Analisi dati esami del 1^o e 2^o anno, su un piccolo campione di studenti laureandi

Marco Calamai Michele De Vita
Matteo Gemignani Elia Mercatanti

Corso di laurea magistrale in informatica

25 gennanio 2018

Dati iniziali

Coorte	Id studente	Test	Voto diploma	Tipo diploma
2013	A	18	80	IT
2013	B	13	67	IT
2013	C	18	78	IT
2014	D	14	66	LS
2014	E	0	82	TC

Dati iniziali

- **Tabella studenti**
  - **Coorte**: L'anno di immatricolazione.
  - **Id studenti**: identificativo univoco di ogni studente, in questo caso una lettera.
  - **Test**: il voto del test di autovalutazione fatto al primo anno.
  - **Voto diploma**: il voto del diploma di maturità.
  - **Tipo diploma**: scuola di provenienza.

Dati iniziali

Id studente	Codice esame	Data esame	Tipo	Voto	Giudizio	Crediti	Descrizione
A	B006807	2015-01-16	S	24		6	ALGEBRA LINEARE
A	B006800	2014-07-17	S	29		12	ALGORITMI E STRUTTURE DATI
A	B006801	2014-06-10	S	30		12	ANALISI I: CALCOLO DIFFERENZIALE ED INTEGRALE
A	B006808	2017-01-23	S	18		6	ANALISI II: FUNZIONI DI PIÙ VARIABILI
A	B006802	2014-06-12	S	25		12	ARCHITETTURE DEGLI ELABORATORI

Dati iniziali

- **Tabella voti**
  - **Id studente**: identificatore univoco di ogni studente, nel nostro caso una lettera
  - **Codice esame**: identificatore univoco di ogni esame
  - **Data esame**: data di registrazione dell'esame
  - **Tipo**: colonna che può contenere S (Sostenuto) o C (Convalidato)
  - **Voto**: voto dell'esame
  - **Giudizio**: nel caso in cui l'esame non preveda un voto in trentesimi questa colonna contiene P (Passato)
  - **Crediti**: numero di crediti assegnati all'esame
  - **Descrizione**: descrizione ampia dell'esame

Problematiche

- Le date dell'esame erano a volte assenti o presenti in formati diversi formati: YYYY-MM-DD e DD/MM/YYYY.
- Giudizio esclusivo sull'esame di Inglese
- Colonna Tipo

Preprocessing

- Sono state standardizzate le date nel formato anglosassone (YYYY-MM-DD), sfruttando la libreria "*dateutil*" di Python, per il parsing delle date in vario formato.
- Sono state eliminate le colonne di giudizio e tipo ritenute poco informative.
- Sono state eliminate le righe riguardanti l'esame di inglese perché non avevano un giudizio in trentesimi
- Per le date mancanti, è stato deciso di stimare i semestri, assegnando a ciascuno di essi il semestre più frequente in cui gli altri studenti hanno svolto lo stesso esame. (Utilizzato poi anche sulle sequenze temporali)

Tabella preprocessata

Id_studente	codice_esame	data_esame	voto	crediti	descrizione	Semestre
A	B006807	2015-01-16	24	6	ALGEBRA LINEARE	2
A	B006800	2014-07-17	29	12	ALGORITMI E STRUTTURE DATI	1
A	B006801	2014-06-10	30	12	ANALISI I: CALCOLO DIFFERENZIALE ED INTEGRALE	1
A	B006808	2017-01-23	18	6	ANALISI II: FUNZIONI DI PIÙ VARIABILI	6
A	B006802	2014-06-12	25	12	ARCHITETTURE DEGLI ELABORATORI	1

File sequenze temporali

- Per analizzare i patern nelle sequenze temporali, sono stati elaborati i dati in due passi per gestirli al meglio con Weka e SPMF.
  1. trasformare la tabella preprocessata in sequenze temporali basandoci sulla colonna "Semestre".
  2. esportare le sequenze temporali nei formati adatti per i due programmi
     - Per Weka è stato creato un file "*.arff*" dove per ogni semestre sono stati indicati sia gli esami sostenuti che quelli non sostenuti.
     - per SPMF è stato creato un file "*.txt*", dove ogni riga contiene: la sequenza temporale di uno studente, il separatore di item (-1) e il separatore di transazioni (-2)

Sequenze temporali

- 1 - ALGEBRA LINEARE
- 2 - ANALISI I: CALCOLO DIFFERENZIALE ED INTEGRALE
- 3 - PROGRAMMAZIONE
- 4 - ANALISI II: FUNZIONI DI PIÙ' VARIABILI
- 5 - FISICA GENERALE
- 6 - MATEMATICA DISCRETA E LOGICA
- 7 - METODOLOGIE DI PROGRAMMAZIONE
- 8 - PROGRAMMAZIONE CONCORRENTE
- 9 - SISTEMI OPERATIVI
- 10 - BASI DI DATI E SISTEMI INFORMATIVI
- 11 - ALGORITMI E STRUTTURE DATI
- 12 - CALCOLO DELLE PROBABILITÀ E STATISTICA
- 13 - ARCHITETTURE DEGLI ELABORATORI

Sequenze temporali

Id studente	Sequenza temporale
A	$< \{11, 2, 13, 3\}, \{1, 7, 8\}, \{4, 10, 12\} >$
B	$< \{11, 2, 3\}, \{7\}, \{4, 10\}, \{1, 13\}, \{6\}, \{9\} >$
C	$< \{11, 3\}, \{10, 7\}, \{13\}, \{8, 9\} >$
D	$< \{2, 3\}, \{1, 11\}, \{4, 13, 12\} >$
...	...
X	$< \{11, 2, 3\}, \{10, 7, 8\}, \{1, 4, 13, 12\} >$
Y	$< \{11, 2, 13, 10, 3, 8\}, \{1\}, \{4, 5, 7, 9\} >$
Z	$< \{11, 2, 13, 6, 3\}, \{1, 12, 7, 8\}, \{4, 10, 5, 9\} >$

File .arff per Weka

```
@relation sequential_example

@attribute sequenceID {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
@attribute algebraLineare {nA,A}
@attribute analisiI {nB,B}
@attribute programmazione {nC,C}
@attribute analisiII {nD,D}
@attribute fisicaGenerale {nE,E}
@attribute matematicaDriscretaELogica {nF,F}
@attribute metodologieDiProgrammazione {nG,G}
@attribute programmazioneConcorrente {nH,H}
@attribute sistemiOperativi {nI,I}
@attribute basiDiDati {nJ,J}
@attribute algoritmiEStruttureDati {nK,K}
@attribute calcoloDellaProbabilitaEStatistica {nL,L}
@attribute architettureDegliElaboratori {nM,M}

@data
1,nA,B,C,nD,nE,nF,nG,nH,nI,nJ,K,nL,M
1,A,nB,nC,nD,nE,nF,G,H,nI,nJ,nK,nL,nM
1,nA,nB,nC,D,nE,nF,nG,nH,nI,J,nK,L,nM
2,nA,B,C,nD,nE,nF,nG,nH,nI,nJ,K,nL,nM
2,nA,nB,nC,nD,nE,nF,G,nH,nI,nJ,nK,nL,nM
2,nA,nB,nC,D,nE,nF,nG,nH,nI,J,nK,nL,nM
2,A,nB,nC,nD,nE,nF,nG,nH,nI,nJ,nK,nL,M
2,nA,nB,nC,nD,nE,F,nG,nH,nI,nJ,nK,nL,nM
2,nA,nB,nC,nD,nE,nF,nG,nH,I,nJ,nK,nL,nM
3,nA,nB,C,nD,nE,nF,nG,nH,nI,nJ,K,nL,nM
3,nA,nB,nC,nD,nE,nF,G,nH,nI,J,nK,nL,nM
```

File .txt per SPMF

```
11 2 13 3 -1 1 7 8 -1 4 10 12 -2
11 2 3 -1 7 -1 4 10 -1 1 13 -1 6 -1 9 -2
11 3 -1 10 7 -1 13 -1 8 9 -2
2 3 -1 1 11 -1 4 13 12 -2
11 -1 2 -2
11 2 3 -1 1 13 -1 10 6 8 -2
11 2 6 3 -1 1 13 12 7 -1 4 10 8 -1 9 -2
11 2 13 6 3 -1 1 8 -1 4 10 12 9 -2
11 2 -1 1 12 8 -1 4 -2
11 2 6 3 -1 1 4 10 12 5 8 -2
11 2 13 6 3 -1 1 12 7 8 -1 4 10 5 9 -2
11 2 13 6 3 -1 1 12 7 8 -1 4 10 5 9 -2
11 3 -1 2 7 -1 4 10 -2
11 2 3 -1 1 13 12 7 8 -1 4 10 6 9 -2
11 2 3 -1 4 -1 1 13 7 8 -2
11 2 13 -1 10 12 3 -1 1 4 6 7 8 9 -2
11 2 3 -1 1 12 -1 4 13 10 7 8 -2
11 2 13 6 3 -1 1 12 7 8 -1 4 10 9 -2
1 11 2 4 13 6 3 -1 12 7 8 -1 10 5 9 -2
11 13 -1 1 3 -1 2 4 10 8 -2
11 4 13 3 -1 1 12 7 8 -1 10 9 -2
11 2 -1 13 7 3 -2
11 2 13 6 3 -1 4 12 7 8 -1 1 10 5 9 -2
11 2 3 -1 10 7 8 -1 1 4 13 12 -2
11 2 13 10 3 8 -1 1 -1 4 5 7 9 -2
11 2 13 6 3 -1 1 12 7 8 -1 4 10 5 9 -2
```

Risultati Weka

```
=== Associator model (full training set) ===

GeneralizedSequentialPatterns
=============================

Number of cycles performed: 12
Total number of frequent sequences: 25716

Frequent Sequences Details (filtered):

- 1-sequences

[1] <{nA}> (26)
[2] <{nB}> (26)
[3] <{B}> (24)
[4] <{nC}> (26)
[5] <{C}> (24)
[6] <{nD}> (26)
[7] <{nE}> (26)
...
[24] <{nE,nG,nI,nJ,K,nL}{nB,nE,nF,nI,nK}> (23)
[25] <{nE,nG,nI,nJ,K,nL}{nE,nF,nI,nJ,nK}> (23)
[26] <{nE,nH,nI,nJ,K,nL}{nB,nE,nF,nI,nK}> (23)
[27] <{nE,nH,nI,nJ,K,nL}{nE,nF,nI,nJ,nK}> (23)
[28] <{nG,nH,nI,nJ,K,nL}{nB,nE,nF,nI,nK}> (23)
[29] <{nG,nH,nI,nJ,K,nL}{nE,nF,nI,nJ,nK}> (23)

- 12-sequences

[1] <{nE,nG,nH,nI,nJ,K,nL}{nB,nE,nF,nI,nK}> (23)
[2] <{nE,nG,nH,nI,nJ,K,nL}{nE,nF,nI,nJ,nK}> (23)
```

Risultati Weka

- Parametri: *minSup* (supporto minimo): 0.9.
- Generate 25716 k-sequenze.
- Si prenda questa 2-sequenza: $< \{C\}\{nC\} >$; questa è un chiaro esempio di scarsa informazione.
- Risultano "frequenti" k-sequenze con item contenenti informazioni riguardanti gli esami non sostenuti

Risultati SPMF

```
1 -1 #SUP: 22
2 -1 #SUP: 24
3 -1 #SUP: 24
4 -1 #SUP: 22
5 -1 #SUP: 7
6 -1 #SUP: 13
7 -1 #SUP: 19
8 -1 #SUP: 21
9 -1 #SUP: 14
10 -1 #SUP: 21
11 -1 #SUP: 26
12 -1 #SUP: 17
13 -1 #SUP: 22
2 -1 1 -1 #SUP: 7
3 -1 1 -1 #SUP: 5
11 -1 1 -1 #SUP: 8
1 12 -1 #SUP: 8
13 -1 1 -1 #SUP: 9
1 13 -1 #SUP: 3
2 3 -1 #SUP: 7
2 -1 4 -1 #SUP: 3
2 -1 7 -1 #SUP: 3
2 -1 10 -1 #SUP: 3
2 -1 12 -1 #SUP: 3
2 -1 13 -1 #SUP: 3
7 8 -1 #SUP: 3
11 -1 12 -1 #SUP: 6
11 13 -1 #SUP: 8
13 -1 12 -1 #SUP: 7
2 -1 1 13 -1 #SUP: 3
11 -1 1 12 -1 #SUP: 5
13 -1 1 12 -1 #SUP: 5
2 3 -1 1 -1 #SUP: 5
11 13 -1 1 -1 #SUP: 7
11 13 -1 12 -1 #SUP: 5
11 13 -1 1 12 -1 #SUP: 4
```

Risultati SPMF

- Parametri: *minSup* (supporto minimo): 0.1.
- Si nota che analizzando soltanto gli esami svolti, risultano molte meno sequenze con supporto basso.

Postprocessing sequenze temporali

- È stato ritenuto interessante analizzare la similarità degli studenti con lo *studente modello*.
- Lo *studente modello* è una sequenza temporale che prevede il sostenimento degli esami nel minimo tempo utile, quindi può essere definito come uno studente "in pari".
- Per vedere la similarità è stato utilizzato il coefficiente di *Jaccard*.
- Gli item degli studenti sono stati trasformati in attributi binari asimmetrici, con un 1 se ha sostenuto il dato esame in quel semestre e 0 se non lo ha sostenuto.

Tabella attributi assimetrici e Studente modello

Id studente	Sequenza binaria asimmetrica
A	<{0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1} {1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0} {0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0}>
B	<{0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0} {0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0} {0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0} {1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1} {0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0} {0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0}>

Studente Modello :

<{0,1,1,0,0,0,0,0,0,1,1,0,1}
{1,0,0,0,0,0,1,1,0,0,0,1,0}
{0,0,0,1,1,0,0,0,1,1,0,0,0}
{0,0,0,0,0,0,0,0,0,0,0,0,0}
{0,0,0,0,0,0,0,0,0,0,0,0,0}
{0,0,0,0,0,0,0,0,0,0,0,0,0}>

Coefficiente di Jaccard

- È stato quindi calcolato il coefficiente di Jaccard per semestre, facendo poi una media dei coefficienti, abbiamo trovato la similarità degli studenti con lo studente ideale.
- È stata calcolata poi un tabella con attributi binari asimmetrici. Nelle colonne sono presenti i pattern sequenziali e nelle righe gli studenti, quindi 1 significa la presenza del pattern nella sequenza temporale dello studente.

Id_studente	<{2};{1}>	<{3};{1}>	<{11};{1}>	<{1;12}>	<{13};{1}>	...	<{11;13};{1;12}>
A	1	1	1	0	1	...	0
B	1	1	1	0	0	...	0
C	0	0	0	0	0	...	0
D	1	1	0	0	0	...	0
E	0	0	0	0	0	...	0

Coefficiente ω

- Definiamo quindi ω come segue:
  - Dati n pattern frequenti, v_i i-esmio elemento della riga (della tabella sopra citata) e σ_i il supporto dell'i-esmio pattern frequente, si definisce il coefficiente ω come segue:

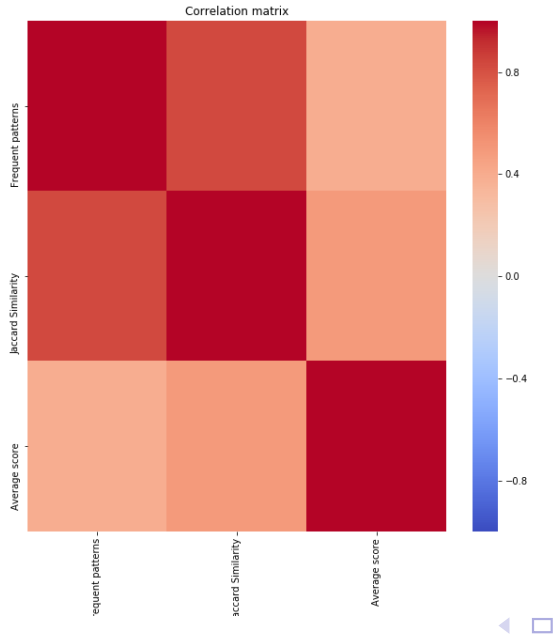
$$\omega = \frac{\sum_{i=0}^{n} v_i \cdot \sigma_i}{\sum_{i=0}^{n} v_i} \qquad \omega \in [0,1]\ \sigma_i \in [3,9]\ v_i \in \{0,1\}$$

- il coefficiente ω rappresenta la somma pesata standardizzata dei pattern che rispetta lo studente, più vicino a 1 più lo studente presenta pattern frequenti nella sua sequenza temporale.

Correlazione

- È stata calcolata la correlazione di Pearson a coppie tra il coefficiente di Jaccard, il coefficiente omega e la media pesata degli studenti.

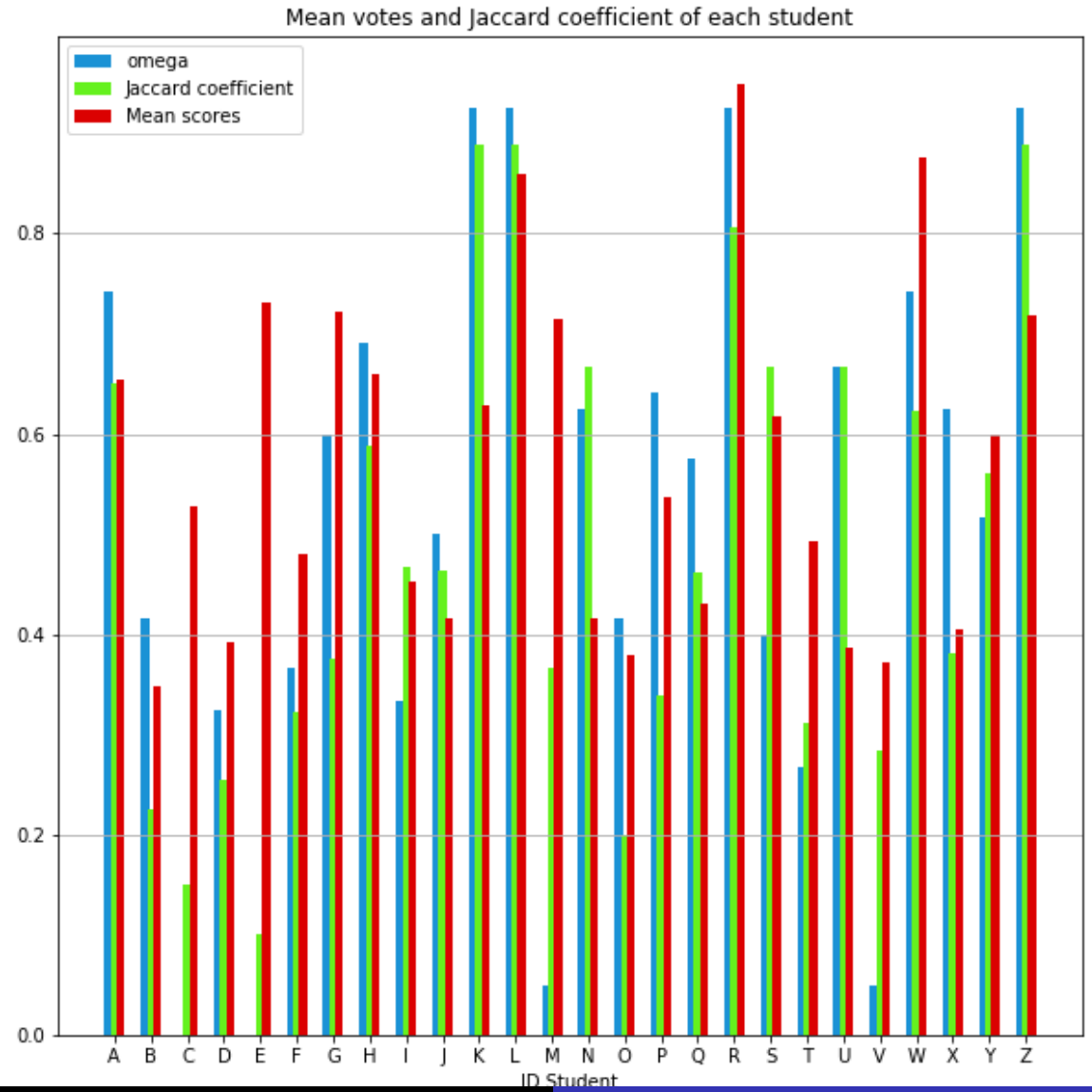
	Jaccard	ω	**Media voti**
Jaccard	1		
ω	0.83	1	
Media voti	0.50	0.40	1



Analisi

- Fra i pattern trovati ne risultato due particolarmente significativi: <{11;13}> e <{1;12}>. Ciò implica che una buona parte degli attuali studenti ha sostenuto questi esami nello stesso semestre.
- L'alta correlazione tra il coefficiente di Jaccard e il coefficiente ω, implica che, chi presenta molti pattern frequenti, allora molto probabilmente sosterrà gli esami nel primo semestre utile e viceversa.
- Chi svolge gli esami in tempo, quindi ha una similarità più alta con lo studente modello, ha una moderata probabilità di avere una media dei voti più alta; questo si nota dalla correlazione positiva tra media dei voti e coefficiente di Jaccard.
- La leggera correlazione tra il coefficiente ω e la media dei voti, porta a non poter supporre che gli studenti con molti pattern frequenti nella loro sequenza di esame abbiano una buona media.

Analisi


Mean votes and Jaccard coefficient of each student
omega
Jaccard coefficient
Mean scores
0.0
0.2
0.4
0.6
0.8
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
ID Student

Creazione Database

Tre tabelle per la gestione dei dati e creazione di viste per facilitare lo studio del dataset tramite clustering

- **students** contenente *student_ id, cohort, test_grade, hs_diploma_grade, hs_diploma_title*
- **courses** contenente *course_id, cfu, description*
- **exams** contenente *student_id, course_id, date, grade, semester*

Voti mancanti al test di ingresso integrati con media complessiva dei risultati al test di tutti gli studenti

Tipi di analisi e viste

- **Due principali tipi di analisi**
  - Sulla carriera e il percorso di ogni studente
  - Sull'andamento dei risultati di ogni esame degli studenti

- **Utilizzando rispettivamente le seguenti viste**
  - **cluster_career** contenente gli attributi *student_id*, *test_grade*, *diploma_grade*, *grade_weighted_avg*, *exams_taken*, *total_cfu*, *years*

student_id	test_grade	hs_diploma_grade	grade_weighted_avg	exams_taken	total_cfu	years
A	18	80	27.0	10	90	4
B	13	67	23.0	10	96	4
C	18	78	25.0	7	69	4
D	14	66	23.0	7	66	2
E	16	82	28.0	2	24	2

Tipi di analisi e viste

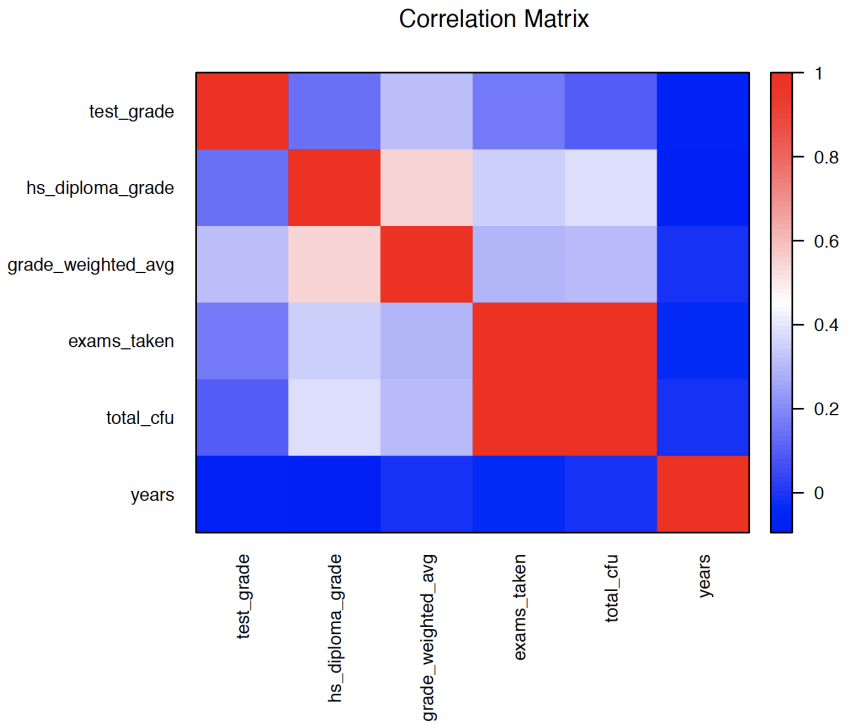
- **cluster_exams** contenente per ogni studente i voti ottenuti ad ogni esame

student_id	B006800	B006801	B006802	B006803	B006804	B006807
A	29	30.0	25.0	25.9	30.0	24.0
B	26	20.0	21.0	18.0	26.0	22.0
C	28	24.7142	22.0	24.7142	26.0	24.7142
D	20	28.0	22.0	23.1428	22.0	21.0
E	28	27.0	27.5	27.5	27.5	27.5

Voti mancanti nella vista *cluster_exams* integrati con la media dei voti dello studente

Data understanding

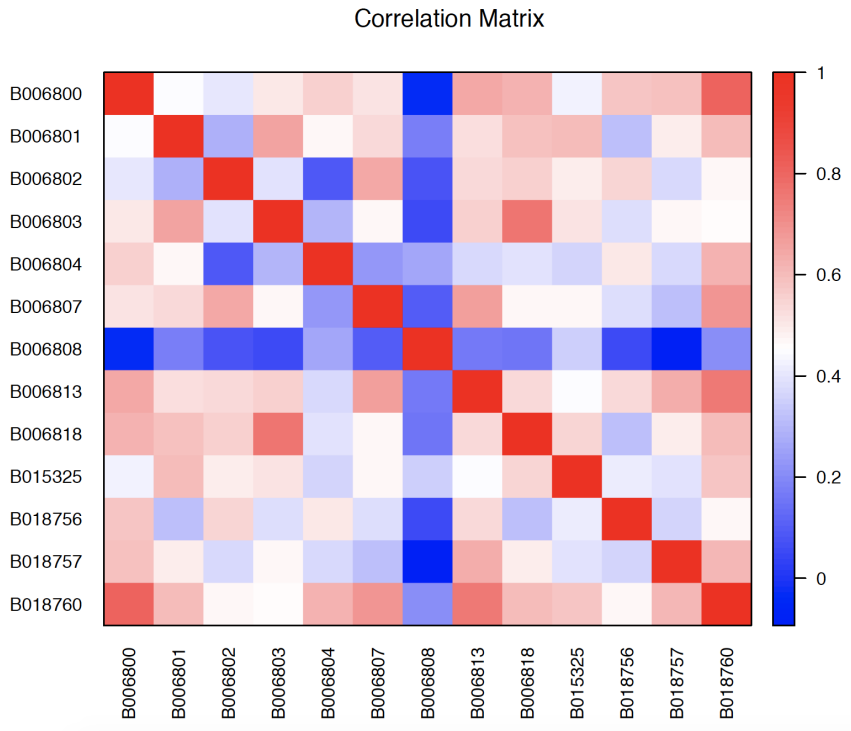
Correlazione sulla vista *cluster_carrer*:



- correlazione di circa 0.545 tra la media dei voti degli esami ed il voto di diploma
- year risulta quasi totalmente non correlato al resto degli attributi

Data understanding

Correlazione sulla vista *cluster_exams*:



- ASD (B006800) CPS (B018760)
- SO (B006818) MDL (B006803)
- BSDI (B006813) CPS (B018760)

Fase preliminare clustering

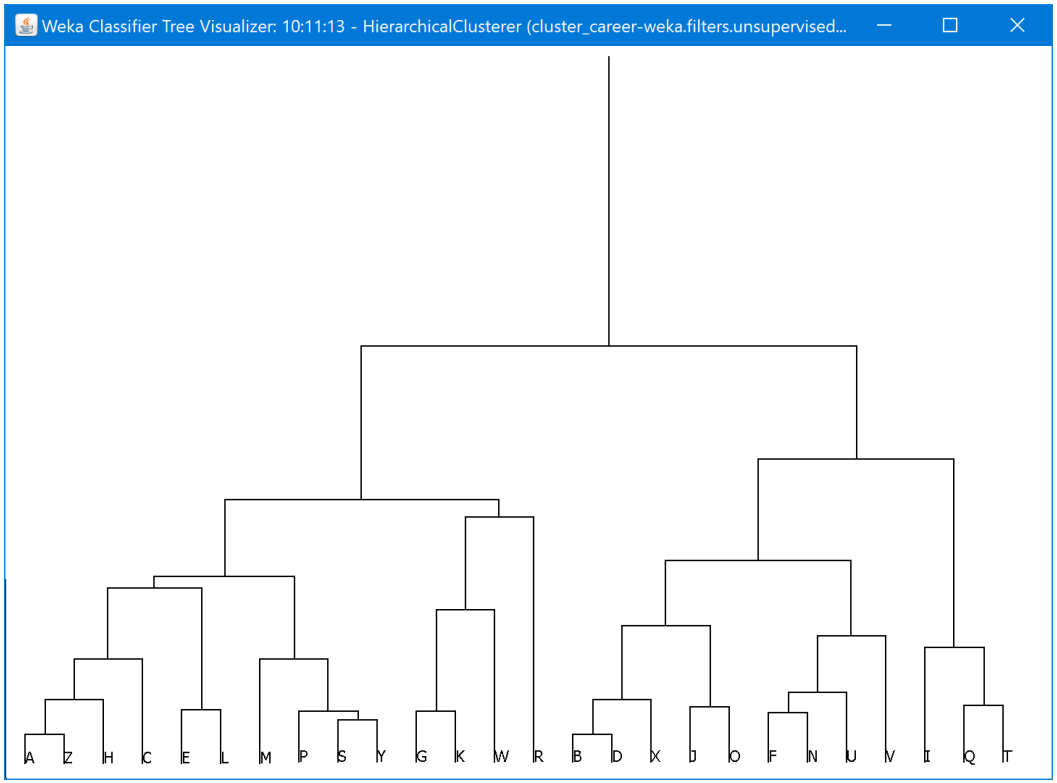
- Collegato Weka al database MySQL
- Normalizzati attributi *cluster_career* in scala 0 - 1
- Attributi cluster_exams voti in scala 18 - 31

Clustering *cluster_career*

- Per indagare il numero di cluster nascosti:
  - **Clustering gerarchico**, eseguito sui tre attributi più correlati: *test_grade, hs_diploma_grade, grade_weighted_avg*
    1. **Complete Link**
    2. **Group Average**
  - **DB-Scan**

Clustering gerarchico Complete Link

Dendogramma risultante:



Dataset diviso in due principali cluster

Clustering gerarchico Group Average e DB-Scan

1. **Gerarchico** con metodo **Group Avarage** risultati simili al metodo **Complete Link**
2. **DB-Scan** non ha restituito risultati soddisfacenti e significativi probabilmente per questioni legate alla dimensioni DataSet ed alla poca densità degli elementi

Applicazione algoritmo K-means

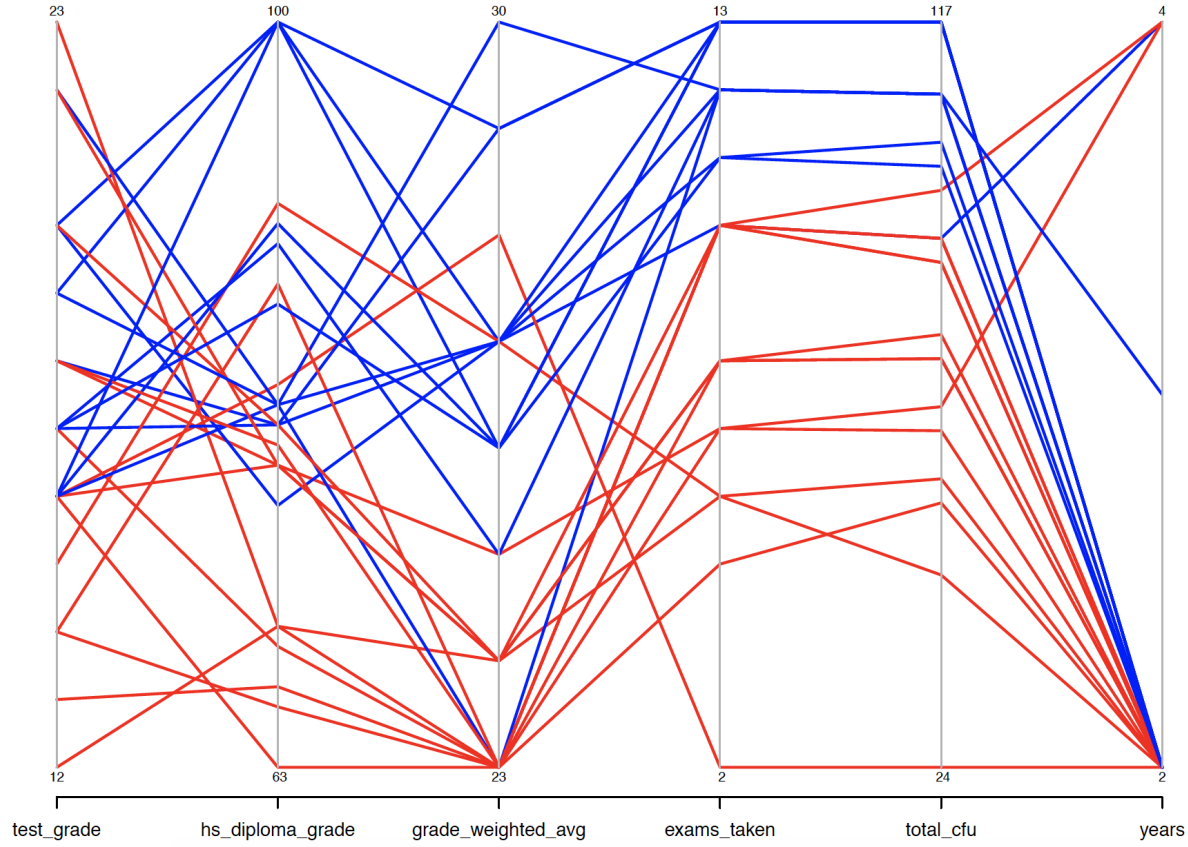
- Due esecuzioni dell'algoritmo sulla vista *cluster_career*:
  1. Considerando solo i primi tre attributi della vista (*test_grade,hs_diploma_grade, grade_weighted_avg*)
  2. Considerando tutti gli attributi della vista *cluster_career* escludendo solo lo *student_id*

Utilizzando in entrambi la distanza Euclidea, specificando due cluster da ricercare e lasciando i valori di default per la generazione casuale dei centroidi

Applicazione ed analisi algoritmo K-means

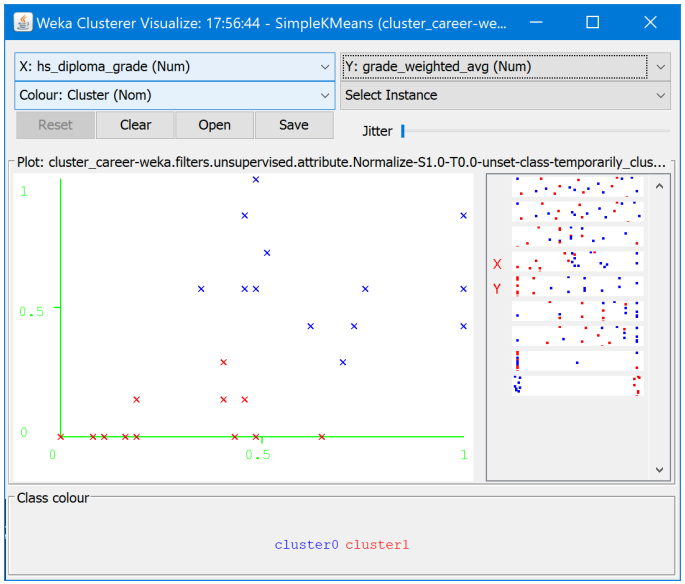
In entrambe le configurazioni di dati, gli studenti vengono divisi in due categorie

- studenti con un miglior percorso indicati in blu
- studenti con un percorso peggiore indicati in rosso



Applicazione ed analisi algoritmo K-means

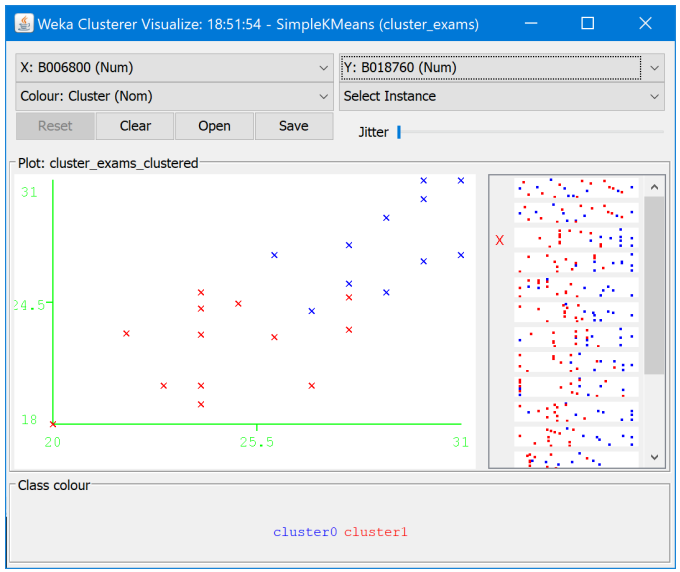
Plot di incrocio tra i due attributi più correlati nelle due esecuzioni dell'algoritmo



- Circa 50% degli studenti assegnata ad ogni cluster
- Differenza principale in alcuni studenti che vengono penalizzati nella seconda operazione di clustering su tutti gli attributi
- SSE prima esecuzione 3.19, seconda esecuzione 8.37

Clustering sugli esami (vista *cluster_exams*)

- Due Cluster
- 50% degli studenti assegnata ad ogni cluster
- SSE 17.96



- Divisi in studenti con sequenza di voti più alta e più bassa
- Assegnamenti ai cluster simili a quelli ottenuti per il clustering sulla carriera degli studenti (vista *cluster_career*)

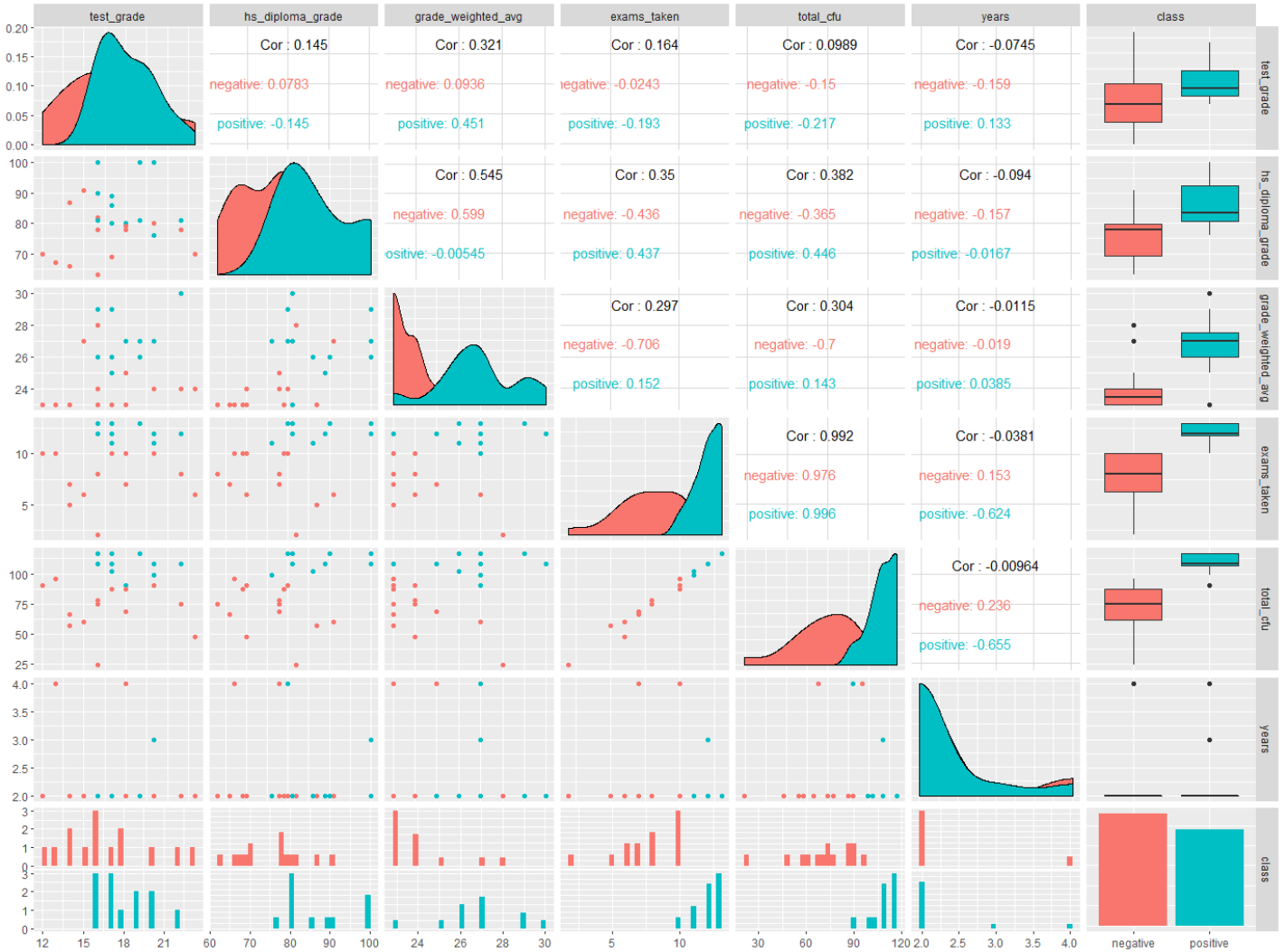
Classificazione degli studenti

- Cercare di prevedere la **classe dello studente:**
  1. *"Positiva"*: carriera soddisfacente
  2. *"Negativa"*: carriera non soddisfacente

- Classi del training set scelte in base al risultato del clustering sulla carriera (vista *cluster_career*)
- Assegnata ad ogni studente una classe
  - *"positive"*: cluster 0
  - *"negative"*: cluster 1
- Previsione basata sugli attributi della vista *cluster_creer* (*voto al test, voto diploma, media dei voti degli esami, numero esami sostenuti, numero CFU acquisiti, years*)

Caratteristiche training set

Grafico riassuntivo delle principali caratteristiche del training set



Classificazione con algoritmo j48

- Importati i dati in Weka e normalizzati gli attributi
- Metodo *Cross-Validation*

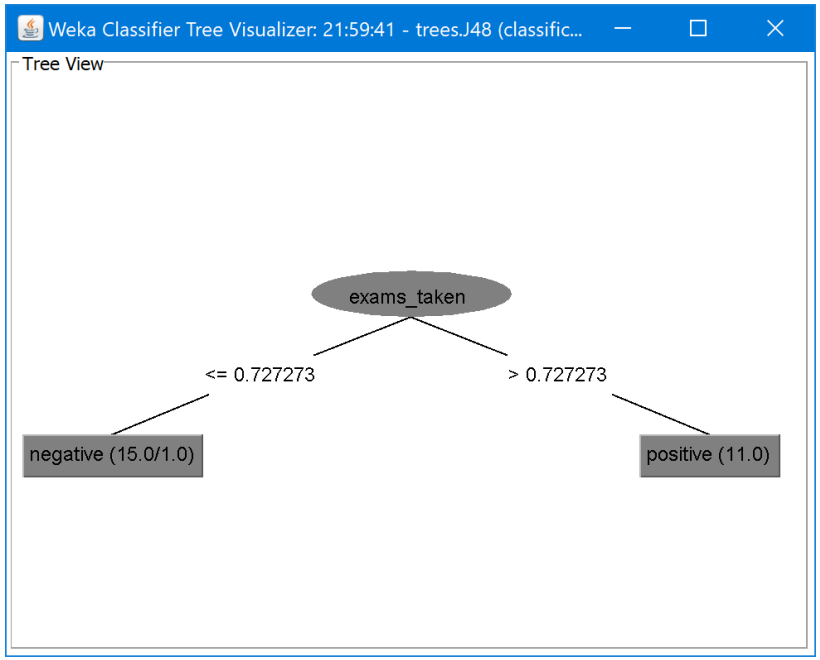
- Matrice di confusione:

		Predicted Class	
Actual Class		*Class*=**Positive**	*Class*=**Negative**
	Class=**Positive**	11	1
	Class=**Negative**	0	14

  - Accuratezza del 96.15%
  - Un solo studente classificato in modo errato

Classificazione con algoritmo j48

Albero di decisione ottenuto:



- Albero di decisione ad un solo livello non molto interessante con split su numero di esami sostenuti
- Risultati simili anche con altri metodi per la generazione del test set
- Risultati penalizzati dalla dimensione ridotta dal dataset e da dati mancanti

Classificazione conclusiva con algoritmo j48

Capire se le precedenti analisi aiutino a prevedere se lo studente riuscirà ad ottenere una carriera soddisfacente o meno

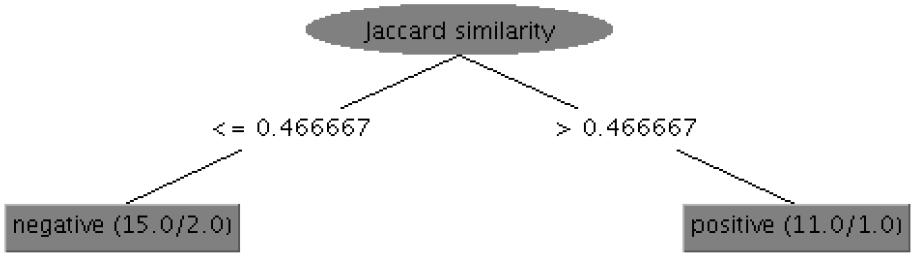
- Cercare di prevedere la **classe dello studente** basandosi su
  - somma pattern frequenti
  - coefficiente di Jaccard
  - media dei voti agli esami (normalizzata 0 - 1)

- Stessi assegnamenti di classe precedenti
- Metodo *Cross-Validation*

Classificazione conclusiva con algoritmo j48

- Accuratezza del 73%
- Albero di decisione ad un solo livello non molto interessante con split sul coefficiente di Jaccard

		Predicted Class	
Actual Class		*Class*=**Positive**	*Class*=**Negative**
	Class=**Positive**	8	4
	Class=**Negative**	3	11



Conclusioni

- La divisione del DataSet in cluster rispetta il grafico dei coefficienti
  - Studenti della classe *"positive"* hanno mediamente valori di ω, Jaccard e media dei voti agli esami superiore allo 0.5
  - Studenti con pattern sequenziali simili allo studente modello mediamente avranno una carriera positiva