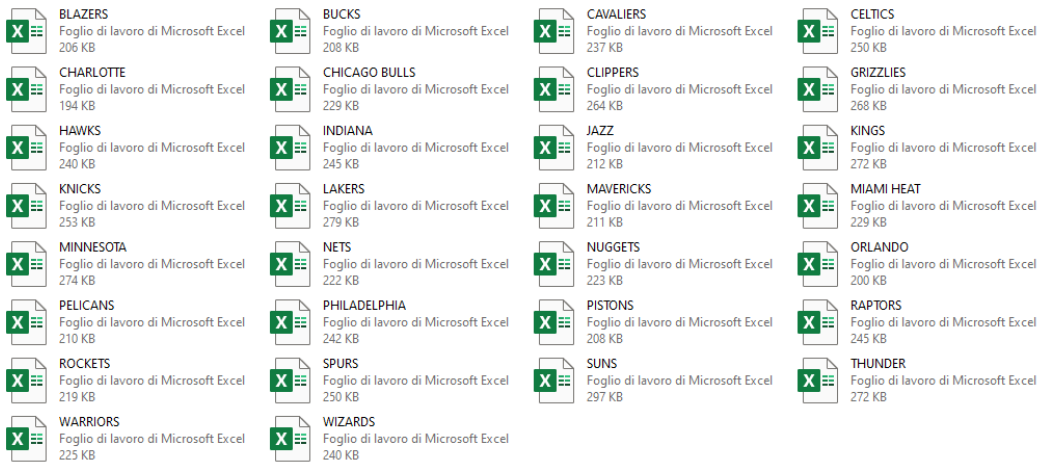
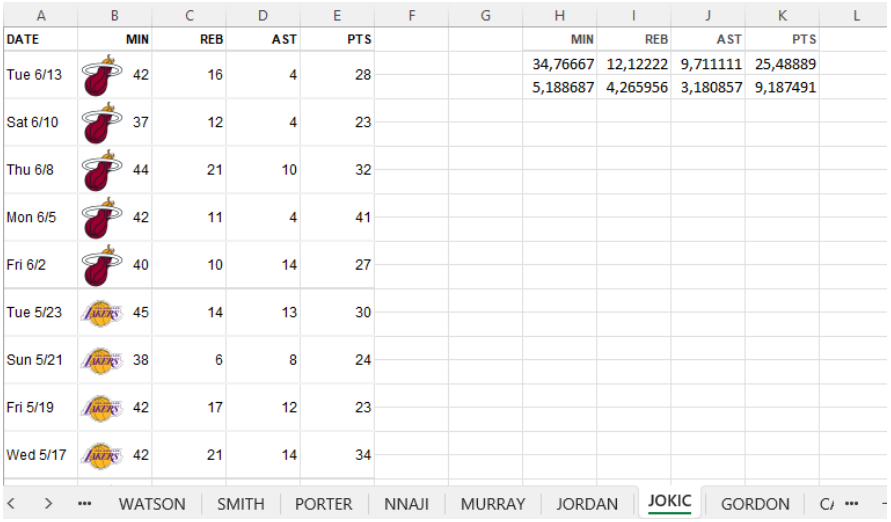


Figura 1: Lista dei file Excel appartenenti a ciascuna squadra NBA

Ogni database di squadra contiene un numero variabile di schede, in base al numero di giocatori presenti nel roaster, ovvero nella lista dei giocatori appartenenti alla società sportiva, ed ogni scheda registra i dati di minuti, assist, punti e rimbalzi relativi a ciascun giocatore, partita per partita, durante la regular season del 2022-2023 (quindi eventualmente 82 partite in caso non ne sia stata saltata

nessuna per infortuni o espulsioni) e durante i playoff e nella finale nel caso in cui la squadra abbia proseguito il proprio percorso, aumentando così il numero di osservazioni totali giocatore per giocatore.



DATE	MIN	REB	AST	PTS			MIN	REB	AST	PTS
							34,76667	12,12222	9,711111	25,48889
							5,188687	4,265956	3,180857	9,187491
Tue 6/13	42	16	4	28						
Sat 6/10	37	12	4	23						
Thu 6/8	44	21	10	32						
Mon 6/5	42	11	4	41						
Fri 6/2	40	10	14	27						
Tue 5/23	45	14	13	30						
Sun 5/21	38	6	8	24						
Fri 5/19	42	17	12	23						
Wed 5/17	42	21	14	34						

WATSON | SMITH | PORTER | NNAJI | MURRAY | JORDAN | JOKIC | GORDON

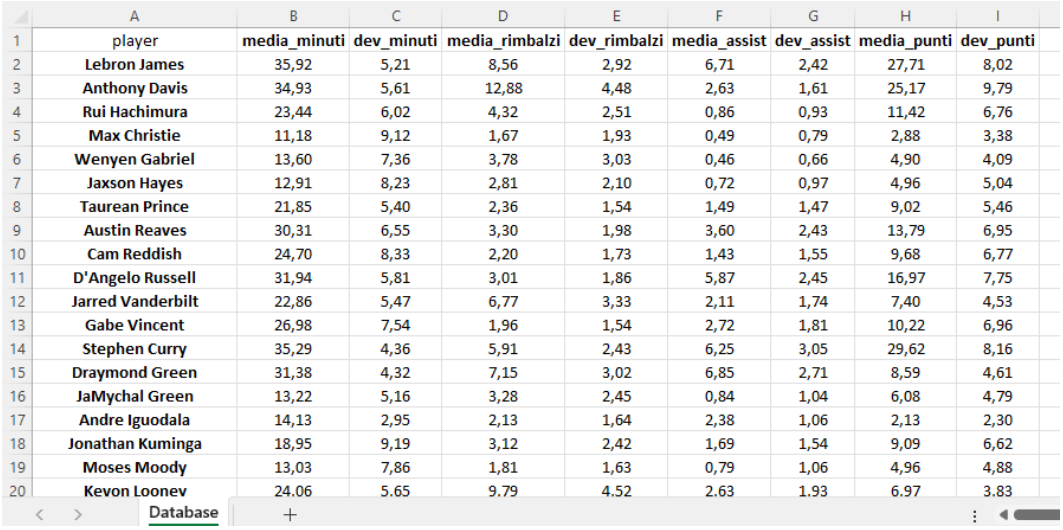
Figura 2: Osservazioni partita per partita di Jokic

Aprendo un file di squadra tra i 30 presenti, nel caso dell'esempio nella Figura 2 i "Denver Nuggets", vincitori della stagione NBA 2022-2023, è possibile osservare in basso le schede di ogni giocatore appartenente alla squadra con i rispettivi cognomi (Watson, Smith, Porter, Nnaji, Murray, Jordan, Jokic, Gordon nella Figura 2), e selezionando una tra queste schede (nell'esempio è stato selezionato Nikola Jokic), è possibile vedere le statistiche di minuti (MIN), rimbalzi (REB), assist (AST) e punti (PTS) per ogni partita giocata da Jokic (tra Regular season, Playoff e Finals presenta 90 partite giocate). Inoltre, nell'area H1:K3, è possibile trovare media e deviazione standard per ciascuna di queste variabili, i cui valori saranno utilizzati nel terzo capitolo.

Per quanto riguarda l'origine dei dati, sono stati estratti dal sito ESPN.com, andando su ciascuna squadra manualmente, selezionando ogni giocatore e copiando i dati relativi a ciascuna partita dalla sezione "game log", considerando solo le osservazioni relative alle variabili citate precedentemente. Prima di procedere con il calcolo degli indici di sintesi, è importante definire il significato di queste variabili:

- MIN (Minuti): Indica il numero di minuti che un giocatore ha trascorso in campo durante un'intera partita.
- REB (Rimbalzi): Registra il numero di volte in cui il giocatore ha recuperato il pallone dopo un tiro sbagliato. Si distinguono in rimbalzi offensivi (dopo un tiro sbagliato dalla stessa squadra) e rimbalzi difensivi (dopo un tiro sbagliato dall'avversario).

- AST (Assist): Rappresenta il numero di passaggi vincenti effettuati dal giocatore per mettere un compagno di squadra in condizione di segnare punti.
- PTS (Punti): Mostra il totale dei punti realizzati dal giocatore durante la partita.



	A	B	C	D	E	F	G	H	I
1	player	media_minuti	dev_minuti	media_rimbalzi	dev_rimbalzi	media_assist	dev_assist	media_punti	dev_punti
2	Lebron James	35,92	5,21	8,56	2,92	6,71	2,42	27,71	8,02
3	Anthony Davis	34,93	5,61	12,88	4,48	2,63	1,61	25,17	9,79
4	Rui Hachimura	23,44	6,02	4,32	2,51	0,86	0,93	11,42	6,76
5	Max Christie	11,18	9,12	1,67	1,93	0,49	0,79	2,88	3,38
6	Wenyen Gabriel	13,60	7,36	3,78	3,03	0,46	0,66	4,90	4,09
7	Jaxson Hayes	12,91	8,23	2,81	2,10	0,72	0,97	4,96	5,04
8	Taurean Prince	21,85	5,40	2,36	1,54	1,49	1,47	9,02	5,46
9	Austin Reaves	30,31	6,55	3,30	1,98	3,60	2,43	13,79	6,95
10	Cam Reddish	24,70	8,33	2,20	1,73	1,43	1,55	9,68	6,77
11	D'Angelo Russell	31,94	5,81	3,01	1,86	5,87	2,45	16,97	7,75
12	Jarred Vanderbilt	22,86	5,47	6,77	3,33	2,11	1,74	7,40	4,53
13	Gabe Vincent	26,98	7,54	1,96	1,54	2,72	1,81	10,22	6,96
14	Stephen Curry	35,29	4,36	5,91	2,43	6,25	3,05	29,62	8,16
15	Draymond Green	31,38	4,32	7,15	3,02	6,85	2,71	8,59	4,61
16	JaMychal Green	13,22	5,16	3,28	2,45	0,84	1,04	6,08	4,79
17	Andre Iguodala	14,13	2,95	2,13	1,64	2,38	1,06	2,13	2,30
18	Jonathan Kuminga	18,95	9,19	3,12	2,42	1,69	1,54	9,09	6,62
19	Moses Moody	13,03	7,86	1,81	1,63	0,79	1,06	4,96	4,88
20	Kevon Looney	24,06	5,65	9,79	4,52	2,63	1,93	6,97	3,83

Database

Figura 3: File riassuntivo di medie e deviazioni standard per ogni giocatore NBA

I dati raccolti sono stati poi aggregati nel file generale (visibile nella Figura 3). Questo file verrà utilizzato nel terzo capitolo e contiene le medie e le deviazioni standard di queste variabili per ciascun giocatore, tenendo conto di un totale di 400 giocatori tra tutte le squadre.

Dopo aver raccolto tutte le informazioni sui giocatori, è possibile analizzare delle statistiche semplici descrittive sui microdati, calcolando gli indici di sintesi e visualizzando grafici per alcuni giocatori.

2.2. ESEMPI DI CALCOLO DEGLI INDICI DI SINTESI

Dopo aver illustrato il processo di realizzazione del database creato, come esempio pratico si va a calcolare questi indicatori chiave per alcuni giocatori iconici nell'era NBA dell'ultimo periodo, come Lebron James, Stephen Curry e Nikola Jokic.

INDICI DI SINTESI LEBRON JAMES (n = 72 partite)				
	MIN	REB	AST	PTS
Moda	37	8	6	21
Mediana	36	8	6	27
Media	35,92	8,56	6,71	27,71
1° quartile	33	7	5	21,75
3° quartile	39	10	9	33
2° decile	33	7	4,2	21
35° percentile	34	8	6	23
Minimo	14	1	3	13
Massimo	48	20	12	48
Campo di variazione	34	19	9	35
Differenza Interquartile	6	3	4	11,25
Differenza Interdecile	12	5	6,9	19
Deviazione Standard	5,17	2,90	2,41	7,96
Varianza	26,74	8,41	5,79	63,43
Coefficiente di variazione	0,14	0,34	0,36	0,29

Tabella 1: Indici di sintesi di Lebron James

Nella Tabella 1 si possono osservare gli indici di sintesi stimati per il primo giocatore, ovvero Lebron James. Di preciso, è stato selezionato il database relativo alla squadra "Los Angeles Lakers", sulla scheda relativa a Lebron James. Potendo osservare i dati per ciascuna partita (72 osservazioni totali) relativi a minuti, rimbalzi, assist e punti, è stato possibile calcolare i principali indici di sintesi, visibili nella figura stessa nella prima colonna.

Per quanto riguarda la moda, è stata calcolata utilizzando la funzione di Excel =MODA(), per la mediana è stata utilizzata la funzione =MEDIANA(), per la media =MEDIA(): Lebron James presenta una moda di 8 rimbalzi, una mediana di 8 rimbalzi, e una media di 8.56 rimbalzi per partita.

In relazione ai quantili, i quartili sono stati calcolati con la funzione =QUARTILE(), mentre i decili e percentili sono stati determinati con la funzione =PERCENTILE() utilizzando l'opportuna percentuale di riferimento. Nello specifico, Lebron James per quanto riguarda il numero di rimbalzi, presenta un primo quartile di 8 rimbalzi, quindi nel 25% delle partite ha effettuato al massimo 8

rimbalzi e nel rimanente 75% ha effettuato almeno 8 rimbalzi, mentre essendo ad esempio il terzo quartile di 10, nel 75% delle partite ha effettuato al massimo 10 rimbalzi e nel rimanente 25% ha realizzato almeno 10 rimbalzi.

Per quanto riguarda il minimo e massimo, sono stati rispettivamente calcolati con la funzione =MIN() e =MAX(), mentre il campo di variazione rappresenta differenza tra questi due valori. Lebron James ha effettuato come valore minimo 1 rimbalzo in una partita, mentre come valore massimo 20 rimbalzi, il campo di variazione risulta quindi essere pari a 19 rimbalzi.

In merito alla differenza interquartile, è stata calcolata come la differenza tra terzo quartile e primo quartile. Rappresenta l'ampiezza dell'intervallo centrale che contiene il 50% delle osservazioni ordinate, mentre la differenza interdecile (differenza tra nono e primo decile) rappresenta l'ampiezza dell'intervallo centrale che contiene l'80% delle osservazioni ordinate. Per quanto riguarda i rimbalzi, Lebron James presenta una differenza interquartile pari a 3 e una differenza interdecile pari a 5 rimbalzi.

Per quanto riguarda la deviazione standard, è stata calcolata utilizzando la funzione =DEV.ST.P(), mentre la varianza con =VAR.P(): Lebron James presenta nei rimbalzi una deviazione standard di 2.90 rimbalzi (varianza pari a 8.41). Il coefficiente di variazione è stato calcolato con il rapporto tra deviazione standard e media, avendo un coefficiente di variazione nei rimbalzi pari a 0.34, si può dire che la deviazione standard è pari al 34% della media e quindi la variabilità dei rimbalzi di Lebron James si può ritenere contenuta.

INDICI DI SINTESI STEPHEN CURRY (n = 69 partite)				
	MIN	REB	AST	PTS
Moda	37	6	5	33
Mediana	36	6	6	30
Media	35,29	5,91	6,25	29,62
1° quartile	33	4	4	24
3° quartile	38	7	8	33
2° decile	33	4	4	22,6
35° percentile	34	5	5	27
Minimo	22	1	1	12
Massimo	43	13	15	50
Campo di variazione	21	12	14	38
Differenza Interquartile	5	3	4	9
Differenza Interdecile	9,4	6,2	7,2	19,2
Deviazione Standard	4,33	2,41	3,03	8,10
Varianza	18,76	5,82	9,17	65,68
Coefficiente di variazione	0,12	0,41	0,48	0,27

Tabella 2: Indici di sintesi di Stephen Curry

Nella Tabella 2 è stato realizzato lo stesso processo analizzando i dati partita per partita di un giocatore chiave dei "Golden State Warriors", Stephen Curry, concentrandoci sul numero di punti segnati per ciascuna partita, con un totale di 69 partite registrate durante la stagione. Stephen Curry presenta una moda di 33 punti, il che significa che 33 è il valore più comune tra le partite giocate. La mediana è di 30 punti, mentre la media è di 29.62 punti.

Per quanto riguarda i quantili, il primo quartile è pari a 24 punti, il che indica che nel 25% delle partite, Stephen Curry ha segnato al massimo 24 punti, e che quindi nel rimanente 75% ha segnato almeno 24 punti, mentre il terzo quartile è pari a 33.

Il valore minimo di punti segnati da Stephen Curry in una partita è stato di 12 punti, mentre il massimo è stato di 50 punti, ottenendo un campo di variazione pari a 38 punti.

Il capitano dei Golden State Warriors presenta una differenza interquartile di 9 punti. La differenza interdecile, che rappresenta l'ampiezza dell'intervallo centrale che include l'80% delle osservazioni ordinate, è di 19.2 punti.

Infine, Stephen Curry nei punti presenta una deviazione standard di 8.10 e una varianza di 65.68, che misurano la dispersione dei dati in tutte le osservazioni. Il coefficiente di variazione per quanto riguarda i punti, è pari a 0.27. Questo significa che, in media, i punti segnati da Curry si discostano di circa 8.10 punti dal suo punteggio medio, mentre il coefficiente di variazione suggerisce che la variabilità dei suoi punti è relativamente bassa rispetto alla media.

INDICI DI SINTESI NIKOLA JOKIC (n = 90)				
	MIN	REB	AST	PTS
Moda	35	12	10	24
Mediana	35	12	10	25,5
Media	34,77	12,12	9,71	25,49
1° quartile	33	10	8	19
3° quartile	38	14,75	12	31
2° decile	32	9	7	18
35° percentile	34	10	9	23
Minimo	20	4	3	4
Massimo	45	27	17	53
Campo di variazione	25	23	14	49
Differenza Interquartile	5	4,75	4	12
Differenza Interdecile	15	11,1	8	25
Deviazione Standard	5,16	4,24	3,16	9,14
Varianza	26,62	18,00	10,01	83,47
Coefficiente di variazione	0,15	0,35	0,33	0,36

Tabella 3: Indici di sintesi di Nikola Jokic

Infine, nella Tabella 3 è possibile esaminare gli indici di sintesi di Nikola Jokic, giocatore dei Denver Nuggets nominato MVP (most valuable player) dell'NBA nella stagione 2021-2022 ed uno dei pretendenti per il titolo MVP della stagione 2022-2023. Per quanto riguarda gli assist, in 90 partite osservate, Jokic presenta una moda di 10, una mediana di 10 e una media aritmetica pari a 9.71 assist per partita.

Per quanto riguarda i quantili, ad esempio, presenta un terzo quartile pari a 12 assist, ciò significa che nel 75% delle partite ha effettuato al massimo 12 assist, mentre nel rimanente 25% ha effettuato almeno 12 assist. Il valore minimo di assist effettuati in una partita di Jokic è pari a 3, mentre ha realizzato come massimo 17 assist, presentando quindi un campo di variazione di 14 assist.

Nikola Jokic presenta una differenza interquartile di 4 assist e una differenza interdecile pari a 8 assist. Infine, negli assist, presenta una deviazione standard di 3.18, una varianza di 10.12 e un coefficiente di variazione pari a 0.33. La deviazione standard suggerisce che gli assist di Jokic variano in media di circa 3.18 assist intorno alla media, mentre il coefficiente di variazione testimonia una variabilità contenuta dei dati.

2.2.1. BOX-PLOT GIOCATORI

Il box-plot è un tipo di grafico utilizzato per visualizzare distribuzioni di dati e identificare caratteristiche chiave delle distribuzioni, come la mediana, i quartili e la presenza di outliers, ovvero valori che si discostano dai restanti dati dell'insieme. Si andrà di seguito a visualizzare i box-plot relativi alle distribuzioni analizzate attraverso gli indici di sintesi, ovvero minuti, rimbalzi, assist e punti per partita per ogni giocatore citato nel paragrafo precedente, ovvero Lebron James, Nikola Jokic e Stephen Curry, dopo aver spiegato come funziona un box-plot e come si realizza.

Un box-plot è caratterizzato da diversi elementi chiave:

- La linea centrale nella scatola rappresenta la mediana dei dati, che divide la distribuzione in due parti uguali.
- La parte inferiore e superiore della scatola mostrano rispettivamente il primo quartile (25% della distribuzione) e il terzo quartile (75%). Questi quartili definiscono la differenza interquartile (IQR).
- Le linee che si prolungano dalla scatola, anche chiamate baffi si estendono per 1.5 volte l'IQR dalla parte superiore e inferiore. I dati al di fuori dei baffi, visualizzati come punti separati, rappresentano gli outliers, ovvero dati inusuali.

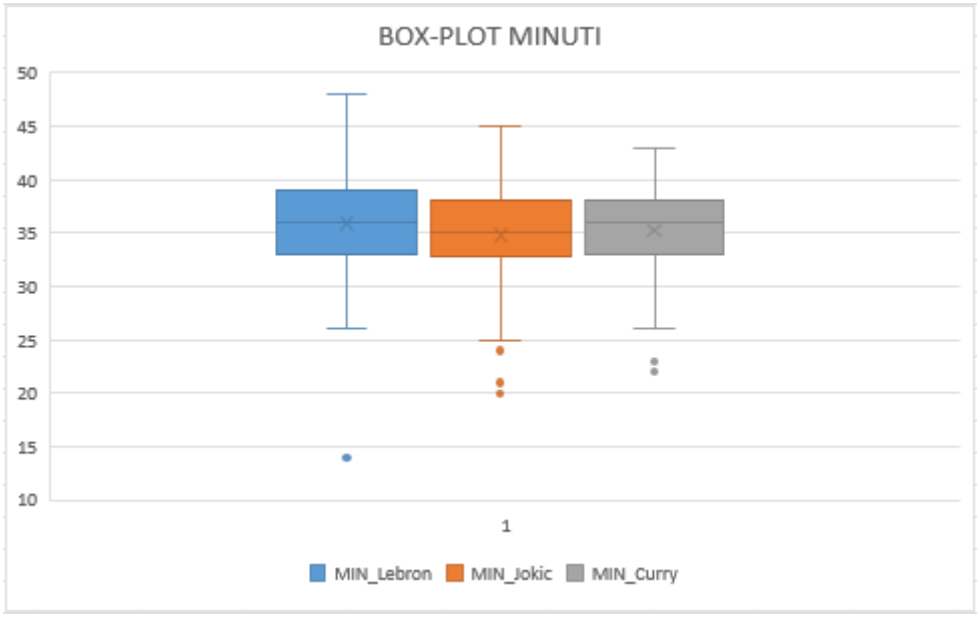


Figura 4: Box-Plot dei minuti di Lebron, Jokic e Curry

Nella Figura 4 è possibile visualizzare il box-plot relativo alla distribuzione di minuti giocati partita per partita dai giocatori menzionati precedentemente. Questo grafico fornisce una rappresentazione visiva della distribuzione relativa ai minuti giocati, mostrando per l'appunto il primo quartile, la mediana, il terzo quartile ed eventuali valori anomali. È importante notare che i valori rappresentati nel box-plot sono gli stessi calcolati nel paragrafo precedente attraverso gli indici di sintesi. In generale, tutti e tre i giocatori hanno una distribuzione di minuti simile, con mediane e quartili simili. Lebron James ha la differenza interquartile (IQR) più ampia (6 minuti), per questo presenta una scatola di dimensioni leggermente maggiori rispetto a Curry e Jokic che hanno entrambi una IQR pari a 5. Sono presenti per tutti e tre i giocatori degli outliers, valori eccezionalmente piccoli in questi casi. Per quanto riguarda Lebron James viene considerato valore inatteso il suo minimo, ovvero 14 minuti. Curry presenta 2 outliers, ovvero il suo minimo (22 minuti) e una partita in cui ha totalizzato 23 minuti, mentre Jokic presenta 3 valori inattesi, ovvero il suo minimo (20 minuti) e due partite in cui ha giocato 21 e 24 minuti.

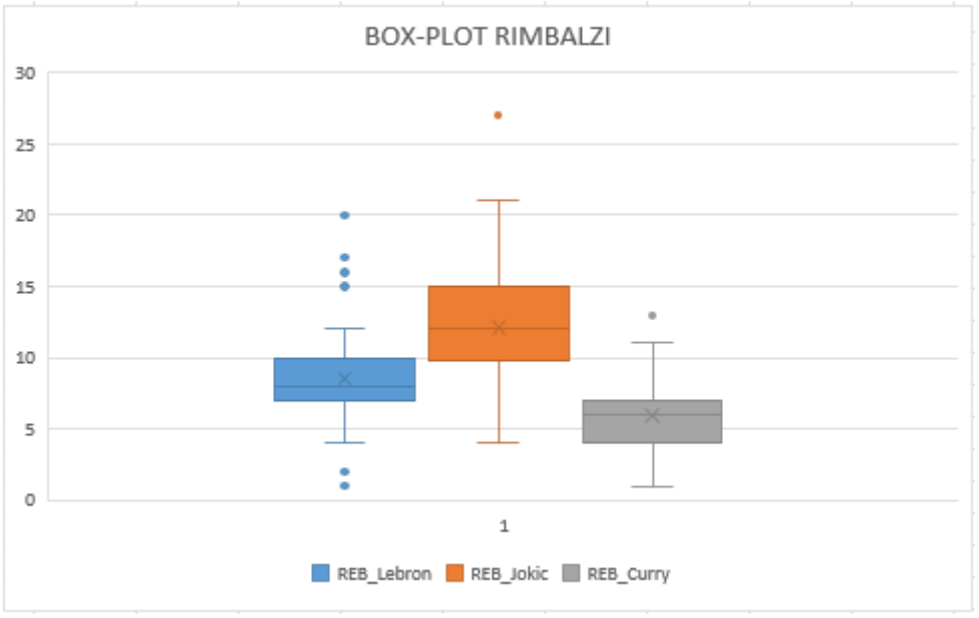


Figura 5: Box-Plot dei rimbalzi di Lebron, Jokic e Curry

Nella Figura 5 è presente il confronto tra i tre box-plot dei giocatori per quanto riguarda le loro distribuzioni di rimbalzi, da cui si possono vedere importanti differenze. Jokic presenta mediana (12 rimbalzi) e quartili di valore più elevato rispetto a Curry e Lebron, presentando una scatola che si colloca in una posizione più elevata. Per quanto riguarda l'IQR, Jokic presenta il range interquartile maggiore (4.75 rimbalzi) e ha una scatola di dimensioni maggiori rispetto agli altri due giocatori, che avendo IQR pari a 3 presentano una scatola più schiacciata. Lebron James è il giocatore che presenta più outliers, valori inattesi sia in positivo che in negativo, in quanto ha realizzato partite con prestazioni straordinarie rispetto alle aspettative per quanto riguarda i rimbalzi (partite in cui ha totalizzato 15,16,17 e 20 rimbalzi) e performance insolitamente sottotono rispetto ai suoi standard (partite in cui ha realizzato 1 e 2 rimbalzi). Stephen Curry presenta una distribuzione di rimbalzi che si colloca in una posizione più bassa rispetto agli altri due giocatori, a dimostrazione del fatto che le sue prestazioni nei rimbalzi sono tendenzialmente inferiori.

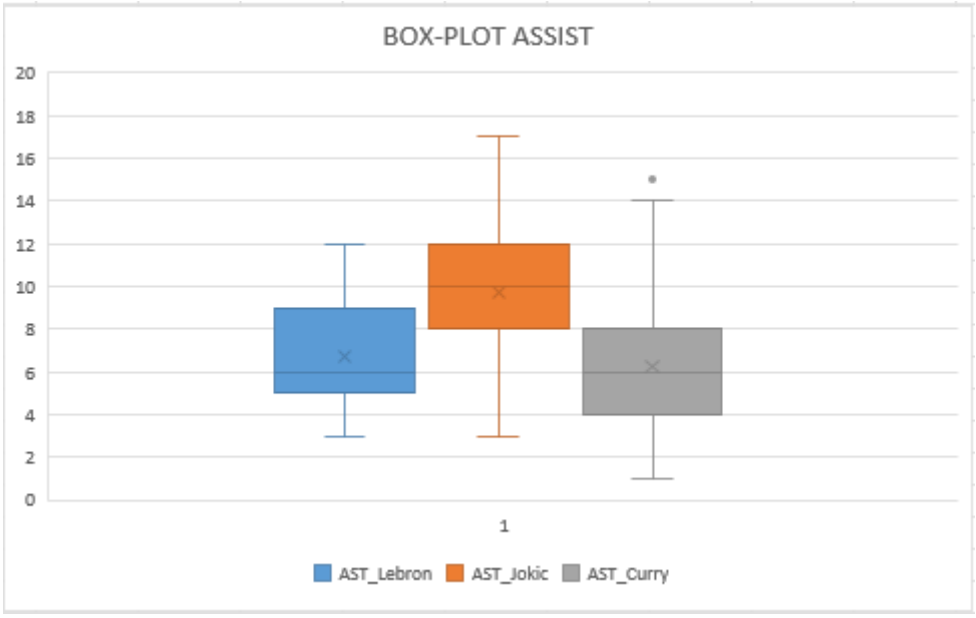


Figura 6: Box-Plot degli assist di Lebron, Jokic e Curry

Nella Figura 6 è presente un confronto tra i box-plot dei tre giocatori per quanto riguarda le loro distribuzioni di assist. Osservando la posizione dei box-plot è possibile notare immediatamente che la scatola di Jokic è posizionata più in alto rispetto a quella di Lebron e di Curry, dimostrando che la distribuzione dei suoi assist abbia tendenzialmente valori più alti. Per quanto riguarda la grandezza delle scatole, tutti e tre i giocatori presentano la medesima IQR (4 assist), quindi presentano le stesse dimensioni. Lebron e Curry presentano la stessa mediana (6 assist), ma osservando primo e terzo quartile è possibile notare che il box-plot di Stephen Curry si colloca verso valori più bassi, in quanto

le sue prestazioni in termini di assist sono tendenzialmente inferiori. Dal grafico non risultano particolari valori inattesi, l'unico outlier corrisponde con il valore massimo totalizzato da Curry, ovvero quando a seguito di una prestazione straordinaria rispetto ai suoi standard ha realizzato 17 assist in una partita.

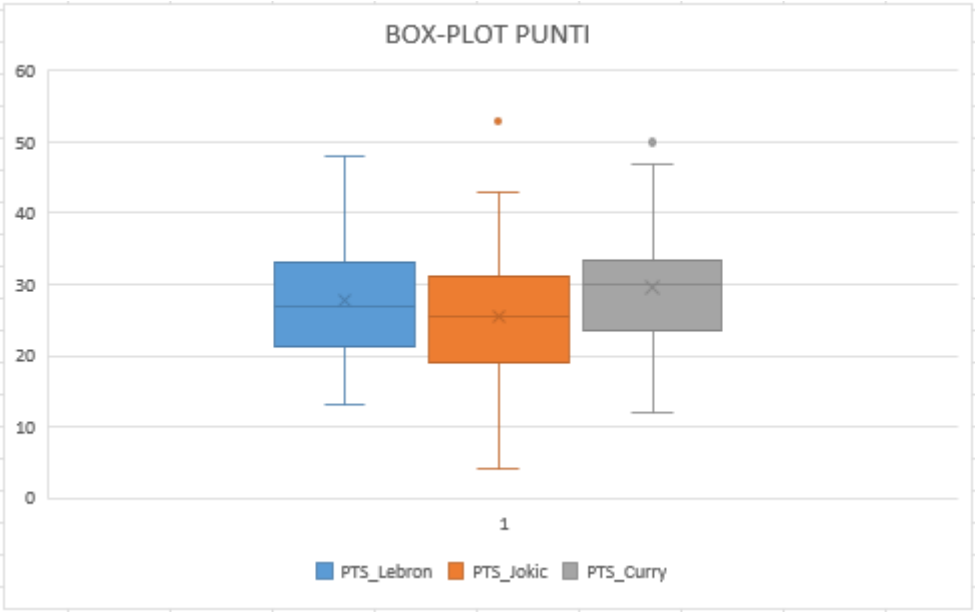


Figura 7: Box-Plot dei punti di Lebron, Jokic e Curry

Nella Figura 7 è presente un confronto tra i box-plot dei tre giocatori per quanto riguarda le distribuzioni di punti totalizzati. Il box-plot di Jokic mostra una mediana (25.5 punti) posizionata lievemente più in basso rispetto a Lebron e Curry, indicando una tendenza a segnare meno punti. Curry ha una distribuzione più concentrata intorno alla mediana rispetto ai due giocatori e presenta per questo, avendo la IQR minore (9 punti), una scatola più schiacciata. Jokic presenta invece la IQR maggiore (12 punti) e la sua scatola è caratterizzata da maggiori dimensioni, mostrando una maggiore variabilità nella parte centrale della distribuzione dei punti. Il valore di outlier che spicca di più riguarda Jokic, con una prestazione oltre alle aspettative in cui ha totalizzato 53 punti in una partita.

2.3. DESCRIZIONE DEL SECONDO DATABASE

Uno degli obiettivi di questa relazione finale è stato quello di individuare un altro database caratterizzato da ulteriori variabili in modo tale da poter condurre analisi statistiche più approfondite, ad esempio esaminando come il ruolo in campo dei giocatori di basket possa influire sulle loro prestazioni in campo.

I dati sono stati estratti dal sito *Basketball Reference* e si riferiscono alle statistiche dei giocatori NBA durante la regular season 2022-2023: per ogni giocatore sono presenti le medie calcolate di minuti, rimbalzi offensivi, rimbalzi difensivi, rimbalzi, assist, palle recuperate, stoppate, palle perse, falli fatti e punti. Questi sono stati sottoposti a operazioni di pulizia, con l'aggiornamento dei nomi delle squadre e dei ruoli in modo che fossero più comprensibili, esplicitando le sigle utilizzate dal sito. Inoltre, sono state selezionate solamente le variabili più rilevanti, per poter condurre analisi sulle medie suddivise per ruolo, in modo tale da poter calcolare ad esempio, considerando le distribuzioni di medie di palle perse di tutti i playmaker, la media generale di palle perse per i playmaker.

Le variabili aggiuntive che sono state considerate rispetto al primo database sono:

- Rimbalzi offensivi: rappresenta il numero di volte durante una partita in cui un giocatore in zona offensiva recupera il pallone dopo un tentativo di tiro errato da parte della propria squadra.
- Rimbalzi difensivi: mostra il numero di volte durante una partita in cui un giocatore in zona difensiva recupera il pallone successivamente ad un tentativo di tiro sbagliato da parte degli avversari.
- Palle recuperate: definisce il numero di volte in cui un giocatore recupera il possesso della palla dopo che la squadra avversaria ha compiuto un errore, in caso ad esempio di intercettazione o palla persa.
- Stoppate: esprime il numero di volte in cui un giocatore respinge il tiro di un avversario, impedendo al pallone di raggiungere il canestro.
- Palle perse: corrisponde al numero di volte in cui un giocatore perde il possesso della palla, a causa di errori come passaggi sbagliati o intercettazioni da parte della squadra avversaria.
- Falli fatti: riflette il numero di infrazioni personali commesse da un giocatore.

Nome	Ruolo	Età	Team	Game Giocati	Minuti	Rimbalzi offensivi	Rimbalzi difensivi	Rimbalzi	Assist	Palle recuperate	Stoppate	Palle perse	Falli fatti	Punti
Aaron Holiday	Playmaker	26	Atlanta Hawks	63	13,4	0,4	0,8	1,2	1,4	0,6	0,2	0,6	1,3	3,9
AJ Griffin	Ala Piccola	19	Atlanta Hawks	72	19,5	0,5	1,6	2,1	1	0,6	0,2	0,6	1,2	8,9
Bogdan Bogdanovic	Guardia Tiratrice	30	Atlanta Hawks	54	27,9	0,4	2,6	3	2,8	0,8	0,3	1,2	1,6	14
Bruno Fernando	Centro	24	Atlanta Hawks	39	10,4	1,4	2,1	3,5	0,8	0,2	0,9	0,6	1,9	3,9
Clint Capela	Centro	28	Atlanta Hawks	65	26,6	4	7,1	11,1	0,9	0,7	1,2	0,8	2,1	12
De'Andre Hunter	Ala Piccola	25	Atlanta Hawks	67	31,7	0,7	3,6	4,3	1,4	0,5	0,3	1,2	3	15,4
Dejounte Murray	Guardia Tiratrice	26	Atlanta Hawks	74	36,4	0,7	4,5	5,2	6,1	1,5	0,3	2,2	1,4	20,5
Garrison Mathews	Guardia Tiratrice	26	Atlanta Hawks	54	12,7	0,2	1,1	1,3	0,5	0,4	0,1	0,4	1,1	4,8
Jalen Johnson	Ala Piccola	21	Atlanta Hawks	70	14,9	0,7	3,3	4	1,2	0,5	0,5	0,6	1,6	5,6
Jarrett Culver	Guardia Tiratrice	23	Atlanta Hawks	10	13,7	1	2,8	3,8	0,6	0,6	0,2	0,7	1,4	4,4
John Collins	Ala Grande	25	Atlanta Hawks	71	30	1,1	5,4	6,5	1,2	0,6	1	1,1	3,1	13,1
Justin Holiday	Guardia Tiratrice	33	Atlanta Hawks	28	14,7	0,1	0,8	0,9	0,9	0,2	0,4	0,4	1,3	4,5
Onyeka Okongwu	Centro	22	Atlanta Hawks	80	23,1	2,7	4,5	7,2	1	0,7	1,3	1	3,1	9,9
Saddiq Bey	Ala Piccola	23	Atlanta Hawks	25	25,2	1,6	3,2	4,8	1,4	0,8	0	0,7	1,5	11,6
Saddiq Bey	Ala Piccola	23	Atlanta Hawks	77	27,6	1,3	3,4	4,7	1,5	0,9	0,2	0,9	1,6	13,8
Trae Young	Playmaker	24	Atlanta Hawks	73	34,8	0,8	2,2	3	10,2	1,1	0,1	4,1	1,4	26,2
Trent Forrest	Playmaker	24	Atlanta Hawks	23	12	0,2	1,4	1,6	1,7	0,3	0,1	0,7	0,7	2,3
Al Horford	Centro	36	Boston Celtics	63	30,5	1,2	5	6,2	3	0,5	1	0,6	1,9	9,8
Blake Griffin	Centro	33	Boston Celtics	41	13,9	1,1	2,6	3,7	1,5	0,3	0,2	0,5	1,8	4,1
Derrick White	Guardia Tiratrice	28	Boston Celtics	82	28,3	0,6	2,9	3,5	3,9	0,7	0,9	1,2	2,2	12,4
Grant Williams	Ala Grande	24	Boston Celtics	79	25,9	1,1	3,5	4,6	1,7	0,5	0,4	1	2,4	8,1
Jaylen Brown	Ala Piccola	26	Boston Celtics	67	35,9	1,2	5,7	6,9	3,5	1,1	0,4	2,9	2,6	26,6
Jayson Tatum	Ala Grande	24	Boston Celtics	74	36,9	1,1	7,7	8,8	4,6	1,1	0,7	2,9	2,2	30,1

Figura 8: Immagine di come si presenta il secondo database considerato

Il database comprende dati relativi a 678 giocatori NBA, indipendentemente dalla quantità di minuti di giocati, ma per evitare distorsioni nelle medie per ruolo dovute a giocatori con tempi di gioco limitati, si andranno ad, consentendo così di valutare le differenze nei dati raccolti. Per comprendere al meglio le valutazioni delle performance dei giocatori in relazione alle loro posizioni in campo è essenziale definire innanzitutto quali siano i ruoli che i giocatori di basket possono ricoprire:

- Playmaker (point guard): noto anche come regista, è il principale gestore del gioco di una squadra di basket, è responsabile della distribuzione della palla, dell'organizzazione dell'attacco e della creazione di opportunità per i compagni di squadra. Come caratteristiche chiave presenta solitamente visione di gioco e capacità di prendere decisioni sul campo.
- Ala piccola (small forward): giocatore versatile che contribuisce sia in attacco che in difesa, giocando solitamente nella zona esterna del campo. Questo ruolo viene assegnato a giocatori in grado di segnare punti, recuperare rimbalzi e difendere efficacemente.
- Guardia tiratrice (shooting guard): giocatore noto per la sua abilità nel tiro a lunga distanza. Questo ruolo è spesso assegnato a cestisti con ottime capacità di tiro e dribbling.
- Centro (pivot): solitamente il giocatore più alto e prestante fisicamente della squadra ed è incaricato di proteggere il canestro sia in difesa che in attacco. Questo ruolo è fondamentale per il gioco vicino al canestro ed è responsabile di prendere rimbalzi, bloccare tiri avversari e segnare punti sotto canestro.
- Ala grande (power forward): giocatore che combina le caratteristiche di un'ala piccola e di un centro, si posiziona generalmente nella parte alta del campo ed è abile nel gestire il pallone. Può essere impiegato sia vicino al canestro che lontano da esso[13].

Ora si procederà con l'analisi statistica legata a questo database, con l'obiettivo di focalizzarsi sull'influenza del ruolo dei giocatori sulle loro performance attraverso grafici di densità e medie per ruolo.

2.3.1. ANALISI PER RUOLO

Dopo aver descritto le variabili chiave all'interno del secondo database ed aver stabilito l'obiettivo dell'analisi da raggiungere in questo paragrafo, che consiste nell'osservare, attraverso le medie di performance, come la posizione in campo di ciascun giocatore incida sulle sue statistiche, è possibile procedere con le prime operazioni statistiche. Per calcolare le medie per ciascun ruolo di rimbalzi

[13] https://www.nike.com/it/a/posizioni-nel-basket

offensivi, rimbalzi difensivi, assist, palle recuperate, stoppate, palle perse, falli fatti e punti è stato sfruttato lo strumento della tabella pivot presente in Excel. Dopo aver selezionato il database, è stata creata la tabella pivot ed è stato inserito il campo "ruolo" nelle righe della tabella pivot. Le variabili d'interesse sono state posizionate nella sezione "valori" ed è stato impostato di riepilogare tali valori per media, in modo che mostrasse le prestazioni medie per ciascun ruolo.

La tabella 4 presenta le medie per ruolo delle variabili menzionate precedentemente, considerando tutti i 678 giocatori presenti nel database.

Ruoli	Numero di osservazioni	Media di RimbOff	Media di RimbDiff	Media di Rimbalzi	Media di Assist	Media di Palle Recuperate	Media di Stoppate	Media di Palle perse	Media di Falli fatti	Media di Punti
Ala Grande	117	1,02	3,16	4,18	1,59	0,55	0,42	1,03	1,74	8,84
Ala Piccola	125	0,74	2,39	3,13	1,54	0,60	0,27	0,91	1,61	8,83
Centro	134	1,60	3,59	5,19	1,31	0,44	0,71	1,00	1,96	8,10
Guardia Tiratrice	168	0,47	1,98	2,45	1,89	0,63	0,23	0,99	1,46	8,64
Playmaker	134	0,50	2,21	2,70	3,68	0,78	0,25	1,41	1,59	9,99
Totale	**678**	**0,84**	**2,62**	**3,47**	**2,01**	**0,60**	**0,37**	**1,07**	**1,66**	**8,87**

Tabella 4: Medie di alcune variabili d'interesse per ruolo di tutti i giocatori NBA

Con riferimento alla descrizione dei ruoli fornita in precedenza, è possibile notare che i playmaker presentano in media un maggior numero di assist rispetto agli altri ruoli, con una media di 3.68 assist a partita. Questo risultato è in linea con il loro ruolo di registi e principali gestori del gioco della squadra, che richiede una grande visione di gioco.

Inoltre, sempre considerando le caratteristiche dei vari ruoli, è possibile osservare che i centri hanno la media più alta di rimbalzi, con una media di 5.19 rimbalzi a partita, suddivisi tra 1.69 offensivi e 3.69 difensivi. Questo dato è influenzato dal fatto che spesso giocano vicino al canestro, sia in fase offensiva che difensiva. Presentano anche la media più alta di stoppate, con una media di 0.71 stoppate a partita, più del doppio rispetto ad altri ruoli, grazie all'altezza ed alla posizione in campo.

Tuttavia, l'analisi attraverso le medie può essere problematica, in quanto il database considera tutti i giocatori presenti, anche coloro che hanno giocato solo pochi minuti, che potrebbero influire negativamente sui valori delle medie in quanto non hanno avuto l'opportunità di avere un impatto significativo sulle prestazioni registrate. Per garantire dati affidabili, si è deciso di effettuare una ulteriore analisi depurata dall'effetto dei minuti. Per fare ciò sono state rapportate le variabili d'interesse ai minuti giocati per ogni giocatore, in modo da eliminare l'effetto dei minuti. Questa procedura ha consentito di ottenere statistiche "per minuto" che riducono l'effetto della disparità nella durata del tempo di gioco. Si può notare la differenza tra la Figura 8 e la Figura 9, che presenta i valori delle variabili d'interesse rapportati ai minuti giocati per ogni giocatore.

Solitamente, i valori “per minuto” vengono moltiplicati per 40 o per 48 nel caso dell’NBA, in modo da ottenere valori “per partita”, per fornire una stima più accurata delle performance dei giocatori per partita.

Il risultato di questa operazione è mostrato nella Figura 9.

Nome	Ruolo	Età	Team	Game Giocati	Minuti	Rimbalzi offensivi	Rimbalzi difensivi	Rimbalzi	Assist	Palle recuperate	Stoppate	Palle perse	Falli fatti	Punti
Aaron Holiday	Playmaker	26	Atlanta Hawks	63	13,4	0,03	0,06	0,09	0,10	0,04	0,01	0,04	0,10	0,29
AJ Griffin	Ala Piccola	19	Atlanta Hawks	72	19,5	0,03	0,08	0,11	0,05	0,03	0,01	0,03	0,06	0,46
Bogdan Bogdanovic	Guardia Tiratrice	30	Atlanta Hawks	54	27,9	0,01	0,09	0,11	0,10	0,03	0,01	0,04	0,06	0,50
Bruno Fernando	Centro	24	Atlanta Hawks	8	5,1	0,16	0,22	0,37	0,02	0,00	0,08	0,12	0,16	0,67
Bruno Fernando	Centro	24	Atlanta Hawks	39	10,4	0,13	0,20	0,34	0,08	0,02	0,09	0,06	0,18	0,38
Clint Capela	Centro	28	Atlanta Hawks	65	26,6	0,15	0,27	0,42	0,03	0,03	0,05	0,03	0,08	0,45
De'Andre Hunter	Ala Piccola	25	Atlanta Hawks	67	31,7	0,02	0,11	0,14	0,04	0,02	0,01	0,04	0,09	0,49
Dejounte Murray	Guardia Tiratrice	26	Atlanta Hawks	74	36,4	0,02	0,12	0,14	0,17	0,04	0,01	0,06	0,04	0,56
Donovan Williams	Guardia Tiratrice	21	Atlanta Hawks	2	2	0,00	0,50	0,50	0,00	0,00	0,00	0,00	0,00	1,00
Frank Kaminsky	Centro	29	Atlanta Hawks	26	6,8	0,01	0,19	0,21	0,12	0,03	0,00	0,04	0,12	0,40
Garrison Mathews	Guardia Tiratrice	26	Atlanta Hawks	9	9,3	0,01	0,12	0,13	0,03	0,01	0,01	0,01	0,08	0,52
Garrison Mathews	Guardia Tiratrice	26	Atlanta Hawks	54	12,7	0,02	0,09	0,10	0,04	0,03	0,01	0,03	0,09	0,38
Jalen Johnson	Ala Piccola	21	Atlanta Hawks	70	14,9	0,05	0,22	0,27	0,08	0,03	0,03	0,04	0,11	0,38
Jarrett Culver	Guardia Tiratrice	23	Atlanta Hawks	10	13,7	0,07	0,20	0,28	0,04	0,04	0,01	0,05	0,10	0,32
John Collins	Ala Grande	25	Atlanta Hawks	71	30	0,04	0,18	0,22	0,04	0,02	0,03	0,04	0,10	0,44
Justin Holiday	Guardia Tiratrice	33	Atlanta Hawks	28	14,7	0,01	0,05	0,06	0,06	0,01	0,03	0,03	0,09	0,31
Onyeka Okongwu	Centro	22	Atlanta Hawks	80	23,1	0,12	0,19	0,31	0,04	0,03	0,06	0,04	0,13	0,43
Saddiq Bey	Ala Piccola	23	Atlanta Hawks	25	25,2	0,06	0,13	0,19	0,06	0,03	0,00	0,03	0,06	0,46
Saddiq Bey	Ala Piccola	23	Atlanta Hawks	77	27,6	0,05	0,12	0,17	0,05	0,03	0,01	0,03	0,06	0,50
Trae Young	Playmaker	24	Atlanta Hawks	73	34,8	0,02	0,06	0,09	0,29	0,03	0,00	0,12	0,04	0,75
Trent Forrest	Playmaker	24	Atlanta Hawks	23	12	0,02	0,12	0,13	0,14	0,03	0,01	0,06	0,06	0,19
Tyrese Martin	Guardia Tiratrice	23	Atlanta Hawks	16	4,1	0,07	0,10	0,17	0,02	0,02	0,00	0,02	0,02	0,32
Vit Krejci	Playmaker	22	Atlanta Hawks	29	5,7	0,04	0,12	0,16	0,11	0,04	0,00	0,04	0,11	0,21

Figura 9: Immagine del database che considera i dati per minuto per i giocatori NBA

Successivamente, applicando lo stesso procedimento precedentemente descritto, è stata utilizzata la tabella pivot per calcolare le medie per ruolo, consentendo così una comparazione più equa delle prestazioni tra giocatori di diversi ruoli. È possibile vedere le medie per ruolo rapportate ai minuti nella Tabella 5.

Ruoli	Media di Rimbalzi offensivi	Media di Rimbalzi difensivi	Media di Rimbalzi	Media di Assist	Media di Palle recuperate	Media di Stoppate	Media di Palle perse	Media di Falli fatti	Media di Punti
Ala Grande	0,05	0,16	0,21	0,07	0,03	0,02	0,05	0,10	0,42
Ala Piccola	0,04	0,12	0,16	0,07	0,03	0,01	0,04	0,08	0,41
Centro	0,09	0,20	0,29	0,07	0,03	0,04	0,06	0,12	0,43
Guardia Tiratrice	0,03	0,11	0,14	0,09	0,04	0,01	0,05	0,08	0,43
Playmaker	0,02	0,10	0,13	0,16	0,04	0,01	0,06	0,08	0,42
Totale	**0,05**	**0,13**	**0,18**	**0,09**	**0,03**	**0,02**	**0,05**	**0,09**	**0,42**

Tabella 5: Medie per minuto per ruolo dei giocatori NBA

Dall’analisi emergono una serie di considerazioni, molte delle quali riflettono quelle effettuate nell’analisi precedente. I centri mantengono la leadership per quanto riguarda stoppate e rimbalzi, sia difensivi che offensivi, grazie alla loro altezza e posizione in campo. Inoltre, si conferma la supremazia dei playmaker per quanto riguarda gli assist, confermandosi come figure chiave nel gioco della squadra.

Da questa analisi si può notare inoltre una lieve superiorità per quanto riguarda i falli fatti per i centri rispetto agli altri ruoli, mentre le medie dei punti per minuto risultano simili per ogni ruolo, indicando l’assenza della distinzione precedente in cui i playmaker registravano un punteggio superiore rispetto agli altri ruoli. È rilevante anche la media dei rimbalzi difensivi delle ali grandi, più elevata rispetto

alla media delle altre posizioni (esclusi i centri che hanno il primato). Per quanto riguarda le restanti variabili, sembra che la posizione in campo dei giocatori non influisca in modo significativo, poiché le medie per minuto risultano molto simili tra i vari ruoli.

2.3.2. DIPENDENZA IN MEDIA

L'analisi della dipendenza in media tra due caratteri, di cui almeno uno deve essere quantitativo, si concentra sull'esaminare come la media del carattere quantitativo varia all'interno di gruppi differenti definiti dall'altro carattere. Si ha dipendenza in media quando le medie dei gruppi sono diverse tra loro. L'indicatore utilizzato per valutare l'intensità della dipendenza in media è il rapporto tra la varianza fra i gruppi e la varianza totale, noto come rapporto di correlazione di Pearson.

Si andrà ora ad analizzare la dipendenza in media tra ciascuna variabile d'interesse per minuto e il ruolo dei giocatori, per vedere se ci sono differenze dipendenti dal ruolo. Per effettuare questa analisi si sfrutta uno script del linguaggio di programmazione R. Si considerano le variabili d'interesse per minuto in quanto in questo modo si valutano efficientemente le performance, che potrebbero altrimenti essere influenzate di periodi di gioco ridotti o differenti.

Si analizza innanzitutto la dipendenza in media dei rimbalzi offensivi per minuto rispetto al ruolo. Una volta caricato il file Excel contenente i dati relativi alle variabili per minuto e impostata la directory di lavoro, si sfrutta questo script:

$$Y = dati\$Rimbalzi.offensivi$$

$$X = dati\$Ruolo$$

$$Medieparziali = tapply(Y, X, mean)$$

$$freq.rel = prop.table(table(X))$$

$$media = mean(Y)$$

Queste righe di codice permettono di calcolare le medie parziali dei rimbalzi offensivi per minuto per ogni ruolo, successivamente si calcolano le frequenze relative delle osservazioni di ciascun ruolo e si calcola la media totale di rimbalzi offensivi, senza considerare le posizioni dei giocatori.

$$var.fra = sum((Medieparziali - media)\text{^}2 * freq.rel)$$

$$Varianzeparziali = tapply(Y, X, function(x) var(x) * (length(x) - 1)/length(x))$$

$$var.nei = weighted.mean(Varianzeparziali, freq.rel)$$

$$Eta2 = var.fra \,/\, (var.fra + var.nei)$$

Si calcola successivamente la varianza fra i gruppi e la varianza nei gruppi e infine il rapporto di correlazione di Pearson, ovvero il rapporto tra varianza fra i gruppi e varianza totale.

Per quanto riguarda i rimbalzi offensivi per minuto, si ottiene un valore di Eta^2 pari a 0.409, il che significa che circa il 40% della varianza totale osservata nei rimbalzi offensivi per minuto può essere spiegata dalla variabilità tra i diversi ruoli dei giocatori.

Lo stesso procedimento è stato realizzato per tutte le variabili d'interesse.

Per i rimbalzi difensivi per minuto è stato calcolato un Rapporto di correlazione di Pearson pari a 0.334, quindi il grado di dipendenza in media del numero di rimbalzi difensivi per minuto dal ruolo del giocatore è discreto.

Per gli assist è stato calcolato un rapporto di correlazione pari a 0.361, il grado di dipendenza in media degli assist per minuto rispetto ai diversi ruoli è moderato e pari circa al 36%.

Per le palle recuperate è stato calcolato un rapporto di correlazione pari a 0.032, il grado di dipendenza in media delle palle recuperate per minuto rispetto al ruolo è contenuto e pari al 3.20% circa. Il fatto che il ruolo influisse relativamente poco sui valori di palle recuperate per minuto si poteva osservare anche nella Tabella 5 precedente, in quanto i vari ruoli presentavano medie simili.

Per le stoppate è stato calcolato un rapporto di correlazione pari a 0.169, il grado di dipendenza in media delle stoppate per minuto rispetto al ruolo è moderato.

Per le palle perse è stato calcolato un rapporto di correlazione pari a 0.037, il grado di dipendenza in media delle palle perse per minuto rispetto al ruolo è contenuto, ma anche questo si poteva notare nella tabella 5, con medie simili tra ruoli.

Per i falli fatti è stato calcolato un rapporto di correlazione pari a 0.118, il grado di dipendenza in media dei falli fatti per minuto è contenuto e pari a circa il 11.80%.

Per i punti è stato calcolato un rapporto di correlazione pari a 0.0023, il grado di dipendenza è il più basso tra tutte le variabili d'interesse per minuto e l'unico approssimabile a zero, come si vedrà anche nel paragrafo successivo con il calcolo dell'ANOVA.

2.3.3. ANOVA

L'analisi della varianza (ANOVA) è un test statistico utilizzato per valutare se ci sono differenze significative tra le medie di tre o più gruppi, la misura principale per l'ANOVA è il p-value associato al test. Questo test è caratterizzato da due ipotesi, l'ipotesi nulla H0 sostiene che non ci siano differenze tra le medie dei gruppi considerati, mentre l'ipotesi alternativa H1, al contrario, suggerisce che esista almeno una media diversa dalle altre.

Se il p-value ottenuto dall'ANOVA è minore di 0.05 (si utilizza solitamente il valore di significatività di alpha uguale a 0.05) si può rifiutare l'ipotesi nulla e ciò significa che è possibile affermare le medie siano statisticamente differenti. Se il p-value è maggiore di 0,05 si conclude che non ci sono abbastanza evidenze per rifiutare l'ipotesi nulla. L'analisi è stata condotta senza verifica delle assunzioni su cui si fonda l'ANOVA, dal momento che, in genere, questo tipo di analisi è robusta rispetto ad allontanamenti dalle ipotesi, soprattutto in presenza di una elevata numerosità campionaria.

Si procede ora con il calcolo dell'ANOVA per quanto riguarda le variabili d'interesse per minuto rispetto al ruolo, utilizzando il linguaggio di programmazione R e iniziando con i rimbalzi offensivi per minuto:

$$Y = dati\$Rimbalzi.offensivi$$

$$X = dati\$Ruolo$$

$$anova(lm(Y \sim X))$$

In questo modo, si esegue un'analisi della varianza sulla relazione lineare Y ~ X, dove Y rappresenta i valori dei rimbalzi offensivi per giocatore, mentre X rappresenta la variabile categorica del ruolo dei giocatori da considerare nell'analisi. In questo modo si vuole testare se ci sono differenze significative tra i ruoli per quanto riguarda i rimbalzi offensivi per minuto. Il risultato ottenuto è visibile nella Figura 10.

```
Analysis of Variance Table

Response: Y
           Df  Sum Sq  Mean Sq F value    Pr(>F)
X           4 0.40783 0.101956  116.78 < 2.2e-16 ***
Residuals 673 0.58757 0.000873
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 10: Risultato dell'ANOVA per i rimbalzi offensivi per minuto rispetto al ruolo

Il valore evidenziato in rosso nella Figura 10 rappresenta il p-value. Nel caso dei rimbalzi offensivi per minuto il p-value è molto piccolo, inferiore a 0.05, quindi si rifiuta l'ipotesi nulla H0 ed è possibile affermare che le differenze per ruolo siano significative.

Questo procedimento è stato realizzato per tutte le variabili d'interesse per minuto. Tutte le variabili (rimbalzi difensivi, assist, palle recuperate, stoppate, palle perse, falli fatti) ad eccezione dei punti presentano un p-value molto piccolo ed inferiore a 0.05. È possibile affermare che per queste variabili esistano differenze significative tra le medie per ruolo, quindi che dipendono dalla diversa posizione in campo di ogni giocatore.

Per quanto riguarda i punti per minuto, è stato calcolato un p-value pari a 0.8156, quindi l'unico maggiore di 0.05. Nel caso dei punti per minuto, quindi si accetta l'ipotesi nulla H0 e si può affermare che non esistano differenze significative tra le medie per ruolo, come si poteva notare anche dalla Tabella 5. Per tutte le variabili per cui si è concluso che le differenze campionarie tra i valori medi di ruolo sono significative, i valori del rapporto di correlazione riportati nel paragrafo precedente consentono di misurare il coefficiente di effect size, ossia la dimensione dell'effetto dell'ANOVA, dato dalla quota di variabilità totale della variabile di interesse attribuibile al ruolo.

2.3.4. GRAFICI DI DENSITÀ ATTRAVERSO RSTUDIO

Inoltre, è stato condotto un lavoro analogo a quello del paragrafo 2.3.1 utilizzando grafici di densità suddivisi per ruolo. Questo lavoro è stato realizzato utilizzando RStudio, un ambiente di sviluppo integrato per il linguaggio di programmazione R, noto per il calcolo statistico e la grafica. L'obiettivo era replicare quanto fatto con i calcoli delle medie attraverso grafici, in modo che le distribuzioni dei dati fossero visualizzate in modo più chiaro, attraverso grafici di densità per le distribuzioni delle medie dei giocatori in base al loro ruolo specifico.

Innanzitutto, su RStudio, è stato necessario impostare la directory di lavoro corrente, ovvero il percorso predefinito da cui un'applicazione avvia le operazioni di lettura e scrittura dei file, a meno che venga specificato un percorso completo per un file:

$$setwd("C:\\Users\\simon\\Desktop\\GIOCATORI\ NBA")$$

In questo caso, il percorso da percorrere per andare a selezionare la cartella di competenza sarà Utente>Simon>Desktop>GIOCATORI NBA, le operazioni avverranno nella cartella "GIOCATORI NBA" presente sul Desktop.

Successivamente, il file del secondo database è stato salvato in formato .csv. Questo passaggio è stato eseguito in modo che, successivamente, in Rstudio fosse possibile aprire e lavorare su quel database specifico attraverso la funzione "read.csv". Per farlo, è stato necessario definire il separatore utilizzato per separare le colonne (";"), il separatore per i numeri decimali (",") ed il nome del file all'interno della directory specifica:

$$nba = read.csv("FILE\_MEDIE.csv", sep = ";", dec = ",")$$

Di seguito, attraverso la funzione "ggplot", è stato possibile realizzare grafici di densità che andassero a confermare i valori osservati nel paragrafo precedente, attraverso una visualizzazione più "immediata" grafica. Per utilizzare la funzione "ggplot" su RStudio, è stato necessario prima installare e caricare il pacchetto "ggplot2", una libreria di grafici per il linguaggio di programmazione R:

$$install.packages("ggplot2")$$

$$library(ggplot2)$$

Infine, è stato finalmente possibile visualizzare graficamente i dati, attraverso grafici di densità per rappresentare come i dati sono distribuiti lungo un'asse, consentendo di identificare la forma della distribuzione e le aree in cui i dati sono maggiormente concentrati. Nell'esempio di seguito si andranno a visualizzare le densità di probabilità per i rimbalzi offensivi e come variano in base al ruolo dei giocatori:

$$ggplot(nba, aes(x = Rimbalzi\_offensivi, fill = Ruolo)) + geom\_density(alpha = 0.25) + scale\_fill\_manual(values = c("darkblue","green","orange","red","black"))$$

In questo esempio, la variabile "nba" rappresenta il database di riferimento da cui estrarre i dati, la funzionalità "aes" permette di specificare le variabili da utilizzare nel grafico. L'asse delle ascisse sarà caratterizzato dai valori dei rimbalzi offensivi, l'asse delle ordinate presenterà la frequenza

specifica di densità, mentre il raggruppamento avviene in base alla variabile "Ruolo". Il "geom_density" specifica il fatto che si vogliono visualizzare le linee di densità di frequenza, mentre "scale_fill_manual" permette di impostare i colori da utilizzare per il riempimento delle aree del grafico per ogni ruolo, in questo caso sono stati utilizzati il blu per l'ala grande, il verde per l'ala piccola, l'arancione per il centro, il rosso per la guardia tiratrice e il bianco per il playmaker. Nella Figura 11 è possibile osservare il risultato di questa azione attraverso RStudio, ovvero la visualizzazione del grafico di densità relativo ai rimbalzi offensivi per ogni ruolo.

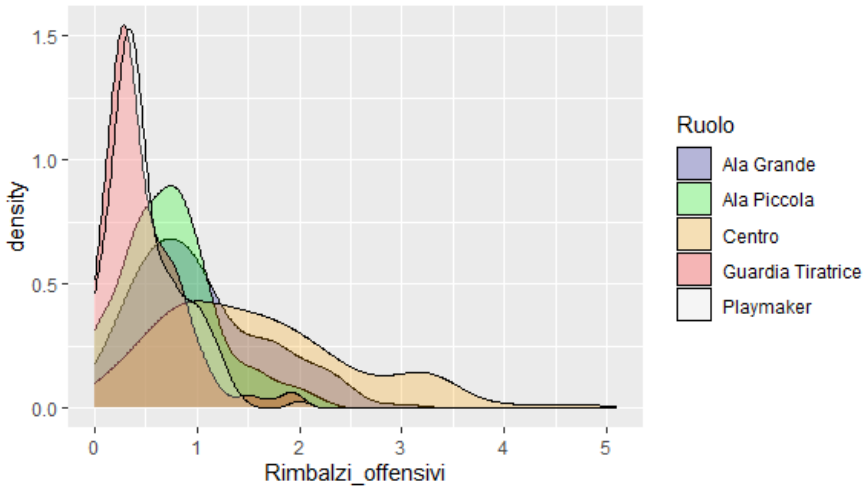


Figura 11: Grafico di Densità di frequenza dei Rimbalzi offensivi, per ruolo

Il grafico presente nella Figura 11 mostra una prevalenza di valori bassi di rimbalzi offensivi per quanto riguarda playmaker e guardie tiratrici, con picchi intorno a 0.40 rimbalzi offensivi circa. Playmakers e guardie tiratrici presentano distribuzioni di rimbalzi offensivi molto simili rispetto agli altri ruoli. Le ali piccole presentano picchi di distribuzione attorno a 1 rimbalzo offensivo a partita, mentre il ruolo che presenta maggiore variabilità di rimbalzi offensivi, come si era visto anche nei paragrafi precedenti, è il centro.

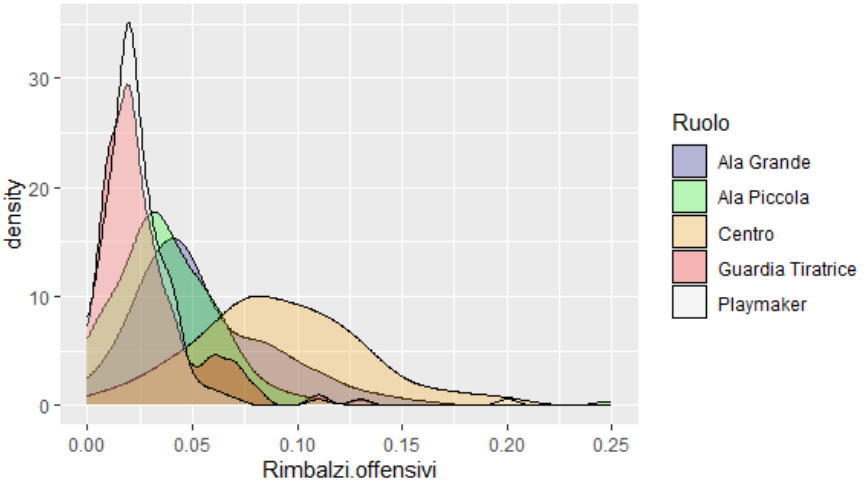


Figura 12: Grafico di Densità di frequenza dei rimbalzi offensivi per minuto, per ruolo

Lo stesso processo è stato realizzato per quanto riguarda il database con i dati depurati dall'effetto dei minuti, per una maggiore equità nei confronti dei giocatori con tempi di gioco diversi. Considerare i dati per minuto fornisce un'indicazione più accurata dell'efficacia del giocatore tenendo conto del tempo in cui è effettivamente in campo.

Nella Figura 12 è presente quindi lo stesso grafico, relativo però ai dati per minuto. Nell'effettivo, la visualizzazione delle distribuzioni per ogni ruolo non cambia molto, l'unica differenza degna di nota si vede nel ruolo del "Centro", con la linea di densità che è andata a formare una curva più spostata verso destra, a dimostrazione del fatto che precedentemente i minuti influivano negativamente sulla distribuzione dei rimbalzi offensivi per i centri.

Lo stesso processo è stato realizzato per tutte le variabili.

Nella Figura 13 è riportato il grafico di densità di frequenza dei rimbalzi difensivi per ogni ruolo.

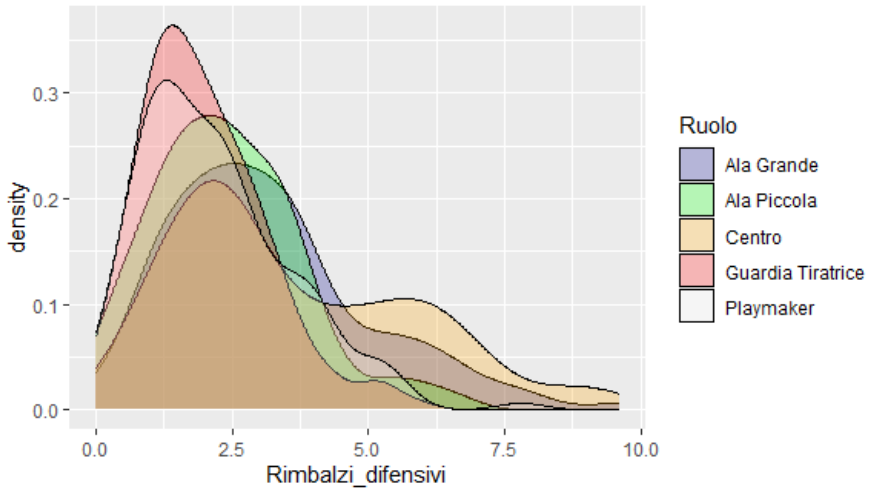


Figura 13: Grafico di Densità di frequenza dei rimbalzi difensivi, per ruolo

Il grafico presente nella Figura 13 mostra distribuzioni approssimativamente simili per quanto riguarda i rimbalzi difensivi di playmaker e guardie tiratrici, così come lo era stato per i rimbalzi offensivi, con la presenza di un picco attorno a 1.3 rimbalzi per partita circa in media. Anche le ali piccoli e le ali grandi presentano curve simili, con la presenza di un picco attorno a 2.5 rimbalzi difensivi circa, anche se la coda della distribuzione delle ali piccole è più breve, indicando una limitata variabilità rispetto a quella delle ali grandi. Ancora una volta, si può notare anche la variabilità nella curva appartenente al centro, che presenta due picchi, uno attorno a 2.5 rimbalzi difensivi e uno di frequenza minore attorno ai 6 rimbalzi difensivi.

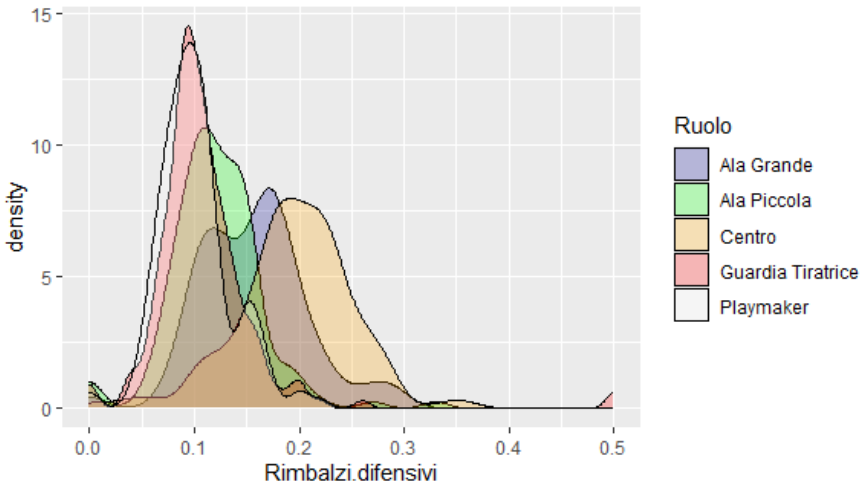


Figura 14: Grafico di Densità di frequenza dei rimbalzi difensivi per minuto, per ruolo

Nella Figura 14 è presente il grafico di densità di frequenza dei rimbalzi difensivi per minuto. Guardie tiratrici e playmaker presentano sempre curve molto simili, mentre si può notare che ali piccole e grandi non presentano più somiglianze per quanto riguarda le distribuzioni, in quanto la distribuzione dell'ala grande presenta due picchi ed è leggermente più spostata verso destra. Questo significa che l'apparente somiglianza era dovuta ad una influenza negativa relativa all'effetto dei minuti. La curva appartenente ai rimbalzi difensivi per minuto per i centri è posizionata più a destra di tutte, in quanto è il ruolo che mostra valori più elevati di rimbalzi difensivi per minuto.

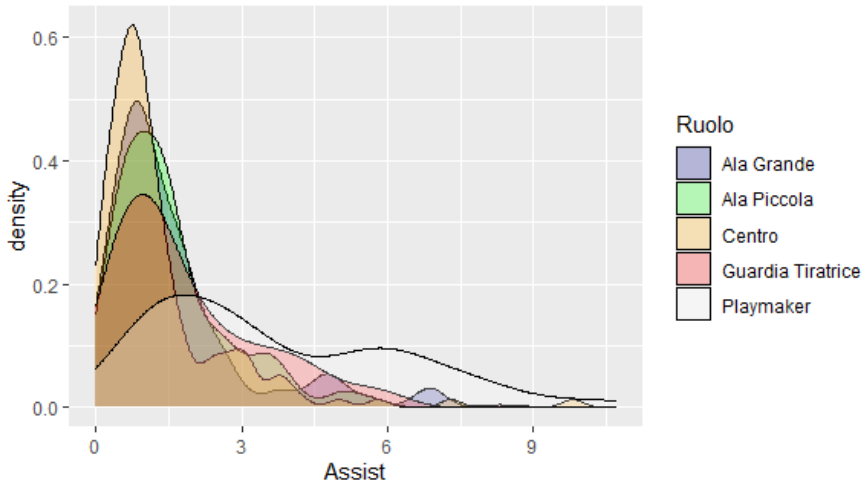


Figura 15: Grafico di Densità di frequenza degli assist, per ruolo

Nel grafico presente in Figura 15 è presente un confronto per quanto riguarda le distribuzioni degli assist per ogni ruolo. Nelle curve di densità che si vengono a creare, la curva dei playmaker presenta due picchi, uno attorno ai 2 assist e uno attorno ai 6 assist ed è caratterizzata da maggiore variabilità

tra i dati. Le altre curve presentano somiglianze per quanto riguarda l'unico picco attorno ad 1 assist, ma presentano variabilità differenti, ad esempio i centri presentano minore variabilità in quanto la curva è più schiacciata, mentre le guardie tiratrici tra gli altri ruoli presentano maggior variabilità.

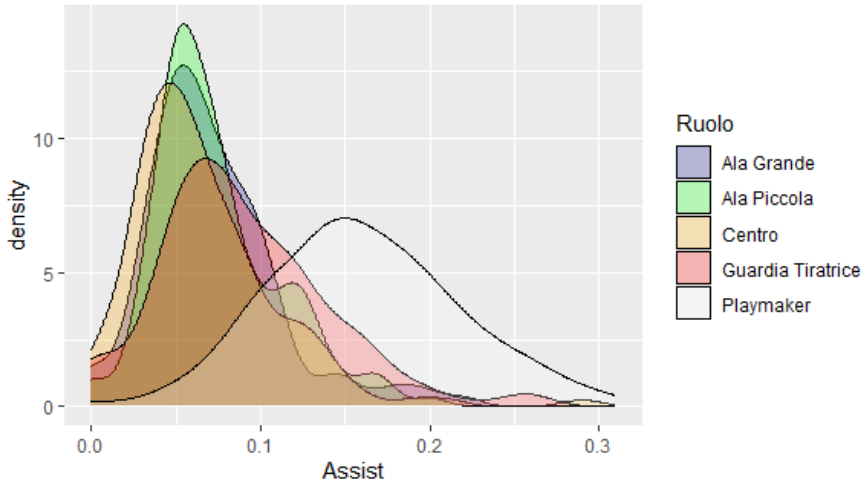


Figura 16: Grafico di Densità di frequenza degli assist per minuto, per ruolo

Per quanto riguarda gli assist per minuto presenti nella Figura 16, si nota immediatamente la forma differente nella curva per quanto riguarda i playmaker, con i dati che precedentemente erano molto influenzati dall'effetto dei minuti giocati. La curva dei playmaker non presenta più due picchi ma uno solo, che si posiziona circa ad un valore doppio rispetto agli altri ruoli. Le distribuzioni degli assist per minuto degli altri ruoli rimangono simili, anche se la distribuzione con minore variabilità non è più quella dei centri ma quella delle ali piccole.

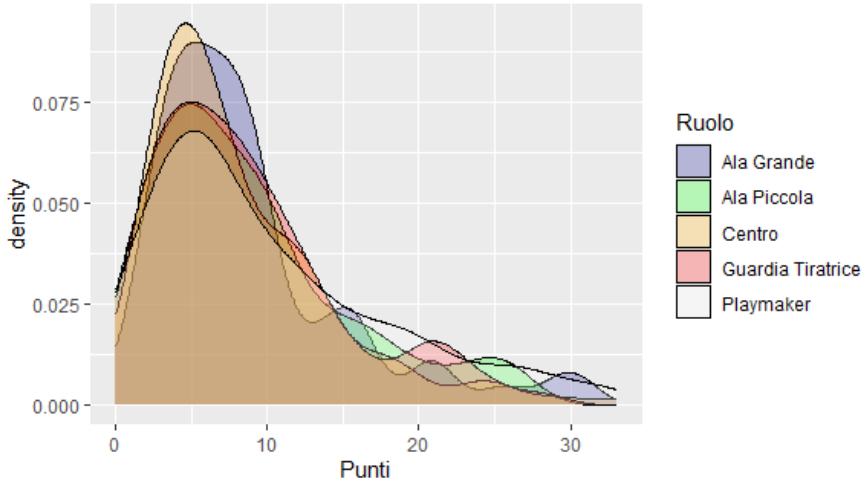


Figura 17: Grafico di Densità di frequenza dei punti, per ruolo

Nel grafico presente nella Figura 17 è presente un confronto tra i ruoli per quanto riguarda le distribuzioni di punti. I vari ruoli presentano distribuzioni di punti molto simili tra loro e tutte con un picco attorno ai 6 punti, senza particolari differenze, come era anche comprensibile stando a quanto dichiarato anche nei paragrafi 2.3.2 e 2.3.3. Nella Figura 18 è invece riportato il confronto per quanto riguarda i punti per minuto.

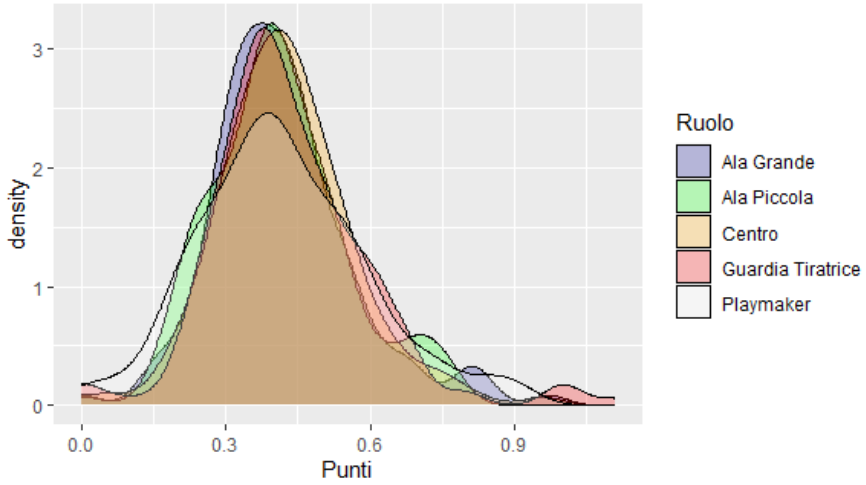


Figura 18: Grafico di Densità di frequenza dei punti per minuto, per ruolo

Come si può notare nel grafico presente nella Figura 18, anche per i punti per minuto i differenti ruoli presentano distribuzioni molto simili caratterizzate da unico picco e simili variabilità, come nella Figura 17.

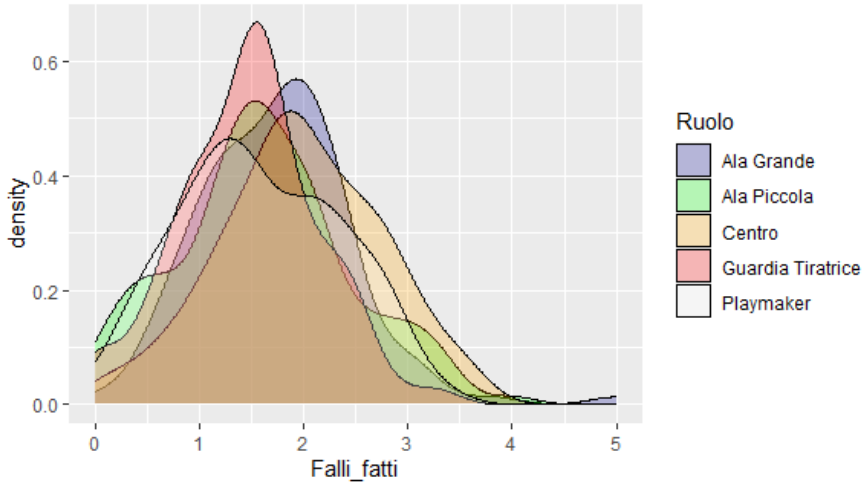


Figura 19: Grafico di Densità di frequenza dei falli fatti, per ruolo

Nella Figura 19 è presente il grafico di densità di frequenza dei falli fatti per ruolo. Le varie curve di densità presentano forme differenti, ma con principale concentrazione nella stessa zona, attorno a 1.5 falli fatti per partita. Il ruolo con meno dispersione nei dati sembrerebbe essere la guardia tiratrice.

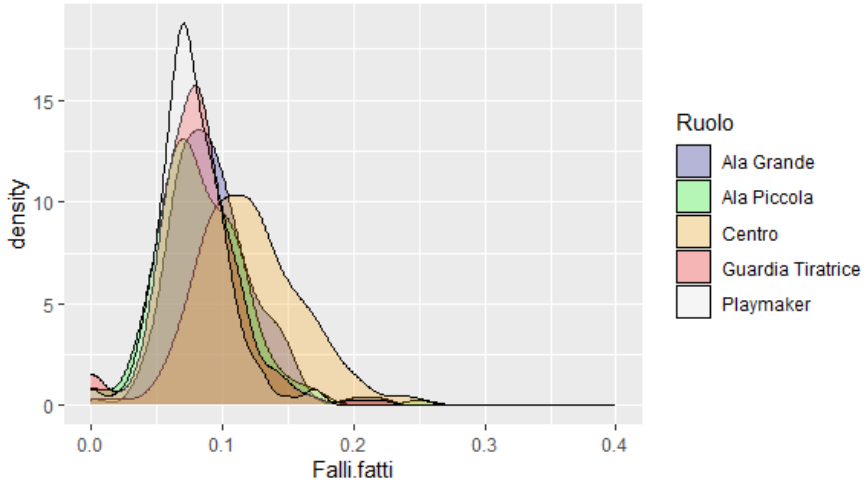


Figura 20: Grafico di Densità di frequenza dei falli fatti per minuto, per ruolo

Nella Figura 20 è presente invece il grafico di densità di frequenza dei falli fatti per minuto. Rispetto alla Figura 19 si può notare uno spostamento verso destra della curva dei centri, a dimostrazione del fatto che l'effetto dei minuti influiva negativamente sulla sua distribuzione. In questo caso, con i dati depurati dall'effetto dei minuti si può notare come la curva caratterizzata da maggiore concentrazione è quella dei playmaker, quando prima era quella delle guardie tiratrici.

In questo capitolo sono stati descritti i due database realizzati e analizzati attraverso analisi statistiche di base. Sono stati calcolati i principali indici di sintesi sui microdati relativi a determinati giocatori specifici per quanto riguarda minuti, rimbalzi, assist e punti fatti e le medie per ruolo con relativi grafici di densità per quanto riguarda rimbalzi offensivi, rimbalzi difensivi, assist, punti, palle recuperate, palle perse, falli fatti e stoppate.

Nel prossimo capitolo, invece, partendo dai microdati relativi ai singoli giocatori sarà proposto un algoritmo, programmato con Python, per visualizzare la probabilità che un determinato giocatore ha di superare una determinata soglia per quanto riguarda minuti, rimbalzi assist o punti, ipotizzando che la distribuzione di ogni giocatore relativamente a ciascuna variabile sia normale, con media e varianza stimate nel campione di dati a disposizione.

CAPITOLO 3: ALGORITMO DI PYTHON

L'obiettivo di questo capitolo è descrivere un algoritmo realizzato per calcolare la probabilità, per un determinato giocatore, di superare una determinata soglia di una variabile di interesse (minuti, rimbalzi, assist e punti).

L'algoritmo è stato realizzato tramite uno script su Python che leggesse il file generale di dati citato nel secondo capitolo, che presenta medie e deviazioni standard per ogni giocatore per ciascuna variabile. L'algoritmo chiede all'utente di specificare il nome di un determinato giocatore di cui si vuole calcolare la probabilità, una variabile di interesse, scelta tra minuti, punti, assist o rimbalzi, e una soglia da superare per quella variabile. Successivamente, l'algoritmo stima la probabilità che il giocatore superi la soglia specificata ipotizzando che la distribuzione dei dati sia normale.

3.1. PYTHON

Per raggiungere l'obiettivo di sviluppare un algoritmo per calcolare la probabilità che un giocatore superasse una determinata soglia per quanto riguarda minuti, assist, rimbalzi o punti, è stato necessario apprendere e utilizzare un determinato linguaggio di programmazione. Tra le varie possibilità, la scelta si è direzionata verso Python, noto per la sua versatilità e relativa semplicità.

Creato nel 1989 da Guido Van Rossum, informatico olandese, il nome "Python" è ispirato al gruppo comico inglese "Monty Python" di cui l'informatico era grande ammiratore. Python è un linguaggio di programmazione multipiattaforma, che significa che può girare su diversi sistemi operativi, ed orientato agli oggetti, che denota il fatto che organizza il codice in oggetti, ovvero entità che contengono variabili o funzioni, che possono interagire tra loro, comunicando informazioni tra loro e semplificando lo sviluppo e la manutenzione del software[14].

Oltre a ciò, Python è open source, gratuito anche quando usato per scopi commerciali e reso disponibile al pubblico, con una grande comunità attiva che fornisce supporto ai programmatori, rendendone l'apprendimento un'opzione più semplice. La sua comunità e la natura open source contribuiscono costantemente all'evoluzione del linguaggio, con un'enorme libreria standard che offre un ampio repertorio di risorse utili. Tra queste ad esempio c'è "Pandas", libreria utilizzata in questo progetto per gestire strutture come i dataframe, fondamentale per manipolare e gestire il file

[14] https://it.wikipedia.org/wiki/Python

Excel contenente il database del file generale, con medie e deviazioni standard di minuti, rimbalzi, assist e punti per ogni giocatore.

Inoltre, questo linguaggio di programmazione è interpretato e interattivo. Interpretato in quanto il codice viene eseguito direttamente da un interprete senza la necessità di generare un file eseguibile in anticipo, in modo tale da poter modificare il codice ed eseguire immediatamente le modifiche senza dover ricompilare l'intero programma. Interattivo in quanto gli utenti interagiscono per l'appunto con l'interprete attraverso la shell (ambiente di sviluppo interattivo), in modo tale da poter eseguire frammenti di codice uno alla volta, ottenendo immediatamente risultati.

Per facilitare la programmazione, è stato necessario installare un ambiente di sviluppo integrato, in modo che il processo di scrittura del codice fosse più semplice e comodo. L'ambiente scelto è stato Pycharm IDE, creato da JetBrains, uno degli ambienti di sviluppo più utilizzati e apprezzati per Python, in quanto la sua versione Community, gratuita e open source, offre funzionalità avanzate come l'auto-completamento intelligente, suggerimenti durante la scrittura e l'ispezione del codice per individuare eventuali errori.

In sintesi, Python è un linguaggio semplice, versatile e potente, con una comunità attiva che contribuisce costantemente al suo sviluppo e alla creazione di nuove funzionalità. Nel seguito, si andrà nel dettaglio a visualizzare come si compone il codice scritto per la realizzazione del progetto legato al calcolo della probabilità per i giocatori NBA, per poi mostrare il risultato finale con esempi.

3.2. ANALISI DELLO SCRIPT DI PYTHON

Lo script finale si presenta come in Figura 21.

```
import pandas as pd
from scipy.stats import norm

file_path = "C:\\Users\\simon\\Desktop\\GIOCATORI NBA\\FILE GENERALE GIOCATORI.xlsx"
df = pd.read_excel(file_path)

nome_giocatore = input("Inserisci il nome del giocatore: ")
variabile_interesse = input("Inserisci la variabile di interesse (minuti, punti, assist, rimbalzi): ")
soglia_da_superare = int(input(f"Inserisci il valore di {variabile_interesse} da superare: "))

giocatore = df[df['player'] == nome_giocatore]
if not giocatore.empty:
    media_variabile = giocatore[f'media_{variabile_interesse}'].values[0]
    dev_std_variabile = giocatore[f'dev_{variabile_interesse}'].values[0]
else:
    print("Giocatore non trovato nel database.")
    exit()

probabilita = 1 - norm.cdf(soglia_da_superare, media_variabile, dev_std_variabile)

print(f"La probabilità che {nome_giocatore} superi {soglia_da_superare} {variabile_interesse} è: {probabilita:.2%}")
input("Premi un tasto per uscire...")
```

Figura 21: Script dell'algoritmo per il calcolo della probabilità

Ora si procederà ad esaminare in modo dettagliato il funzionamento di tale script, analizzando ogni riga in modo approfondito.

$$import\ pandas\ as\ pd$$

Si inizia con l'importazione di una delle librerie chiave per l'estrazione e la gestione dei dati, ovvero "Pandas". Attraverso la funzione import è possibile accedere alle funzionalità fornite da tale libreria, che sarà necessaria per la manipolazione e analisi dei dati presenti nel file Excel generale delle medie e deviazioni standard di minuti, rimbalzi, assist e punti di ogni giocatore NBA per quanto riguarda la stagione 2022-2023 tra Regular Season e Playoff. La libreria viene rinominata come "pd", in modo tale da attribuire un nome di riferimento più immediato per facilitarne l'utilizzo e risparmiare tempo.

$$from\ scipy.stats\ import\ norm$$

Permette di importare la funzione "norm" dal modulo "stats" di "Scipy", che è una libreria scientifica caratterizzata da algoritmi e strumenti matematici. La funzione "norm" si riferisce alla distribuzione Normale, che è una distribuzione di probabilità continua spesso utilizzata per descrivere variabili causali. È difficile che una variabile assuma esattamente la stessa forma di una distribuzione teorica come ad esempio la distribuzione normale, ma essa spesso può essere utilizzata come approssimazione per il calcolo della probabilità, quando simile.

Il grafico presente nella Figura 22 mostra la funzione di densità di una variabile casuale Normale X, di media μ e varianza σ^2.

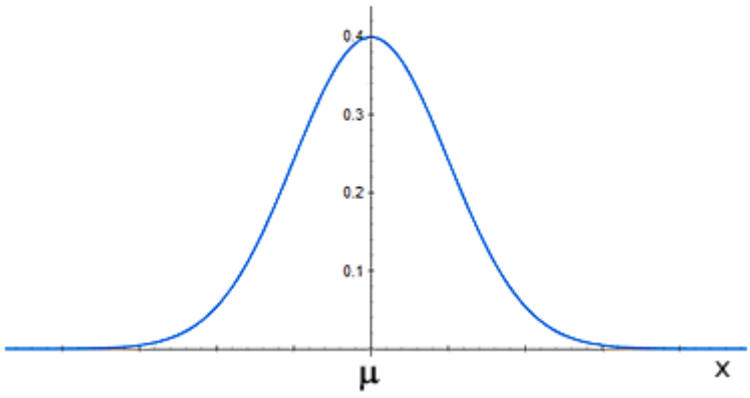


Figura 22: Funzione di densità di una variabile casuale Normale con media μ e varianza σ^2

Questa distribuzione mostra una curva simmetrica a forma di campana, con l'asse orizzontale che rappresenta i valori delle misurazioni e l'asse verticale che indica la densità di frequenza di tali valori.

La formula matematica della funzione di densità f(x) di una variabile casuale normale con media μ e varianza σ^2 può apparire piuttosto complessa:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Tuttavia, a livello pratico per il progetto di questa relazione finale è sufficiente conoscere due parametri: la deviazione standard (o la varianza) e la media. La media determina il punto centrale della distribuzione, mentre una deviazione standard più o meno elevata indica una maggiore o minore variabilità dei dati, quindi una curva più o meno "schiacciata" verso il centro.

L'obiettivo di questo progetto è cercare di definire per determinati giocatori NBA distribuzioni di minuti, assist, punti o rimbalzi approssimabili ad una distribuzione normale, e di conseguenza, avendo media e deviazioni standard, essere in grado di poter calcolare la probabilità, attraverso una funzione che verrà definita successivamente, che il giocatore in questione superi la soglia prestabilita di quella variabile di interesse in una partita. Questo però verrà definito meglio in seguito, ora si continua a definire lo script realizzato.

$$file\_path = "C:\\Users\\simon\\Desktop\\GIOCATORI\ NBA\\FILE\ GENERALE\ GIOCATORI.xlsx$$

$$df = pd.read\_excel(file\_path)$$

Questa parte di codice si occupa di leggere il file Excel e caricare i dati in un dataframe attraverso la libreria Pandas precedentemente importata. Il "file_path" funziona come visto nel secondo capitolo quando è stata impostata la directory di lavoro, è una variabile che contiene il percorso completo del file excel da leggere, presente sul desktop, nella cartella "GIOCATORI NBA" e con il nome "FILE GENERALE GIOCATORI". Attraverso la directory di lavoro impostata, è possibile grazie alla funzione "pd.read_excel()" di Pandas leggere il contenuto del file indirizzato dal percorso e poterne manipolare il dataframe, che nominiamo "df" per utilizzare un nome di riferimento più immediato.

Nello specifico, si andrà a leggere il database generale descritto nel secondo capitolo contenente le medie e le deviazioni standard per ciascuna variabile (minuti, rimbalzi, assist e punti) per ogni giocatore NBA, calcolate sulle partite della regular season e playoff/finali 2022-2023, in caso la squadra del giocatore in analisi abbia proseguito il cammino durante le fasi eliminatorie.

Nella Figura 23 è presente una raffigurazione del database a cui si sta facendo riferimento, descritto nel secondo capitolo.

	A	B	C	D	E	F	G	H	I
1	player	media_minuti	dev_minuti	media_rimbalzi	dev_rimbalzi	media_assist	dev_assist	media_punti	dev_punti
2	Lebron James	35,92	5,21	8,56	2,92	6,71	2,42	27,71	8,02
3	Anthony Davis	34,93	5,61	12,88	4,48	2,63	1,61	25,17	9,79
4	Rui Hachimura	23,44	6,02	4,32	2,51	0,86	0,93	11,42	6,76
5	Max Christie	11,18	9,12	1,67	1,93	0,49	0,79	2,88	3,38
6	Wenyen Gabriel	13,60	7,36	3,78	3,03	0,46	0,66	4,90	4,09
7	Jaxson Hayes	12,91	8,23	2,81	2,10	0,72	0,97	4,96	5,04
8	Taurean Prince	21,85	5,40	2,36	1,54	1,49	1,47	9,02	5,46
9	Austin Reaves	30,31	6,55	3,30	1,98	3,60	2,43	13,79	6,95
10	Cam Reddish	24,70	8,33	2,20	1,73	1,43	1,55	9,68	6,77
11	D'Angelo Russell	31,94	5,81	3,01	1,86	5,87	2,45	16,97	7,75
12	Jarred Vanderbilt	22,86	5,47	6,77	3,33	2,11	1,74	7,40	4,53
13	Gabe Vincent	26,98	7,54	1,96	1,54	2,72	1,81	10,22	6,96
14	Stephen Curry	35,29	4,36	5,91	2,43	6,25	3,05	29,62	8,16
15	Draymond Green	31,38	4,32	7,15	3,02	6,85	2,71	8,59	4,61
16	JaMychal Green	13,22	5,16	3,28	2,45	0,84	1,04	6,08	4,79
17	Andre Iguodala	14,13	2,95	2,13	1,64	2,38	1,06	2,13	2,30
18	Jonathan Kuminga	18,95	9,19	3,12	2,42	1,69	1,54	9,09	6,62
19	Moses Moody	13,03	7,86	1,81	1,63	0,79	1,06	4,96	4,88
20	Kevon Looney	24,06	5,65	9,79	4,52	2,63	1,93	6,97	3,83

Database

Figura 23: File riassuntivo con medie e deviazioni standard per ogni giocatore NBA

$$nome\_giocatore = input\ (\text{"Inserisci il nome del giocatore: "})$$

La funzione input() di Python consente di ricevere per l'appunto un input dall'utente. La stringa tra parentesi (in questo caso, "Inserisci il nome del giocatore:") rappresenta il messaggio che viene visualizzato dalla persona con cui il codice sta interagendo, fornendo informazioni sul valore da inserire. L'utente attraverso l'input deve fornire con precisione nome e cognome del giocatore NBA, identico a come contenuto nella colonna A del file Excel, altrimenti l'algoritmo non sarà in grado di riconoscerlo. Anche per questo motivo, ovvero la situazione di errore di digitazione, è emersa la necessità di trovare una possibile alternativa, che verrà mostrata nel paragrafo 3.3.2. In questo caso, il valore o la stringa inserita attraverso l'input viene assegnata alla variabile "nome_giocatore". Con questo passaggio, il sistema viene a conoscenza del nome del giocatore NBA su cui attuare il calcolo della probabilità e su cui estrarre i dati dal file Excel generale.

$$variabile\_interesse = input\ (\text{"Inserisci la variabile di interesse (minuti, punti, assist, rimbalzi): "})$$

Analogamente alla riga precedente, Python attende un input dall'utente che verrà inserito nella variabile "variabile_interesse", mentre la stringa tra parentesi (in questo caso "Inserisci la variabile di interesse (minuti, punti, assist, rimbalzi)" rappresenta il messaggio che viene visualizzato. Con questo passaggio, il codice sta chiedendo all'utente di specificare la variabile su cui si desidera

calcolare la probabilità, in modo che grazie a questa informazione possa focalizzarsi sulla variabile di interesse per il giocatore inserito nella riga precedente, estraendone i valori di media e deviazione standard.

$$soglia\_da\_superare = int(input(f\text{"Inserisci il valore di \{variabile\_interesse\} da superare: "}))$$

In questo caso ci sono alcune differenze rispetto ai passaggi precedenti, nonostante la funzione input() richieda comunque all'utente di inserire un determinato valore. La stringa, di formato "f", consente l'inclusione di variabili all'interno della stringa, al posto di {variabile_interesse} verrà inserito il valore della variabile di interesse inserito dall'utente nel passaggio precedente. Inoltre, la funzione int() denota il fatto che si richiede di inserire un valore numerico, che verrà assegnato all'oggetto "soglia_da_superare". Con questo passaggio, il sistema richiede di inserire il valore della soglia da superare per il calcolo della probabilità.

In sintesi, con queste tre righe di codice, si richiede all'utente di inserire il nome del giocatore, la variabile di riferimento e la soglia da superare. In questo modo, il sistema viene a conoscenza dei dati da estrarre dal file Excel, facendo riferimento ai valori di media e deviazione standard per il giocatore e variabile inseriti negli input. Questo passaggio si definisce più precisamente con le righe di codice presentate qui di seguito, in cui si vede in che modo il programma estrae i dati.

$$giocatore = df[df['player'] == nome\_giocatore]$$

Questa riga di codice esegue un'operazione di indicizzazione, necessaria per l'estrazione dei dati all'interno del dataframe "df" nominato precedentemente, per ottenere le informazioni necessarie per il calcolo della probabilità sul giocatore identificato per mezzo dell'input fornito dall'utente (nome_giocatore). Con df['player'] == nome_giocatore, il sistema estrae i valori appartenenti al giocatore nel momento in cui il nome nella colonna "player" del database corrisponde allo stesso dell'input inserito. Nella variabile "giocatore" si crea il nuovo dataframe con i valori risultanti dall'operazione di filtraggio.

$$if\ not\ giocatore.empty:$$

$$media\_variabile = giocatore\ [f'media\_\{variabile\_interesse\}'].values[0]$$

$$dev\_std\_variabile = giocatore\ [f'dev\{variabile\ interesse\}'].values[0]$$

In Python la funzione "if" è utilizzata per eseguire una determinata azione solo se una certa condizione è vera, in caso contrario viene realizzata la parte di codice anticipata dall'istruzione "else". In questo caso specifico, la funzione "if not" verifica che il dataframe estratto precedentemente e salvato nella variabile "giocatore" non sia vuoto. Se la variabile contiene dei valori, il sistema accede alle colonne specifiche nel database per estrarre la media (media_variabile) e la deviazione standard (dev_std_variabile) della variabile di interesse specificata dall'utente attraverso l'input precedente (variabile_interesse).
La funzione ". values[0]" permette di estrarre il valore della colonna corrispondente, ovvero il primo valore (e anche unico, essendo filtrato per nome del giocatore) della colonna, che viene considerata come una vera e propria lista di valori. Le informazioni sulle medie e deviazioni standard per la variabile d'interesse verranno utilizzate successivamente per il calcolo della probabilità.

$$else:$$

$$print("\text{Giocatore non trovato nel database.}")$$

$$exit()$$

Se la condizione è falsa, ovvero se nella variabile "giocatore" non è presente alcun dato, viene eseguita l'azione anticipata da "else", ovvero stampa un messaggio che avvisa l'utente che il giocatore non è stato trovato nel database. Dopo aver mostrato questo messaggio, chiude l'esecuzione del programma con exit(), evitando che vengano compiute ulteriori azioni successive.

$$probabilita = 1 - norm.cdf\ (soglia\_da\_superare, media\_variabile, dev\_std\_variabile)$$

Questa riga di codice ha l'obiettivo di calcolare la probabilità che una variabile casuale distribuita normalmente (tra quelle di interesse citate precedentemente, ossia minuti, assist, rimbalzi, punti), superi una certa soglia. Per quanto riguarda "norm.cdf" il "norm" si riferisce al modulo appartenente alla distribuzione normale all'interno di "scipy.stats", mentre "cdf" fa riferimento alla funzione di

ripartizione, che fornisce la probabilità che la variabile casuale assuma un valore minore o uguale alla soglia stabilita. Per quanto riguarda gli argomenti utilizzati nella funzione cdf, si usano "soglia_da_superare", che è il valore da superare, ottenuto per mezzo di un input proveniente dall'utente, "media_variabile" e "dev_std_variabile" che sono la media e la deviazione standard della variabile di interesse calcolate. Calcolando il complemento a 1 della probabilità ottenuta, si è in grado di calcolare la probabilità di superare la soglia, e non di essere al di sotto.
È fondamentale notare che questa operazione assuma che la variabile di interesse segua una distribuzione normale, che è caratteristica necessaria per poter calcolare la probabilità con questo algoritmo. Se la distribuzione della variabile di interesse per il giocatore non è approssimabile ad una normale, l'uso della distribuzione cumulativa della normale potrebbe non riflettere accuratamente la probabilità di superare la soglia, quindi sarebbe più appropriato utilizzare approcci diversi, ad esempio considerando altre forme distributive.

$$print(f"La\ probabilità\ che\ \{nome\_giocatore\}\ superi\ \{soglia\_da\_superare\}\ \{variabile\ interesse\}\ \text{è:}\ \{probabilita:.2\%\}")$$

Questa riga di codice conclusiva permette di mostrare una stringa formattata in modo da incorporare le variabili calcolate, ovvero "nome_giocatore", "soglia_da_superare", "variabile_interesse" e "probabilita", ad esempio "nome_giocatore" verrà sostituita con il nome del giocatore selezionato.
Il ":.2%" successivo alla variabile probabilita indica che si vuole visualizzare la probabilità calcolata con due cifre decimali e convertirla in percentuale.

$$input(\text{premi un tasto per uscire...})$$

Questa riga è spesso utilizzata per concludere uno script, per mantenere la finestra aperta dopo che il programma è stato eseguito, in modo che l'utente possa leggere il messaggio di output prima che si chiuda automaticamente. Il programma mostra quindi un messaggio all'utente e si chiude non appena viene premuto un tasto casuale. Si andrà ora a verificare come funziona lo script, con esempi pratici.

3.3. ESEMPI PRATICI

Prima di applicare la funzione norm.cdf per calcolare la probabilità di superare una determinata soglia, è essenziale che la distribuzione di riferimento sia effettivamente normale, o almeno approssimabile, per questo è necessario valutare attentamente la natura delle distribuzioni delle variabili di interesse.

Per verificare il funzionamento dell'algoritmo, si utilizzeranno le distribuzioni per partita di due cestisti NBA che hanno giocato con costanza durante la stagione 2022-2023, ovvero Fred VanVleet, playmaker degli Houston Rockets che durante la stagione ha giocato per i Toronto Raptors, e Nikola Vucevic, centro dei Chicago Bulls. Per il primo si analizzeranno i punti totalizzati, mentre per il secondo i rimbalzi realizzati. Prima di tutto si verifica la conformità di queste distribuzioni di minuti e punti ad una forma approssimabile ad una distribuzione normale, per cercare di ottenere risultati predittivi più affidabili.

3.3.1. ESEMPIO FRED VANVLEET

Per verificare l'approssimazione alla Normale della distribuzione dei punti totalizzati da Fred VanVleet, playmaker attualmente agli Houston Rockets, è stata effettuata un'analisi utilizzando il linguaggio di programmazione R. Innanzitutto è stato generato un istogramma (Figura 24) per visualizzare la distribuzione dei dati contenente i valori dei minuti delle 72 partite giocate da VanVleet. Per essere simile dovrebbe avere una forma simmetrica rispetto al suo centro in modo da avere media e mediana vicine, la media calcolata dei punti totalizzati da VanVleet è di 19.34, mentre la mediana è di 18. Per quanto riguarda l'indice di asimmetria, è stato calcolato un indice di asimmetria attraverso la funzione "skewness" di R pari a 0.1680175, che indica che la distribuzione dei dati presenta una leggera coda nella parte destra, ovvero una leggera prevalenza di valori più alti nella distribuzione. Tuttavia, questo valore di asimmetria può essere considerato come relativamente basso. Inoltre, dovrebbe presentare una curva a forma di campana, con maggiore frequenza di osservazioni nella parte centrale, diminuendo allontanandosi dal centro.

$$hist(dati, main = "Istogramma\ punti\ VanVleet", xlab\ =\ "Minuti", col = "lightblue",$$

$$border = "black"\ , freq = FALSE)$$

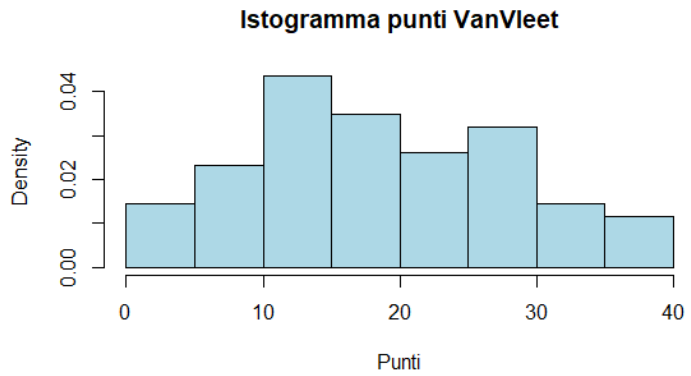


Figura 24: Istogramma della distribuzione di densità dei punti di Fred VanVleet

Ulteriormente, è stato creato un QQ-plot (Figura 25), un grafico che viene spesso utilizzato per verificare se un insieme di dati segue una distribuzione teorica, come ad esempio la distribuzione normale. per valutare la normalità dei dati, confronta i quantili osservati con quelli teorici di una distribuzione Normale, ovvero i valori che ci si aspetterebbe in base alla distribuzione di riferimento.

$$qqnorm(dati)$$

Se i dati seguissero perfettamente la distribuzione teorica, i punti nel QQ-plot dovrebbero allinearsi su una retta, mentre le deviazioni dalla retta indicano quanto i quantili osservati si discostano dai quantili teorici.

$$qqline(dati)$$

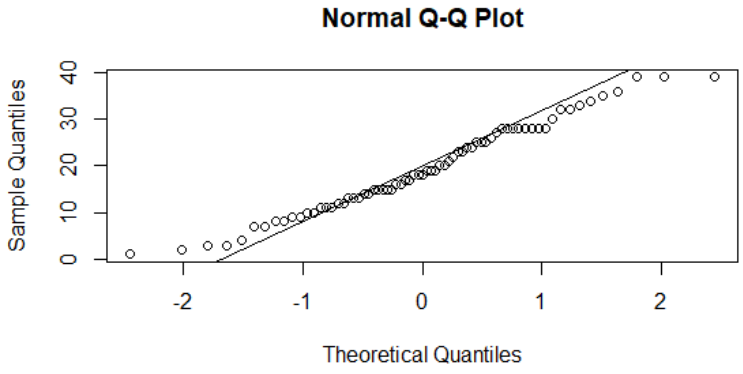


Figura 25: confronto tra QQ-plot e retta per i punti di Fred VanVleet

Per quanto riguarda questo caso, nella Figura 25 si può notare come i quantili osservati nella parte centrale della distribuzione empirica coincidono, nella sostanza, con i quantili teorici della

distribuzione Normale, mentre i valori alle estremità della distribuzione empirica presentano discrepanze rispetto a quanto ci si aspetterebbe da una distribuzione Normale.

Infine, è stato realizzato un grafico di densità di kernel (linea blu, Figura 26) per visualizzare la distribuzione empirica in modo alternativo rispetto all'istogramma. È possibile, grazie a questa tipologia di grafico, comprendere la struttura dei dati e la somiglianza della distribuzione osservata a modelli teorici.

$$plot(density(dati), main = "Grafico\ di\ Densità", xlab = "Minuti", ylab = "Densità",$$

$$col = "darkblue")$$

Successivamente è stata sovrapposta una curva normale (linea rossa, Figura 26). La curva rossa rappresenta la funzione di densità di una variabile casuale Normale caratterizzata dalla stessa media e deviazione standard delle osservazioni reali, in modo da poter effettuare un paragone.

$$curve(dnorm(x, mean = mean(dati), sd = sd(dati)), add = TRUE, col = "red")$$

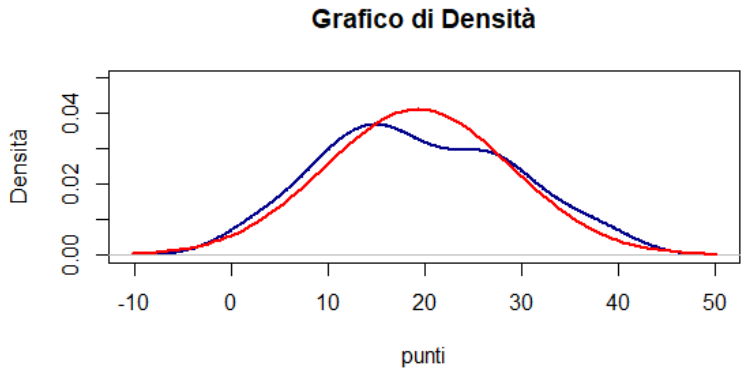


Figura 26: Confronto tra grafico di densità di kernel e funzione di densità di una variabile casuale Normale con stessa μ e σ dei punti di Fred VanVleet

Successivamente, al fine di valutare con un test di ipotesi la normalità della distribuzione dei dati, è stato condotto un test di Shapiro-Wilk. L'ipotesi nulla del test è che i dati provengano da una distribuzione Normale, quindi per confermare la normalità si cerca un valore di p-value superiore a 0.05. Se il p-value ottenuto dal test di Shapiro-Wilk è maggiore di 0.05, ciò suggerisce che si deve

accettare l'ipotesi nulla, indicando che i dati possono essere considerati approssimativamente distribuiti normalmente.

$$shapiro.test(dati)$$

```
        Shapiro-Wilk normality test

data:  dati
W = 0.97698, p-value = 0.2322
```

Figura 27: Risultato del test di Shapiro-Wilk per la distribuzione di punti di Fred VanVleet

Come si può notare nella Figura 27, è stato calcolato un p-value pari a 0.2322, quindi non si può rifiutare l'ipotesi nulla ed è possibile affermare che i dati relativi alla distribuzione di punti di Fred VanVleet possano essere considerati distribuiti Normalmente.

Questi passaggi hanno fornito una panoramica sulla natura della distribuzione dei punti totalizzati giocati da Fred VanVleet durante la stagione 2022-2023 ed è possibile affermare che la distribuzione è più o meno approssimabile ad una Normale. Si può quindi procedere e calcolare la probabilità di superare una determinata soglia di minuti utilizzando l'algoritmo proposto in questa relazione finale.

È possibile quindi aprire il programma realizzato, che richiede immediatamente un primo input relativo al nome del giocatore NBA da voler considerare per il calcolo come si può vedere nella Figura 28. In questo modo, il sistema nell'effettivo andrà a filtrare i valori di media e deviazione standard di tutte le variabili d'interesse del giocatore selezionato nell'input.

Figura 28: primo input che si presenta all'utente relativo al nome del giocatore NBA

Successivamente, una volta inserito il nome del giocatore (in questo caso Fred VanVleet) e premuto il tasto invio, apparirà un secondo input, nel quale inserire la variabile di interesse da considerare, come si può notare nella Figura 29. In questo modo, il sistema andrà a selezionare, dopo aver filtrato precedentemente per il giocatore, solamente i valori di media e deviazione standard della variabile

interessata (punti), che saranno necessari per specificare la distribuzione normale su cui si basa il calcolo della probabilità, per mezzo della funzione norm.cdf.

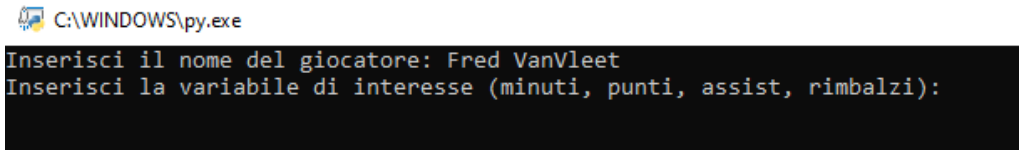


Figura 29: secondo input che si presenta all'utente relativo alla scelta della variabile

Dopo aver inserito la variabile di interesse che si vuole considerare (in questo caso punti) e premuto il tasto invio, apparirà il terzo e ultimo input da inserire, come si può osservare nella Figura 30, ovvero quello riguardante la soglia da superare per il calcolo della probabilità.

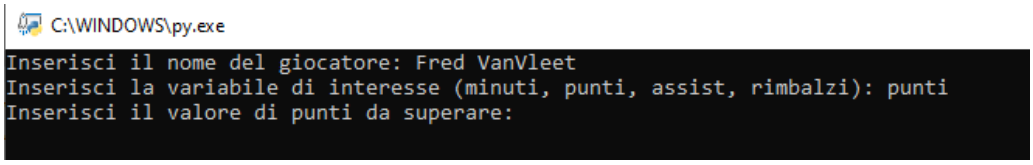


Figura 30: terzo input che si presenta all'utente relativo al valore da superare

Infine, dopo aver inserito il valore della soglia da superare (in questo caso 15 punti) e premuto invio, apparirà un messaggio ad indicare il risultato ottenuto, come si vede nella Figura 27. Supponendo che la distribuzione dei punti totalizzati da VanVleet per partita sia approssimativamente normale (necessario in quanto per il calcolo si sfrutta la funzione norm.cdf invertita), si ottiene un risultato 67.32%. Ciò significa che c'è un'alta probabilità (67.32%) che VanVleet totalizzi 15 punti in una partita, secondo la distribuzione approssimativamente Normale dei suoi punti realizzati.

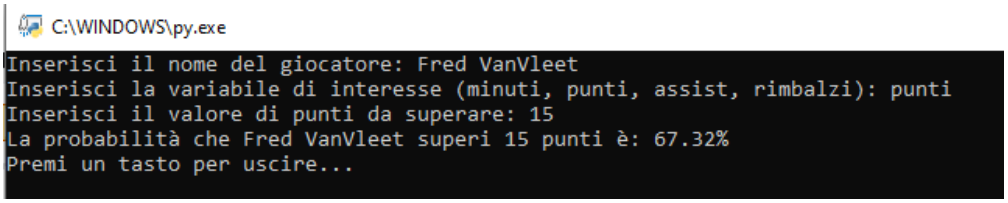


Figura 31: Calcolo della probabilità dell'esempio relativo ai punti di Fred VanVleet

In caso di errore durante l'inserimento del nome del giocatore, ad esempio digitando "Fred VanVlet" anziché "Fred VanVleet", comparirà un messaggio di avviso per informare l'utente che il giocatore non è stato trovato nel database, come illustrato nella Figura 32. Questo avviso è stato inserito per garantire che l'utente sia consapevole di eventuali errori di digitazione nel nome del giocatore.

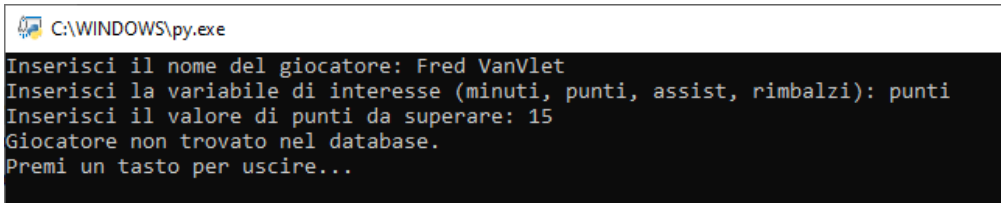


Figura 32: Esempio di errore nel digitare il nome del giocatore NBA

3.3.2. INTERFACCIA GRAFICA CON IA

L'obiettivo iniziale relativo alla realizzazione di uno script per calcolare la probabilità di un giocatore NBA di superare una determinata soglia per quanto riguarda minuti, rimbalzi, assist o punti è stato raggiunto grazie allo script di Python. Tuttavia, riconoscendo che l'aspetto estetico del programma risultava essere poco accattivante e abbastanza approssimativo è stata realizzata un'interfaccia grafica con l'aiuto dell'intelligenza artificiale di ChatGPT alla quale sono stati modificati colori e dimensioni dei caratteri di testo, ed ha portato, come si vede in Figura 33 ad un'interfaccia esteticamente accattivante per migliorare la praticità dell'utente, mantenendo lo script precedente come riferimento centrale.

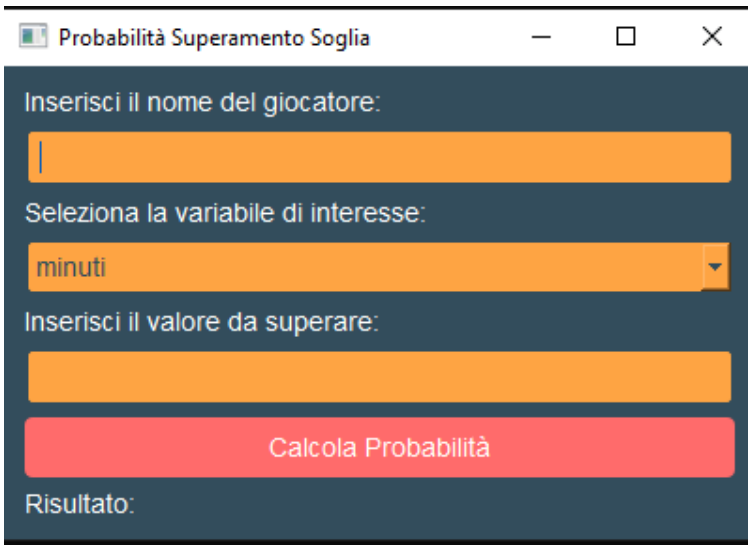


Figura 33: Interfaccia grafica realizzata grazie a ChatGPT

In aggiunta, per migliorare la praticità del programma e ridurre gli errori di inserimento, sono state avanzate determinate richieste da integrare all'aspetto estetico del programma, come ad esempio l'implementazione dell'inserimento automatico dei nomi dei giocatori durante la digitazione (ad esempio, come si vede nella Figura 34, digitando solamente "Leb" appare un suggerimento sotto la casella di inserimento, con il nome "Lebron James") e l'aggiunta di un menù a tendina per semplificare all'utente la selezione della variabile di interesse (Figura 35). Ulteriormente, è stato richiesto di includere un pulsante "calcola probabilità", per consentire all'utente di eseguire il calcolo con un solo click e visualizzare immediatamente il risultato della probabilità, con lo stesso messaggio dello script precedente.

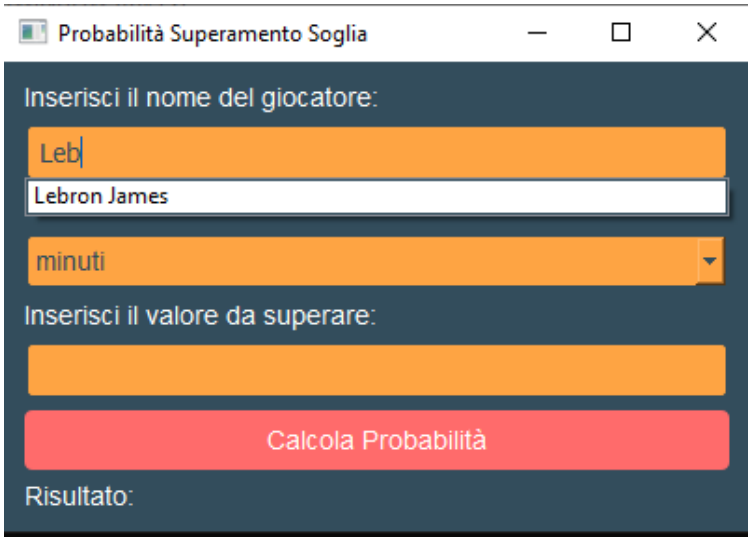


Figura 34: Esempio di inserimento automatico del nome del giocatore NBA

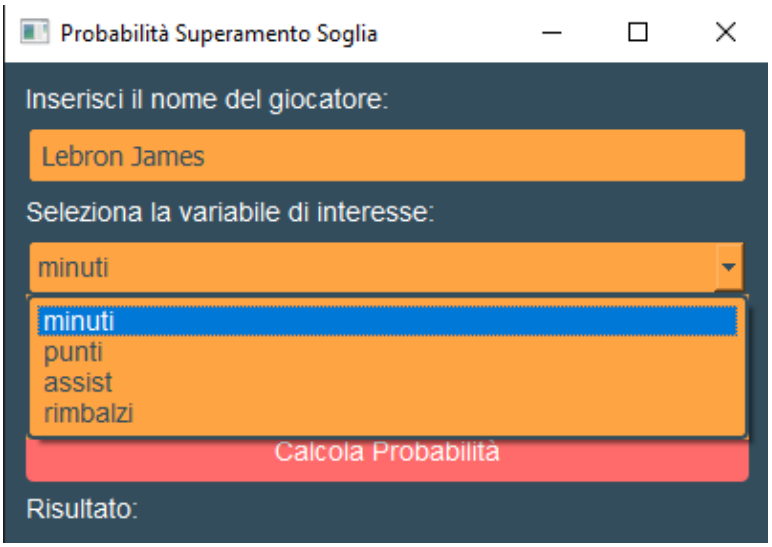


Figura 35: Menù a tendina con le variabili di interesse

3.3.3. ESEMPIO NIKOLA VUCEVIC

È possibile quindi sfruttare la nuova interfaccia grafica per visualizzare ed eseguire il calcolo di probabilità per il secondo giocatore citato precedentemente, ovvero Nikola Vucevic, centro dei Chicago Bulls, facendo riferimento questa volta alla distribuzione di rimbalzi per partita. Nello specifico, si andrà a determinare, la probabilità che Vucevic superi una determinata soglia di rimbalzi. Per il calcolo però, è necessario verificare, come fatto in precedenza con Fred VanVleet, che la distribuzione dei rimbalzi per partita del giocatore sia almeno approssimativamente Normale.

Per valutare l'approssimazione ad una distribuzione normale della distribuzione di rimbalzi realizzati per partita da Nikola Vucevic, caratterizzata da 82 osservazioni, sono stati realizzati gli stessi passaggi

attuati precedentemente con l'esempio su Fred VanVleet, utilizzando il linguaggio di programmazione R.

Innanzitutto, è stato creato un istogramma, osservabile nella Figura 36, per visualizzare la distribuzione dei dati cercando di evidenziare la presenza di una forma a campana e simmetrie rispetto al suo centro, con maggiore frequenza di osservazioni nella parte centrale e progressiva diminuzione allontanandosi dal centro. La mediana è pari a 11 rimbalzi, mentre la media è pari a 11.0122 rimbalzi. Per quanto riguarda l'indice di asimmetria calcolato, è pari a 0.1865428. Questo significa che presenta una coda destra lievemente più lunga.

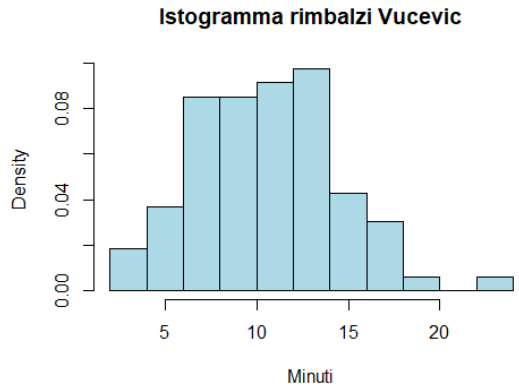


Figura 36: Istogramma della distribuzione di densità dei rimbalzi di Nikola Vucevic

Successivamente, così come per VanVleet, è stato realizzato un QQ-plot anche per la distribuzione dei rimbalzi di Vucevic, presente nella Figura 37, per confrontare i quantili osservati con quelli teorici di una distribuzione normale. Il QQ-plot suggerisce una buona approssimazione dei dati alla distribuzione normale.

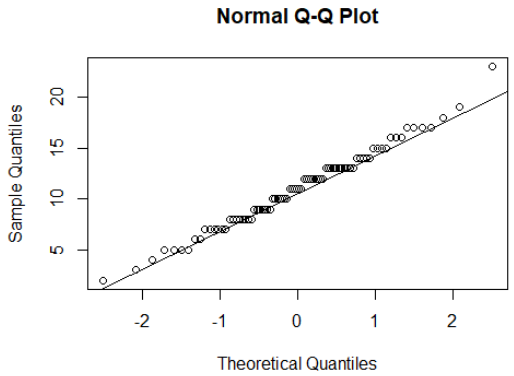


Figura 37: Confronto tra QQ-plot e retta per rimbalzi di Nikola Vucevic

Successivamente, è stato generato un grafico di densità di kernel (Figura 38) a cui è stata sovrapposta la funzione di densità della variabile casuale Normale con stesse media e varianza osservate nel campione.

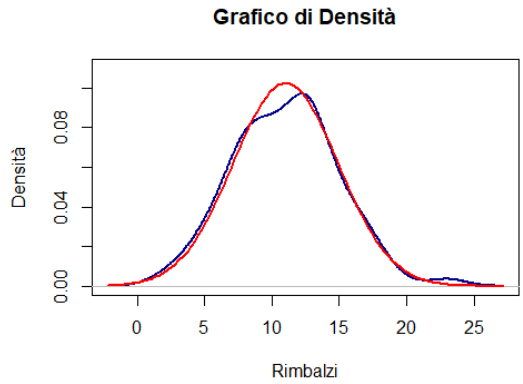


Figura 38: Confronto tra grafico di densità di kernel e funzione di densità di una variabile casuale Normale con stessa μ e σ dei rimbalzi di Nikola Vucevic

Infine, al fine di confermare la normalità della distribuzione dei dati, è stato condotto un test di Shapiro-Wilk.

```
        Shapiro-Wilk normality test

data:  dati
W = 0.98839, p-value = 0.6751
```

Figura 39: Risultato del test di Shapiro-Wilk per la distribuzione di rimbalzi di Nikola Vucevic

Come si può notare nella Figura 39, è stato calcolato un p-value pari 0.6751, quindi non si può rifiutare l'ipotesi nulla ed è possibile affermare che i dati relativi alla distribuzione di Nikola Vucevic possono essere considerati distribuiti Normalmente. Di conseguenza, è possibile utilizzare la funzione norm.cdf per calcolare la probabilità di superare una determinata soglia di rimbalzi, consentendo un risultato approssimativamente preciso. In particolare, si calcola nella Figura 40 la probabilità per Nikola Vucevic di superare durante una partita 10 rimbalzi.

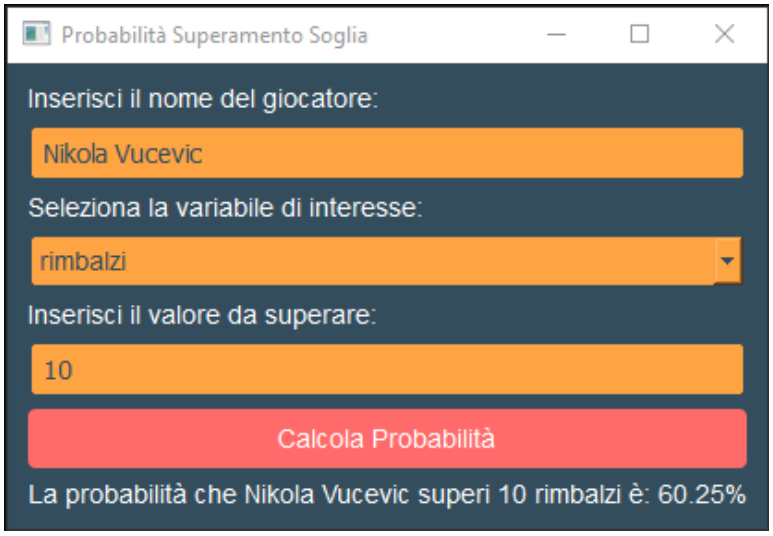


Figura 40: Calcolo della probabilità dell'esempio relativo ai rimbalzi di Nikola Vucevic

Siccome la distribuzione dei rimbalzi realizzati da Vucevic per partita può essere considerata approssimativamente Normale, l'utilizzo dell'algoritmo consente di ottenere il risultato cercato: C'è una probabilità abbastanza elevata (60.25%) che Nikola Vucevic totalizzi più di 10 rimbalzi in una partita.

In sintesi, questo capitolo ha avuto come obiettivo l'illustrazione dell'algoritmo realizzato attraverso Python per calcolare la probabilità che un giocatore superi un determinato valore per quanto riguarda variabili chiave come minuti, punti, assist e rimbalzi. Nello specifico, il programma estrae le informazioni dal file generale descritto nel secondo capitolo contenente medie e deviazioni standard per ciascuna variabile e calcola la probabilità attraverso la funzione di ripartizione della variabile casuale Normale, quindi è stato necessario verificare il fatto che le distribuzioni delle osservazioni considerate fossero approssimativamente normali.
Per verificare l'efficacia dello script, sono stati analizzati due casi specifici: la distribuzione dei punti di Fred VanVleet e la distribuzione dei rimbalzi di Nikola Vucevic.

CONCLUSIONE

In conclusione, questa relazione finale ha raggiunto gli obiettivi prefissati inizialmente. Nel primo capitolo è stata illustrata l'importanza della statistica e dei dati nel contesto sportivo, con riferimento alla pallacanestro e alle sue dinamiche. Dopo avere descritto le principali regole e funzionamenti di questo sport, si sono evidenziate le influenze che i dati possono avere sulle prestazioni e sulle strategie delle squadre.

Nel secondo capitolo, l'obiettivo legato alla descrizione dei database realizzati per questa relazione e all'analisi statistica dei giocatori e all'influenza della posizione in campo sulle prestazioni è stato soddisfatto attraverso metodi quali medie, grafici di densità, analisi della dipendenza in media rispetto al ruolo e analisi della varianza (ANOVA).

Nel terzo capitolo il focus si è spostato sull'obiettivo principale della relazione, ovvero lo sviluppo di un algoritmo in Python per calcolare la probabilità che un giocatore superi una specifica soglia di prestazione, considerando variabili come minuti, rimbalzi, assist o punti. L'algoritmo ha permesso l'estrazione e la manipolazione dei dati dei giocatori NBA da analizzare, individuati attraverso input dell'utente. Questi dati sono stati poi stati sfruttati dal complementare della funzione "norm.cdf " di Python, che fa riferimento alla funzione di ripartizione e fornisce la probabilità che la variabile casuale assuma un valore maggiore della soglia stabilita.

Tuttavia, va sottolineato che l'algoritmo presenta alcune limitazioni e mancanze. Per rendere il processo più completo potrebbe essere utile implementare all'interno dell'algoritmo un test di normalità (per esempio quello di Shapiro-Wilk) per accertarsi immediatamente della Normalità della distribuzione su cui si sta facendo il calcolo di probabilità ed evitare di effettuare calcoli non accurati e interpretazioni errate. Il problema principale che emerge è la necessità di garantire la normalità della distribuzione considerata, sottolineando l'importanza di valutare l'assunzione di Normalità prima di procedere con il calcolo della probabilità. Ulteriori implementazioni invece per quanto riguarda l'aspetto estetico possono essere ad esempio la realizzazione dell'algoritmo sotto forma di applicazione, oppure l'implementazione dei profili selezionati dei giocatori più completi, con foto e dati temporali relativi anche ad altre stagioni.

BIBLIOGRAFIA

Delmastro M., Nicita A. (2019), *Big data. Come stanno cambiando il nostro mondo*, il Mulino.

De Mauro A. (2022), *Data analytics per tutti: imparare ad analizzare, visualizzare e raccontare i dati.* Apogeo Editore.

Lewis M. (2003), *Moneyball: The Art of Winning an Unfair Game.* W.W Norton & Company.

Piccolo D. (2010), *Statistica*. Il Mulino.

SITOGRAFIA

https://www.andreapacchiarotti.it/archivio/big-data.html - ultima consultazione 30.11.2023

https://www.purestorage.com/it/knowledge/big-data/big-data-vs-traditional-data.html - ultima consultazione 30.11.2023

https://www.themarketingfreaks.com/2019/11/big-data-cosa-sono-la-storia-le-caratteristiche-le-analisied-esempi/ - ultima consultazione 30.11.2023

https://www.bigdata4innovation.it/big-data/5-v-dei-big-data-cosa-sono-quale-ruolo-rivestono/ - ultima consultazione 30.11.2023

https://www.culturedigitali.org/analytics-e-sport-dati-prestazioni-sportive/ - ultima consultazione 30.11.2023

https://it.wikipedia.org/wiki/Pallacanestro - ultima consultazione 30.11.2023

https://it.wikipedia.org/wiki/NBA - ultima consultazione 30.11.2023

https://towardsdatascience.com/nba-data-analytics-changing-the-game-a9ad59d1f116 - ultima consultazione 30.11.2023

https://www.nike.com/it/a/posizioni-nel-basket - ultima consultazione 30.11.2023

https://www.corriere.it/tecnologia/22_giugno_04/gallinari-belinelli-dai-sensori-metaverso-vi-raccontiamo-nuovo-basket-dd27ce8a-e0f0-11ec-a138-4bfa3d154041.shtml - ultima consultazione 30.11.2023