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

Tommaso, Ceccarini
tommaso.ceccarini1@stud.unifi.it

Filippo, Mameli
filippo.mameli@stud.unifi.it

Giugno 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'universit 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 che gli studenti in questione non avendo voti in trentesimi non ci sono pari 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 database tramite l'operazione di import riportata nel Codice 1.

Codice 1: Creazione della table

```
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
```

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

Codice 2: Update della tabella

```
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';
```

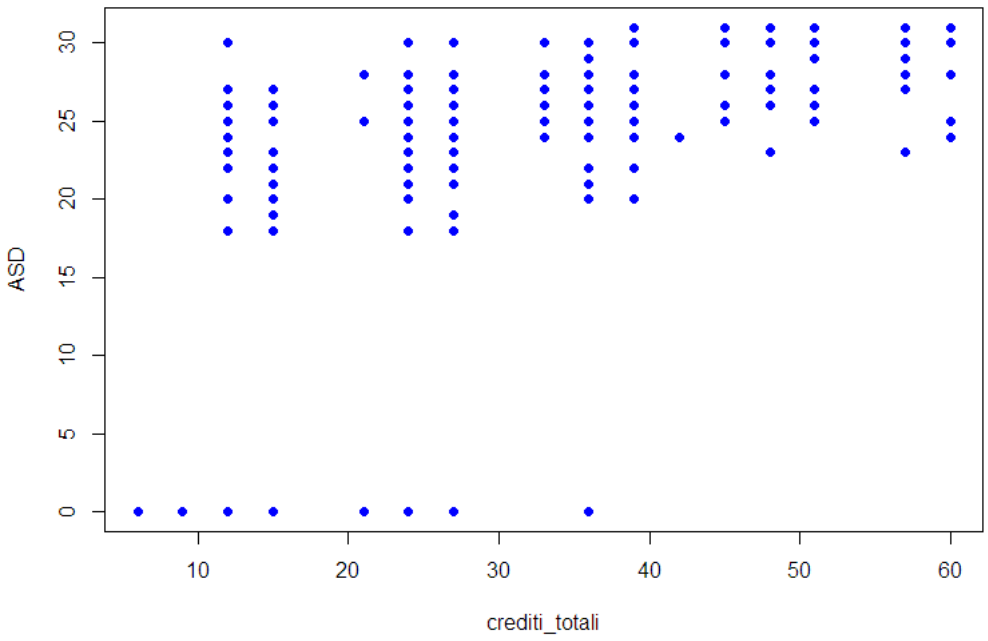
3 Analisi dei dati

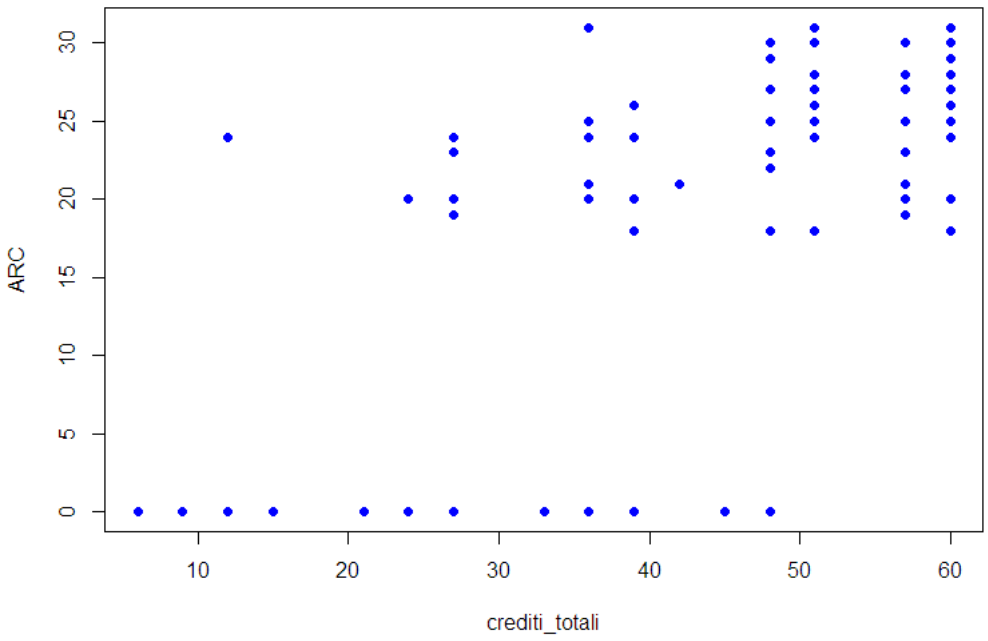
Per analizzare i dati a disposizione abbiamo utilizzato il linguaggio R per determinare la correlazione tra i diversi attributi. In Tabella 1 sono riportati i valori di correlazione.

Tabella 1: 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

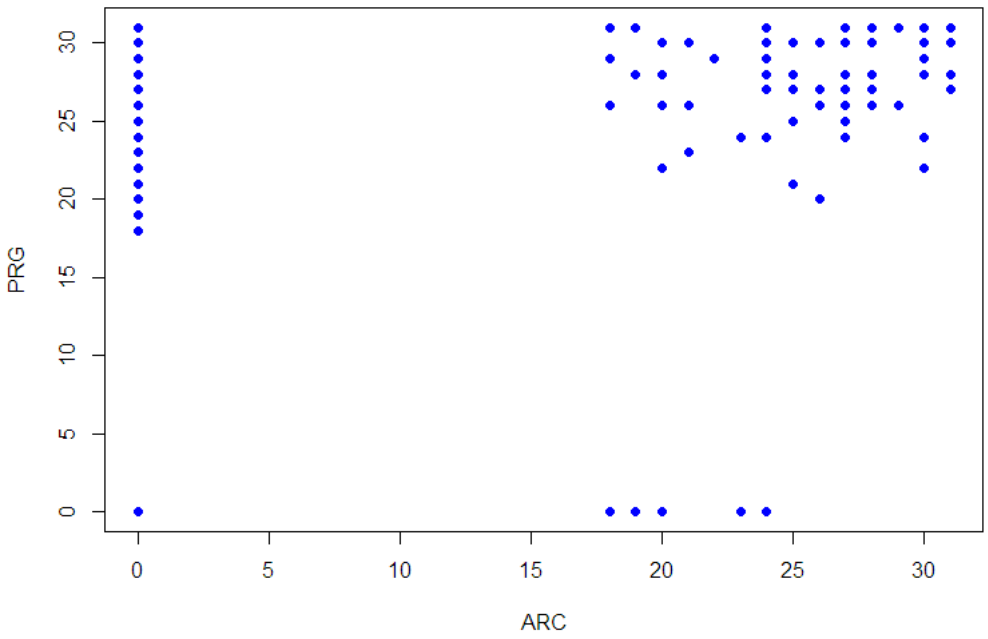
Come possibile notare dalla tabella, ci sono alcune ovvie correlazioni che coinvolgono gli attributi crediti totali e crediti con voto. Per quanto riguarda l'attributo corte si pu notare come questo abbia una scarsa correlazione con gli altri attributi. Le correlazioni pi evidenti e significative sono quelle che coinvolgono l' attributo crediti totali (e quindi anche crediti con voto) con il voto che gli studenti hanno ottenuto per i diversi esami sostenuti. Infatti il valore di queste correlazioni 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 pu essere interpretato in maniera abbastanza intuitiva, infatti mostra che se uno studente ha sostenuto pi esami(e quindi ha pi 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.


ASD
0
5
10
15
20
25
30
10
20
30
40
50
60
crediti_totali

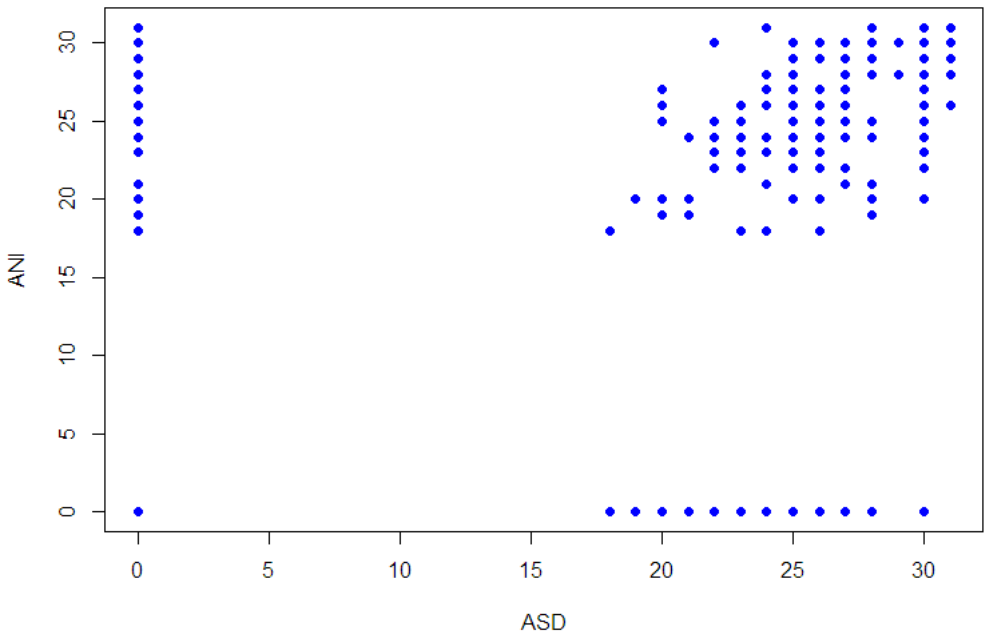


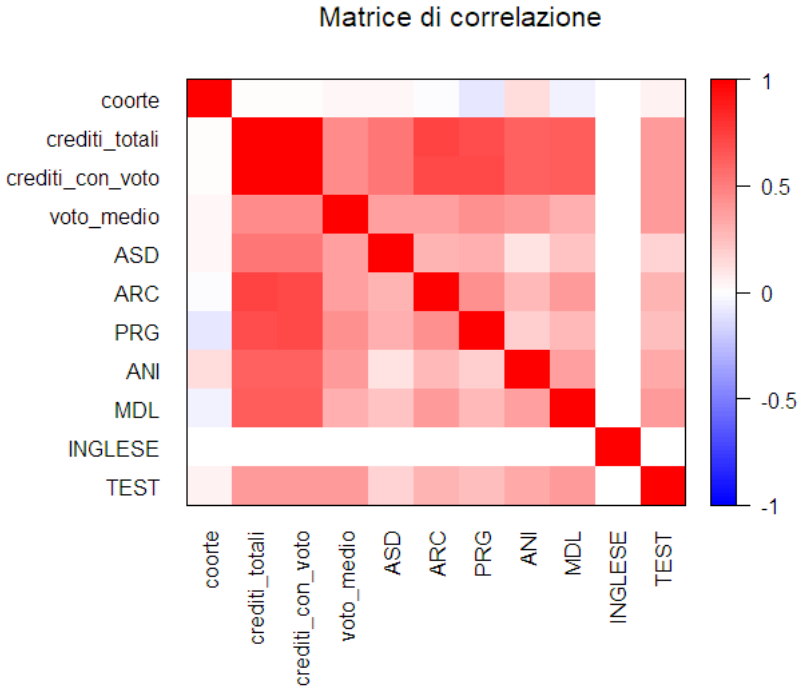
L'attributo test correlato maggiormente con l'attributo voto medio. Si pu notare inoltre che il valore di correlazione significativamente pi alto di quello che l'attributo test ha con i voti dei singoli esami, con l'unica eccezione di MDL. Questi valori lasciano suggerire che il punteggio conseguito 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 possibile notare che il valore pi alto quello tra Architetture degli elaboratori e Programmazione che circa pari a 0.43. Mentre invece 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 tra ARC e PRG e ASD e Analisi 1



In figura X viene riportata la matrice di correlazione.


ANI
ASD
0
5
10
15
20
25
30


Matrice di correlazione
coorte
crediti_totali
crediti_con_voto
voto_medio
ASD
ARC
PRG
ANI
MDL
INGLESE
TEST
1
0.5
0
-0.5
-1

In figura Y viene riportato lo scatter plot dei attributi del dataset.

Per applicare gli algoritmi di clustering stato utilizzato il software Weka. Tuttavia prima di applicare tali algoritmi stato necessiaria 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

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 stato gi precedentemente detto c' una forte correlazione tra i crediti totali e architetture pari a 0.73 mentre tra crediti totali e Programmazione pari a 0.70.

 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.

Tabella 2: Cluster k = 2 SSE 51.35

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

Come si pu notare dalle cordinate 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 risultata pari a 51.53

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

In questo caso il valore di SSE pari a 14.8. inoltre possibile constatare come l'algoritmo di kmean abbia messo in evidenza tre tipologie ben distinte di studenti:

- Gli studenti appartenenti al cluster 0 sono gli studenti "migliori" avendo sostenuto la quasi totalit degli esami del primo anno alla fine della

Tabella 3: Cluster k = 3 SSE 14.85

	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%)

sessione estiva e riportando delle ottime valutazioni per quanto riguarda gli esami di Architetture e 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.

Tabella 4: Cluster k = 2 SSE 12.4

	voto medio	Test	Istanze
0	0.38	0.49	134 (42%)
1	0.70	0.67	182 (58%)

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

Tabella 5: Cluster k = 3 SSE 9.6

	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%)

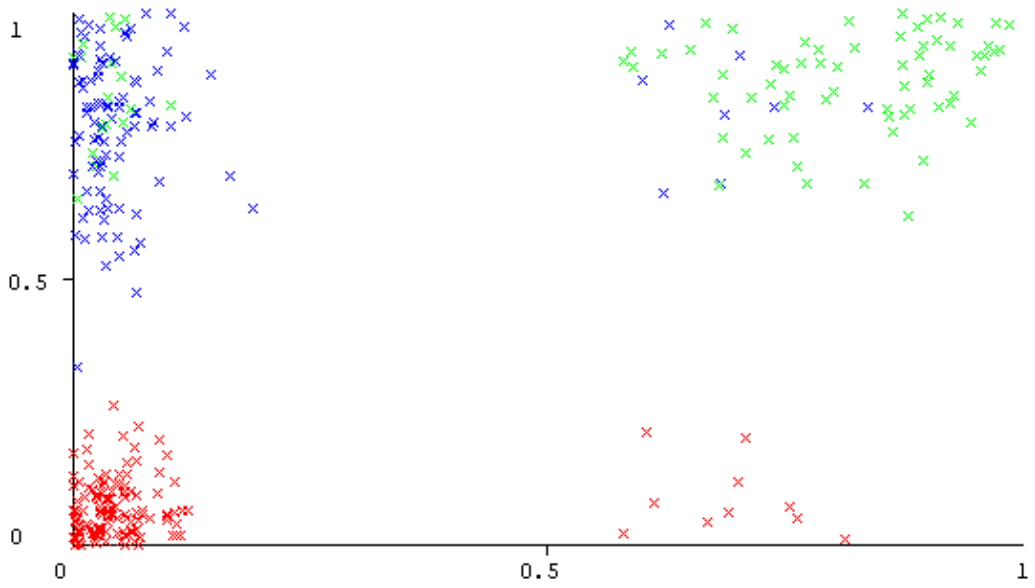
In Tabella 5 sono riportate le coordinate dei centroidi relativi all'esecuzione dell'algoritmo con un numero di cluster pari a 3.

In questo caso 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 migliore e quelli che invece hanno una media pi 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 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 in questo caso 9.6

Infine l'ultima analisi che abbiamo deciso di condurre quella relativa ai voti conseguiti dagli studenti del primo anno durante la sessione estiva. In questo caso l'algoritmo di clustering stato inizializzato con un valore di k=3. In Tabella W sono riportate le coordinate dei centroidi al termine dell'algoritmo. 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 in pi non hanno sostenuto Programmazione
- gli studenti che hanno sostenuto tutti gli esami e con un buona votazione.

In Figura 4 riportato lo scatter plot relativo ai voti di Architetture degli Elaboratori e Programmazione che sono maggiormente correlati. Il valore del SSE in questo caso pari a 106.19.



5 Conclusioni

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

- l'esame di Architetture degli Elaboratori risulta essere l'esame pi 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 profitto 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 pi basso hanno anche una media pi bassa. Tuttavia presente un significato 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