Statistica computazionale Progetto:

Alessandro Zanotta 827320

APRILE 22, 2020

Abstract

Il progetto si è svolto in merito all'analisi di un data set scaricato dal forum KAGGLE con dati riguardanti il gioco per consolle FIFA20.
FIFA 20 è un gioco di calcio creato da EASPORTS.

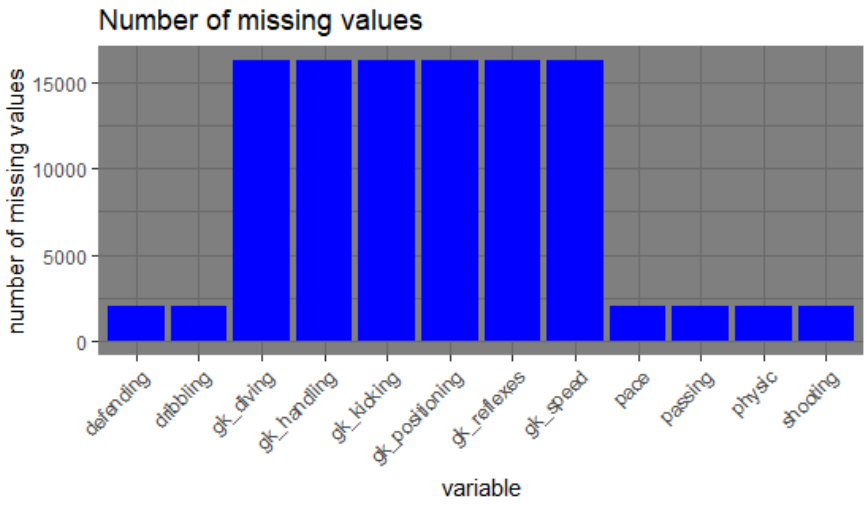
1 Introduzione

Scopo della mia analisi sarà classificare i giocatori in uno dei quattro ruoli: portiere, difensore, centrocampista o attaccante.
L'obiettivo dello studio, condotto sulle osservazioni composte dagli attributi abilità per ciascun giocatore, è la formulazione di un modello che ci permetta di classificare il giocatore in 4 gruppi, che rappresentano il ruolo del giocatore. La domanda che mi ha guidato durante lo svolgimento del lavoro, è la seguente: è possibile trovare un modello che, considerando le abilità dei giocatori, mi consenta di determinarne il ruolo occupato nel gioco? In questa relazione illustrerò i passaggi che hanno portato all'elaborazione dei metodi di classificazione dei giocatori nei 4 diversi ruoli.

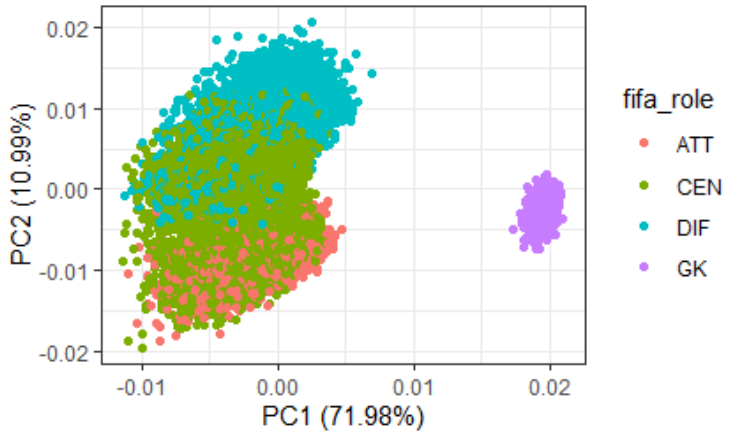
2 Materiali e metodi

MATERIALI

Il dataset su cui ho lavorato presentava 18278 osservazioni con 104 variabili. Tra queste ultime 33 non erano di mio interesse in quanto riferite ad aspetti non significativi per la mia analisi. Delle 71 variabili di mio interesse 24 si riferivano ai progressi del giocatore in un determinato ruolo, le restanti 47 ad attributi fisici e tecnici del giocatore. Il dataset nel complesso si presentava poco "pulito", erano presenti missing in alcuni attributi, ad ogni giocatore erano assegnati tre ruoli separati da una virgola sotto l'attributo "players_positions" e inoltre le 24 variabili riferite al progresso del giocatore nel corso gioco si presentavano come factor con al loro interno una somma tra virgolette (come ad esempio "89+2"), tali variabili erano fondamentali per la buona riuscita della task.



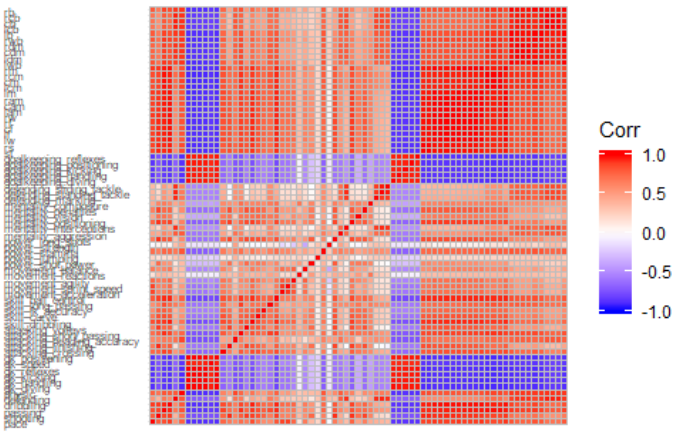
La presenza di missing values non era casuale e dipendeva dal ruolo che il giocatore occupa. Come possiamo dedurre dal grafico la maggior parte dei missing values si concentrava in attributi riferiti ai portieri (GK). Nel trattamento dei valori mancanti, sia per le osservazioni riferite ai portieri che per quelle riferite ai giocatori di movimento, ho imputato il numero zero. I giocatori di movimento avranno abilità da portieri nulle e all'inverso i portieri avranno le abilità dei giocatori di movimento pari a 0. Le 26 variabili che si presentavano come "factor" e con valori numerici compresi tra virgolette erano fondamentali per l'analisi perché mostravano il potenziale che ciascun giocatore avrebbe conseguito in un determinato ruolo nel corso del gioco, motivo per cui nel loro trattamento ho deciso di sommare i due numeri tra le virgolette facendoli diventare un unico numero. Un ulteriore problema presente nel dataset è stato che ogni giocatore per la variabile "player_positions" presentava da uno a tre ruoli separati da una virgola. Ho dovuto ricondurre il primo ruolo (il ruolo principale, nel caso fossero stati due o tre) a una delle 4 classi che mi interessava discriminare.



METODI

PRE-PROCESSING

Dato l'elevato numero di variabili del mio dataset ho applicato due diversi metodi per la riduzione della dimensionalità: l'analisi delle componenti principali e l'analisi delle correlazioni. Confrontando poi i risultati dei metodi di classificazione sui due dataset ridotti alla ricerca del modello più performante. Tramite l'analisi delle componenti principali ho selezionato le prime 8 componenti, che spiegano il 95% della varianza. Come si può notare dal grafico delle prime due componenti principali la classe che risulta essere meglio discriminata dalle altre è quella riferita ai portieri (GK), ciò è coerente con quella che è anche la realtà del gioco del calcio essendo il portiere l'unico giocatore non di movimento. Più difficoltà sembrerebbero esserci invece nel discriminare i centrocampisti (CEN) sia dagli attaccanti (ATT) che dai difensori (DIF). Sono infatti presenti centrocampisti con spiccate doti offensive e spiccate doti difensive. Le classi dei difensori e attaccanti risultano distinguibili in maniera chiara tra di loro, infatti i due ruoli necessitano di doti fisiche e tecniche differenti.



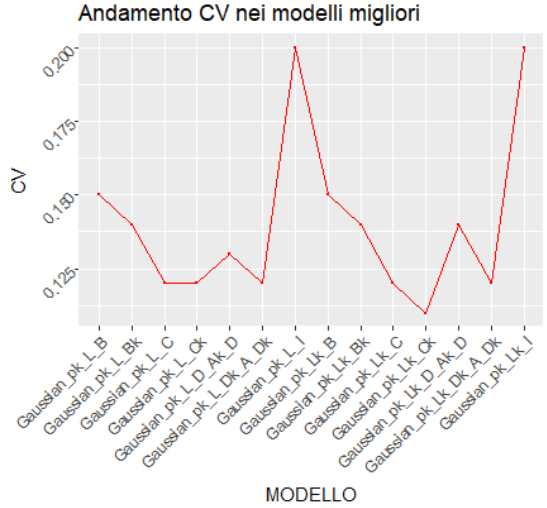
Attraverso l'analisi delle correlazioni ho eliminato 56 delle 71 variabili quantitative a mia disposizione (presenti nel grafico qui a lato). Come era prevedibile le variabili riferite al potenziale del giocatore in un determinato ruolo erano altamente correlate con le abilità fisiche e tecniche caratterizzanti il ruolo stesso, per tale motivo le 26 variabili che indicavano il potenziale del giocatore nel ruolo in questa analisi sono state scartate. A testimonianza di tale osservazione occorre ricordare che la selezione è avvenuta eliminando le variabili che presentavano tra di loro una correlazione maggiore di 0,8 (in valore assoluto). Nella scelta tra due variabili altamente correlate è stata eliminata la variabile con correlazione media più alta con le altre variabili.

CLASSIFICAZIONE

Per classificare le osservazioni in uno dei 4 gruppi ho utilizzato i metodi di classificazione basati sulle misture. Avendo a disposizione le variabili riferite alle etichette ho utilizzato un approccio di tipo supervised con l'utilizzo di modelli EDDA ed MDA. Come criterio di scelta tra i modelli EDDA ed MDA ho deciso di utilizzare il CV, perché ho ritenuto importante privilegiare di più l'aspetto predittivo del modello. Esso consiste nella minimizzazione del classification error rate sul training set. Il training set è composto dall'80% delle osservazioni a mia disposizione mentre il test set dal restante 20%. Andiamo a vedere quindi le performance sui due diversi dataset ottenuti dall'analisi delle componenti principali e dalla riduzione di dimensionalità tramite l'analisi delle correlazioni.

PCA EDDA

Con l'utilizzo dei modelli EDDA sul dataset composto dalle componenti principali il miglior modello è risultato essere il "GAUSSIAN_pk_Lk__Ck", che corrisponde alla QUADRATIC DISCRIMINANT ANALYSIS.



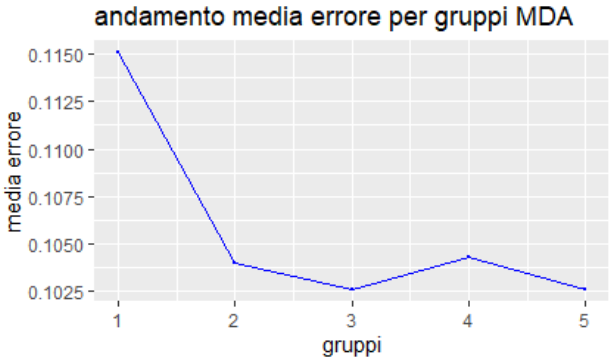

```
          Reference
Prediction  ATT  CEN  DIF   GK
       ATT  498  105    0    0
       CEN   64 1280  116    0
       DIF    3  171 1036    0
       GK     0    0    0  382

Overall Statistics

               Accuracy : 0.8744
                 95% CI : (0.8632, 0.885)
    No Information Rate : 0.4257
    P-Value [Acc > NIR] : < 2.2e-16
```

Un risultato coerente con i grafici visti in precedenza in quanto questo modello risulta

avere tra tutti i modelli le linee di confine più flessibili per la classificazione.
Su quest' ultimo abbiamo ottenuto un'accuracy pari all'87%. Come era prevedibile la classe più semplice da discriminare è stata quella dei portieri mentre la più complicata è stata la classe dei centrocampisti.

PCA MDA

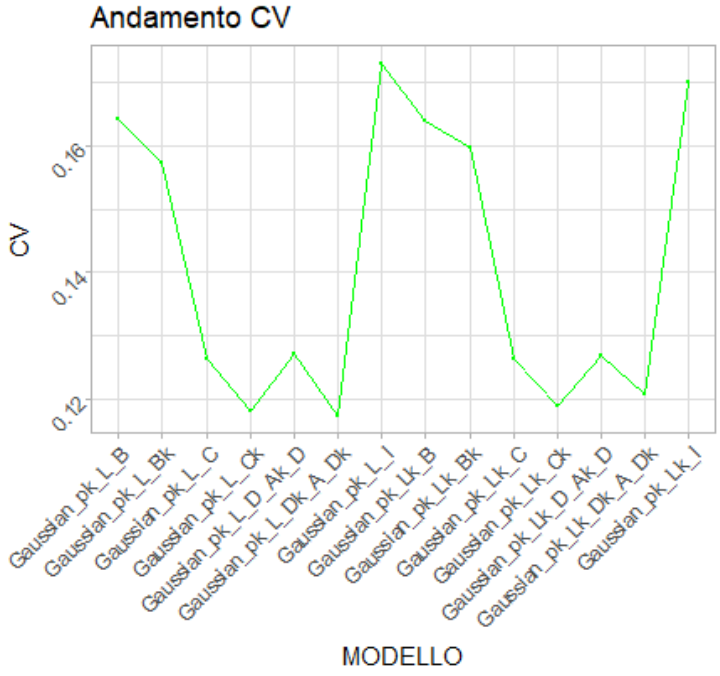
Con l'utilizzo della mixture discriminant analysis sul dataset delle componenti principali ho ottenuto un'accuracy pari all' 89,88%. Basandomi sul risultato ottenuto in precedenza con il modello EDDA ho deciso di scegliere il numero ottimale di componenti per le misture di ogni classe tra i modelli di tipo "VVV" la cui matrice di varianze e covarianze cambia per ogni componente (formata in questo caso a sua volta da misture) del modello. Il numero di componenti per mistura che minimizza la media dell'errore come si evince dal grafico risulta essere pari a 5.




```
          Reference
Prediction  ATT  CEN  DIF   GK
       ATT  483   53    0    0
       CEN   80 1366   98    0
       DIF    2  137 1054    0
       GK     0    0    0  382

Overall Statistics

               Accuracy : 0.8988
                 95% CI : (0.8885, 0.9084)
    No Information Rate : 0.4257
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8519
```

CORRELAZIONE EDDA

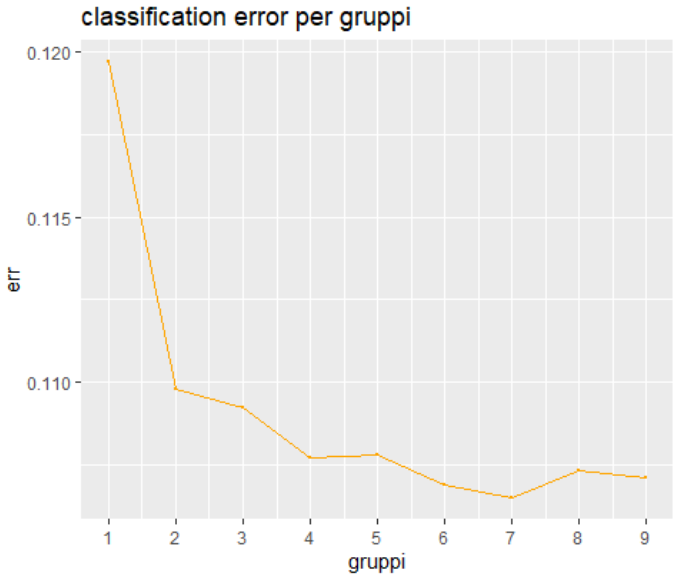
Dopo aver ridotto la dimensionalità tramite la correlazione ho ottenuto un dataset con 15 variabili. Con l'utilizzo dei modelli EDDA sul training set in base al CV ho ottenuto come modello il "GAUSSIAN_pk_L_Dk_A_Dk" che presenta volume fisso(L), forma fissa (A) e orientation (Dk) variabile. Tale modello mi fornisce un accuracy sul test pari all'87,99%.




```
          Reference
Prediction  ATT  CEN  DIF   GK
       ATT  440   91    1    0
       CEN   60 1310  140    0
       DIF    2  145 1061    0
       GK     0    0    0  405

Overall Statistics

               Accuracy : 0.8799
                 95% CI : (0.8689, 0.8903)
    No Information Rate : 0.423
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8245
```

CORRELAZIONE MDA

Utilizzando i modelli MDA ho ottenuto in termini di accuracy un punteggio pari all'89,14%.




```
          Reference
Prediction  ATT  CEN  DIF   GK
       ATT  483   74    1    0
       CEN   80 1349  107    0
       DIF    2  133 1044    0
       GK     0    0    0  382

Overall Statistics

               Accuracy : 0.8914
                 95% CI : (0.8808, 0.9013)
    No Information Rate : 0.4257
    P-Value [Acc > NIR] : < 2.2e-16
```

La scelta come specificato in precedenza anche in questo caso è ancora avvenuta tra modelli di tipo "VVV". Il numero di gruppi ottimale per ciascuna componente, in termini di errore medio risulta essere pari a 7.

3 Discussione

Il dataset presentava ingenti dimensioni. L'analisi esplorativa ha evidenziato la presenza di dati poco puliti: missing values e variabili non pronte per essere subito trattate. Attraverso un'accurata pulizia dei dati sono riuscito a risolvere tali problematiche. Ho ottenuto dei buoni risultati in termini di accuracy. Il migliore, per pochi centesimi, utilizzando la MDA sul dataset delle componenti principali, tale procedimento porta sì al migliore risultato in termini previsivi, ma dall'altro lato porta con sé anche una scarsa interpretabilità delle variabili. Lo stesso metodo applicato sulle variabili scelte in termini di correlazione porta ad un risultato con accuracy poco minore ma con maggiore interpretabilità delle variabili. L'obiettivo di classificare i giocatori nei 4 diversi ruoli posto all'inizio dell'analisi è stato raggiunto con buoni risultati. Sicuramente approcci di tipo predittivo porterebbero a performance migliori in termini di accuracy.