UNIVERSITÀ DEGLI STUDI DI FIRENZE

CURRICULUM DATA SCIENCE

DATA MINING AND ORGANISATION

Analisi carriere studenti iscritti al I anno del corso di laurea in Informatica

Studenti
TOMMASO CECCARINI
tommaso.ceccarini1@stud.unifi.it
FILIPPO MAMELI
filippo.mameli@stud.unifi.it

10 luglio 2018

Indice

1 Descrizione dati iniziali

Il dataset che abbiamo analizzato contiene dati sulle carriere accademiche degli studenti del corso di laurea di informatica dell'università degli studi di Firenze e il loro voto conseguito al test di ingresso. In particolare, le informazioni presenti sono:

- Coorte: Anno di immatricolazione
- Crediti totali: Numero crediti complessivi dello studente
- Crediti con voto: Numero di crediti assegnati allo studente per esami con votazione in trentesimi (tutti tranne Inglese)
- Voto medio: Media pesata dei voti degli esami sostenuti

In seguito ci sono sei coppie di attributi che indicano:

- Nome dell'esame
- Data in cui lo studente ha sostenuto l'esame

Gli esami sono Algoritmi e strutture dati (ASD), Architetture degli elaboratori (ARC), Programmazione (PRG), Analisi I (ANI), Matematica discreta e logica (MDL) e Inglese.

- Punteggio conseguito al test di ingresso.

Per quanto riguarda gli attributi relativi ai voti degli esami i valori che questi assumono sono: il voto conseguito dallo studente nel caso in cui abbia sostenuto l'esame oppure zero se lo studente non ha sostenuto l'esame durante la sessione estiva del proprio anno di immatricolazione.

2 Gestione dei dati

Abbiamo deciso di effetture le seguenti operazioni sul dataset:

- eliminazione degli studenti che hanno sostenuto solo inglese
- riportare tutti gli attributi relativi alle date degli esami nel formato YYYY-MM-DD

Il motivo per cui abbiamo deciso di eseguire l'operazione del primo punto è che gli studenti in questione non avendo voti in trentesimi non ci sono parsi particolarmente significativi per analisi.

Per quanto riguarda il secondo punto invece, nel caso in cui uno studente non avesse sostenuto un particolare esame, erano presenti valori pari a zero in alcuni casi e valori pari a "0000-00-00" in altri. Abbiamo quindi deciso di trattare questa incosisistenza dei dati ponendo i valori pari a "0000-00-00".

Per effetturare queste due operazioni abbiamo importato il dataset in un database creando per prima una tabella, che è stata popolata in seguito importando il file CSV con i comandi riportati in Codice 1.

```
CREATE TABLE 'studenti' (
  'coorte' int(11),
  'crediti_totali' int(11),
  'crediti_con_voto' int(11),
  'voto_medio' int(11),
  'ASD' int(11),
  'data_ASD' text,
  'ARC' int(11),
  'data_ARC' text,
  'PRG' int(11),
  'data_PRG' text,
  'ANI' int(11),
  'data_ANI' text,
  'MDL' int(11),
  'data_MDL' text,
  'INGLESE' int(11),
  'data_INGLESE' text,
  'TEST' int(11)
) ENGINE=InnoDB

LOAD DATA INFILE 'studenti.csv' INTO TABLE studenti
  FIELDS TERMINATED BY ',' ENCLOSED BY '"'
  LINES TERMINATED BY '\r\n'
  IGNORE 1 LINES;
```

Codice 1: Creazione della table

Per quanto riguarda la gestione dell incosisistenza relativa alle date degli esami abbiamo risolto il problema specificando le query ripotate nel Codice 2.

```
update dmo.studenti set data_ARC = '0000-00-00' where data_ARC='0';
update dmo.studenti set data_ASD = '0000-00-00' where data_ASD='0';
update dmo.studenti set data_PRG = '0000-00-00' where data_PRG='0';
update dmo.studenti set data_ANI = '0000-00-00' where data_ANI='0';
update dmo.studenti set data_MDL = '0000-00-00' where data_MDL='0';
update dmo.studenti set data_INGLESE = '0000-00-00' where
    data_INGLESE = '0';
```

Codice 2: Update della tabella

3 Analisi dei dati

Per analizzare i dati a disposizione abbiamo utilizzato il linguaggio R per determinare la correlazione tra i diversi attributi. In Tabella 1 sono riportati i valori di correlazione.

	coorte	crediti totali	crediti con voto	voto medio	ASD	ARC	PRG	ANI	MDL	INGLESE	TEST
coorte	1	0.013343	0.01821	0.03655	0.03581	-0.01609	-0.0822	0.13386	-0.04033	NA	0.04126
crediti_totali	0.01334	1	0.99522	0.44571	0.52984	0.72508	0.69882	0.61015	0.62789	NA	0.38433
crediti_con_voto	0.01821	0.99522	1	0.44838	0.52957	0.71955	0.70879	0.61593	0.62654	NA	0.39025
voto_medio	0.03655	0.44571	0.44838	1	0.36900	0.36427	0.43085	0.39777	0.31828	NA	0.39428
ASD	0.03581	0.52984	0.52957	0.36900	1	0.29321	0.31192	0.10116	0.23775	NA	0.16149
ARC	-0.0160	0.72508	0.71955	0.36427	0.29321	1	0.43166	0.27541	0.39622	NA	0.29979
PRG	-0.0822	0.69882	0.70879	0.43085	0.31192	0.43166	1	0.19585	0.27295	NA	0.24356
ANI	0.13386	0.61015	0.61593	0.39777	0.10116	0.27541	0.19585	1	0.36333	NA	0.32378
MDL	-0.0403	0.62789	0.62654	0.31828	0.23775	0.39622	0.27295	0.36333	1	NA	0.38777
INGLESE	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
TEST	0.04126	0.384332	0.39025	0.39428	0.16149	0.29979	0.2435	0.32378	0.38777	NA	1

Tabella 1: Correlazione

Come è possibile notare dalla tabella, ci sono alcune ovvie correlazioni, come quella che coinvolge l'attributo crediti totali e crediti con voto. Per quanto riguarda l'attributo coorte si può notare come questo abbia una scarsa correlazione con tutti gli altri attributi.

Le correlazioni più evidenti e significative sono quelle che coinvolgono l' attributo crediti totali (e quindi anche crediti con voto) con il voto che gli studenti hanno ottenuto nei diversi esami sostenuti. Infatti il valore

di queste correlazioni è compreso tra 0.53 per quanto riguarda l'attributo di crediti totali e quello di Algoritmi e strutture dati e 0.73 per quanto riguarda Crediti totali e il voto di Architetture degli elaboratori. Tale risultato può essere interpretato in maniera abbastanza intuitiva, infatti mostra che se uno studente ha sostenuto più esami (e quindi ha più crediti) ha in generale ottenuto dei voti migliori.

Nelle figure sono riportati gli scatterplot degli attributi relativi a crediti totali con i voti di Architetture degli elaboratori e Algoritmi e strutture dati.

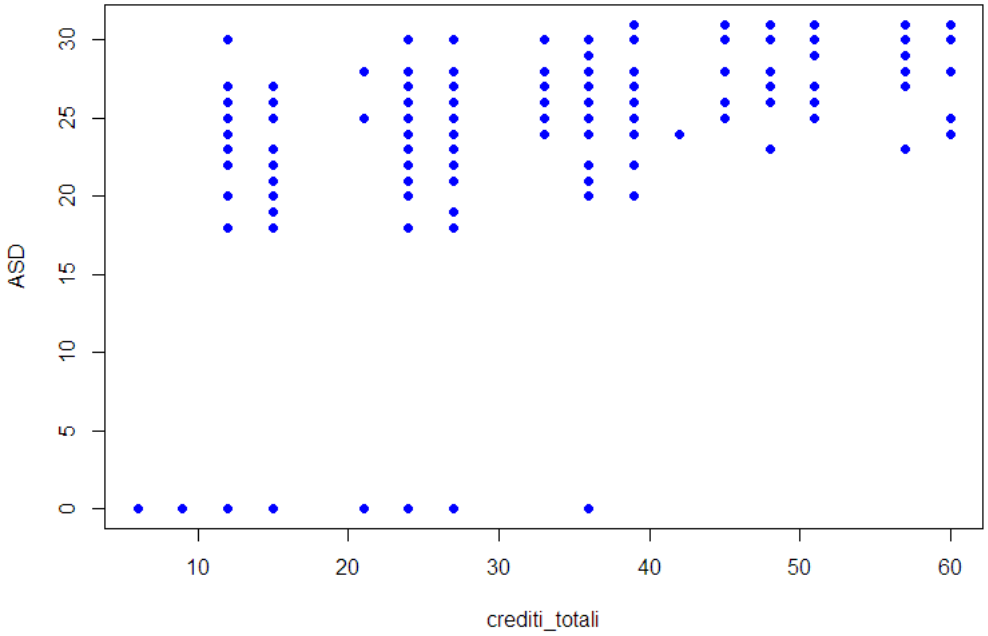


Figura 1: Scatterplot tra Crediti totali e Algoritmi e strutture dati

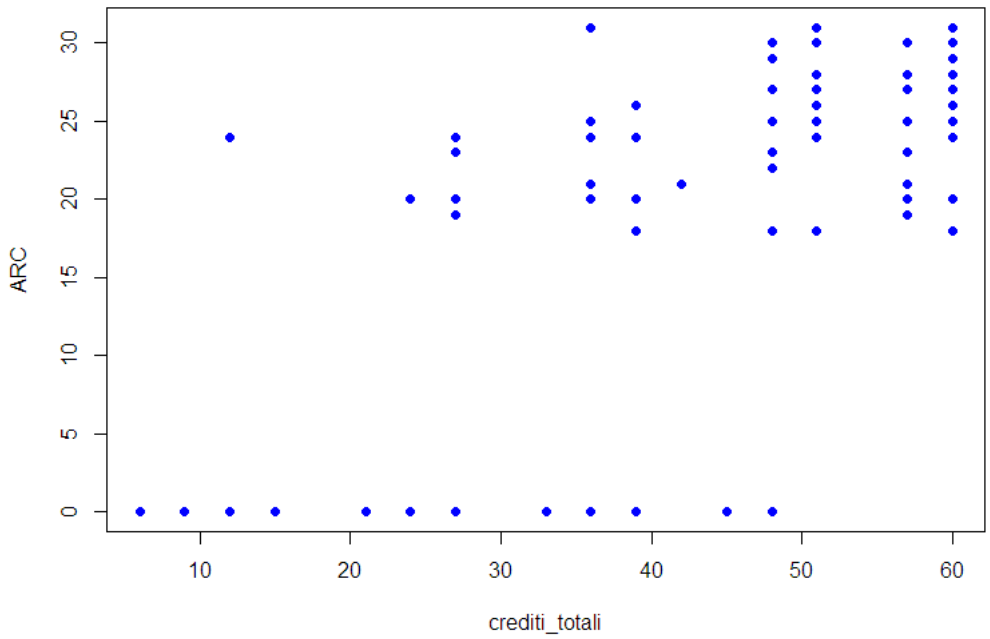
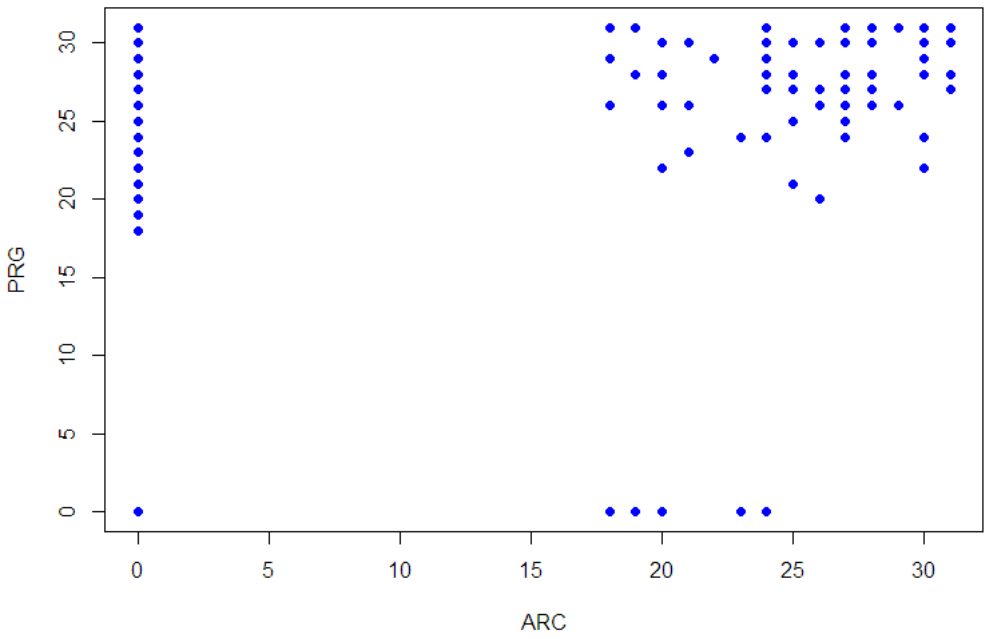


Figura 2: Scatterplot tra Crediti totali e Architetture degli elaboratori

L'attributo test è correlato maggiormente con l'attributo voto medio. Si può notare inoltre che il valore di correlazione è significativamente più alto di quello che l'attributo test ha con i voti dei singoli esami, con l'unica eccezione di MDL. Questi valori lasciano suggerire che il punteggio conseguito al test d'ingresso di ogni studente sia un buon parametro per valutarne l'andamento generale piuttosto che i voti di un singolo esame. Inoltre l'attributo test mostra una discreta correlazione con i crediti totali (e quindi anche con i crediti con voto).

Per quanto riguarda le correlazioni tra i voti dei diversi esami è possibile notare che il valore più alto è quello tra Architetture degli elaboratori e Programmazione che è circa pari a 0.43. Mentre si ha una correlazione quasi nulla, ossia circa 0.10, tra il voto di Algoritmi e strutture dati e il voto di Analisi 1.

Nelle figure sono riportati gli scatterplot tra Architetture degli elaboratori e Programmazione e tra Algoritmi e strutture dati e Analisi I.



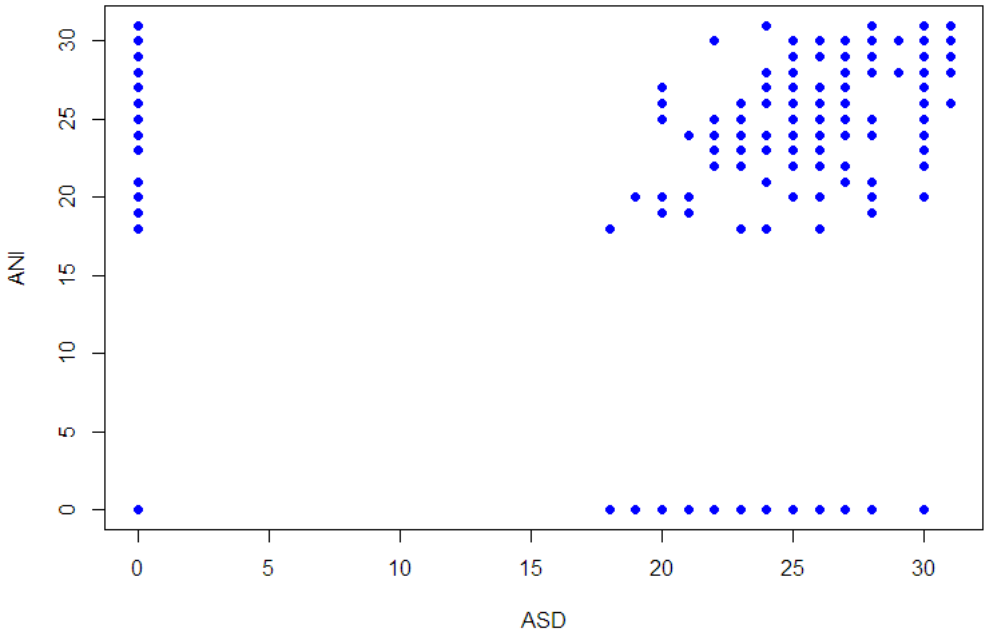


Figura 3: Scatterplot tra Architetture degli elaboratori e Programmazione e tra Algoritmi e strutture dati e Analisi I

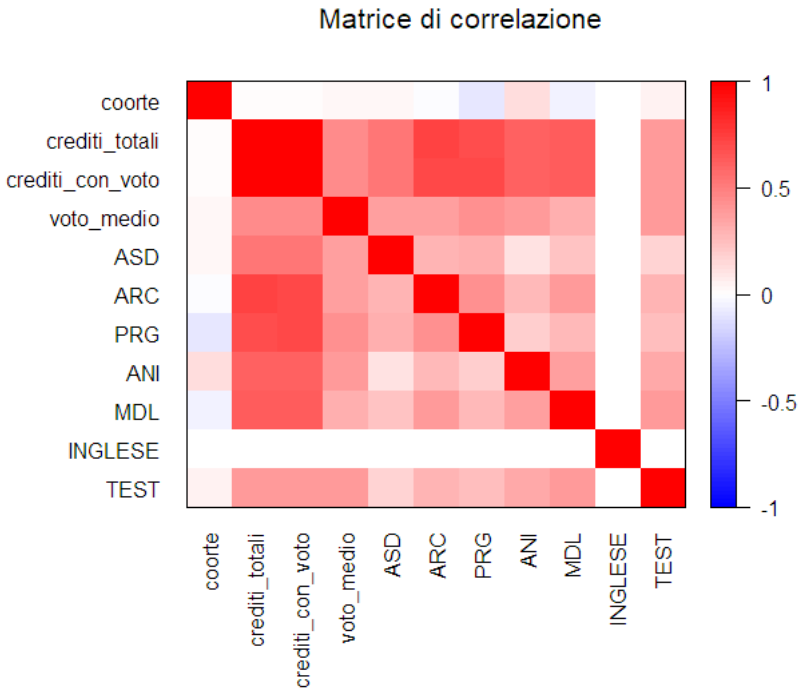


Figura 4: Matrice di correlazione

In Figura 4 viene riportata la matrice di correlazione. Per applicare gli algoritmi di clustering è stato utilizzato il software Weka. Tuttavia prima di applicare tali algoritmi è stato necessaria un ulteriore fase di preprocessing nella quale sono stati normalizzati tutti gli attributi del dataset in una scala di valori compresi tra zero e uno in modo da evitare problemi dovuto alle diverse scale di valori dei attributi.

4 Clustering

In questo Capitolo verranno effettuate alcune analisi di clustering utilizzando alcuni degli algoritmi piu' noti per risolvere questo problema. In particolare, nel caso dell'algoritmo di Kmeans viene stabilito preventivamente il numero dei cluster possibili utilizzando valori ritenuti sensati di volta in volta. Per la valutazione del clustering ottenuto si rimanda al

Capitolo 5 in cui viene analizzata la validità dei risultati ottenuti e viene determinato il numero di cluster ottimali per ogni analisi effettuata. Come prima analisi mostriamo quella relativa ai tre attributi del dataset maggiormente correlati (a meno di correlazioni ovvie) ossia crediti totali, Architetture e Programmazione. Infatti come è stato già precedentemente detto c'è una forte correlazione tra i crediti totali e architetture pari a 0.73 mentre tra crediti totali e Programmazione pari a 0.70.

È stato utilizzato l'algoritmo di Kmeans implementato in Weka specificando inizialmente un numero di cluster pari a due, lasciando i valori di default per la generazione dei centroidi.

In Tabella 2 sono riportate le coordinate dei centroidi relativi all'esecuzione dell'algoritmo con un valore di k pari a 2.

	Crediti totali	ARC	PRG	Istanze
0	0.65	0.32	0.85	183 (58%)
1	0.27	0.05	0	133 (42%)

Tabella 2: Cluster con ARC e PRG con k = 2 SSE 51.35

Come si può notare dalle coordinate dei centroidi questa prima esecuzione dell'algoritmo suddivide il dataset in due gruppi piuttosto distinti. Il valore della somma degli errori al quadrato(SSE) in questa esecuzione è risultata pari a 51.35

In Tabella 3 sono riportate le coordinate dei centroidi relativi all'esecuzione dell'algoritmo con un valore di k pari a 3.

	Crediti totali	ARC	PRG	Istanze
0	0.88	0.82	0.89	73 (23%)
1	0.27	0.05	0	133 (42%)
2	0.50	0	0.81	110 (35%)

Tabella 3: Cluster con ARC e PRG con k = 3 SSE 14.85

In questo caso il valore di SSE è pari a 14.85. È inoltre possibile constatare come l'algoritmo di kmeans abbia messo in evidenza tre tipologie ben distinte di studenti:

- Gli studenti appartenenti al cluster 0 sono gli studenti "migliori" avendo sostenuto la quasi totalità degli esami del primo anno alla fine della sessione estiva e riportando delle ottime valutazioni per quanto riguarda gli esami di Architetture e di Programmazione;
- La seconda categoria di studenti (cluster 1) sono gli studenti "peggiori" che hanno sostenuto pochi esami e nel caso specifico delle materie considerate hanno conseguito valutazioni basse o non hanno sostenuto l'esame;
- Infine gli studenti appartenenti all'ultimo cluster sono gli studenti che hanno sostenuto Programmazione con un buon voto ma non hanno fatto l'esame di Architetture.

Come si evince andando ad analizzare i tre cluster in particolare notando le diverse combinazioni dei valori assunti dai centroidi dei voti, si capisce come sia assente la categoria di studenti che ha sostenuto con profitto l'esame di Architetture, ma non ha sostenuto l'esame di Programmazione, lasciando quindi intendere che se uno studente ha sostenuto l'esame di Architetture allora generalmente ha sostenuto con una buona valutazione l'esame di Programmazione.

La seconda analisi che è stata condotta riguarda gli attributi TEST e voto_medio. E' stato scelto di analizzare questi due attributi congiuntamente in quanto, come è stato detto nel capitolo precedente, l'attributo voto_medio presenta una buona correlazione con l'attributo TEST.

In Tabella 4 sono riportate le coordinate dei centroidi relativi all'esecuzione dell'algoritmo con un numero di cluster pari a 3.

	voto medio	Test	Istanze
0	0.36	0.41	85 (27%)
1	0.75	0.66	146 (42%)
2	0.45	0.67	85 (27%)

Tabella 4: Cluster con Voto_medio e Test con k = 3 SSE 9.6

In questo caso è possibile notare come l'algoritmo di k-means determini due cluster ben definiti che suddividono il dataset tra gli studenti che hanno una media complessiva maggiore e un voto al test d'ingresso alto

e quelli che invece hanno una media più bassa e hanno conseguito punteggio basso al test di ingresso. Questi gruppi determinati sono coerenti con la correlazione che esiste tra i due attributi che tuttavia non è particolarmente elevata (diversamente dagli attributi presi in considerazione nell'analisi precedente). Infatti, oltre ai primi due cluster che identificano gli studenti "migliori" e quelli "peggiori" esiste un terzo cluster di studenti che hanno conseguito un punteggio al test d'ingresso decisamente positivo, ma non hanno mantenuto una media dei voti altrettanto buona. SSE del clustering in questo caso è 9.6.

Infine l'ultima analisi che abbiamo deciso di condurre è quella relativa a tutti i voti conseguiti dagli studenti del primo anno durante la sessione estiva. In questo caso l'algoritmo di clustering è stato inizializzato con un valore di k=3. In Tabella 5 sono riportate le coordinate dei centroidi al termine dell'algoritmo.

	ASD	ARC	PRG	ANI	MDL	Istanze
0	0.73	0.05	0.81	0.43	0.02	100 (32%)
1	0.65	0.05	0	0.50	0.12	133 (42%)
2	0.91	0.65	0.89	0.88	0.60	83 (26%)

Tabella 5: Cluster di tutti i voti con k = 3 SSE 106.19

In questo caso sono stati determinati i profili di tre diversi gruppi di studenti:

- gli studenti che hanno conseguito una buona votazione negli esami di Algoritmi e Strutture Dati e Programmazione, una votazione discreta all'esame di Analisi I e che non hanno sostenuto Matematica discreta e Logica e Architetture degli elaboratori;
- gli studenti con le stesse caratteristiche del cluster precedente, ma che hanno sostenuto Programmazione
- gli studenti che hanno sostenuto tutti gli esami e con un buona votazione.

In Figura 5 è riportato lo scatter plot relativo ai voti di Architetture degli Elaboratori e Programmazione che sono maggiormente correlati. Il valore del SSE in questo caso è pari a 106.19.

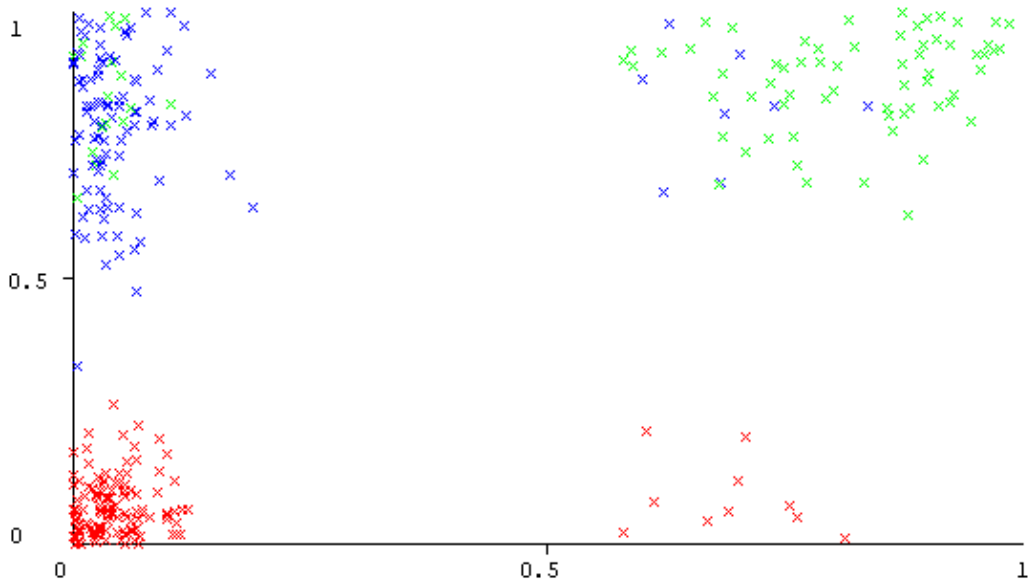


Figura 5: Scatter plot relativo ai cluster dei voti di Architetture degli Elaboratori e Programmazione

5 Valutazione del clustering e model selection

Nel Capitolo 4 l'algoritmo di Kmeans è stato eseguito scegliendo preventivamente il numero di cluster possibili (tipicamente 2 o 3) basandosi esclusivamente sull'intuizione e quindi senza avere garanzie circa la bontà e correttezza dei risultati ottenuti. In questo Capitolo viene valutata la validità delle analisi di clustering effettuate e viene utilizzata una procedura basata sul SSE e che tenta di determinare il valore ottimale di k con cui inizializzare l'algoritmo di Kmeans in modo da migliorarne la validità. Per ciascuno degli aspetti inizializzati vengono quindi eseguite le seguenti operazioni:

1. determinazione dei valori del SSE in funzione di k;
2. scelta di k_{opt} come il più piccolo k tale per cui il valore di SSE "smette di diminuire";
3. confronto del valore di correlazione ottenuta tra la matrice di incidenza e quella delle distanze per i valori di $k = 2$,$k = 3$ e k_{opt}.

Per quanto riguarda il Punto 2 è necessario approssimare il valore di k_{opt} scegliendo, ad esempio, il primo valore di k per cui si verifica una variazione nel SSE minore di una quantità fissata ε. Per quanto concerne il Punto 3 si ha che un valore inferiore della correlazione (sperabilmente negativo) indica un miglior risultato di clustering poiché, idealmente, punti appartenenti allo stesso cluster (quindi con valore di incidenza 1) dovrebbero trovarsi a una distanza minore. Quindi, per i tre attributi crediti totali, architetture e programmazione su cui è stata fatta la prima analisi con l'algoritmo di Kmeans si ottiene il grafico $k - SSE$ riportato in Figura 6.

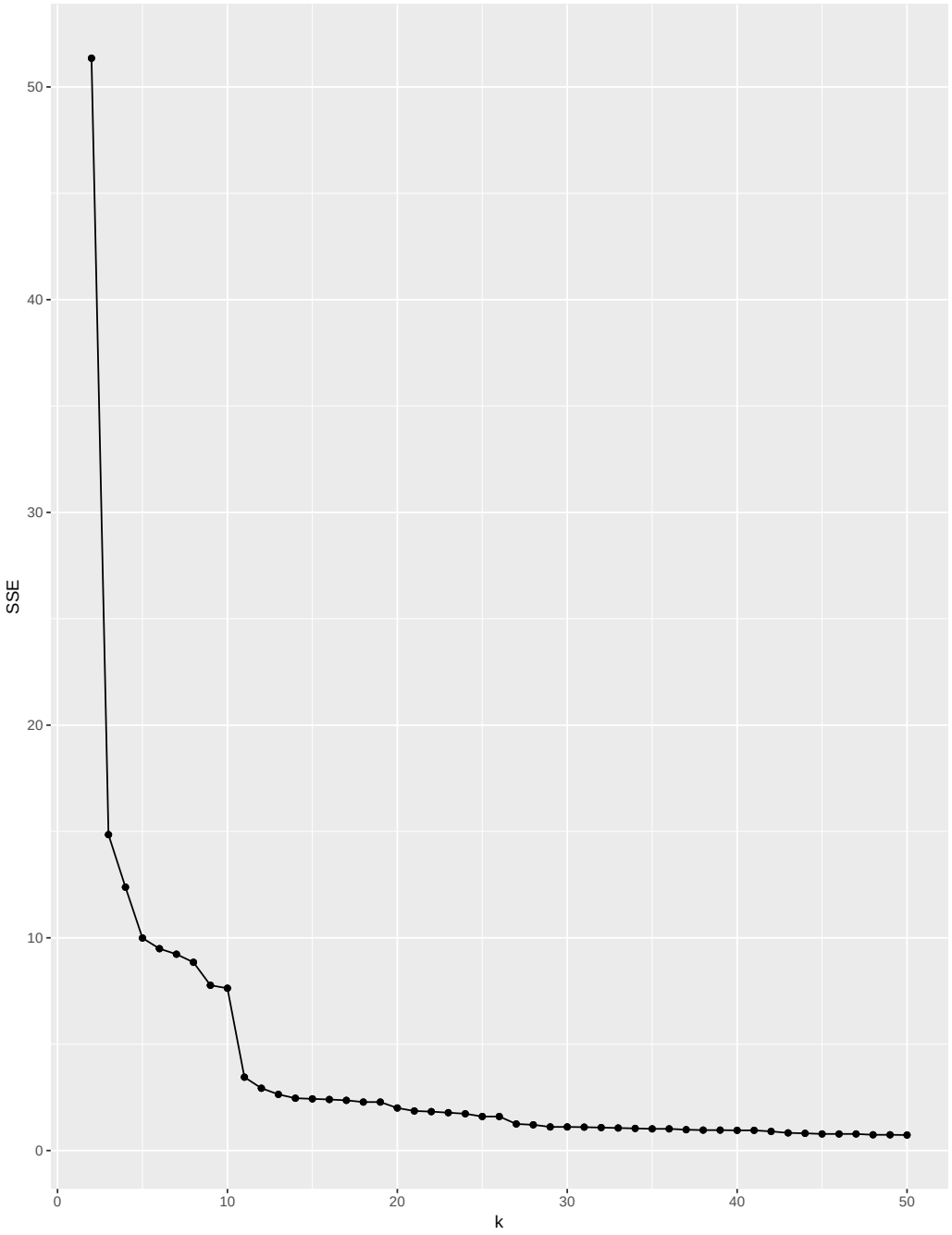


Figura 6: Andamento del valore del SSE in funzione del valore di k.

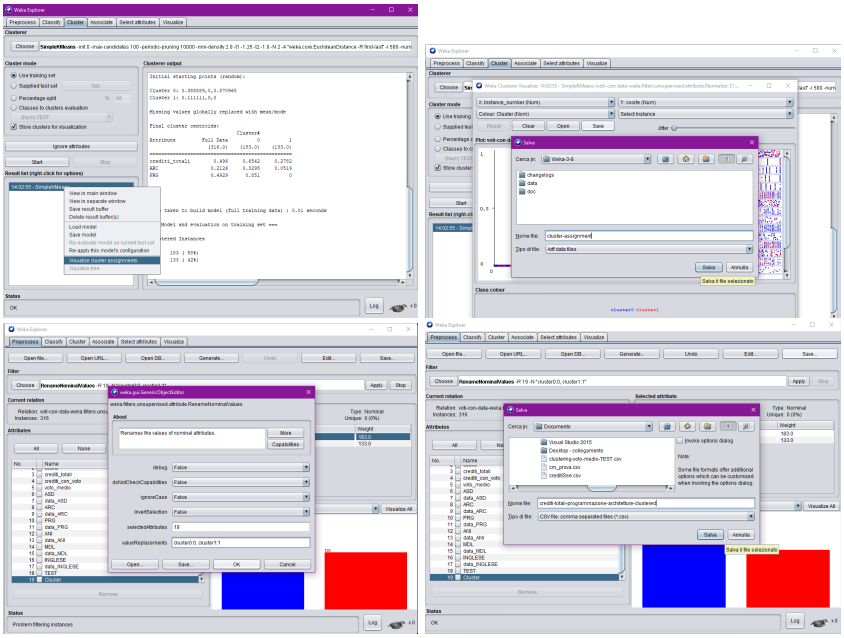


Figura 7: creazione ed esportazione del dataset con indicazione del cluster di appartenenza in Weka.

Scegliendo il valore minimo di $\varepsilon = 0.01$ (dato che è stato scelto di memorizzare i valori di SSE fino alla seconda cifra decimale) si ottiene un valore di $k_{opt} = 18$. Quindi, per determinare il valore di correlazione tra matrice di incidenza e matrice delle distanze è necessario esportare preventivamente da Weka il dataset munito di un attributo aggiuntivo che indichi il cluster di appartenenza di ciascun record. In Figura 7 viene mostrato come creare ed esportare un dataset con Weka aggiungendo per ogni record il riferimento al cluster di appartenenza a seguito dell'esecuzione dell'algoritmo di Kmeans per gli attributi crediti totali, architetture e programmazione con $k = 2$.

Nel Codice 3 viene mostrato importare il dataset comprensivo degli attributi crediti totali, architetture, programmazione e cluster con R, mentre il Codice 4 calcola la Matrice di incidenza dei cluster, la Matrice delle distanze per i punti e, infine, la correlazione tra le due matrici. Come si evince dal Codice, per poter calcolare il valore della correlazione è stato necessario linearizzare preventivamente le due matrici utilizzando l'istruzione `as.vector` di `R`.

```
library(readr)
crediti_totali_prg_arc_clustered <- read_csv("git/
   clusteringStudentiInformatica/crediti_totali-prg-
   arc-clustered.csv",
    col_types = cols(ANI = col_skip(), ASD = col_skip
       (),
        INGLESE = col_skip(), Instance_number = col_
           skip(),
        MDL = col_skip(), TEST = col_skip(),
        coorte = col_skip(), crediti_con_voto = col_
           skip(),
        data_ANI = col_skip(), data_ARC = col_skip(),
        data_ASD = col_skip(), data_INGLESE = col_skip
           (),
        data_MDL = col_skip(), data_PRG = col_skip(),
        voto_medio = col_skip()))
View(crediti_totali_prg_arc_clustered)
```

Codice 3: Importazione degli attributi crediti totali, architetture, programmazione e cluster.

```
#Matrice di incidenza
C = matrix(nrow = 316,ncol = 316)
for(i in 1:316){
    for(j in 1:316){
        if(crediti_totali_programmazione_architetture_
           clustered[i,4]==crediti_totali_
           programmazione_architetture_clustered[j,4])
           {
            C[i,j] = 1
        }else{
            C[i,j] = 0
        }
    }
}
#Matrice distanza
D = as.matrix(dist(crediti_totali_programmazione_
   architetture_clustered[,1:3],method = 'euclidean',
   diag = TRUE,upper = TRUE))
```

```
c = as.vector(t(C))
d = as.vector(t(D))

cor(c,d,method="pearson")
```

Codice 4: Calcolo Matrice di incidenza dei cluster, delle distanze e correlazione tra le due matrici.

Il valore di correlazione ottenuto in questo caso è pari a -0.687. Ripetendo il clustering per gli stessi attributi con $k = 3$ e $k = 18$ si ottiene, rispettivamente un valore della correlazione tra le due matrici di -0.854 e -0.489. Diversamente da quanto atteso, il valore calcolato per k_{opt} non presenta un valore di correlazione inferiore rispetto agli altri due clustering, bensì risulta essere il valore peggiore tra quelli verificati con il metodo della correlazione.

6 Conclusioni

In riferimento ai risultati ottenuti con le tecniche di clustering è possibile trarre le seguenti conclusioni riguardo le carriere degli studenti:

- l'esame di Architetture degli Elaboratori risulta essere l'esame più difficile per gli studenti del primo anno, infatti la maggioranza non riesce a sostenerlo nel corso della sessione estiva. Tuttavia, generalmente gli studenti che sostengono con profitto tale esame riescono a sostenere con un buon voto anche gli altri;

- la maggior parte degli studenti che ottengono un buon punteggio al test d'ingresso mantengono una buona media mentre quelli che hanno ottenuto un punteggio più basso hanno anche una media più bassa. Tuttavia è presente un significativo gruppo di studenti che pur avendo ottenuto un buon punteggio al test di ingresso non riescono ad avere una media altrettanto buona;

- La maggior parte degli studenti riesce a sostenere nel corso della sessione estiva gli esami di Algoritmi e Strutture Dati, Analisi I e in alcuni casi l'esame di Programmazione con risultati altalenanti. Mentre

generalmente gli esami di Architetture degli Elaborati e Matematica Discreta e Logica non vengono sostenuti dagli studenti al termine del loro primo anno.

inoltre, in ciascuna delle analisi eseguite risultavano esserci almeno 3 gruppi di studenti che presentavano caratteristiche significativamente differenti.

Elenco delle figure

Elenco delle tabelle