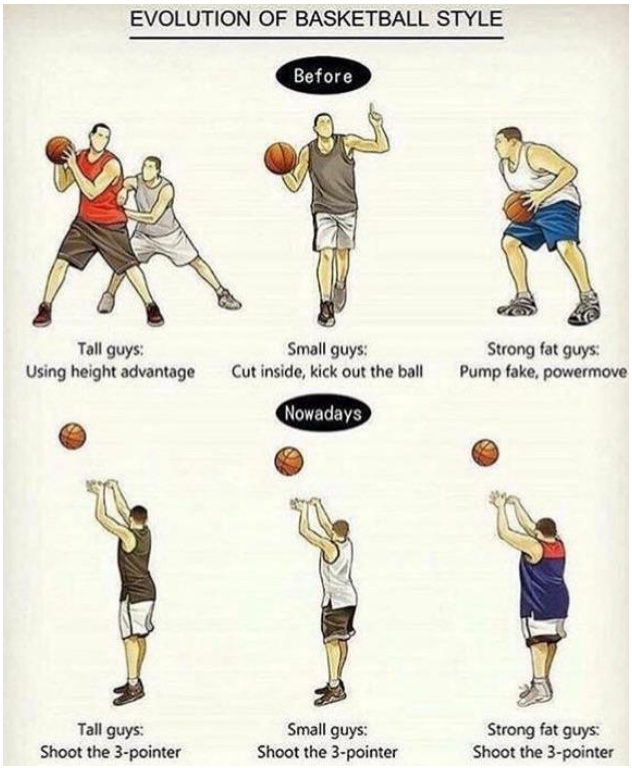
BASKETBALL ROLES ANALYSIS

JULIUS MALIWAT 864520 --- GIACOMO RABUZZI 864452 --- ANDREA ROBBIANI 865118

INTRODUZIONE

È possibile classificare i giocatori di pallacanestro per ruolo in base alle caratteristiche dei tiri che effettuano nell'arco della stagione? È proprio questa la domanda a cui questo progetto vuole dare risposta, grazie ai dati di tiro che sono stati raccolti dall'NBA dalla stagione 1997-1998 ad oggi. Il primo obiettivo del progetto è infatti quello di implementare alcuni modelli di classificazione probabilistica generativa (in particolare EDDA models e MDA) per cercare di classificare al meglio i giocatori per posizione sulla base di alcuni dati che concernono il tiro, in particolare la distanza da canestro e altre caratteristiche dei tiri tentati. Il secondo obiettivo è invece quello di verificare se, come viene ripetuto da molti cronisti e "addetti ai lavori" della pallacanestro, sia effettivamente vero che il gioco del basket, soprattutto in NBA, si stia sempre più orientando alla definitiva scomparsa dei ruoli tradizionali mutando gradualmente nel tempo; a questo scopo si dividerà il dataset in 3 "ere" e si cercherà di capire come performano modelli di classificazione allenati su una determinata era e testati sulle altre, per verificare l'esistenza di un qualche tipo di trasformazione.



DATI

I dati utilizzati sono stati ottenuti tramite un'operazione di unione di 27 differenti dataset riguardanti le caratteristiche dei tiri tentati dai giocatori NBA in 27 differenti stagioni sportive (i dati coprono solo la stagione regolare per evitare sbilanciamenti nel numero di partite potenzialmente giocabili) a partire dall'annata 1997-1998. Questi sono reperibili presso il sito Basketball Reference nella sezione shooting data per ogni stagione di interesse. In particolare, è stata fatta poi una scelta di selezione delle variabili che eliminasse tutte quelle che avessero a che fare con la qualità del giocatore al tiro, ossia tutte quelle features che includessero in qualche modo la percentuale di tiri riusciti su tiri tentati; questa scelta è stata dettata dal fatto che l'obiettivo dell'analisi non è quello di valutare la bravura dei giocatori per ruolo ma quello di classificarli indipendentemente dal livello individuale. Il dataset selezionato contiene allora le seguenti variabili:

- Player: nome del giocatore
- POS: posizione/ruolo
- MP: minuti giocati
- 0\3: percentuale di tiri tentati nell'area che dista da canestro meno di 3 piedi sui tiri tentati totali
- 3\10: percentuale di tiri tentati nell'area che dista da canestro tra 3 e 10 piedi sui tiri tentati totali

- 10\16: percentuale di tiri tentati nell'area che dista da canestro tra 10 e 16 piedi sui tiri tentati totali
- 16\3P: percentuale di tiri tentati nell'area che va dai 16 piedi alla linea del tiro da 3 punti sui tiri tentati totali
- 3P: percentuale di tiri tentati oltre la linea da 3 punti sui tiri tentati totali
- %_ASTD_2P: percentuale dei tiri riusciti da 2 punti che sono scaturiti grazie a un assist di un compagno sui tiri totali riusciti da 2 punti
- %_ASTD_3P: percentuale dei tiri riusciti da 3 punti che sono scaturiti grazie a un assist di un compagno sui tiri totali riusciti da 3 punti
- DUNKS: percentuale di schiacciate tentate sui tiri totali tentati
- %3PA_CORNER: percentuale di tiri da 3 tentati dall'angolo sul totale dei tiri da 3 tentati
- YEAR: stagione di gioco

L'analisi si è concentrata a questo punto sui valori mancanti: siccome più del 15% delle osservazioni per la variabile %_ASTD_3P avevano valore mancante, si è deciso quindi di optare per l'eliminazione (ha senso farlo perché è una variabile poco significativa, il concetto di assist sul tiro da 3 punti è spesso soggetto a interpretazioni personali); l'altra variabile con valori mancanti era %3PA_CORNER (poco più del 5% di NA) e data l'importanza nella discriminazione (in particolare per le ali) si è deciso di procedere all'imputazione della mediana condizionata per ruolo e anno di gioco.

Sempre relativamente alle variabili si è deciso di ricodificare la variabile di classificazione, ossia POS, in modo da rendere i ruoli più uniformi dividendoli in 3 modalità: guardia (G), ala (F), centro (C). Nel dataset originale, infatti, i ruoli presenti erano i 5 più classici con in aggiunta un numero esiguo di giocatori che assumevano ruoli ibridi.

Infine, si è optato per adottare un filtro sulle osservazioni del dataset eliminando tutti quei giocatori che in una determinata stagione avessero giocato meno di 1000 minuti, ossia quei giocatori marginali che vengono spesso usati dalle squadre NBA per fare rifiatare i titolari durante la stagione regolare, ma che hanno di fatto poco spazio e opportunità all'interno della stagione complessiva.

ANALISI ESPLORATIVA

Il primo passo dell'analisi esplorativa è stato quello di rappresentare la struttura di correlazione tra variabili (figura 1); in merito si possono notare alcuni elementi interessanti: vi è una forte correlazione positiva tra schiacciate e quota di tiri ravvicinati, come ci si potrebbe aspettare. Decidiamo comunque di conservare entrambe le variabili siccome possono essere utili a identificare giocatori diversi (la schiacciata è spesso operata da giocatori alti e forti come i centri, mentre il tiro ravvicinato può essere utilizzato anche da guardie e giocatori bassi in situazioni di penetrazione). Al contrario tra la quota di tiri da 3/10 piedi e 0/3, quindi in generale ravvicinati, e quota dei tiri da 3 vi è correlazione

Figura 1: Correlazione tra le variabili

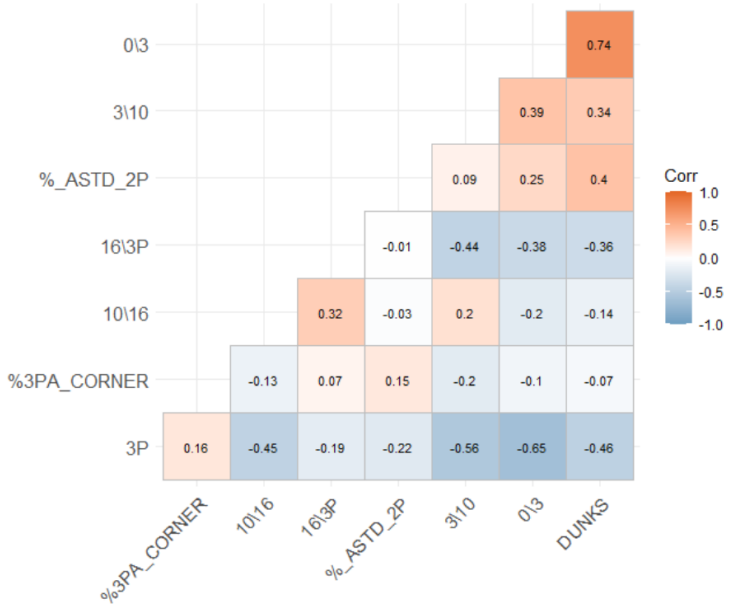
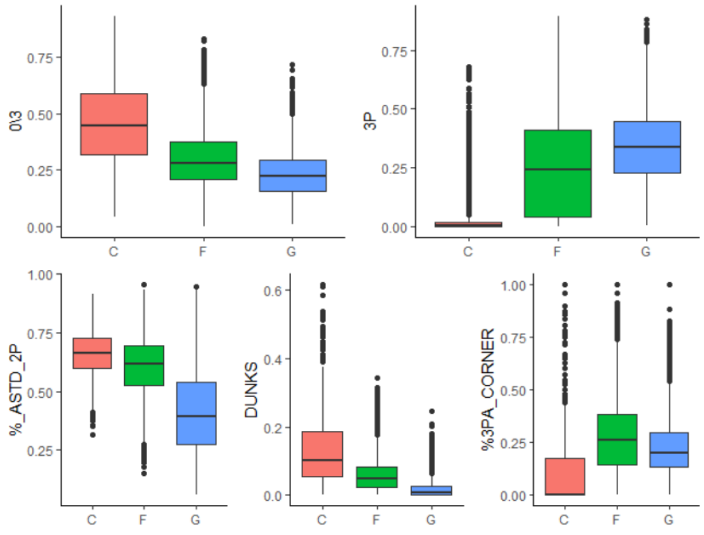


Figura 2: Boxplot delle variabili in base al ruolo

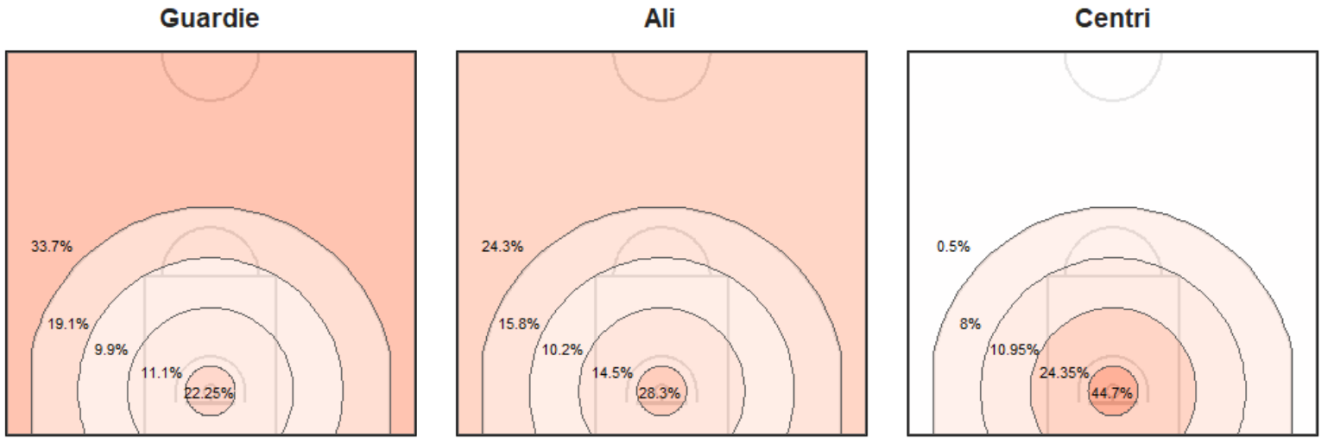


negativa; ciò è plausibile siccome i tiratori da vicino (centri) tireranno plausibilmente meno da distanze elevate (più frequenti per le guardie o ali).

Si è poi proseguito l'analisi esplorativa calcolando alcune statistiche descrittive e rappresentando con alcuni strumenti grafici le distribuzioni delle variabili condizionatamente alle classi (boxplot in figura 2) per cercare di interpretare a priori alcune possibili caratteristiche intrinseche per i vari ruoli. Si può facilmente notare come il tiro ravvicinato (0\\3) sia prerogativa dei centri e solo marginalmente delle ali. Al contrario il tiro dalla lunga distanza (3P), ossia oltre la linea dei 3 punti, diventa centrale per le guardie ma anche per le ali, quest'ultime sembrano avere una variabilità molto alta per questo tipo di tiro. È anche evidente come centri e ali si avvalgano spesso di assistenza nei tiri che segnano (%_ASTD_2P), mentre le guardie, che sono di solito i principali costruttori di gioco della squadra, hanno percentuali più basse per i tiri da 2 assistiti sui tiri a segno. Come ci si può aspettare le schiacciate (DUNKS) costituiscono una modalità di tiro particolarmente comune per i centri e il tiro da 3 dall'angolo (%3PA_CORNER) è invece prerogativa di guardie e soprattutto ali.

Si è deciso anche di rappresentare, tramite una mappa geografica del campo di gioco, la distribuzione dei tiri condizionatamente ai diversi ruoli (figura 3, la percentuale e la colorazione rappresentano la mediana per ruolo). Anche in questo caso vengono confermate le sensazioni già accennate: i centri tirano molto poco dalla lunga distanza e concentrano le proprie scelte di tiro vicino a canestro; al contrario le ali, e ancora maggiormente le guardie, prediligono il tiro da 3 punti. Anche in questo caso si nota una maggiore somiglianza tra il ruolo di guardia e ala rispetto ai centri.

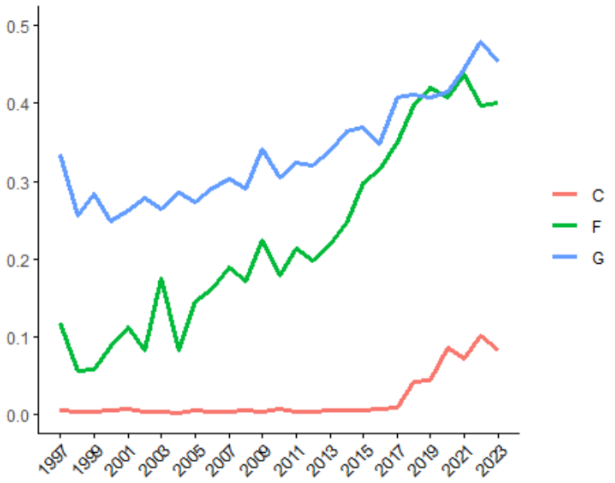
Figura 3: Distribuzione dei tiri tentati per ruolo



La sezione di analisi esplorativa prosegue con l'analisi della serie storica del tiro da 3 punti nella lega condizionatamente ai ruoli (figura 4), in modo da comprendere quale sia stata l'evoluzione dei vari ruoli nel tempo. Si può infatti notare come progressivamente l'andamento, sempre crescente, della percentuale di tiri da 3 punti su tiri totali tentati per guardie e ali sia convergente; sembra, infatti,

che le differenze siano andate riducendosi, almeno per quanto riguarda questa variabile. Inoltre, è altresì evidente come a partire dal 2017 anche i centri abbiano cominciato a considerare sempre più il tiro da 3 come un'arma offensiva utilizzabile. Sembra essere quindi confermata l'ipotesi che sia in atto una trasformazione del gioco della pallacanestro con una graduale riduzione delle differenze tra ruoli.

Figura 4: Mediana 3P nel tempo



CLUSTER ANALYSIS

Infine, si è deciso di concludere l'analisi esplorativa con l'implementazione di alcune tecniche di clustering, affinché sia possibile verificare se, a priori (senza le etichette note), questi algoritmi siano in grado di evidenziare una suddivisione in gruppi analoga a quella nota e d'interesse. L'algoritmo delle k-medie ha dato risultati confortanti con un'accuracy del 63.44 %, anche se dopo un'attenta analisi dei risultati è possibile evidenziare come la classe Ala (F) sia decisamente più problematica nella individuazione rispetto alle altre. Si è poi implementato una seconda tecnica di clustering, ossia model based cluster analysis; si è proceduto selezionando il modello (tra i 14 possibili) che minimizzasse l'ICL, che è risultato essere la forma 'VVV', che non comporta vincoli di alcun tipo sulla struttura di variabilità delle componenti. Allo stesso tempo si è imposto il numero di gruppi pari a 3, poiché noto a priori. Questo secondo metodo di clustering fornisce un'accuracy leggermente più elevata pari al 65.74 % (anche se pure in questo caso la classe delle ali risulta problematica). Si riportano nella tabella a seguire alcuni indici di bontà di classificazione per questo secondo metodo più performante.

Entropia relativa	R^2 determinante	Incertezza mediana
0.1548586	0.7445802	0.01341737

Può essere, inoltre, interessante valutare la "distanza" tra raggruppamenti per verificare se vi siano gruppi più vicini e similari. Per fare ciò si è calcolata la distanza di Kullback-Leibler simmetrizzata:

Distanza centri-guardie	2530.125
Distanza ali-centri	1321.266
Distanza guardie-ali	18.936

Si può effettivamente notare, come già evidenziato in precedenza (anche graficamente), che i gruppi di guardie e ali sono effettivamente molto più vicini delle altre coppie, è quindi più difficile discriminare i due ruoli.

MODEL BASED CLASSIFICATION

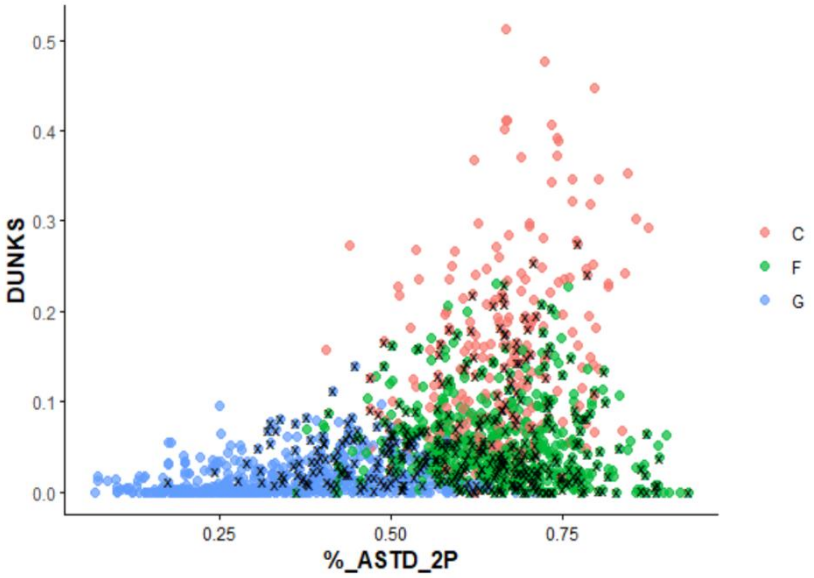
Per operare classificazione si è diviso il dataset in training (80%) e test set (20%), in modo da poter allenare i modelli scelti e testare le relative performance, tramite il MER (misclassification error rate), avvalendosi quindi della cross-validation. Il primo modello implementato è stato EDDA (che assume normalità della distribuzione delle features per classe); poiché la tecnica comporta casualità nella divisione nelle V-fold, il modello è stato allenato più volte (anche cambiando il numero di V considerate); nei vari tentativi effettuati il modello migliore è sempre risultato essere EEE, che

implica vincoli nella struttura di variabilità della mistura (matrici di varianze e covarianze uguali per tutte le componenti). Una volta scelto il modello sui dati di training si è proceduto allora alla fase di testing confrontando le etichette stimate con quelle note a priori, l'accuracy ottenuta è risultata essere 71.29%. Nessuna delle 3 classi sembra essere particolarmente problematica (la sensitivity per ogni ruolo non scende mai al di sotto della soglia del 65%); è riportata in seguito la confusion matrix:

		TRUE		
		C	F	G
PRED	C	187	72	1
	F	80	340	135
	G	0	105	449

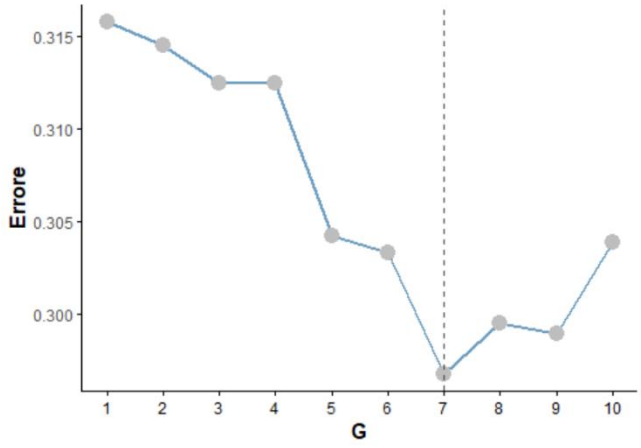
Visualizzando graficamente la classificazione (figura 5, si è operata una proiezione sulle due variabili che rendessero più evidenti i raggruppamenti), si nota che le unità mis-classificate (segnate con una **x**) si concentrano nella zona più centrale del grafico (in particolare per le ali), dove la distinzione tra unità statistiche è meno chiara.

Figura 5: Classificazione unità statistiche sulle variabili "DUNKS" e "%_ASTD_2P"



A questo punto è stato implementato un secondo modello di classificazione: MDA (mixture discriminant analysis). A differenza del modello precedente in questo caso la distribuzione delle variabili non è assunta come normale all'interno di ogni gruppo, ma si suppone sia una mistura di normali. Si è imposto che ogni mistura di normali per classe seguisse un modello EEE (struttura di variabilità comune a ogni componente interna delle misture) e si è scelto il valore G (numero di componenti per ogni mistura, supposto uguale per tutti) tramite CV, provando con valori di G e V da 1 a 10. La fase di training ha restituito come modello migliore quello per cui ognuna delle 3 componenti della mistura è a sua volta una mistura di 7 ulteriori normali (figura 6). Il test ha fornito un'accuracy del 71.37%, quindi leggermente superiore alla classificazione con modello EDDA. Anche in questo caso la classe delle ali sembra essere più problematica delle altre nella classificazione, come d'altronde

Figura 6: Cross-Validation Error per numero di cluster



anche in base all'analisi esplorativa ci si sarebbe potuto aspettare. Riportiamo la confusion matrix:

		TRUE		
		C	F	G
PRED	C	195	96	6
	F	66	272	69
	G	6	149	510

EVOLUZIONE DEI RUOLI

Infine, si è proceduto con l'ultima analisi al fine di verificare l'esistenza di una graduale evoluzione dei ruoli nella pallacanestro. Si sono allenati 3 diversi modelli MDA (con procedura analoga a quella precedente) uno per ogni era di 9 anni ciascuna e si è testato ognuno sulle ere restanti; ci si aspetta che modelli allenati e testati su ere più lontane performino in maniera peggiore in termini di accuracy rispetto a modelli allenati e testati su ere adiacenti, se effettivamente ci fosse uno sviluppo del gioco in atto. Alla luce delle nostre analisi è effettivamente confermato che vi sia un'evoluzione in corso; a tal proposito riportiamo in seguito una tabella che mostra come ogni classificatore allenato su una determinata era performi (in termini di accuracy) nel test sulle altre.

		TESTING ERA		
		1996/2005	2005/2014	2014/2023
TRAINING ERA	1996/2005		70.7%	68.3%
	2005/2014	65.0%		71.7%
	2014/2023	58.3%	67.1%	

Si nota in particolare come, oltre alla sopracitata evoluzione dei ruoli, la classificazione sia particolarmente poco efficace nella prima era. Non a caso, infatti, nella narrazione sportiva si sottolinea spesso come l'evoluzione maggiore del gioco sia avvenuta proprio in quegli anni, quando i centri hanno cominciato a giocare in posizioni del campo anche più lontane da canestro e il tiro da 3 è diventata sempre più la soluzione offensiva preferita da guardie e ali.

CONCLUSIONI

In conclusione, si può affermare che l'analisi svolta ha centrato gli obiettivi prefissati: sono stati efficacemente implementate due tecniche di classificazione per discriminare i giocatori NBA dal 1996 al 2023 per ruolo in base alla distribuzione e alle caratteristiche dei tiri tentati. Entrambe le tecniche hanno riportato risultati discreti con un'accuracy superiore al 70%. Ciò che è risultato evidente è stata la difficoltà nella discriminazione dei ruoli guardia e ala, che risultano spesso sovrapponibili.

Anche il secondo obiettivo ha dato risultati soddisfacenti, mostrando come ci sia effettivamente stata un'evoluzione nell'interpretazione dei vari ruoli nel corso degli anni. Sembra in particolare che il momento di svolta sia stato tra la prima e la seconda era (quindi nei primi anni 2000), quando il gioco ha subito trasformazioni radicali. Infatti, i modelli allenati sulle ere più recenti risultano poco efficaci nella classificazione dei giocatori della prima era.