FOOTBALL TEAMS: MODEL BASED - CLASSIFICATION & CLUSTERING

Addis Gaia - 864410 Dell'Elba Lidia - 876217 Furfaro Serranò Antonio - 870770

1. Descrizione del Dataset

Il database relazionale "Football Database" contiene 7 tabelle relative alle principali statistiche (su squadre e calciatori) dei migliori 5 campionati di calcio europei dal 2014 al 2020.
Fonte database: *https://www.kaggle.com/datasets/technika148/football-database*

Il dataset oggetto d'analisi *stats_summary* è stato creato ad hoc gestendo 4 delle 7 tabelle disponibili.
Si rimanda al codice per una descrizione più dettagliata della fase di data engineering.
stats_summary è formato da 686 osservazioni (squadre di calcio) e 13 variabili.

VARIABILE	DESCRIZIONE
Team Name Season	Nome della squadra + anno del campionato
League Name	Nome del campionato: Bundesliga (GER) , La Liga (SPA) , Ligue 1 (FRA), Premier League (ENG) , Serie A (ITA)
Tot Points	Totale dei punti
Tot Goals	Totale dei goal
Expected Tot Goals	Stima del totale dei goal
Mean Shots Accuracy	Media della precisione del tiro (Shots Accuracy = Shots On Target / Shots)
Tot Deep	Passaggi totali completati (esclusi i cross) entro 18 metri dalla porta avversaria
Mean PPDA	Media dell'indicatore di pressing pdda (PPDA = rapporto tra il numero di passaggi effettuati dalla squadra che imposta e il numero di azioni difensive (tackle, intercetti e falli) compiute dalla squadra che aggredisce senza palla)
Tot Fouls	Totale dei falli
Tot Corners	Totale dei corner
Tot Cards	Totale dei cartellini (rossi + gialli)
Position	Posizione in classifica
Stand Pos	Fascia della classifica (HS = alta classifica, MHS = medio-alta classifica, MLS = medio-bassa classifica, LS = bassa classifica)

Tabella 1: Variabili utilizzate nell'analisi

2. Obiettivi

L'obiettivo dell'analisi è duplice: da un lato, utilizzando i dati relativi alle prestazioni offensive e difensive della squadra durante la stagione, si intende determinare in quale fascia della classifica terminerà il campionato (alta, medio alta, medio bassa, bassa); dall'altro, si desidera individuare dei cluster nascosti interpretando i risultati ottenuti grazie alle informazioni raccolte.

3. Analisi Esplorativa

L'esplorazione del dataset verte a valutare le correlazioni tra le variabili, le distribuzioni condizionate alla classe e non, ed il range di variazione attraverso i boxplot condizionati.

Analizzando la correlazione tra le variabili si nota che il coefficiente di correlazione fra la variabile *tot_goals* e la variabile *x_tot_goals* (expected tot goals) è pari a 0,92; pertanto rimuoviamo dal dataset *x_tot_goals* che risulta ridondante. Inoltre, si nota che le variabili *tot_cards* e *tot_fouls* risultano le meno correlate con tutte le altre, mentre *tot_goals* e *tot_points* è la coppia di variabili più correlata. Questo risultato non è sorprendente, infatti le squadre che segnano di più tendenzialmente vincono anche più partite (cioè accumulano più punti).

Attraverso l'analisi degli istogrammi delle variabili quantitative (differenziate per *stand_pos*) con sovrapposte le stime di densità non parametriche, si evince che le variabili più discriminanti rispetto alla fascia della classifica sono: *tot_points*, *tot_goals*, *mean_shots_accuracy*, *tot_deep* e *tot_corners*. Questo risultato è confermato dai boxplot condizionati a *stand_pos* (*Figura 1*), dove si osserva, inoltre, che la classe HS risulta essere quella più distante dalle altre.

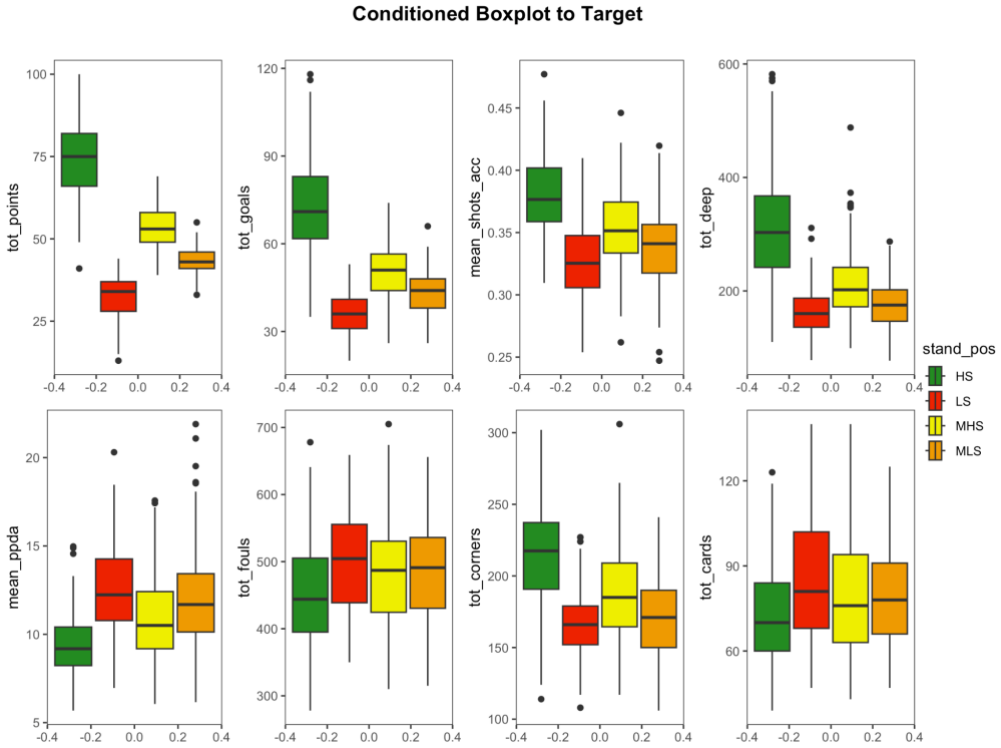


Figura 1: Boxplot Condizionati alla variabile target "stand pos"

Inoltre, è possibile notare la differenza nei range di variazione delle variabili in esame, la quale normalmente condurrebbe ad una standardizzazione dei dati. In questo caso, però, abbiamo deciso di non procedere in tal senso, in quanto le funzioni utilizzate per la classificazione e per il clustering effettuano una standardizzazione dei dati al loro interno oppure non la necessitano.

Attraverso l'help (digitando *mclust.options()*) abbiamo constatato che le funzioni di classe *mclust* adoperano una standardizzazione per i dati in input con la tecnica SVD (default di *R*).
La funzione *mixmodLearn*, invece, non richiede la standardizzazione dei dati. Per verificarlo abbiamo standardizzato i dati e rieseguito il codice, ed il modello risultante, così come l'accuracy e le unità erroneamente classificate sono rimaste invariate, a differenza ovviamente dei parametri stimati, che seguono l'ordine di grandezza dei dati di input.

4. Classificazione

Ai fini della classificazione verrà inserita tra le features la variabile *tot_points*. La quantità di punti di una squadra ha ottima capacità discriminativa rispetto a *stand_pos*, ma non implica necessariamente il posizionamento in

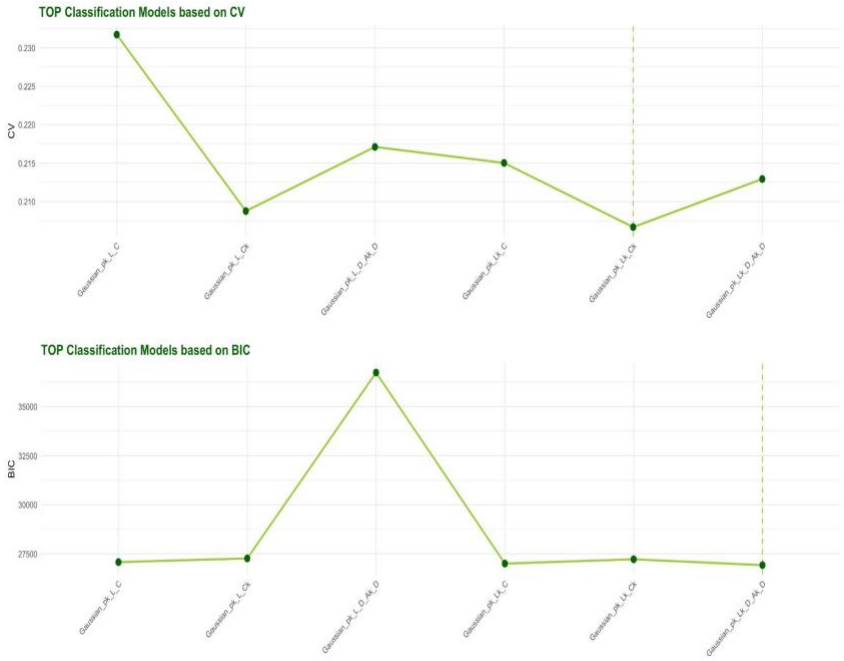
classifica. Ciò significa che squadre che hanno lo stesso numero di punti possono trovarsi in posizioni differenti in classifica a seconda dell'anno o della lega in cui competono.

Dopo aver svolto le analisi preliminari è possibile procedere con la classificazione. Applicheremo due modelli: **EDDA** ed **MDA**.

Innanzitutto, si divide il dataset in training e test set rispettivamente al 70% e 30%, e si valuta poi il bilanciamento delle classi. Dopo aver osservato la frequenza relativa delle osservazioni per ogni classe all'interno dei due set, si conclude che in quest' ultimi le classi sono bilanciate e quindi non ci troviamo in ambito di campionamento retrospettivo.

Per avere un'idea sui gruppi che il classificatore distinguerà meglio o peggio calcoliamo la *Symmetrized Kullback-Leibler Divergence* per valutare le distanze tra i gruppi. Da questa analisi è emerso che i gruppi più distanti sono quelli individuati dalle etichette HS ed MLS, mentre i meno distanti sono quelli individuati dalle etichette MHS ed MLS (le due classi centrali).

4.1) MODEL BASED CLASSIFICATION - EDDA



Tra i possibili modelli **EDDA** sono stati considerati soltanto i modelli con mixing weights non fissi dato che il campionamento non è stato di tipo retrospettivo. Confrontando i valori assunti dal *CV* (Cross Validation) il modello da scegliere risulta essere VVV (QDA), invece guardando ai valori del *BIC* (Bayesian Information Criterion) il modello migliore è VVE (il *BIC* privilegia modelli meno complessi). Selezioniamo il modello scelto con il criterio del *CV* e quindi VVV. (*Figura 2*)

Figura 2: Valori di CV e BIC per i migliori modelli

Nella *Tabella 2* si riportano i risultati della classificazione per il test set ottenuta allenando il classificatore sul training set con il modello QDA.

MODEL NAME	ACCURACY	CV	BIC
VVV / QDA	**79,23%**	**0,203**	**27222**

Tabella 2: Risultati per modello EDDA migliore

Dalla *Figura 3* è possibile notare che il classificatore fa più fatica ad allocare correttamente le unità appartenenti alle classi MHS ed MLS, le quali, come detto in precedenza, risultano essere le più simili. Per le suddette unità, infatti, l'incertezza risulta essere maggiore rispetto all'incertezza associata alle unità appartenenti alle classi HS ed LS (classi estreme). Di conseguenza, la maggior parte delle unità classificate erroneamente appartengono alle classi MHS ed MLS. La scelta delle variabili da rappresentare è ricaduta su quelle che, come visto nei boxplot, permettono di discriminare meglio le classi.

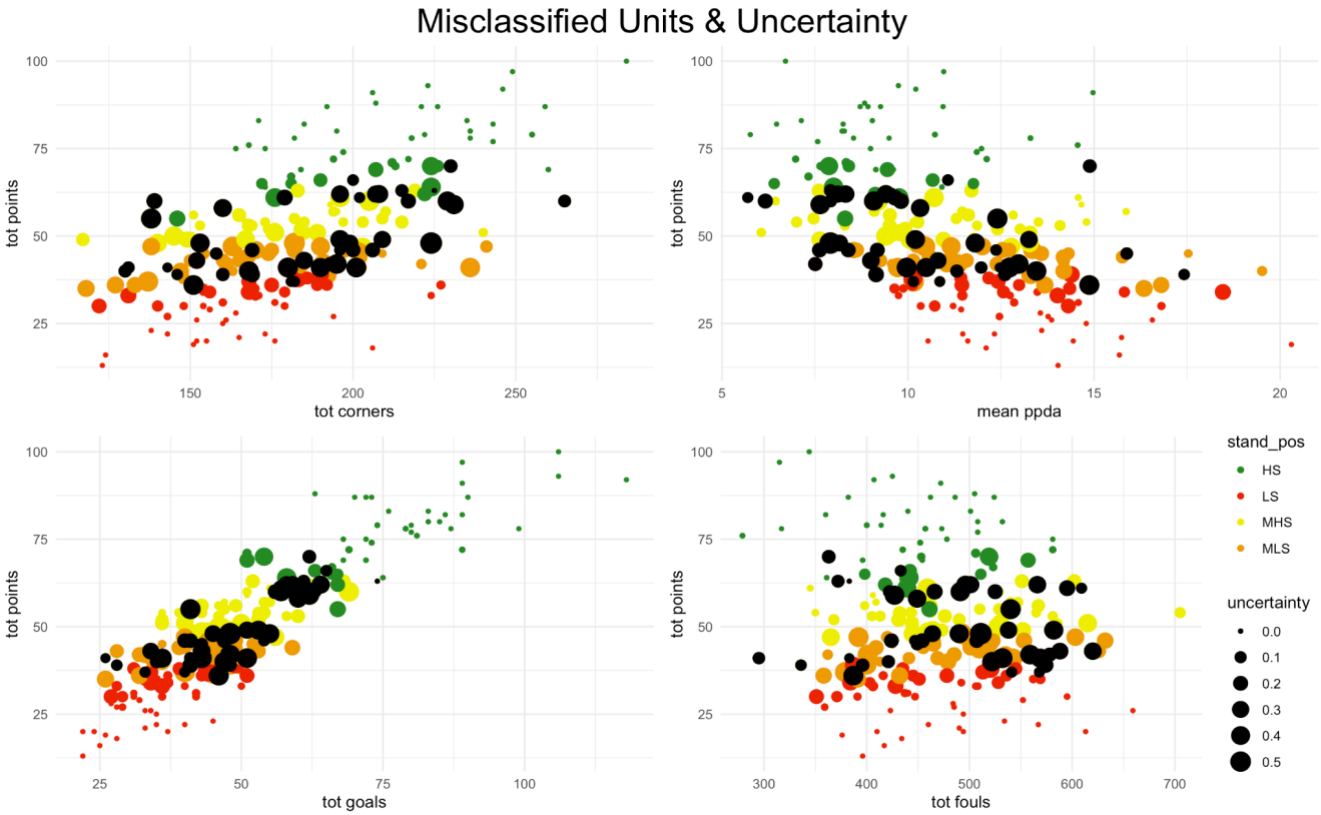
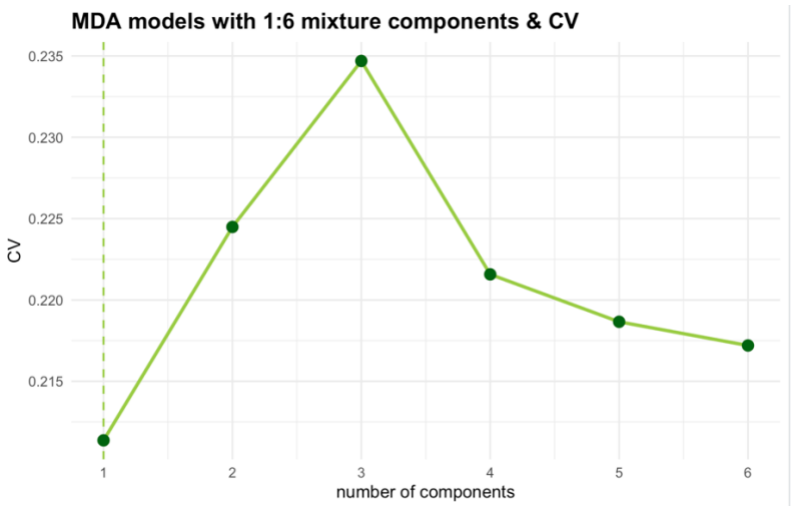


Figura 3: Scatterplot colorati in base alle classi assegnate, in nero i punti classificati erroneamente. La grandezza dei punti è proporzionale all'incertezza

4.2) MODEL BASED CLASSIFICATION - MDA



Applicando **MDA** e selezionando i modelli sia in base al *BIC* che al *CV* (*Figura 4*), il miglior modello risulta avere una sola componente per ogni mistura interna, ovvero si ritorna alla classificazione mediante **EDDA**.

Figura 4: Valori CV per modelli MDA con misture da 1 a 6 componenti

5. Clustering

Lo scopo del clustering in questa analisi è quello di raggruppare le squadre a seconda delle loro statistiche stagionali in classi non assegnate a priori ed interpretarne i risultati.

5.1) MODEL-BASED CLUSTERING - modello VVE,3

In base al criterio del *BIC* viene selezionato il modello VVE con 3 gruppi. Si scelgono *tot_goals* e *mean_ppda* come variabili da mostrare nel grafico poiché risultano utili per interpretare i cluster. La variabile *tot_goals* rappresenta una statistica offensiva (gol segnati), mentre *mean_ppda* una statistica difensiva (pressing). Ogni gruppo sarà quindi caratterizzato da un livello di abilità di gioco sancito dall'attacco e dalla difesa.

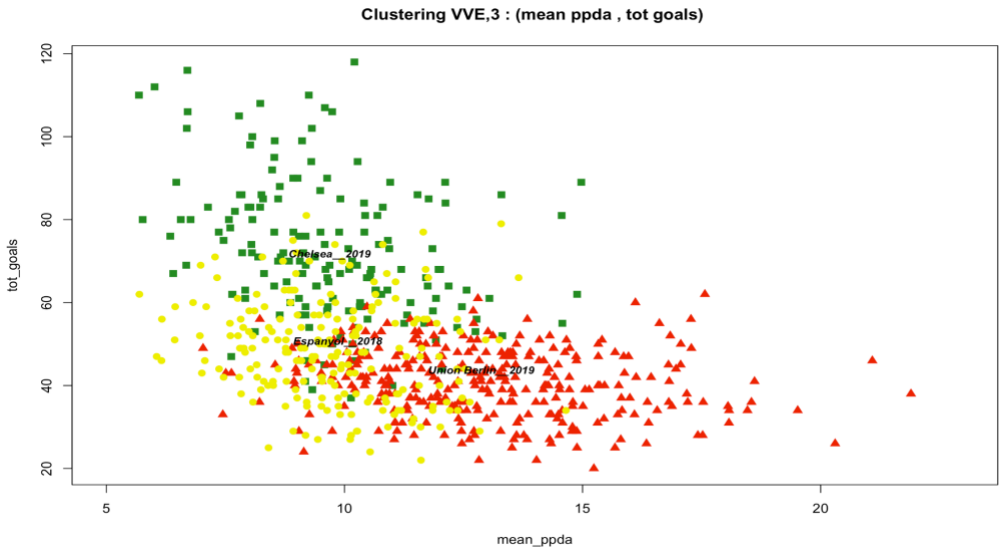
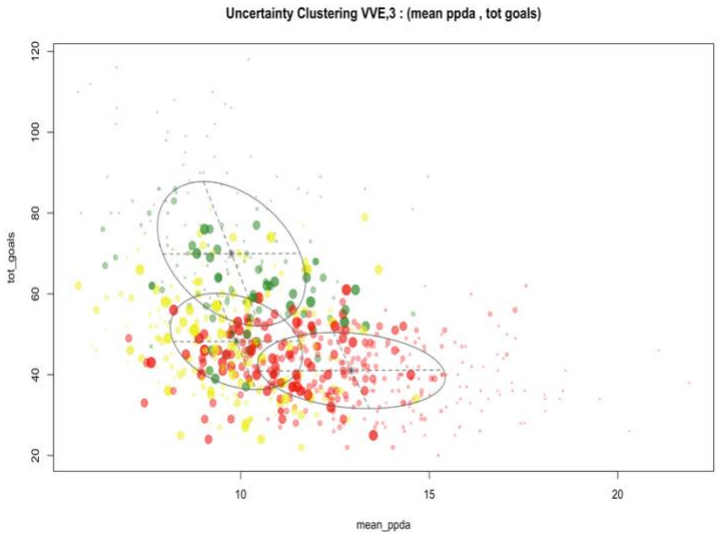


Figura 5: Scatterplot delle unità divise in 3 gruppi secondo il modello VVE

Dalla *Figura 5* è possibile dedurre che: il gruppo verde rappresenta le squadre di alto livello (forte attacco e forte difesa), il gruppo giallo le squadre di medio livello (modesto attacco e modesta difesa) e il gruppo rosso le squadre di basso livello (scarso attacco e scarsa difesa). L'aggiunta dei nomi delle squadre (per sapere come abbiamo individuato tali unità si rimanda al codice) supporta l'interpretazione dei gruppi. Il Chelsea del 2019 è arrivato terzo qualificandosi in Champions League e ha disputato un ottimo campionato sia dal punto di vista offensivo che difensivo. L' Espanyol del 2018 è invece riuscito a qualificarsi in Europa League conquistando la settima posizione del campionato spagnolo. L'Union Berlin del 2019 si è invece posizionato nella parte destra del tabellone non qualificandosi in nessuna competizione europea, ma consolidando il suo posto in Bundesliga per l'anno successivo.



Dalla *Figura 6* si riscontra una maggiore incertezza nelle unità a cavallo tra i gruppi di medio e basso livello, mentre le unità assegnate al gruppo di alto livello riportano meno incertezza. Questo potrebbe essere dovuto al fatto che le squadre più competitive si differenziano maggiormente dalle altre.

Figura 6: VVE a 3 gruppi: Rappresentazione dell'incertezza per ogni unità statistica

Dalla *Tabella 3*, come ci si poteva aspettare, distanza è maggiore tra il gruppo rappresentante le squadre di alto livello e quello rappresentante le squadre di basso livello. La distanza minore è, invece, tra il gruppo rappresentante le squadre di medio livello e quello rappresentante squadre di basso livello. Queste misure ci suggeriscono che le squadre di medio e basso livello sono più simili tra loro, mentre le squadre di alto livello si differenziano di più dagli altri 2 gruppi, proprio come avevamo intuito dalla *Figura 6.*
Sia l'R^2 (calcolato con il determinante) che l'*Entropia* indicano una buona classificazione.

MODEL NAME	VVE, 3
R SQUARED	0,848
ENTROPY	0,322
KLs (1,2)	15,057
KLs (1,3)	11,083
KLs (2,3)	7,288

Tabella 3: Indicatori di bontà dei cluster: R^2, EN, KLs

5.2) MODEL-BASED CLUSTERING - modello EVE,2

Utilizzando il criterio *ICL* viene invece selezionato un modello EVE con 2 sole componenti.

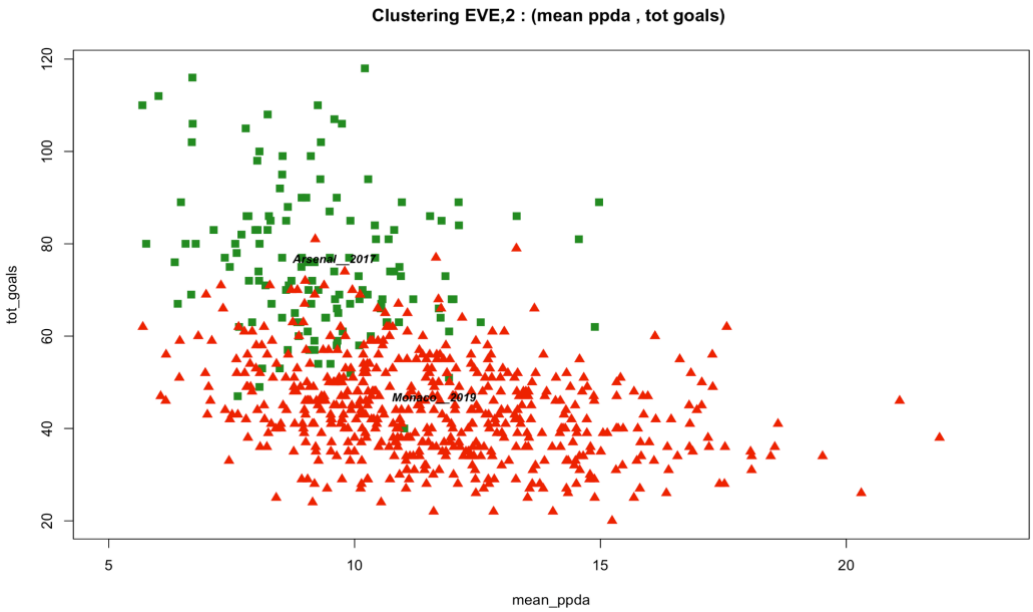
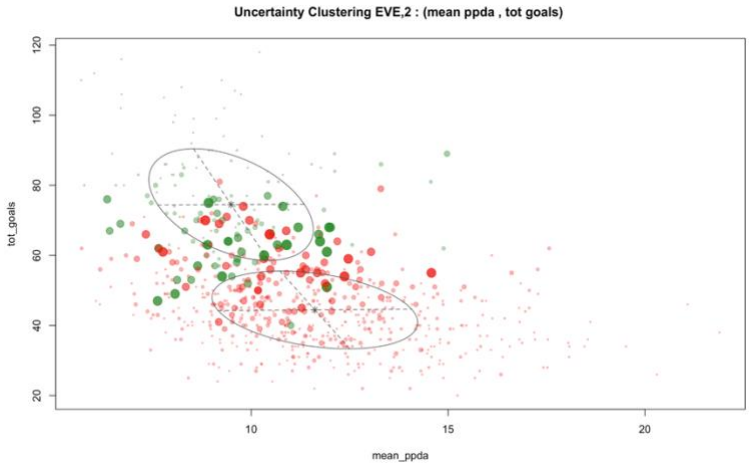


Figura 7: Scatterplot delle unità divise in 2 gruppi secondo il modello EVE

Dalla *Figura 7* è possibile dedurre che: il gruppo verde rappresenta le squadre di alto livello, (forte attacco, forte difesa) e il gruppo rosso le squadre di medio-basso livello (mediocre/scarso attacco, mediocre/scarsa difesa). Rispetto al precedente cluster a 3 gruppi sembra quasi che le squadre di altissimo livello (Champions League / Europa League) siano state isolate, mentre le squadre di medio e basso livello unite. L'aggiunta dei nomi delle squadre supporta ancora una volta l'interpretazione dei gruppi: l'Arsenal del 2017 si è qualificato in Europa League, mentre il Monaco del 2019 si è piazzato a metà classifica.



Si nota che, rispetto alla *Figura 6*, ci sono meno unità con alta incertezza; ma anche in questo caso, ovviamente, le unità con maggiore incertezza sono quelle che si trovano in una posizione intermedia.

Figura 8: EVE a 2 gruppi: Rappresentazione dell'incertezza per ogni unità statistica

A primo impatto questa clusterizzazione appare migliore della precedente, sia valutando l'R^2 che l'*Entropia.*

MODEL NAME	EVE, 2
R SQUARED	0,998
ENTROPY	0,121
KLs (1,2)	11,266

Tabella 4: Indicatori di bontà dei cluster: R^2, EN, KLs

In conclusione, nonostante la clusterizzazione EVE,2 risulta essere la migliore, si preferisce la clusterizzazione VVE,3 perché permette una migliore interpretazione dei gruppi. Infatti, raggruppare in un unico cluster sia le squadre di medio che basso livello risulta essere troppo approssimativo.