UNIVERSITÀ
DEGLI STUDI
DI TORINO

INNOVAZIONE SOCIALE, COMUNICAZIONE E NUOVE TECNOLOGIE

Corso: **Gestione e condivisione di basi di dati e conoscenza (2021-2022)**

PROGETTO FINALE

Nome: **Elena Sartori**
Matricola: **975134**

1. DATABASE

Progetto**: Palestra**

Descrizione progetto e Glossario delle Entità

Il progetto in questione considera l'attività di una palestra in tre anni: 2017, 2018, 2019 (la datazione è stata anticipata per non dover tenere in considerazione I periodi chiusura dovuti all'emergenza sanitaria).

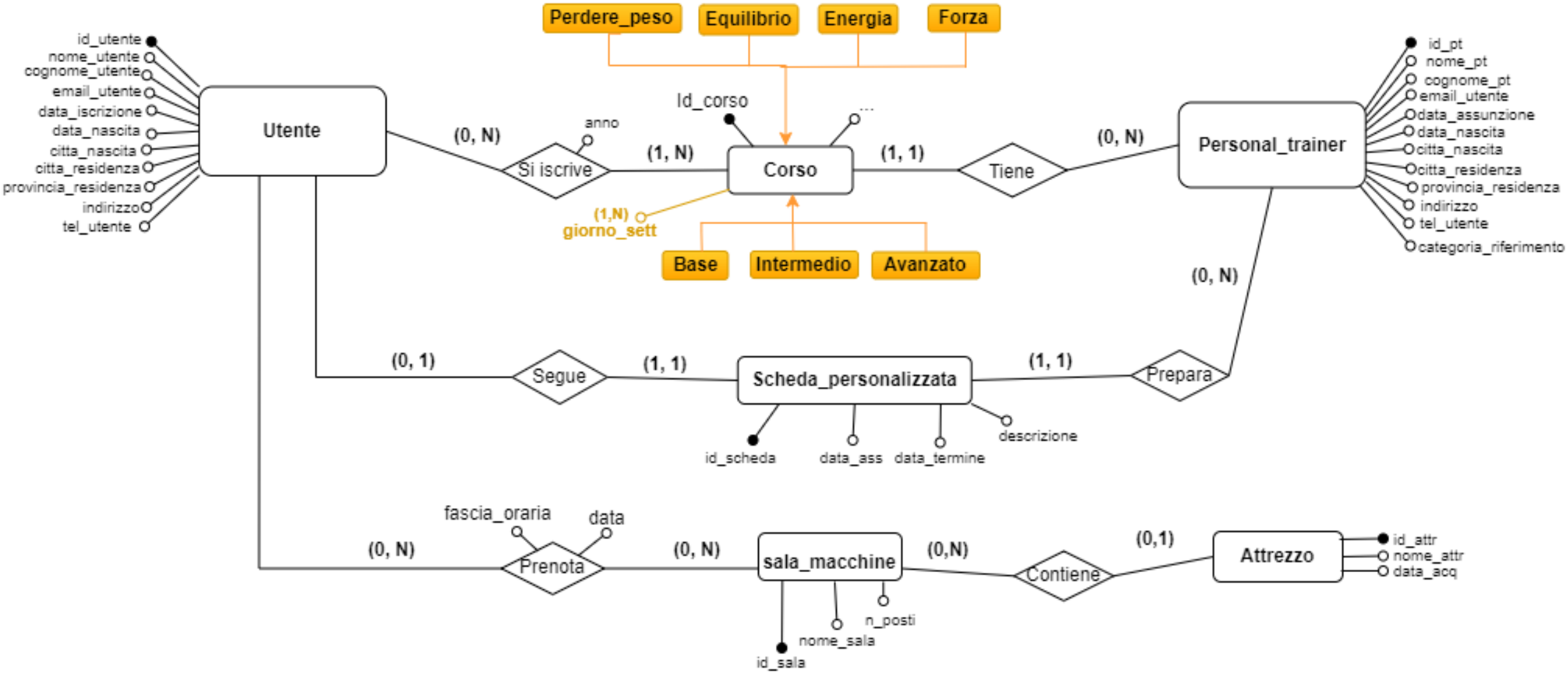
La palestra considerata è formata da una grande sala dedicata allo svolgimento dei corsi e tre sale macchine più piccole in cui gli utenti si possono allenare individualmente.

Gli **utenti** hanno la possibilità di iscriversi ad uno o più **corsi** tenuti da **personal trainer** professionisti o possono prenotare una delle **sale macchine** disponibili per allenarsi individualmente con gli **attrezzi** che contengono.

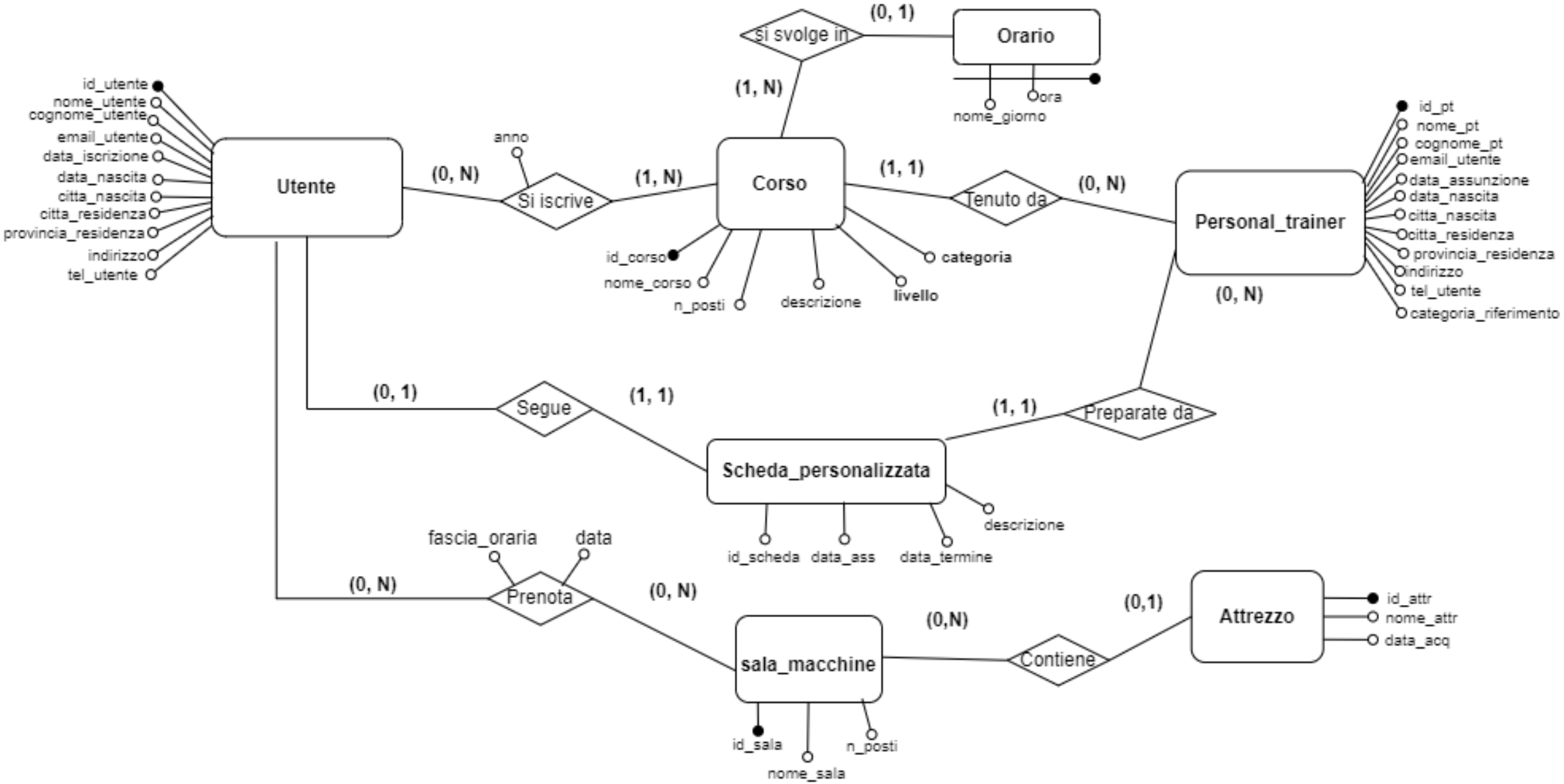
Se lo desiderano possono richiedere una **scheda personalizzata** preparata in base alle esigenze individuali da uno dei personal trainer.

Entità	Descrizione	Relazioni
Utente	Persona che frequenta la palestra, può iscriversi ai corsi, prenotare l'accesso ad una delle sale macchine in una delle fasce orarie stabilite, seguire una scheda di allenamento personalizzata creata da un personal trainer.	corso, scheda personalizzata, sala macchine
Personal trainer	Figura professionale che lavora nella palestra. Può tenere dei corsi, oppure può redigere le schede di allenamento personalizzate per gli utenti.	scheda personalizzata, corso
Corso	Attività sportiva di vario tipo tenuta da un personal trainer, una o più volte alla settimana, in sessioni della durata fissa di un'ora. Ogni corso si riferisce ad una categoria (Forza, Perdere peso, Equilibrio, Energia), e ha specificato il livello di difficoltà (Base, Intermedio, Avanzato).	Utente, personal trainer, orario
Sala macchine	Sala che contiene gli attrezzi per l'allenamento individuale, gli utenti possono prenotarla in diverse fasce orarie (8.00-12.00, 12.00-17.00, 17.00- 22.00). Ha una capienza massima (15 o 20 posti).	Utente, attrezzo
Scheda personalizzata	Scheda di allenamento individuale, nella descrizione il personal trainer inserisce gli esercizi e i commenti che l'utente deve seguire.	Utente, Personal trainer
Attrezzo	L'entità comprende macchinari di vario tipo con cui è possibile allenare diverse parti del corpo.	Sala macchine
Orario	Comprende giorno della settimana e ora inizio in cui si svolgono i corsi .	Corso

1.1 Modello concettuale del database (non ristrutturato)



1.2 Modello concettuale database (ristrutturato)

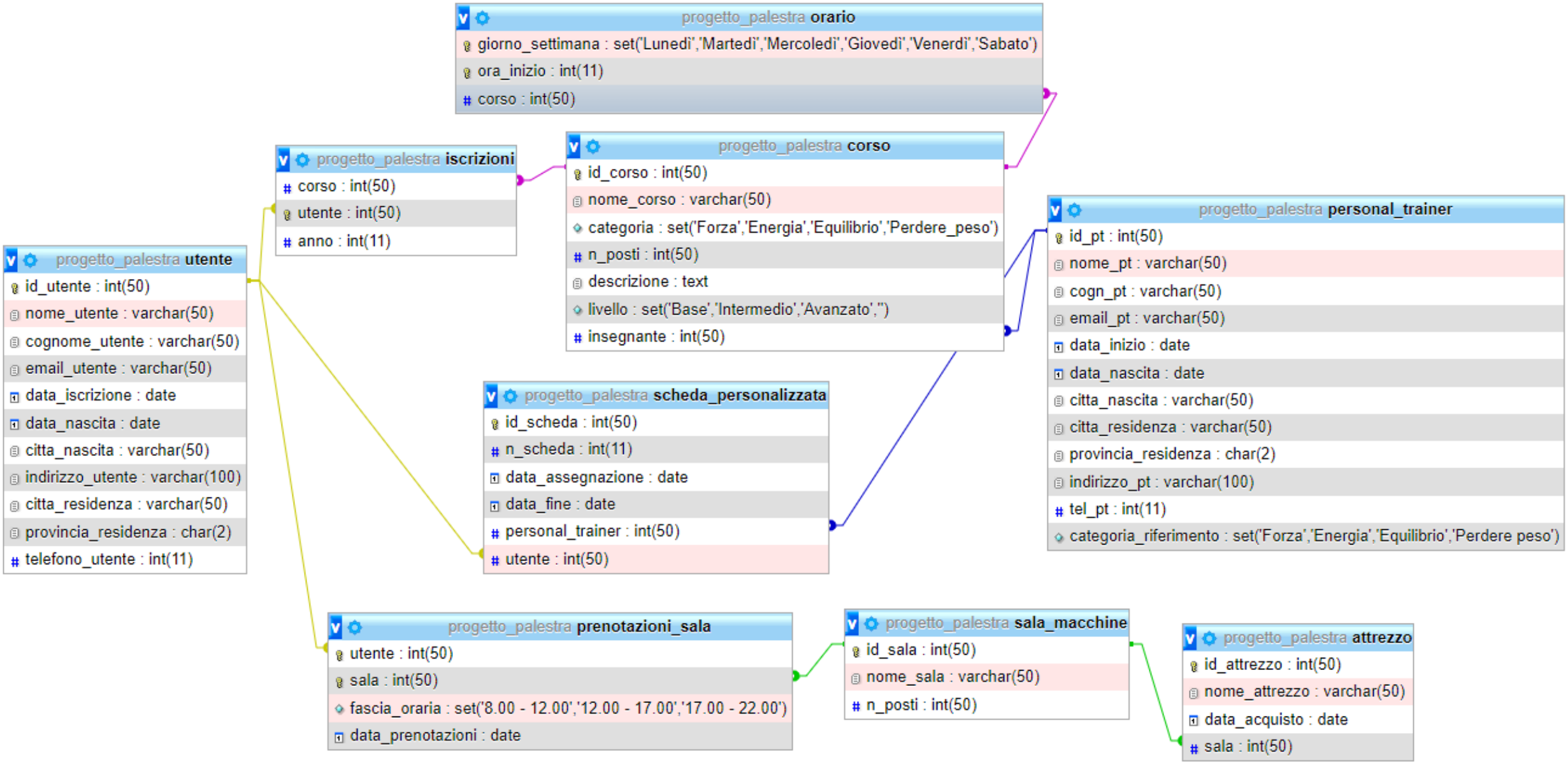


1.3 Modello logico

Utente (id_utente, nome_utente, cognome_utente, email_utente, data_iscrizione, data_nascita, citta_nascita, indirizzo_residenza, prov_residenza, tel_utente)
Corso (id_corso, nome_corso, categoria*, livello*, n_posti, descrizione, insegnante)
Iscrizioni (utente, corso, anno)
Personal_trainer (id_personal, nome_pt, cognome_pt, email_pt, data_assunzione, data_nascita, citta_nascita, citta_residenza, prov_residenza, indirizzo, tel_pt, categoria_riferimento)
Scheda_personalizzata (id_scheda, n_scheda, data_ass, pt, id_utente)
Sala_macchine (id_sala, nome_sala, cap_max)
Attrezzo (id_attr, nome_attr, data_acq, sala)
Prenotazioni_sala (sala, utente, data, fascia_oraria*)
Orario (id_corso, giorno_sett, ora)

*Note:
Categorie possibili (Corso): Forza, Equilibrio, Energia, Perdere peso
Livelli (Corso): Base, Intermedio, Avanzato
Fascia_oraria (Sala_macchine): 8.00-12.00, 12.00-17.00, 17.00-22.00

1.4 Designer del Data Base



2. DATA WAREHOUSE

Progetto: **Palestra**

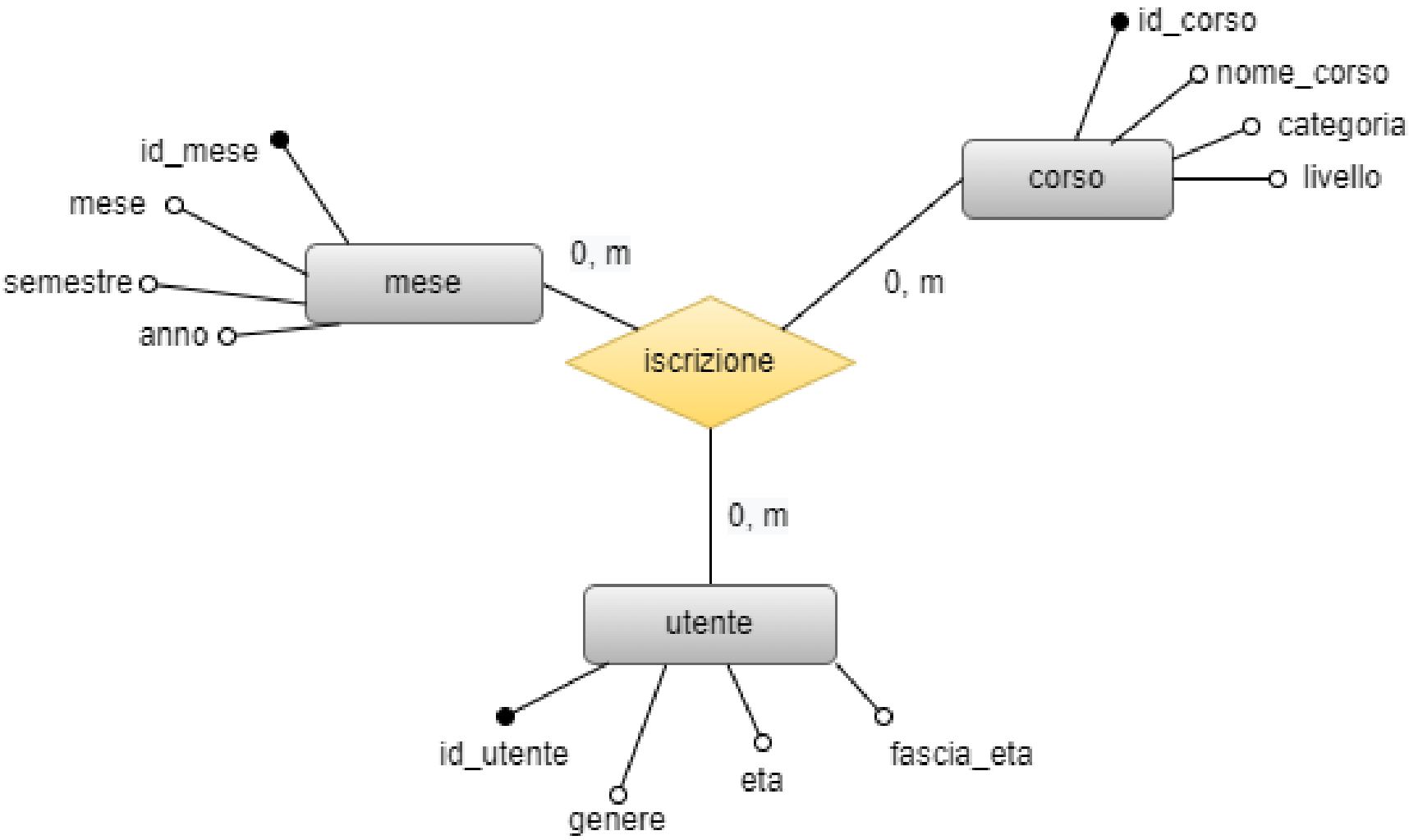
Descrizione Data Warehouse

In questa sezione si è creato un data warehouse per fare analisi sulla distribuzione degli utenti nei vari corsi e categorie di corso. Il **fatto di interesse** attorno a cui si struttura il data warehouse è quindi l'**ISCRIZIONE**.

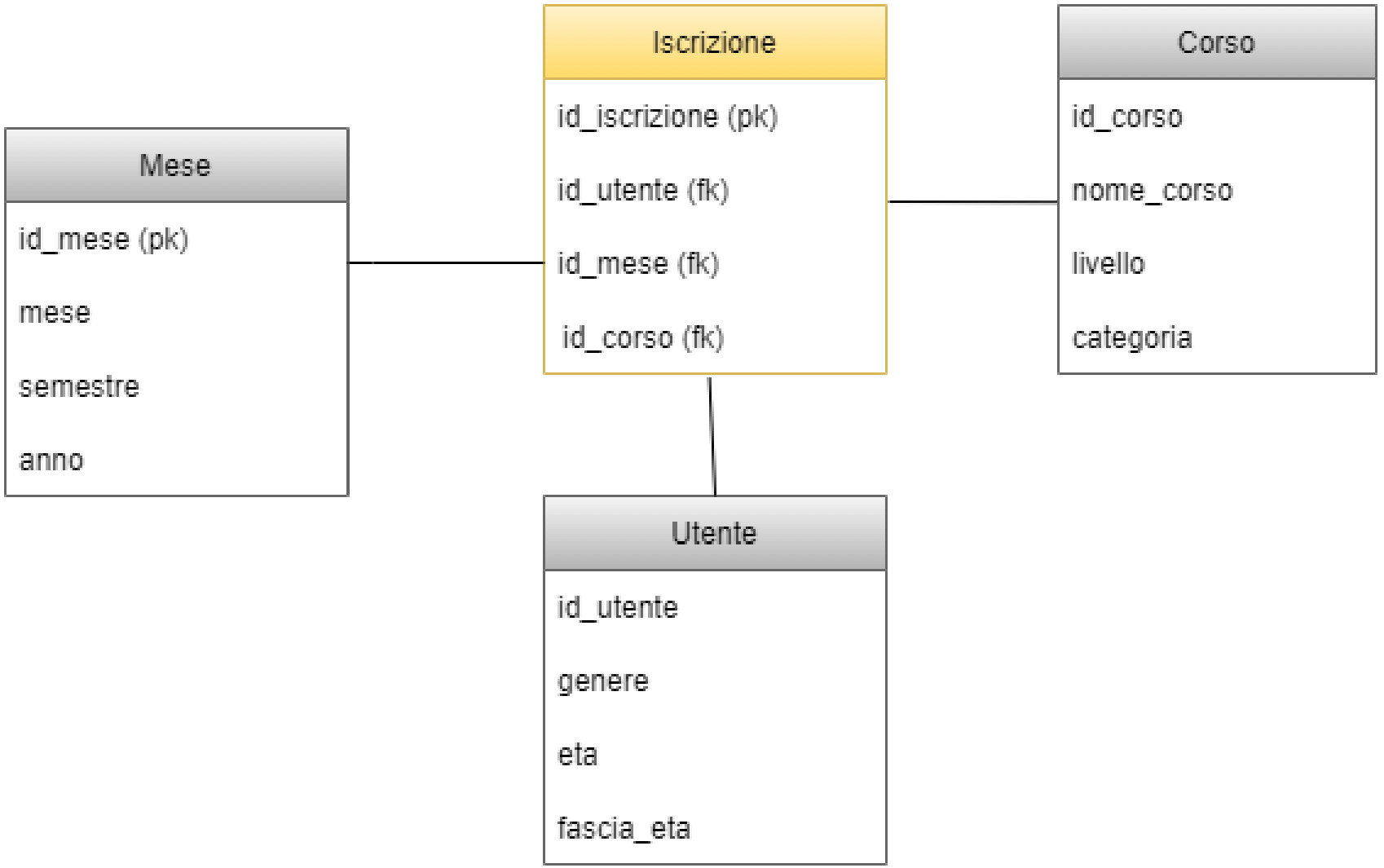
Le **dimensioni** sono invece costituite dagli **utenti**, di cui vengono tenuti in considerazione genere, età e fascia d'età; il **mese** di iscrizione, di cui si registra anche semestre e anno; e ovviamente il **corso**, di cui oltre al nome si considerano anche livello e categoria.

2.1 Modello concettuale

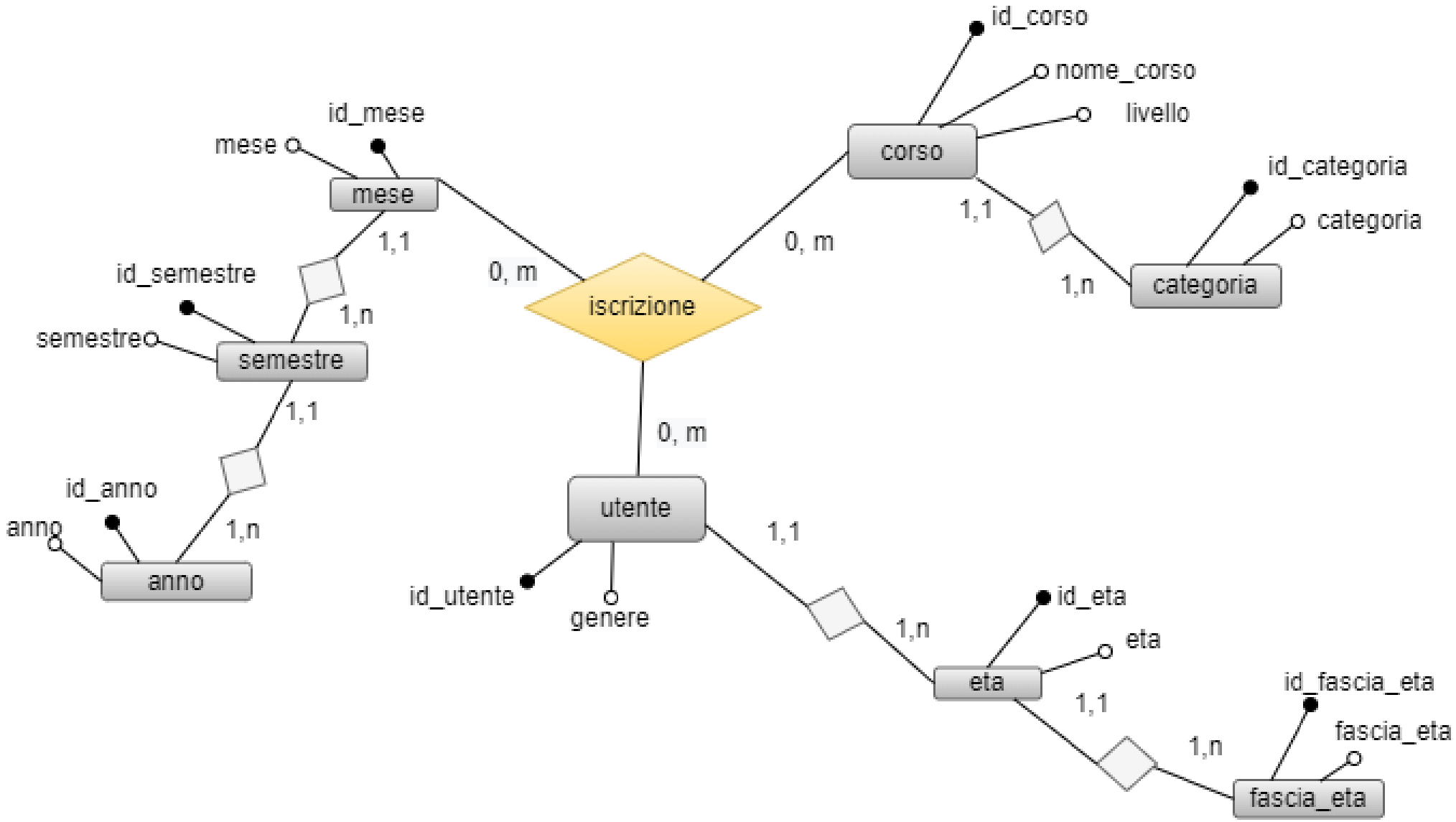
Schema a stella



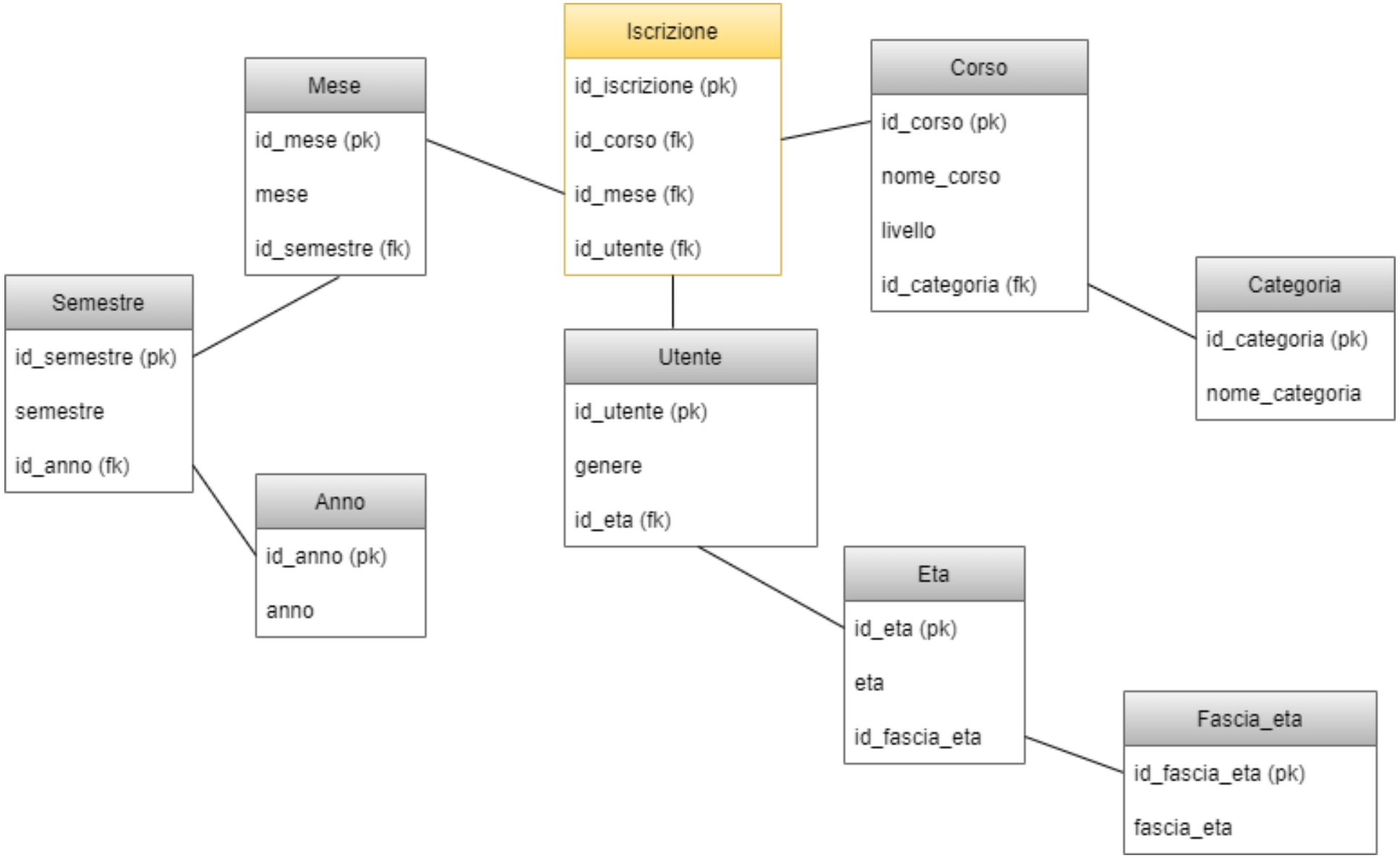
Schema a stella implementato



Schema concettuale: Schema a fiocco di neve



Schema a fiocco di neve implementato

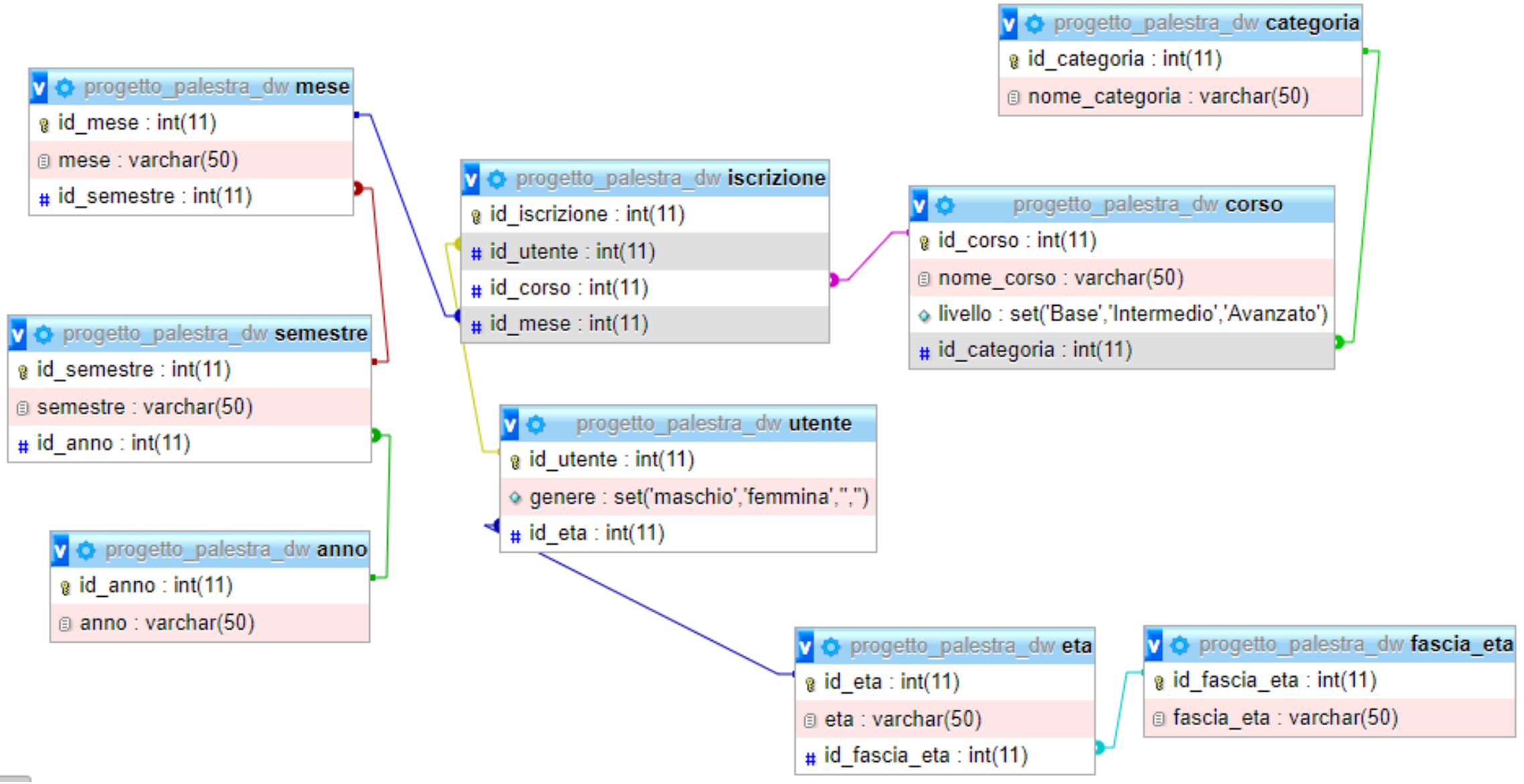


2.2 Modello logico

2.2 Modello logico (dello schema a fiocco di neve)

Iscrizione (id_iscrizione, id_utente, id_corso, id_mese)
Utente (id_utente, genere, id_eta)
Eta (id_eta, eta, id_fascia_eta)
Fascia_eta (id_fascia_eta, fascia_eta)
Corso (id_corso, nome_corso, livello, id_categoria)
Categoria (id_categoria, nome_categoria)
Mese (id_mese, mese, id_semestre)
Semestre (id_semestre, semestre, **id_anno**)
Anno (id_anno, anno)

Designer del Data Warehouse



2.3 Analisi OLAP con tabelle Pivot di Excel

Query esportazione

```
SELECT iscrizione.id_iscrizione, corso.nome_corso, categoria.nome_categoria,
 utente.id_utente, utente.genere, eta.eta, fascia_eta.fascia_eta, anno.anno
 semestre.semestre, mese.mese
FROM iscrizione JOIN corso ON iscrizione.id_corso=corso.id_corso JOIN categoria
   ON corso.id_categoria=categoria.id_categoria JOIN utente ON utente.id_utente
   =iscrizione.id_utente JOIN eta ON utente.id_eta=eta.id_eta JOIN fascia_eta ON
   eta.id_fascia_eta=fascia_eta.id_fascia_eta JOIN mese ON iscrizione.id_mese=
   mese.id_mese JOIN semestre ON mese.id_semestre=semestre.id_semestre  JO
   IN anno ON semestre.id_anno=anno.id_anno
```

Andamento delle iscrizioni ai corsi a seconda del periodo (anno/semestre)

Conteggio di id_iscrizione	Periodo (anno/ semestre)									
	2017		2017 Totale	2018		2018 Totale	2019		2019 Totale	Totale complessivo
Corso	S1	S2		S1	S2		S1	S2		
Atletica	6	2	8	3	1	4	4	1	5	17
CrossFit	3		3	2	1	3	5	5	10	16
Fitness	2		2					2	2	4
Kick boxing	1		1		1	1	5		5	7
Pilates	1	2	3	1	1	2	6	1	7	12
Rugby	5		5	1	1	2	4	2	6	13
Yoga	2		2	2	2	4	2	5	7	13
Zumba	2		2				6		6	8
Totale complessivo	22	4	**26**	9	7	**16**	32	16	**48**	**90**

Operazione di **ROLL UP**: mostrare l'andamento delle iscrizioni a seconda dell'anno

Conteggio di id_iscrizione	Periodo			
	2017	**2018**	**2019**	**Totale complessivo**
Corsi				
Atletica	8	4	5	17
CrossFit	3	3	10	16
Fitness	2		2	4
Kick boxing	1	1	5	7
Pilates	3	2	7	12
Rugby	5	2	6	13
Yoga	2	4	7	13
Zumba	2		6	8
Totale complessivo	26	16	48	90

Operazione di **DRILL DOWN**: mostrare l'andamento delle iscrizioni a seconda del mese (anno 2017)

Conteggio di id_iscrizione	Periodo								
	2017								2017 Totale
	S1				S1 Totale	S2		S2 Totale	
Corsi	gennaio	febbraio	aprile	giugno		settembre	novembre		
Atletica	6				6	2		2	8
CrossFit	2	1			3				3
Fitness	1		1		2				2
Kick boxing			1		1				1
Pilates		1			1	1	1	2	3
Rugby	4		1		5				5
Yoga			1	1	2				2
Zumba			2		2				2
Totale complessivo	13	2	6	1	22	3	1	4	26

Tutte le iscrizioni ai corsi per fascia d'età

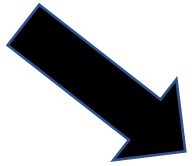
Conteggio di id_iscrizione	Fascia d'età				
Corso	adulti	anziani	bambini	giovani	Totale complessivo
Atletica			8	9	17
CrossFit	5			11	16
Fitness			1	3	4
Kick boxing	1			6	7
Pilates	5			7	12
Rugby			6	7	13
Yoga	6			7	13
Zumba	4	1		3	8
Totale complessivo	21	1	15	53	90

Operazione di **SLICE**: tutte le iscrizioni ai corsi fatte dai bambini

Conteggio di id_iscrizione	Fascia d'età	
Corso	**bambini**	**Totale complessivo**
Atletica	8	8
Fitness	1	1
Rugby	6	6
Totale complessivo	15	15

Operazione di **DICE**: tutte le iscrizioni ai corsi fatte dai bambini nei corsi di categoria "Energia"

Conteggio di id_iscrizione	Fascia d'età				
categorie	**adulti**	**anziani**	**bambini**	**giovani**	**Totale**
Energia			14	16	30
Equilibrio	11			14	25
Forza	6			17	23
Perdere_peso	4	1	1	6	12
Totale complessivo	21	1	15	53	90



Conteggio di id_iscrizione	Fascia età	
Categoria	**bambini**	**Totale complessivo**
Energia	14	14
Totale complessivo	14	14

Operazione di **PUSH**: Media di età delle persone iscritte a corsi nella categoria "Equilibrio"

Media di età	Genere		
Categorie	**femmina**	**maschio**	**Totale complessivo**
Energia	12,66666667	16,04761905	15,03333333
Equilibrio	**27,16666667**	**28,63157895**	**28,28**
Forza	20,85714286	22,375	21,91304348
Perdere_peso	37	20,88888889	24,91666667
Totale complessivo	21,36	21,95384615	21,78888889

3. DATA MINING: CLUSTER

Clustering

In questa sezione si è voluta analizzare la suddivisione degli iscritti nelle varie categorie di corsi ("Energia", "Forza", "Equilibrio", "Perdere peso") tenendo conto del loro genere e della loro età.

Informazioni generali clustering

Instances: **90**
Attributes: **3**
nome_categoria
genere
eta
Test mode: **evaluate on training data**

Misura di distanza usata: **Distanza euclidea**
Algoritmo: **k-means**
Numero di cluster: **4**
Seed: **10**
Scarto quadratico: **24.483343634694986**
Numero di iterazioni: **3**

Cluster ottenuti

Attribute	Full Data	0	1	2	3
	(90.00)	(20.0)	(45.0)	(12.0)	(13.0)
nome_categoria	Forza	Energia	Forza	Perdere_peso	Equilibrio
genere	maschio	femmina	maschio	femmina	maschio
eta	20.1667	9.65	20.3333	22.5833	33.5385

Femmina, 9 anni, ENERGIA

Femmina, 22 anni, PERDERE PESO

Maschio, 20 anni, FORZA

Maschio, 33 anni, EQUILIBRIO

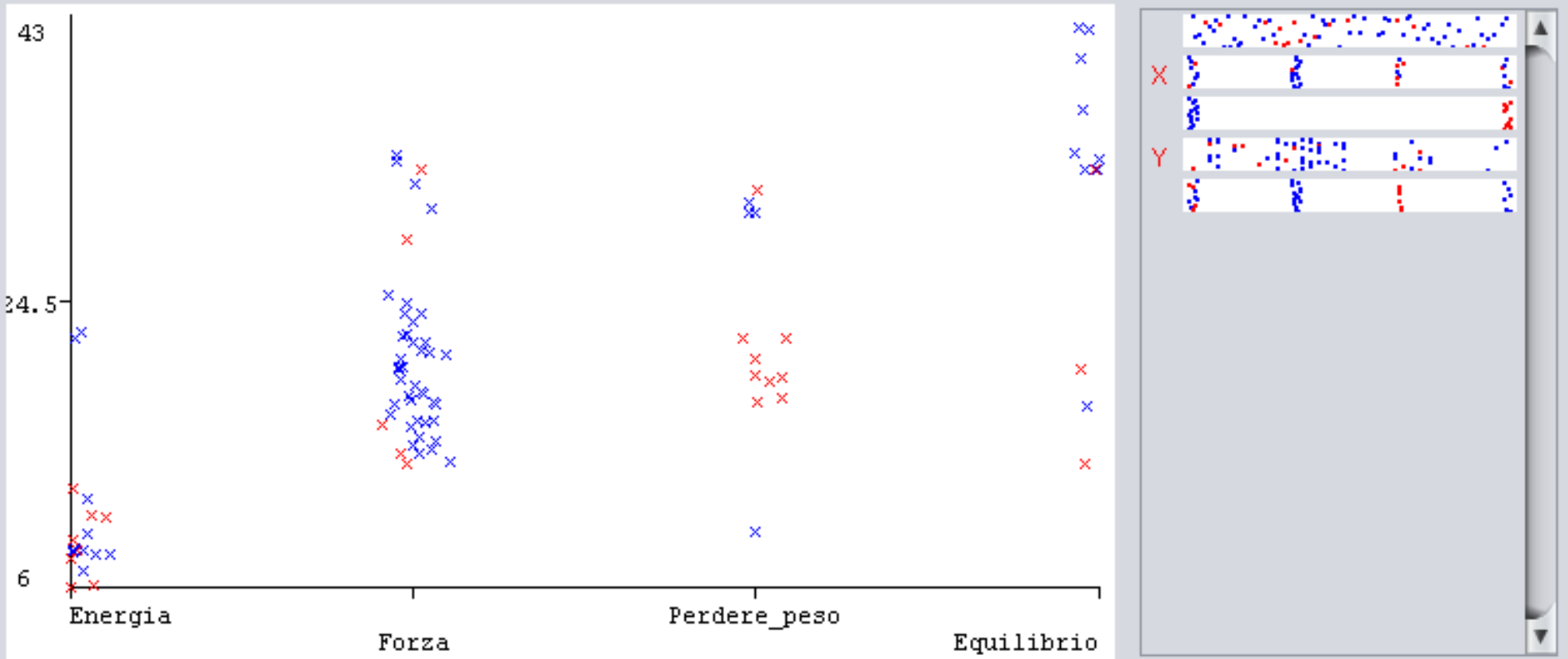
Attribute	Full Data	0	1	2	3
	(90.0)	(20.0)	(45.0)	(12.0)	(13.0)
nome_categoria	Forza	Energia	Forza	Perdere_peso	Equilibrio
Energia	19.0 (21%)	17.0 (85%)	2.0 (4%)	0.0 (0%)	0.0 (0%)
Forza	46.0 (51%)	2.0 (10%)	42.0 (93%)	2.0 (16%)	0.0 (0%)
Perdere_peso	13.0 (14%)	0.0 (0%)	1.0 (2%)	9.0 (75%)	3.0 (23%)
Equilibrio	12.0 (13%)	1.0 (5%)	0.0 (0%)	1.0 (8%)	10.0 (76%)
genere	maschio	femmina	maschio	femmina	maschio
maschio	65.0 (72%)	9.0 (45%)	44.0 (97%)	0.0 (0%)	12.0 (92%)
femmina	25.0 (27%)	11.0 (55%)	1.0 (2%)	12.0 (100%)	1.0 (7%)
eta	20.1667	9.65	20.3333	22.5833	33.5385
	+/-8.6365	+/-2.4121	+/-4.9955	+/-5.6159	+/-6.2931

X: nome_categoria (Nom) | Y: eta (Num)

Colour: genere (Nom) | Select Instance

Reset | Clear | Open | Save | Jitter

Plot: iscrizione (3)-weka.filters.unsupervised.attribute.Remove-R1-2,4,7-10_clustered



Class colour

maschio femmina

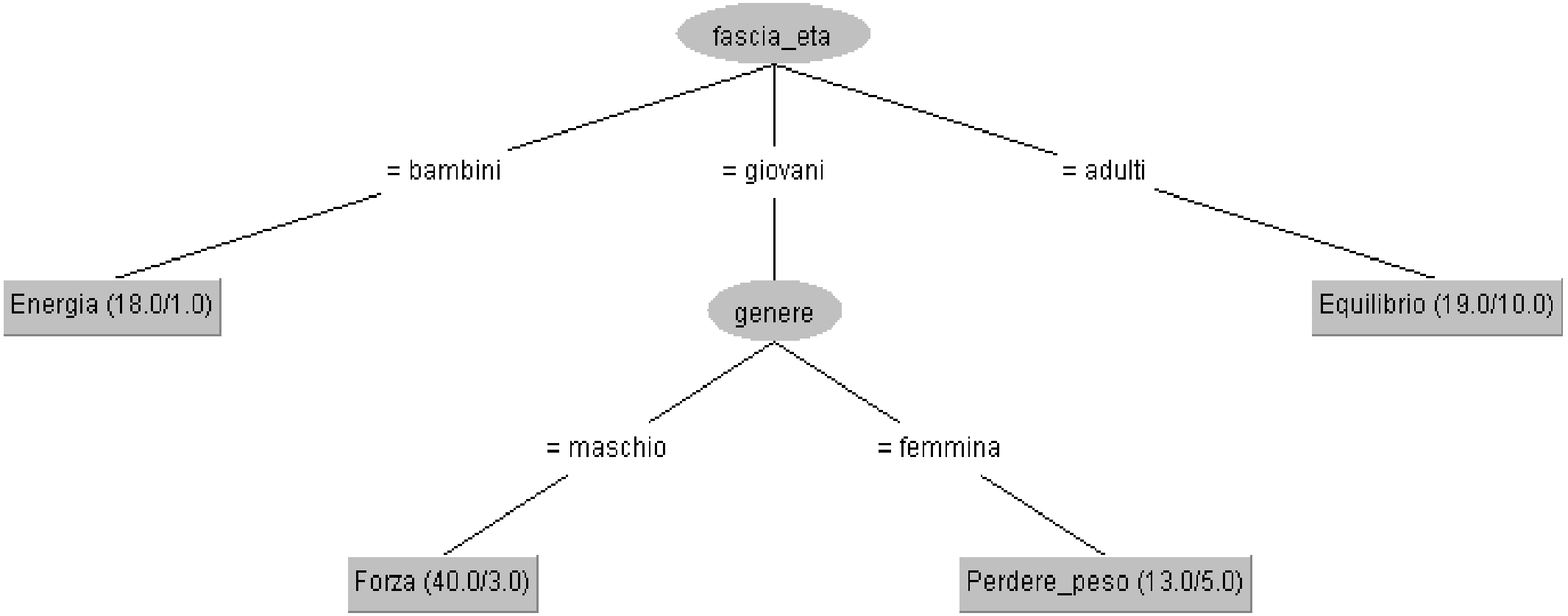
4. DATA MINING: ALBERI DECISIONALI

Albero decisionale

In questa sezione si è voluta fare una classificazione per capire a quale categoria di corso si possono iscrivere gli utenti date alcune variabili di input: genere, età, mese di iscrizione.

Nota: il mese di iscrizione non è stato ritenuto come significativo e discriminante nella scelta della categoria dal software che ha creato l'albero decisionale, e non è stato quindi tenuto in considerazione.

Albero decisionale generato da Weka



Informazioni sull'albero di decisione

Informazioni sull'albero decisionale	
Variabili di input	genere, fascia_eta, mese, nome_categoria
Variabili di output	nome_categoria
Tecnica di valutazione	Training set

```
Correctly Classified Instances          71               78.8889 %
Incorrectly Classified Instances        19               21.1111 %
```

Precision	Recall	F-Measure
0,944	0,895	0,919
0,925	0,804	0,860
0,615	0,615	0,615
0,474	0,750	0,581
0,824	0,789	0,800

Descrizione della matrice di confusione (Confusion Matrix)

```
 a  b  c  d   <-- classified as
17  2  0  0 |  a = Energia
 0 37  3  6 |  b = Forza
 1  0  8  4 |  c = Perdere_peso
 0  1  2  9 |  d = Equilibrio
```

La matrice di confusione da un'indicazione della tipologia di classificazione che è stata fatta eventualmente in modo scorretto. Ad ogni valore che può assumere la foglia viene associata una lettera. Le istanze classificate sotto le lettere mostrano le istanze classificate correttamente e scorrettamente. Alcune istanze in questo esempio presentano una certa percentuale di classificazione erronea. Nella prima riga ad esempio sono presenti due istanze scorrette. In generale i valori scorretti sono quelli che si discostano dalla diagonale.

Descrizione albero

```
fascia_eta = bambini: Energia (18.0/1.0)
fascia_eta = giovani
|   genere = maschio: Forza (40.0/3.0)
|   genere = femmina: Perdere_peso (13.0/5.0)
fascia_eta = adulti: Equilibrio (19.0/10.0)
```

L'albero mostra che se la fascia d'età è "bambini", la categoria di corsi scelta è "Energia"; se la fascia d'età è "giovani" occorre guardare al genere, i maschi sceglieranno corsi di categoria "Forza", le femmine corsi di categoria "Perdere peso"; per gli "adulti invece i corsi scelti rientrano nella cateogoria "Equilibrio".

Confronto con performance di un NAIVE BAYES

J-48

Correctly Classified Instances	71	78.8889 %
Incorrectly Classified Instances	19	21.1111 %

Precision	Recall	F-Measure
0,944	0,895	0,919
0,925	0,804	0,860
0,615	0,615	0,615
0,474	0,750	0,581
0,824	0,789	0,800

Naive Bayes

Correctly Classified Instances	64	71.1111 %
Incorrectly Classified Instances	26	28.8889 %

Precision	Recall	F-Measure
0,944	0,895	0,919
0,702	0,870	0,777
0,500	0,385	0,435
0,400	0,167	0,235
0,684	0,711	0,685

In questo caso la percentuale di istanze classificate correttamente è maggiore con l'utilizzo dell'algoritmo J-48, che va quindi preferito al Naive Bayes. Anche la Precision, la Recall e la F-measure presentano valori più vicini ad 1, e quindi migliori, con J-48.

5. TEXT MINING

Classificazione di un corpus di tweet

Text mining (Weka)

In questa sezione sono presentati i report di 3 diversi esperimenti realizzati su un corpus di tweet utilizzando tecniche di text mining (software weka). Ogni esperimento mostra l'utilizzo di diversi algoritmi, tecniche di validazione e parametri.

	NOME DATASET	TIPO DI VALIDAZIONE	ALGORITMO	STOPWORDS	METODO DI TOKENIZZAZIONE	ACCURACY	F-MEASURE AVG	F-MEASURE CLASS 1	PRECISION CLASS 1	RECALL CLASS 1	5 ATTRIBUTI TOKENIZER
Esperimento 1	Haspeede_TW-train-windows	5-fold	Naive Bayes	no	NGramTokenizer min 1, max 2	73.8%	0,746	0,657	0,571	0,774	#LIDL, alla, @corriere, Casa, Campo
Esperimento 2	Haspeede_TW-train-windows	Split 20/80	Naive Bayes	no	Character NGram Tokenizer min 2, max 5	72.8333 %	0,735	0,634	0,566	0,719	Finti, gia, gov, pago, porte
Esperimento 3	Haspeede_TW-train-windows_1	Split 20/80	J-48	si	WordTokenizer	73.3333 %	0,763	0,556	0,400	0,909	Cedric, #welfare, #, Earth