Obiettivo:

L'obiettivo di questa analisi nasce dall'idea di sfruttare le statische di utilizzo delle diverse zone del campo da gioco per caratterizzare le squadre in base alle aree di forza e di debolezza, sia in attacco che in difesa. Osservare i punti di gioco e gli sbilanciamenti in numero di azioni o esiti di queste, può essere utile per preparare meglio il proprio gioco in vista della prossima partita contro una determinata squadra, di cui a sua volta si conosce la disposizione sul campo.

Raccolta dati:

La raccolta e il pre processing dei dati seguono questi passi:

1. Definizione di un campionato tramite la scelta di una lega e di una stagione.
2. Raccolta degli eventi appartenenti alle partite di quel campionato.

   La tipologia di dato necessario per l'analisi è stata identificata all'interno del dataset come un entry di tipo *attempt,* (colonna event_type='Attempt'). Ad un evento di questo tipo è associata un'occasione offensiva da parte di una squadra, questa contiene due importanti valori categorici:

   - Location: definisce la zona d'esecuzione all'interno del campo

Id	Descrizione		Id	Descrizione		Id	Descrizione
1	Attacking half		7	Difficult angle on the left		13	Very close range
2	Defensive half		8	Difficult angle on the right		14	Penality spot
3	Centre of the box		9	Left side of the box		15	Outside the box
4	Left wing		10	Left side of the six yard box		16	Long range
5	Right wing		11	Right side of the box		17	More than 35 yards
6	Difficult angle and long range		12	Right side of the six yard box		18	More than 40 yards

   - Shot Outcome: definisce l'esito dell'azione

Id	Descrizione		Id	Descrizione		Id	Descrizione
1	Is Goal		3	Off Target		5	Hit the br
2	On Target		4	Blocked		6	Penality spot

   Per quanto riguarda l'attributo location, esso ha 18 modalità + 1 ('Not recorded'), che definiscono 18 zone della metà campo offensiva della squadra a cui è associata l'occasione. Essendo molto granulari e alcune di queste davvero scarse in quanto a numero di azioni eseguite, le 18 categorie verranno accorpate in 6 macro aree:

Id	Descrizione	Locations		Id	Descrizione	Locations		Id	Descrizione	Locations
1	Centre of the box	3, 13, 14		3	Right Wing	5, 8, 11		5	Right side of the box	12
2	Left Wing	4, 7, 9		4	Left side of the box	10		6	Long range	15, 16, 17, 18

   Le originali modalità 1 e 2 dell'attributo location ('Attacking half' e 'Defensive half') sono state scartate in quanto nessuna occasione è associta ad esse. Le occasioni associate alla location con modalita 19 ('Not recorded') sono state anch'esse scartate in quando poco numerose, e non utili ai fini di questa indagine.
3. Infine, per ognuna delle squadre sono calcolate le frequenze di eventi esguiti, distribuite sulle 6 location e per tipo di esito.

Si ottiene quindi un dataset in cui ogni entry rappresenta una partita, che associa ai due team in gara, le seguenti colonne di interesse:

Attributo	**Descrizione**
n_attempts<T>	Totale azioni eseguite dal team T
n_goal<T>	Totale goal eseguiti dal team T
n_loc<L>_tot_att_<T>	Numero di azioni eseguite nella location L dal team T
n_loc<L>_is_goal_<T>	Numero di goal eseguiti nella location L dal team T
n_loc<L>_shot_on_target_<T>	Numero di tiri nello specchio della porta eseguiti nella location L dal team T
n_loc<L>_shot_off_target_<T>	Numero di tiri fuori dallo specchio della porta eseguiti nella location L dal team T
n_loc<L>_shot_hit_bar_<T>	Numero di tiri sulla traversa eseguiti nella location L dal team T
n_loc<L>_shot_blocked_<T>	Numero di tiri intercettati eseguiti nella location L dal team T

dove L varia tra 1 e 6, e T varia logicamente tra 1 e 2.

4. Le partite vengono divise quindi in due set, il primo contenente le partite di andata, il secondo contenente quelle del girone di ritorno. L'idea è quella di sfruttare la base di dati creatasi durante l'andata per poi utilizzarla nella preparazione della squadra in vista delle partite di ritorno. Le analisi verrano comunque svolte anche sul set del girono di ritorno per un eventuale confronto futuro.
5. Per ogni partita di ogni set, ad esempio il set di andata, vengono suddivisi, per ogni team, le azioni di attacco (attempts propri) e le azioni di difesa (attempts dell'avversario).
6. Si generano a questo punto 4 dataset distinti, rispettivamente per l'attacco e la difesa, sia in andata che a ritorno, ognuno di questi contiene un'entry per squadra, con associate le frequenze totali degli eventi, sempre divisi per location ed esito d'esecuzione. Su ognuno di questi, verranno analizzate le location e i loro utilizzi da parte delle squadre.

Clustering:

Preso un dataset tra i quattro sopra citati, la procedura di clustering si basa sull'utilizzo di alcuni indici che rappresentano misure di probabilità che si verifichino eventi con determinati esiti in determinate location. Come campionato di riferimento sul quale è stato sviluppato il workflow, e di cui vengono mostrati successivamente i risultati è stato scelto, senza particolari ragioni, il campionato inglese del 2016.

Per ogni location k, con k da 1 a 6, per ognuna delle squadre, vengono definiti i seguenti indici:

Indice	Calcolo	Descrizione
$P(L_k)$	$\frac{n\_loc\_k\_tot\_att}{n\_attempts}$	Probabilità che si sviluppi un'azione nella location k.
$P(On \mid L_k)$	$\frac{n\_loc\_k\_shot\_on\_target}{n\_loc\_k\_tot\_att}$	Probabilità che dato lo sviluppo di un'azione nella location k il suo esito si on target.
$P(Blk \mid L_k)$	$\frac{n\_loc\_k\_shot\_blocked}{n\_loc\_k\_tot\_att}$	Probabilità che dato lo sviluppo di un'azione nella location k, questa venga bloccata.
$P(G \mid OnL_k)$	$\frac{n\_loc\_k\_is\_goal}{n\_loc\_k\_shot\_on\_target}$	Probabilità che data un'azione nella location k con esito on target, questa sia anche un goal.
$P(L_k \mid G)$	$\frac{P(G \mid OnL_k)*P(L_k)}{P(GoalL_k)}$ dove $P(GoalL_k) = \sum_{k=1}^{6} P(G \mid OnL_k) \cdot P(L_k)$	Probabilità che dato un goal, questo venga eseguito nella location k.

Oltre a queste, vengono calcolati anche altri indici, ma questi vengono ritenuti quelli che portano un maggiore significato e interpretabilità nello stimare l'efficacia di una squadra in una certa location.

Con questi cinque attributi, una volta normalizzati in un range [0, 1], vengono avviate sei diverse procedure di clustering, una per location, con lo scopo di formare gruppi di squadre differenti per efficacia e per disposizione dei valori degli attributi scelti in quella zona. All'interno del workflow Knime ogni procedura di clustering confronta 4 diversi metodi di aggregazione:

- Hierarchical
- K-Means
- Fuzzy c-means
- K-Medoids

Il numero di cluster scelto è univoco tra clustering nelle diverse location, pur non essendo una scelta del tutto ottimale a livello della singola clusterizzazione, questo permette di confrontare e interpretare i livelli delle squadre con un ottica più globale, considerando l'insieme delle location nel complesso. Ad ogni modo il numero di cluster scelto è pari a 3, il quale si è dimostrato essere il numero ottimale nella maggioranza dei casi, dato anche il numero ridotto di elementi da raggruppare (20 squadre). Il numero di cluster è comunque direttamente impostabile globalmente nella sezione *GLOBAL VARIABLES* del workflow Knime.

A supporto di questa scelta e della scelta dei metodi di cluster vengono fornite diverse misure di valutazione delle aggregazioni ottenute, sia interne che esterne.

Misure interne:

1. Connectivity: utilizzata per hierarchical, K-Means e K-Medoids con 3, 4 e 5 cluster
2. Silhouette: utilizzata per K-Means e K-Medoids con 3, 4 e 5 cluster

3. Cophenetic: utilizzata per paragonare i metodi di linkage per nella costruzione dei cluster gerarchici.

Misure esterne:

Vengono calcolati due diversi indici esterni:

Indice	Calcolo	Descrizione
G	n_loc_k_is_goal	Numero di goal eseguiti nella location k
GL	$\frac{\text{n\_loc\_k\_is\_goal}}{\text{n\_loc\_k\_tot\_att}}$	Probabilità che data un'azione all'interno della location k, questa sfoci in un goal

Entrambi vengono normalizzati tra 0 e 1 e discretizzati a 5 possibili valori, ottenuti dividendo il range in 5 intervalli di uguale dimensione. Come per il numero di cluster, anche il numero di partizioni è impostabile globalmente nella sezione *GLOBAL VARIABLES* del workflow Knime

Di seguito sono mostrati i valori delle misure interne che evidenziano come il numero di cluster globalmente ottimale sia 3. I valori sono riferiti alle clusterizzazioni effettuate per la difesa nel girone di andata del campionato inglese 2016, con metodo gerarchico (H), K-Means (K), e K-Medoids clustering (M).

Location andata difesa	Connectivity									Silhouette						Cophenetic
	3 Cluster			4 Cluster			5 Cluster			3 Cluster		4 Cluster		5 Cluster		H complete linkage
	H	K	M	H	K	M	H	K	M	K	M	K	M	K	M	
L1	17	19	19	21	22	22	23	27	25	.23	.23	.27	.27	.26	.23	.56
L2	7	17	17	17	22	22	22	29	25	.30	.26	.31	.30	.25	.39	.65
L3	14	16	20	17	17	21	24	24	26	.27	.20	.28	.21	.27	.22	.70
L4	10	10	13	16	18	20	20	22	24	.40	.38	.38	.37	.35	.36	.82
L5	10	10	11	16	16	16	18	22	24	.33	.31	.33	.33	.31	.38	.77
L6	15	16	15	18	26	24	25	30	29	.22	.20	.23	.24	.25	.20	.52

Scelto il numero di cluster, si passa poi alla scelta dell'algoritmo di aggregazione tra i quattro sopra citati. Di seguito vengono mostrate le misure di validazione esterne calcolate confrontando il risultato di diversi algoritmi con le partizioni definite dall'attributo GL sopra definito. A differenza dell'attributo G, le partizioni realizzate tramite l'attributo GL "fittano" generalmente meglio rispetto ai cluster ottenuti, i suoi valori si sono mostrati maggiormente simili tra squadre dello stesso cluster e differenti tra squadre di cluster diverso. Inoltre considerando nel calcolo, non solo il numero di goal, ma anche il rapporto di questi con il totale delle azioni eseguite, sembra essere una misura più utile e accurata nel rappresentare sia la capacità di gestione del gioco sia quella di concretizzazione effettiva di una squadra. Le misure esterne si basano sui calcoli Rand (R), Jaccard (J), Fowlkes-Mallows (FM) e F-Statistic (F).

Location andata difesa	Hierachical				K-Means				Fuzzy K-Means				K-Medoids			
	R	J	FM	F	R	J	FM	F	R	J	FM	F	R	J	FM	F
L1	.69	.25	.41	.22	.67	.25	.42	.21	.61	.20	.36	.12	.61	.20	.36	.12
L2	.59	.24	.43	.19	.72	.30	.47	.30	.74	.34	.52	.36	.64	.24	.41	.19
L3	.60	.20	.34	.06	.70	.32	.48	.28	.64	.23	.37	.12	.71	.39	.58	.38
L4	.92	.83	.91	.85	.92	.83	.91	.85	.67	.35	.52	.28	.70	.43	.60	.37
L5	.71	.42	.60	.39	.71	.42	.60	.39	.72	.43	.61	.40	.71	.42	.60	.39
L6	.70	.26	.43	.24	.70	.26	.43	.24	.70	.26	.43	.24	.70	.26	.43	.24

Per ogni location viene scelto un algoritmo di clustering (verde scuro), tra quelli con misure esterne maggiori (verde chiaro), al fine di raggiugere un compromesso ottimale sia tra misure interne che esterne. Ad esempio per la location 1, l'algoritmo gerarchico e k-means ottengono, rispetto agli altri, misure esterne maggiori ma comunque simili tra loro, viene però scelto il metodo gerarchico in quanto produce una minore connessione tra gruppi. Analogamente, per la location numero 3, sebbene il metodo K-Medoids fornisca misure esterne

leggermente maggiori al K-means, viene comunque scelto quest'ultimo in quanto produce minore connessione e un coefficiente di silhouette maggiore.

Infine, data la numerosità delle clusterizzazioni eseguite, si è reso necessario sviluppare un metodo automatico che, per ogni location, eseguisse l'ordinamento per qualità di prestazione dei tre gruppi di squadre. Per fare questo, viene sfruttato l'attributo GL utilizzato in precedenza per la valutazione tramite misure esterne. Per ognuno dei 3 cluster viene computata la media dell'attributo GL, quindi il cluster con media più bassa è considerato quello contenente le squadre più scarse in quella location, al contrario, quello con media maggiore è considerato il cluster a cui appertengono le squadre più forti all'interno di quella location.

Risultati ottenuti e utilizzo dei cluster:

Al termine di questa procedura ad ogni team sono associati 12 attributi, 6 relativi alla capacità offensiva nelle location di attacco, 6 relativi alla capacità difensiva nelle location di difesa. Ognuno di questi ha un valore intero compreso tra 0 e numero di cluster-1 (in questo caso 2), dove 0 indica scarsa abilità mentre 2 indica maggiore bravura nel conludere o meno azioni in quella location, rispetto alle altre squadre del campionato. E' quindi possibile analizzare i prototipi dei cluster per singola location ed eseguire successivamente dei paragoni tra squadre, realizzando mappe, di seguito mostrate, che evidenziano queste caratteristiche sull'area del campo di gioco.

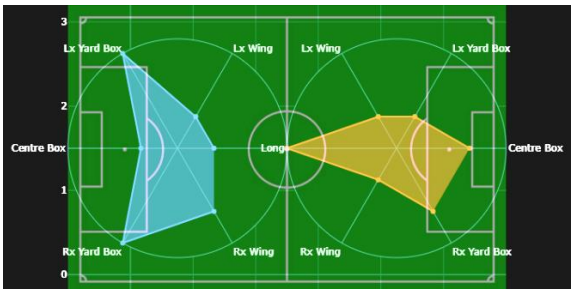
- Descrizione Prototipi:
  Di seguito vengono mostrate le proprietà dei prototipi dei tre differenti cluster realizzati per quanto riguarda l'attacco all'interno location 1. I cluster sono già ordinati secondo l'attributo GL come descritto precedentemente

Cluster 0 – Prototipo		
Attributi		
P(L)	.53	Aston Villa, Liverpool, West Brom, Arsenal, Manchester City
P(On|L)	.24	
P(Blk|L)	.68	
P(G|OnL)	.18	
P(L|G)	.27	
Squadre in difficoltà in questa location, seppur abbiano una discreta capacità di intraprendere l'azione [P(L)], spesso queste vengono bloccate [P(Blk|L)], o non rientrano in conclusioni mirate [P(On|L)], quando invece lo sono, hanno comunque scarse possibilità di tramutarsi in goal [P(G|OnL)].		

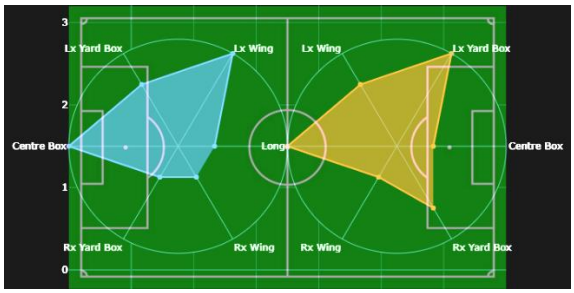
Cluster 1 – Prototipo		
Attributi		
P(L)	.55	Chelsea, Norwithc City, Swansea, Southampton, Newcastle, Stoke City, Watford, Manchester Utd
P(On|L)	.25	
P(Blk|L)	.54	
P(G|OnL)	.50	
P(L|G)	.76	
Squadre di abilità media in questa location, hanno una buona capacità di intraprendere azioni in questa location, ma, come nel cluster 0, raramente queste arrivano nello specchio della porta [P(On|L) e P(Blk|L)], quando riescono ad arrivare però, allora hanno discrete change di fare goal [P(G|OnL)].		

Cluster 2 – Prototipo		
Attributi		
P(L)	.22	Crystal Palace, Everton, Leicster City, West Ham, Tottenham
P(On|L)	.53	
P(Blk|L)	.29	
P(G|OnL)	.63	
P(L|G)	.36	
Squadre superiori nell'utilizzo della location, queste sfruttano un minor numero di volte questa location [P(L)], ma quando lo fanno hanno probabilità nettamente superiori rispetto agli altri due gruppi di non essere intercettati [P(Blk|L)] e di concludere concretamente [P(On|L) e P(G|OnL)].		

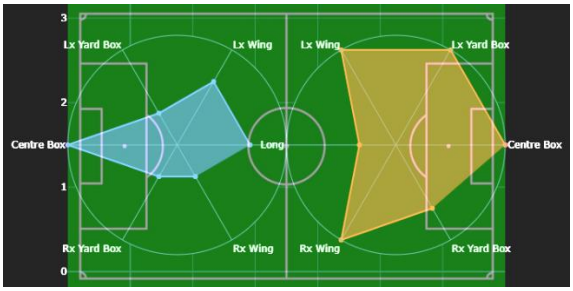
- Mappa di abilità difensiva e offensiva
  Di seguito vengono presentate alcune delle mappe ottenibili che evidenziano la capacità difensiva, in azzurro, e la capacità offensiva, in arancione, sul campo da gioco.



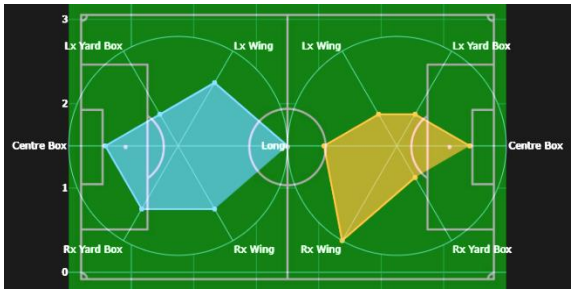
Chelsea



Manchester City



Leicster City

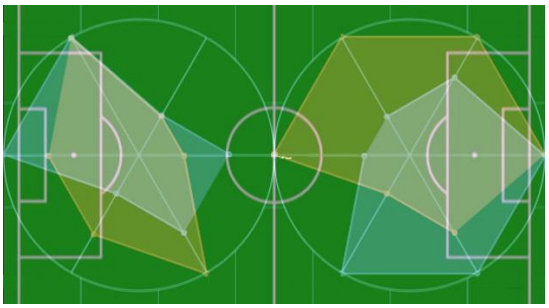


Swansea

Limiti, conclusioni, sviluppi futuri:

I cluster ottenuti permettono di definre e confrontare gli approcci offensivi e difensivi delle squadre nelle diverse zone del campo. Le misure di validazione sia interne che esterne non sono del tutto ottimali, può capitare che in alcuni cluster, considerati "scarsi", siano presenti squadre considerate storicamente capaci. Tuttavia il significato delle feature scelte non punta ad ottenere una classifica ordinata di squadre forti o deboli ma metodi e conseguenti risultati delle diverse tipologie di azioni eseguite. Di conseguenza l'utilizzo dei risultati non sta nel predirre la vittoria o sconfitta, ma consiste nella comprensione di come programmare un'allenamento mirato in vista della prossima partita, il quale possa rafforzare i propri punti deboli difensivi e addestrare l'attacco in modo ottimale, conoscendo rispettivamente i pregi e i difetti della squadra avversaria. Di seguito, viene mostrato come sovrapponendo le mappe di due squadre, si possano individuare zone più favorevoli di altre in cui tentare l'attacco, di coseguenza l'altra squadra saprà dove deve preparare maggiormente la difesa.

Tottenham vs. Arsenal



Si vede come la difesa del Tottenham (azzurro a sinistra), copra in buona parte la capacità d'attacco dell'Arsenal, al contrario, la difesa dell'Arsenal (azzurro a destra) risulta più sbilanciata rispetto all'attacco del Tottenham (arancione a destra), quest'ultimo risulta più capace sulla lato sinistro, area dove la difesa dell'Arsenal sembra essere più carente.

Infine si potrebbe considerare una raccolta dati maggiormente mirata a questo scopo. Attualmente il dataset propone il dato della location in formato categorico, pensare invece di poter disporre di coordinate numeriche che identificano la posizione precisa all'interno del campo da gioco, potrebbe permettere sia di svolgere la stessa indagine modo accurato, sia di individuare veri e propri schemi di costruzione di azioni che risultino più e meno efficaci di altri.