ELEMENTI DI BASI DI DATI E DATA MINING

LA PALLAVOLO

A.S. 2024/2025
Alessia Lucente
960769

TESTO DESCRITTIVO

La pallavolo è uno sport formato da squadre.

Un giocatore appartiene ad una squadra e si tiene conto del suo Codice Fiscale, nome, cognome, data di nascita, altezza, età e del **ruolo** che ricopre all'interno della squadra. I ruoli dei giocatori sono: palleggio, opposto, centrale, banda e libero.

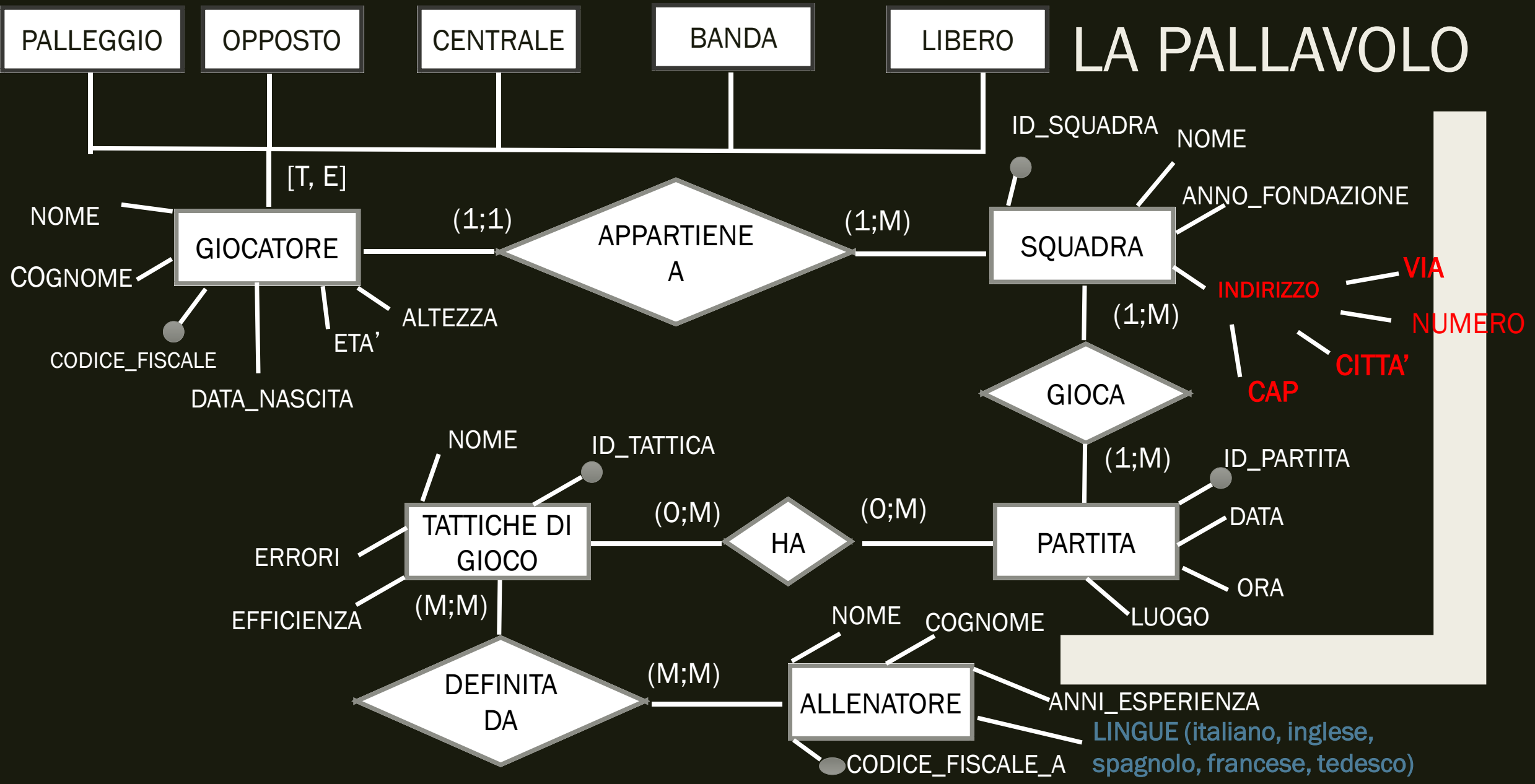
Ogni squadra è composta da giocatori; di ogni squadra si tiene traccia dell'Id squadra, del nome, dell'anno di fondazione e dell'**indirizzo** (via, numero, città e cap).

La squadra gioca almeno una partita; di essa si memorizzano l'Id partita, la data, l'ora e il luogo.

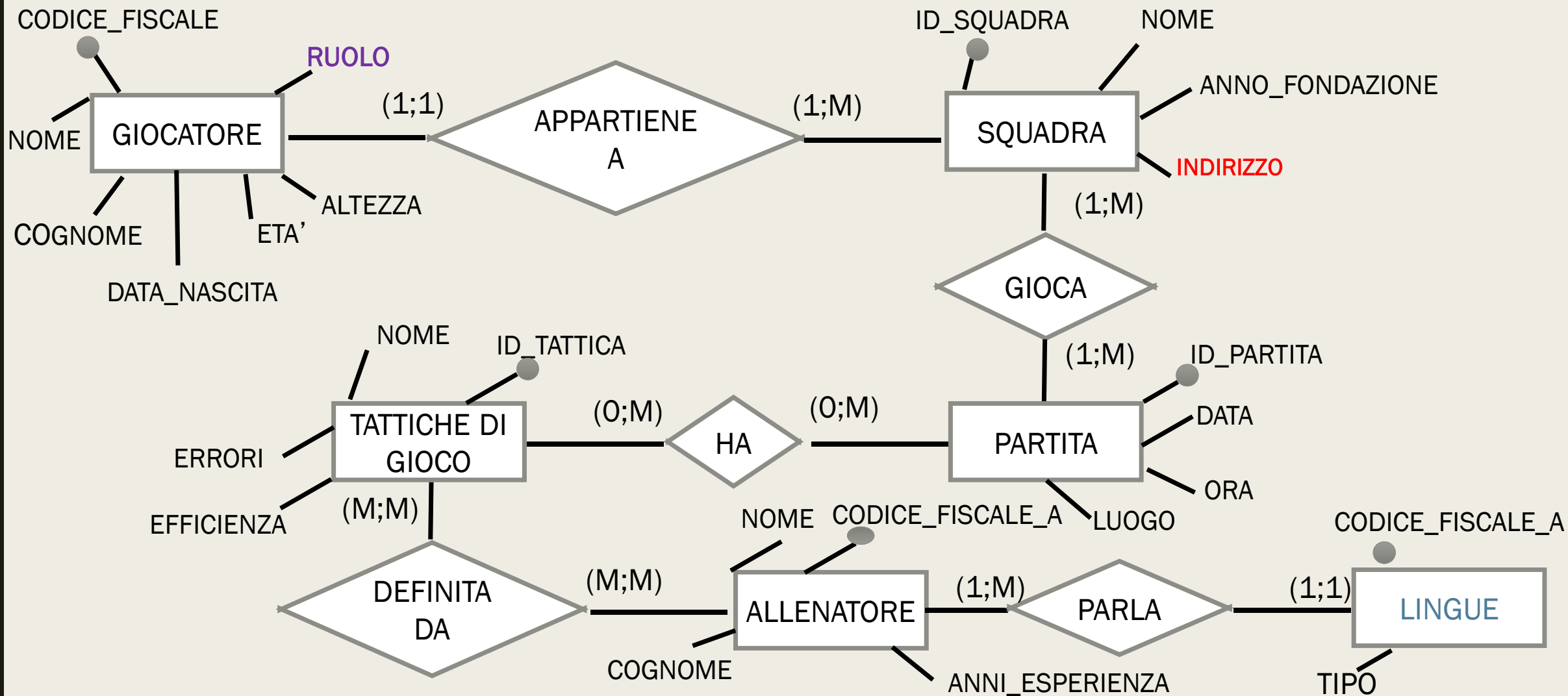
Ogni partita potrebbe avere una tattica di gioco o più, ma non è detto. Di una tattica si sa il Id tattica, il nome, gli errori e l'efficienza.

La tattica di gioco è definita dall'allenatore di cui si tiene traccia del suo Codice Fiscale, nome, cognome, anni di esperienza e delle **lingue parlate** (italiano, inglese, spagnolo, francese e tedesco).

MODELLO CONCETTUALE NON RISTRUTTURATO: LA PALLAVOLO



MODELLO CONCETTUALE RISTRUTTURATO



Nel modello concettuale non ristrutturato *nell'entità giocatore* è presente una gerarchia TOTALE (non ci sono altri elementi possibili oltre quelli indicati) ed ESCLUSIVA (un elemento di un insieme non può essere un elemento dell'altro) formata da:

- Palleggio
- Opposto
- Centrale
- Banda
- Libero

Nel modello concettuale ristrutturato tolgo le sottoclassi; aggiungo nel padre un attributo "RUOLO" che comprenda i figli soppressi. Gli attributi delle sottoclassi vengono uniti a quelli dell'entità padre.

Nel modello concettuale non ristrutturato *l'entità squadra* ha un attributo composto INDIRIZZO formato da:

- Via
- Numero
- Città
- CAP

Nel modello concettuale ristrutturato elimino i sotto-attributi di INDIRIZZO.

Nel modello concettuale non ristrutturato *l'entità allenatore* presenta un attributo multivalore LINGUE (italiano, inglese, spagnolo, francese, tedesco). Nel modello concettuale ristrutturato l'attributo **LINGUE** diventa una nuova entità collegata esternamente all'entità di partenza; l'attributo multi valore è rappresentato con un attributo mono-valore TIPO (che contiene le tipologie di lingue parlate dall'allenatore: italiano, inglese, spagnolo, francese, tedesco) con cardinalità nuova 1:1.

MODELLO LOGICO

GIOCATORE (CODICE_FISCALE, NOME, COGNOME, DATA DI NASCITA, ETA', ALTEZZA, RUOLO, ID_SQUADRA)

SQUADRA (ID_SQUADRA, NOME, ANNO_FONDAZIONE, INDIRIZZO)

GIOCA (ID_SQUADRA; ID_PARTITA)

PARTITA (ID _PARTITA, DATA, ORA, LUOGO)

HA (ID_PARTITA; ID_TATTICHE)

TATTICHE_DI_GIOCO (ID TATTICHE, NOME, ERRORI, EFFICIENZA)

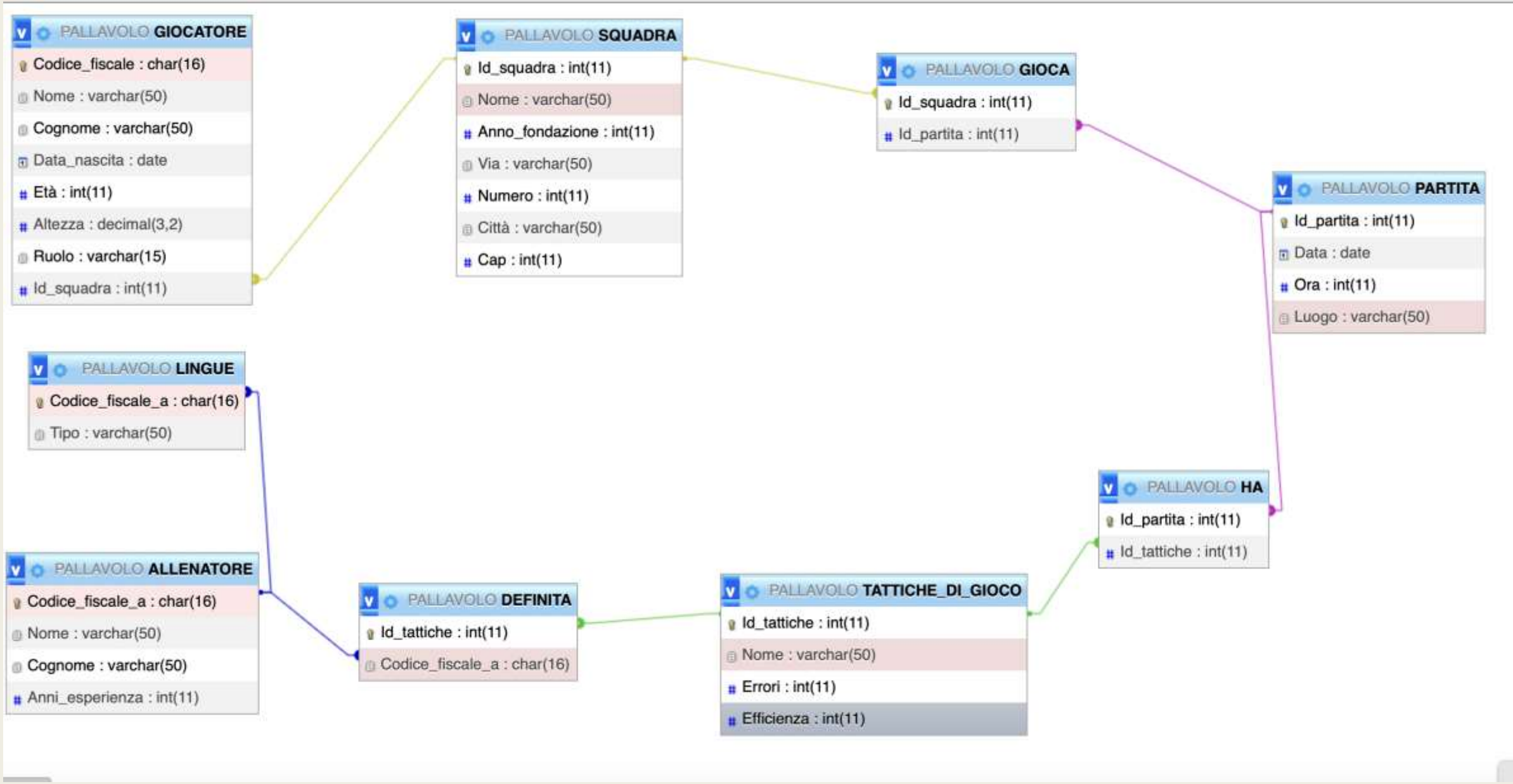
DEFINITA (ID_TATTICHE; CODICE_FISCALE_A)

ALLENATORE (CODICE_FISCALE_A, NOME, COGNOME, ANNI_ESPERIENZA)

LINGUE (CODICE FISCALE A, TIPO)

- La tabella **GIOCATORE** contiene all'interno la sua chiave primaria CODICE_FISCALE e la chiave esterna ID_SQUADRA della tabella SQUADRA poiché la cardinalità massima tra l'entità GIOCATORE e l'entità SQUADRA è 1:M e la cardinalità minima è 1:1, di conseguenza la tabella di raccordo non va fatta e bisogna mettere la chiave esterna nella tabella con cardinalità massima 1 (GIOCATORE).
- La tabella **SQUADRA** contiene la chiave primaria ID_SQUADRA; viene fatta la tabella di raccordo GIOCA poiché la cardinalità massima è M:M. GIOCA contiene le chiavi esterne ID_SQUADRA e ID_PARTITA.
- La tabella **PARTITA** contiene la chiave primaria ID_PARTITA; viene fatta la tabella di raccordo HA poiché la cardinalità massima è M:M. La tabella HA contiene le chiavi esterne ID_PARTITA e ID_TATTICHE.
- La tabella **TATTICHE_DI_GICOO** contiene la chiave primaria ID_TATTICHE; viene fatta la tabella di raccordo DEFINITA poiché la cardinalità massima è M:M.
- La tabella **DEFINITA** contiene al suo interno le chiavi esterne ID_TATTICHE e CODICE_FISCALE_A.
- La tabella **ALLENATORE** contiene la chiave primaria CODICE_FISCALE_A; non si fa la tabella di raccordo poiché la cardinalità massima tra l'entità ALLENATORE e l'entità LINGUE è 1:M.
- La tabella LINGUE contiene la chiave esterna CODICE_FISCALE_A poiché la cardinalità massima tra le due entità è 1:M e la cardinalità minima tra le due entità è 1:1, quindi metto la chiave esterna nella tabella con cardinalità massima 1 (LINGUE).

DESIGNER DATABASE



QUERY:

- una query con raggruppamento e operatori aggregati (con JOIN)

OBBIETTIVO

Per ogni squadra, mostra il numero totale di partite giocate, la media dell'efficienza tattica, e il numero totale di errori tattici.

Mostro le righe 0 - 4 (5 del totale, La query ha impiegato 0.0016 secondi.)

```
SELECT SQUADRA.Nome AS NomeSquadra, COUNT(DISTINCT PARTITA.Id_partita) AS NumeroPartiteGiocate, SUM(TATTICHE_DI_GIOCO.Errori) AS TotaleErrori,
AVG(TATTICHE_DI_GIOCO.Efficienza) AS MediaEfficienza FROM SQUADRA JOIN GIOCA ON SQUADRA.Id_squadra = GIOCA.Id_squadra JOIN PARTITA ON GIOCA.Id_partita
= PARTITA.Id_partita JOIN HA ON PARTITA.Id_partita = HA.Id_partita JOIN TATTICHE_DI_GIOCO ON HA.Id_tattiche = TATTICHE_DI_GIOCO.Id_tattiche GROUP BY
SQUADRA.Id_squadra, SQUADRA.Nome ORDER BY NumeroPartiteGiocate DESC;
```

Profiling [Modifica inline] [Modifica] [Spiega SQL] [Crea il codice PHP] [Aggiorna]

Mostra tutti | Numero di righe: 25 Filtra righe: Cerca nella tabella

Opzioni extra

NomeSquadra	NumeroPartiteGiocate ▼ 1	TotaleErrori	MediaEfficienza
Aquile della Rete	1	3	70.0000
Draghi del Servizio	1	4	60.0000
Leoni del Volley	1	0	99.0000
Thunder Spikers	1	3	70.0000
Stelle del Net	1	2	80.0000

Mostra tutti | Numero di righe: 25 Filtra righe: Cerca nella tabella

- una query con matching approssimato (con JOIN)

OBBIETTIVO:
trovare i giocatori il cui nome contenga almeno una delle prime cinque lettere del nome della loro squadra.

```
SELECT GIOCATORE.Nome AS NomeGiocatore, GIOCATORE.Cognome AS CognomeGiocatore, SQUADRA.Nome AS NomeSquadra FROM GIOCATORE JOIN SQUADRA ON
GIOCATORE.Id_squadra = SQUADRA.Id_squadra WHERE GIOCATORE.Nome LIKE CONCAT('%', SUBSTRING(SQUADRA.Nome, 1, 1), '%') OR GIOCATORE.Nome LIKE CONCAT('%',
SUBSTRING(SQUADRA.Nome, 2, 1), '%') OR GIOCATORE.Nome LIKE CONCAT('%', SUBSTRING(SQUADRA.Nome, 3, 1), '%') OR GIOCATORE.Nome LIKE CONCAT('%',
SUBSTRING(SQUADRA.Nome, 4, 1), '%') OR GIOCATORE.Nome LIKE CONCAT('%', SUBSTRING(SQUADRA.Nome, 5, 1), '%');
```

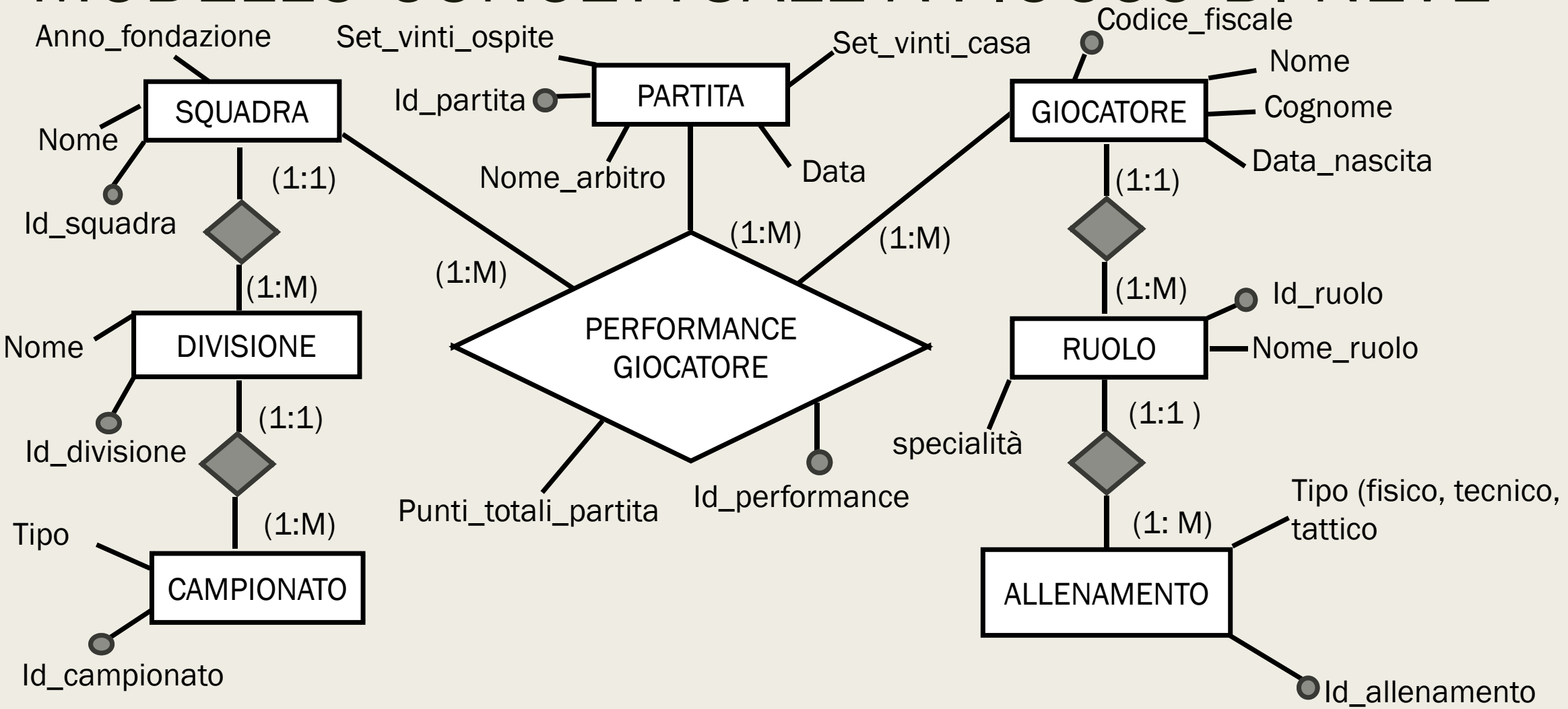
Profiling [Modifica inline] [Modifica] [Spiega SQL] [Crea il codice PHP] [Aggiorna]

Mostra tutti | Numero di righe: 25 Filtra righe: Cerca nella tabella Ordina per chiave: Nessuno

Opzioni extra

NomeGiocatore	CognomeGiocatore	NomeSquadra
Luca	Conti	Thunder Spikers
Ambra	Delvecchio	Aquile della Rete
Elia	Fiumicelli	Stelle del Net
Elisa	Galli	Leoni del Volley
Ginevra	Montaguti	Draghi del Servizio
Chiara	Moretti	Aquile della Rete
Andrea	Ricci	Draghi del Servizio
Federica	Rinaldi	Stelle del Net

MODELLO CONCETTUALE A FIOCCO DI NEVE



La tabella dei fatti è **PERFORMANCE_GIOCATORE** con un suo ID_performance e l'attributo punti_totali_partita; l'attributo punti_totali_partita è un attributo quantitativo, cioè ne descrive l'aspetto. La tabella dei fatti è collegata alla dimensione **PARTITA** che ha come attributi: ID_partita, nome_arbitro, set_vinti_casa, set_vinti_ospite, data.

Dalla medesima tabella dei fatti PERFORMANCE_GIOCATORE partono due gerarchie; la prima è composta dalle seguenti dimensioni con i relativi attributi:

- Dimensione **GIOCATORE** di cui si sa il codice_fiscale, nome, cognome e la data di nascita.
- Dimensione **RUOLO** di cui si conosce il proprio ID_ruolo e il nome.
- Dimensione **ALLENAMENTO** con il proprio ID_allenamento e il tipo di allenamento che può essere fisico, tecnico, tattico, mentale o coordinativo.

La seconda gerarchia è formata dalle seguenti dimensioni con i relativi attributi:

- Dimensione **SQUADRA** di cui si sa l'ID_squadra, il nome e l'anno di fondazione.
- Dimensione **DIVISIONE** con l'ID_divisione e il nome.
- Dimensione **CAMPIONATO** con l'ID_campionato e il nome.

Sono collegate tra loro con cardinalità massima M:M, mentre la cardinalità minima è indifferente (può essere 0 o 1).

MODELLO LOGICO

PERFORMANCE_GIOCATORE (ID_PERFORMANCE, PUNTI_TOTALI, ID_SQUADRA, CODICE_FISCALE, ID_PARTITA)

PARTITA (ID_PARTITA, SET_VINTI_CASA, SET_VINTI_OSPITE, NOME_ARBITRO, DATA)

SQUADRA (ID_SQUADRA, NOME, ANNO_FONDAZIONE, ID_DIVISIONE)

DIVISIONE (ID_DIVISIONE, NOME, ID_CAMPIONATO)

CAMPIONATO (ID_CAMPIONATO, TIPO)

GIOCATORE (CODICE_FISCALE, NOME, COGNOME, DATA_NASCITA, ID_RUOLO)

RUOLO (ID_RUOLO, NOME_RUOLO, ID_ALLENAMENTO)

ALLENAMENTO (ID_ALLENAMENTO, TIPO)

La tabella dei fatti **PERFORMANCE_GIOCATORE** ha come chiave primaria ID_performance e come chiavi esterne l'ID_squadra, codice_fiscale, ID_partita.

La dimensione **PARTITA** ha come chiave primaria ID_partita.

La dimensione **SQUADRA** ha come chiave primaria ID_squadra e come chiave esterna l'ID_divisione.

La dimensione **DIVISIONE** ha come chiave primaria l'ID_divisione e come chiave esterna l'ID_campionato.

La dimensione **CAMPIONATO** ha come chiave primaria l'ID_campionato.

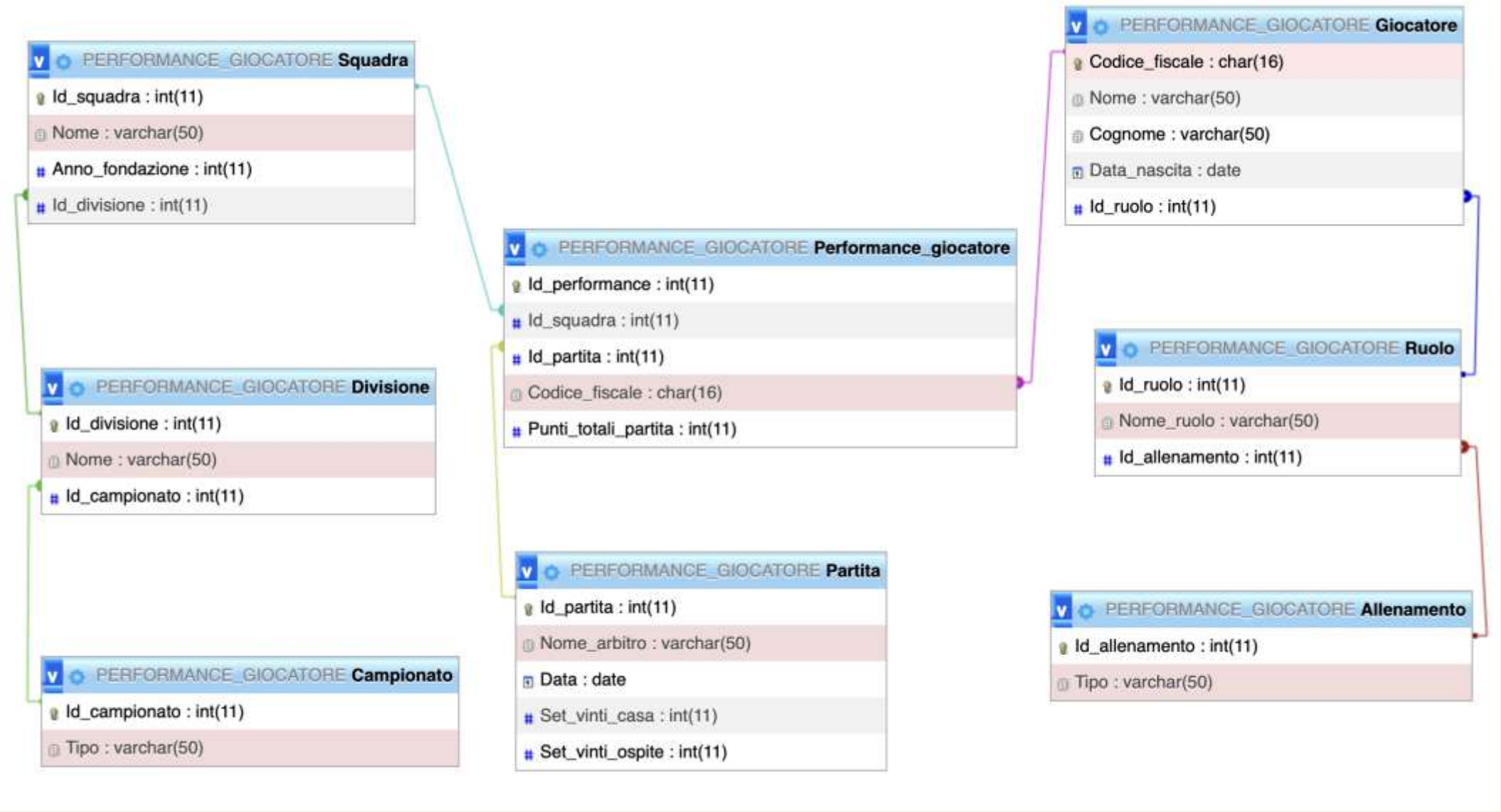
La dimensione **GIOCATORE** ha come chiave primaria codice_fiscale e come chiave esterna l'ID_ruolo.

La dimensione **RUOLO** ha come dimensione primaria l'ID_ruolo e come chiave esterna l'ID_allenamento.

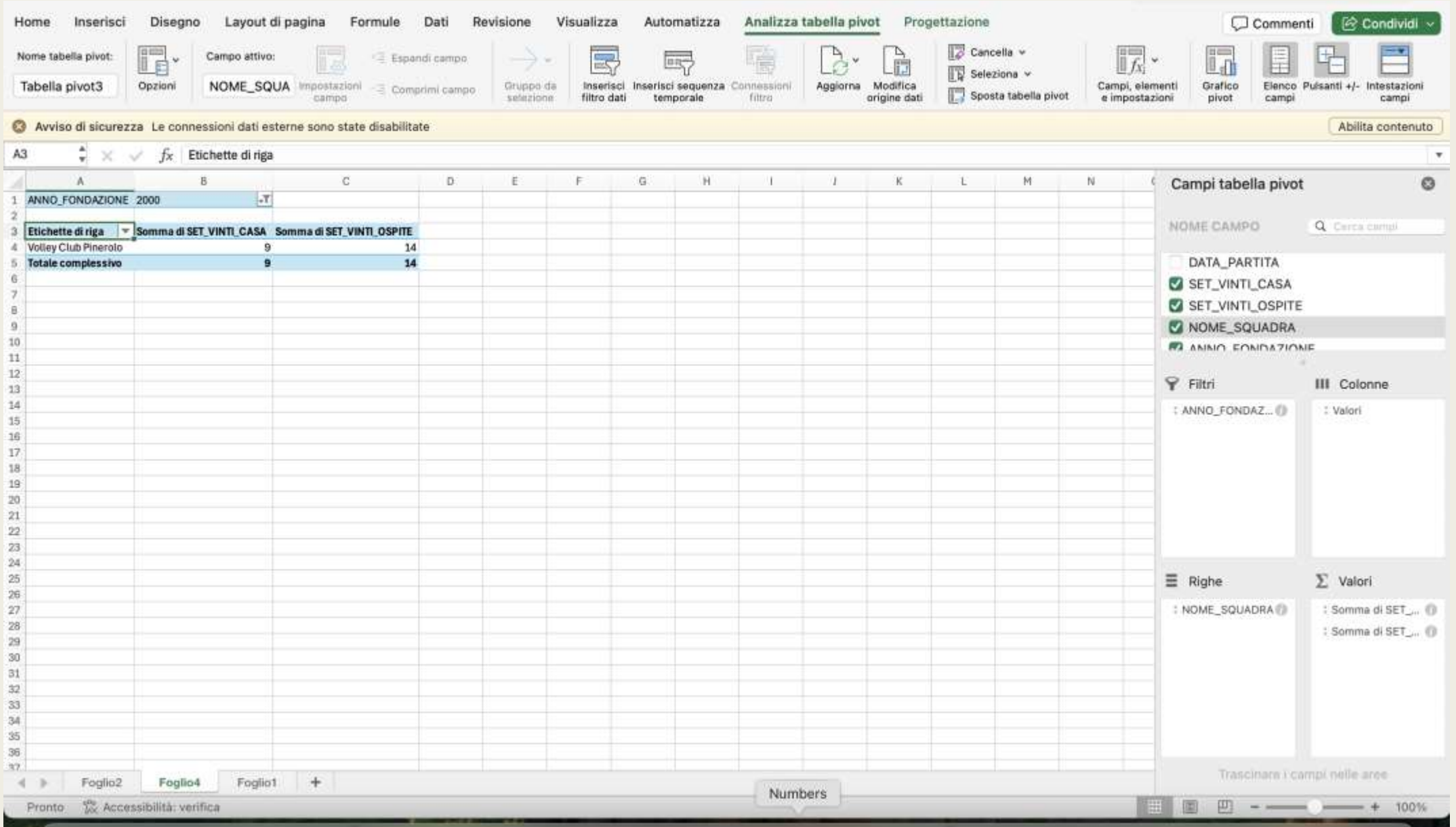
La dimensione **ALLENAMENTO** ha come chiave primaria l'ID_allenamento.

Le chiavi esterne sono state inserite nelle dimensioni con cardinalità minima 1 dato che le cardinalità massime sono M:M e la chiave esterna va inserita nella dimensione che ha cardinalità minima 1.

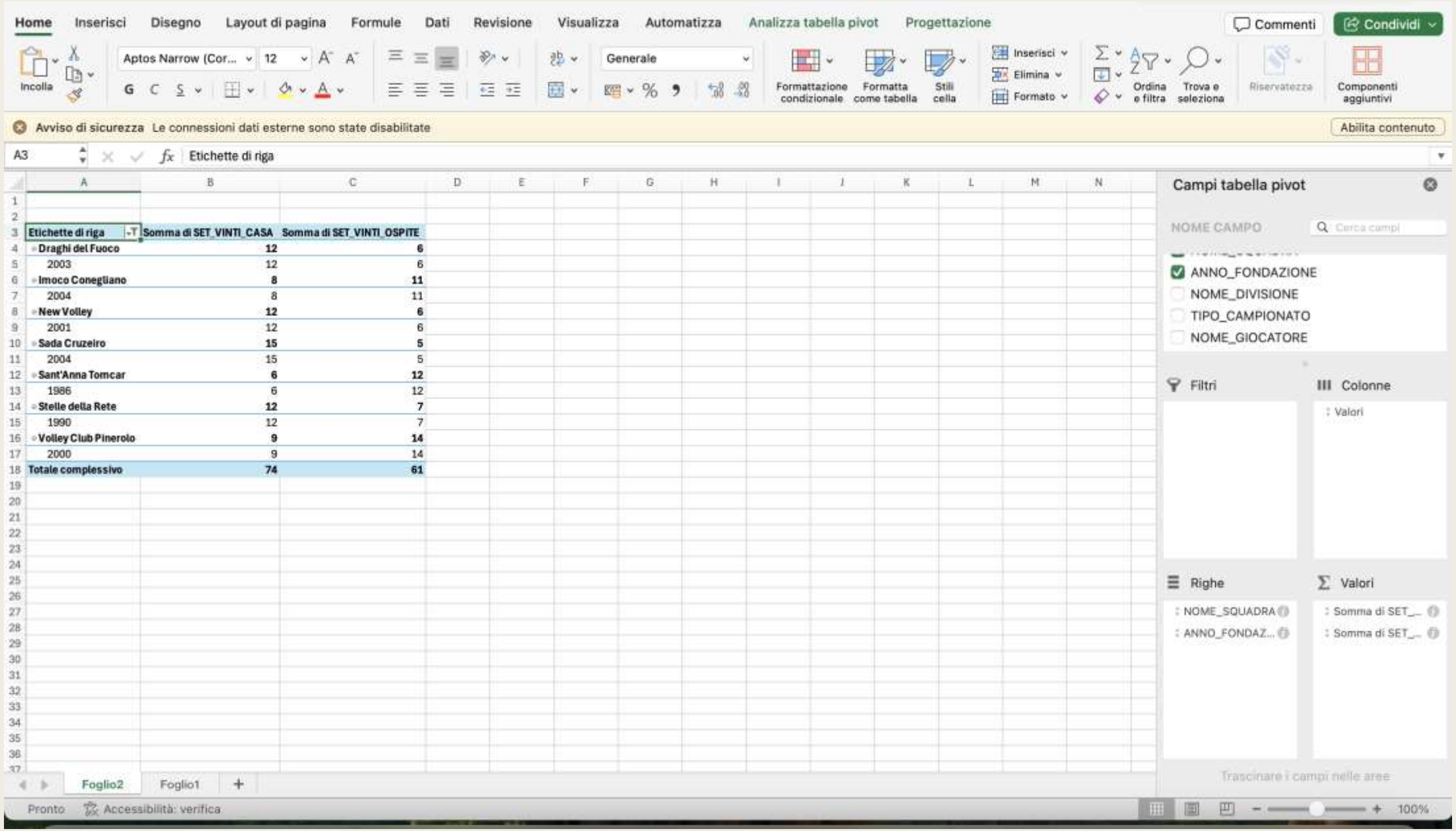
DESIGNER DATABASE MODELLO CONCETTUALE A FIOCCO DI NEVE



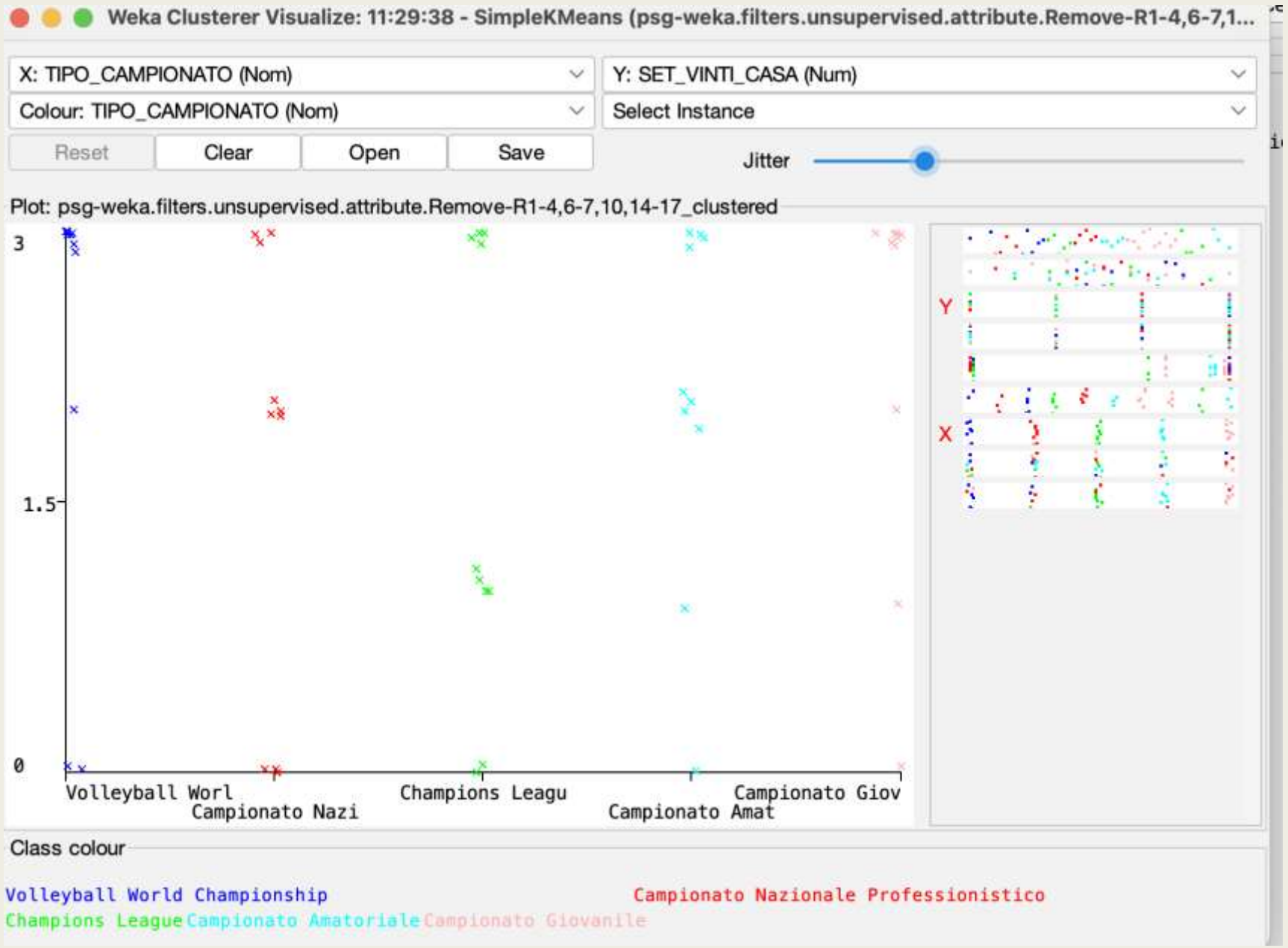
- **Slice:** analizzare i dati relativi a un singolo anno di fondazione (*ANNO_FONDAZIONE = 2000*), visualizzando solo le somme dei set vinti dalle squadre fondate in quell'anno.



- **Drill-down**: visualizzare i dati di una squadra basati sull’anno di fondazione dopo il 1980



CLUSTER


Weka Clusterer Visualize: 11:29:38 - SimpleKMeans (psg-weka.filters.unsupervised.attribute.Remove-R1-4,6-7,1...
X: TIPO_CAMPIONATO (Nom)
Y: SET_VINTI_CASA (Num)
Colour: TIPO_CAMPIONATO (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: psg-weka.filters.unsupervised.attribute.Remove-R1-4,6-7,10,14-17_clustered
3
1.5
0
Volleyball Worl
Campionato Nazi
Champions Leagu
Campionato Giov
Campionato Amat
Y
X
Class colour
Volleyball World Championship
Campionato Nazionale Professionistico
Champions League
Campionato Amatoriale
Campionato Giovanile

I cluster sono distribuiti in modo separato lungo l'asse TIPO_CAMPIONATO, mostrando che questa variabile influisce molto sulla separazione dei dati. L'asse SET_VINTI_CASA aggiunge una dimensione numerica, distinguendo i cluster in base al numero di set vinti in casa.

Nella configurazione mostrata:
•**X:** TIPO_CAMPIONATO (Nominale)
•**Y:** SET_VINTI_CASA (Numerico)

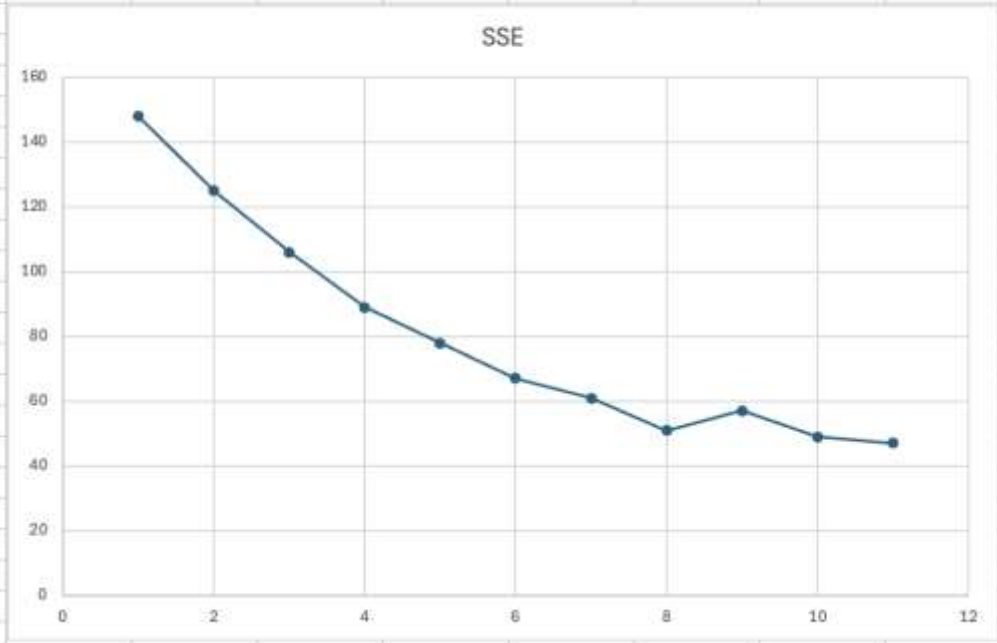
- **Misura di distanza usata:** euclidea
- **Algoritmo usato:** k-means
- **Numero di cluster:** 5
- **Seed:** 10
- **Scarto quadratico:** 78.87056779647445

I cluster più significativi sono:
- Cluster di colore rosso → contiene il valore Campionato Nazionale Professionistico indicando il numero di set vinti in casa. Questo cluster potrebbe rappresentare i dati con maggior successo dato che si distribuisce maggiormente verso valori più alti sull'asse SET_VINTI_CASA.
- Cluster di colore azzurro → contiene il valore Campionato Giovanile indicando il numero di set vinti in casa. Distribuito in modo intermedio, rappresenta squadre con successi moderati.

Il cluster **rosso** è significativo perché rappresenta il vertice competitivo, mentre il cluster **azzurro** è importante per il suo ruolo intermedio e per la sua chiara separazione dagli altri.

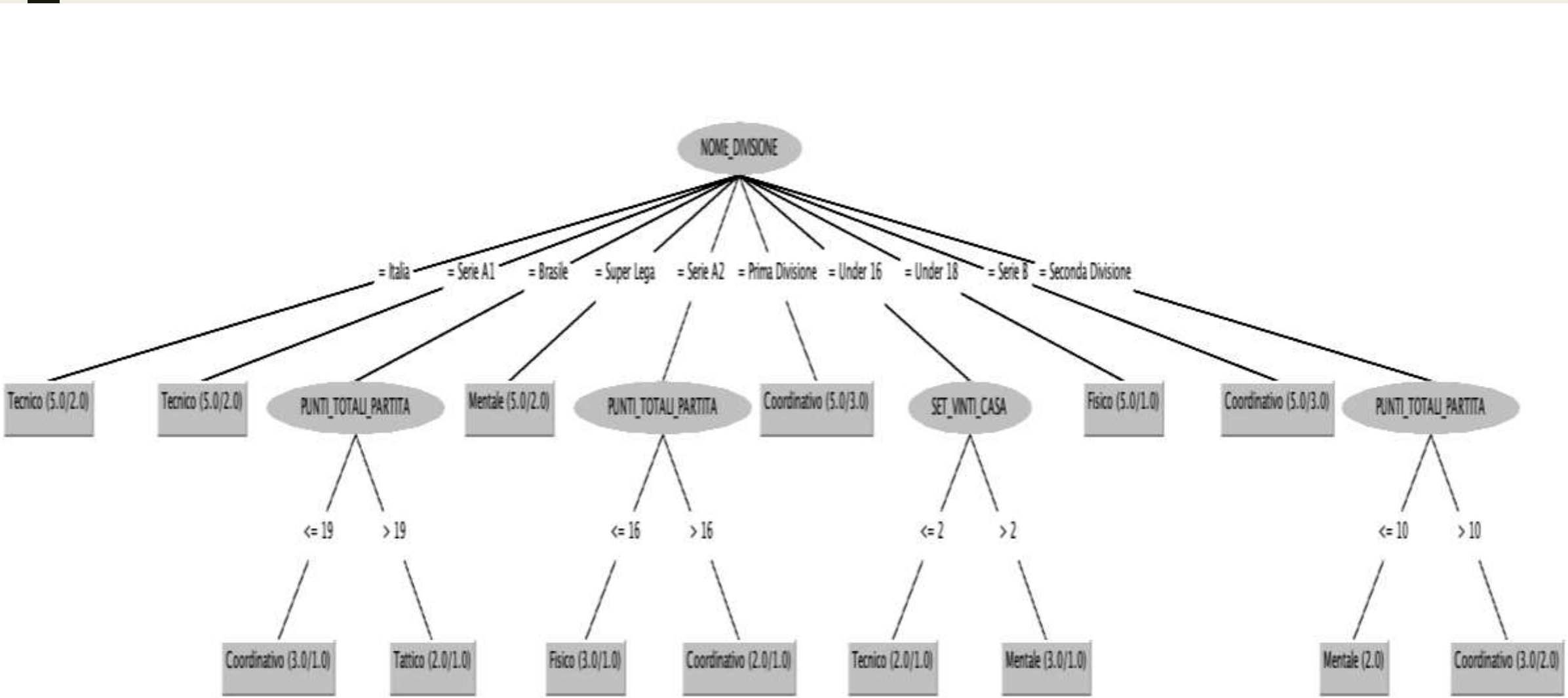
GRAFICO DEL GOMITO

K	SSE
1	148
2	125
3	106
4	89
5	78
6	67
7	61
8	51
9	57
10	49
11	47



Il numero di cluster utilizzato è 5 perché è il punto del grafico in cui la riduzione dell'errore inizia a diventare meno significativa.
Intorno a K=5 sembra esserci un gomito visibile, indicando che da quel punto in poi l'aggiunta di ulteriori cluster non migliora significativamente la qualità del clustering.

ALBERO DECISIONALE



- **Variabili di input:**
  - Nome_divisione
  - Punti_totali_partita
  - Set_vinti_casa

Variabili di output:

- Tattico
- Mentale
- Fisico
- Coordinativo
- Tecnico

- **Tecnica di valutazione:** training set

- **Misure di precision e recall:**

Precision	Recall	F-Measure
0,583	0,700	0,636
0,444	0,727	0,552
0,700	0,538	0,609
0,500	0,143	0,222
0,750	0,667	0,706
0,601	0,580	0,565

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,700	0,125	0,583	0,700	0,636	0,539	0,876	0,529	Tecnico
0,727	0,256	0,444	0,727	0,552	0,406	0,815	0,477	Coordinativo
0,538	0,081	0,700	0,538	0,609	0,502	0,866	0,666	Mentale
0,143	0,023	0,500	0,143	0,222	0,212	0,824	0,352	Tattico
0,667	0,049	0,750	0,667	0,706	0,648	0,953	0,725	Fisico
0,580	0,115	0,601	0,580	0,565	0,474	0,866	0,564	

- **Descrizione matrice di confusione**

```
=== Confusion Matrix ===

a b c d e   <-- classified as
7 2 1 0 0 | a = Tecnico
2 8 0 0 1 | b = Coordinativo
1 4 7 0 1 | c = Mentale
2 3 1 1 0 | d = Tattico
0 1 1 1 6 | e = Fisico
```

TECNICO:
7→ sono i veri positivi
2 → sono i falsi positivi
COORDINATIVO:
8→ sono i veri positivi
2→ sono i falsi positivi
MENTALE:
7→ sono i veri positivi
4→ falsi positivi
TATTICO:
1→ veri positivi
2--> falso positivo

FISICO:
6→ veri positivi
1→ falso positivo

- **Descrizione albero con regole**

IF NOME_DIVISIONE = Italia
THEN Classe = Tecnico (5.0/2.0)

IF NOME_DIVISIONE = Serie A1
THEN Classe = Tecnico (5.0/2.0)

IF NOME_DIVISIONE = Brasile **AND** PUNTI_TOTALI_PARTITA ≤ 19
THEN Classe = Coordinativo (3.0/1.0)

IF NOME_DIVISIONE = Brasile **AND** PUNTI_TOTALI_PARTITA > 19
THEN Classe = Tattico (2.0/1.0)

IF NOME_DIVISIONE = Super Lega
THEN Classe = Mentale (5.0/2.0)

IF NOME_DIVISIONE = Serie A2 **AND** PUNTI_TOTALI_PARTITA ≤ 16
THEN Classe = Fisico (3.0/1.0)

IF NOME_DIVISIONE = Serie A2 **AND** PUNTI_TOTALI_PARTITA > 16
THEN Classe = Coordinativo (2.0/1.0)

IF NOME_DIVISIONE = Prima Divisione
THEN Classe = Coordinativo (5.0/3.0)

IF NOME_DIVISIONE = Under 16 **AND** SET_VINTI_CASA ≤ 2
THEN Classe = Tecnico (2.0/1.0)

IF NOME_DIVISIONE = Under 16 **AND** SET_VINTI_CASA > 2
THEN Classe = Mentale (3.0/1.0)

IF NOME_DIVISIONE = Under 18
THEN Classe = Fisico (5.0/1.0)

IF NOME_DIVISIONE = Serie B
THEN Classe = Coordinativo (5.0/3.0)

IF NOME_DIVISIONE = Seconda Divisione **AND** PUNTI_TOTALI_PARTITA ≤ 10
THEN Classe = Mentale (2.0)

IF NOME_DIVISIONE = Seconda Divisione **AND** PUNTI_TOTALI_PARTITA > 10
THEN Classe = Coordinativo (3.0/2.0)

- **Descrizione albero a parole :**

L’albero decisionale inizia con il nodo radice **NOME_DIVISIONE**, che rappresenta la variabile principale usata per distinguere i vari rami:

- ITALIA,
- SERIE A1,
- BRASILE,
- SUPER LEGA,
- SERIE A2,
- PRIMA DIVISIONE,
- UNDER 16,
- UNDER 18,
- SERIE B,
- SECONDA DIVISIONE.

Per ogni ramo, si possono incontrare nodi successivi, che dipendono da altre variabili:

1. **PUNTI_TOTALI_PARTITA**, che misura i punti totali in una partita.
2. **SET_VINTI_CASA**, che indica il numero di set vinti in casa.

Alla fine di ogni percorso si arriva a una **foglia**, che rappresenta il valore predetto della variabile target.
Le foglie sono: TECNICO, MENTALE, COORDINATIVO, FISICO, TATTICO.

Se il nome divisione è ITALIA la tipologia di allenamento sarà tecnico.
Se il nome divisione è SERIE A1 l'allenamento sarà tecnico.
Se il nome divisione è BRASILE si va a vedere i punti totali della partita: se essi sono minori o uguali a 19 l'allenamento sarà coordinativo, se sono maggiori di 19 sarà tattico.
Se il nome divisione è SUPER LEGA la tipologia di allenamento sarà mentale.
Se il nome divisione è SERIE A2 guardo i punti totali della partita: se sono minori o uguali a 16 l'allenamento sarà fisico, se sono maggiori di 16 sarà coordinativo.
Se il nome della divisione è PRIMA DIVISIONE l'allenamento sarà coordinativo.
Se il nome divisione è UNDER 16 guardo i set vinti casa: se sono minori o uguali a 2 l'allenamento sarà tecnico, se sono maggiori di 2 sarà mentale.
Se il nome della divisione è UNDER 18 l'allenamento sarà fisico.
Se il nome della divisione è SERIE B l'allenamento sarà coordinativo.
Se il nome della divisione è SECONDA DIVISIONE guardo i punti totali partita: se sono minori o uguali a 10 l'allenamento sarà mentale, se sono maggiori di 10 sarà coordinativo.

Confronto performance con naive bayese:

Con j48 troviamo il 58% delle istanze classificate correttamente.

```
Correctly Classified Instances          29               58      %
Incorrectly Classified Instances        21               42      %
Kappa statistic                          0.4659
Mean absolute error                      0.2096
Root mean squared error                  0.3237
Relative absolute error                 66.1009 %
Root relative squared error             81.3364 %
Total Number of Instances               50
```

Con il naive bayese troviamo il 54% delle istanze classificate correttamente

```
Correctly Classified Instances          27               54      %
Incorrectly Classified Instances        23               46      %
Kappa statistic                          0.4195
Mean absolute error                      0.2348
Root mean squared error                  0.3394
Relative absolute error                 74.0603 %
Root relative squared error             85.272  %
Total Number of Instances               50
```

Tra i due scegliamo l'algoritmo con j48 perché ha il valore delle istanze corrette più alto rispetto al naive bayese.