

# Clustering studenti informatica

---

Tommaso Ceccarini, Filippo Mameli

21 agosto 2018

# Introduzione

---

## Il dataset

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.

- 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

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

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

- Punteggio conseguito al test di ingresso.

# Gestione dei dati

---

Le principali operazioni effettuate sul dataset sono:

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

Le principali operazioni effettuate sul dataset sono:

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

# Creazione 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,  
  ...  
  '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;
```



# Update 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';
```

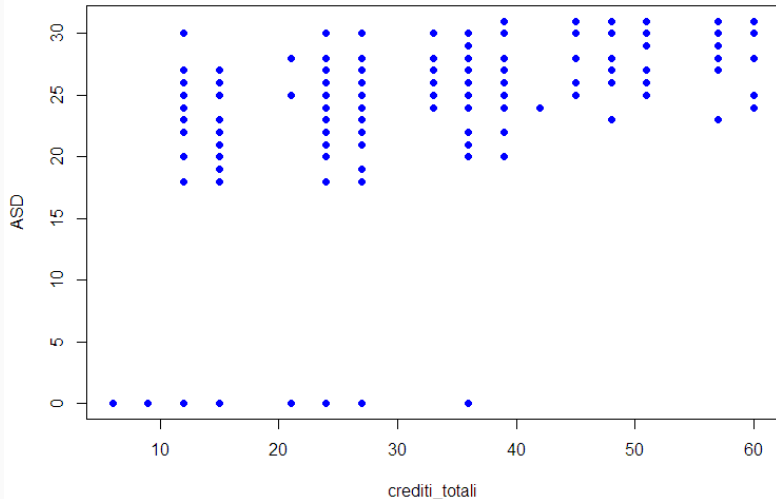
# **Analisi dei dati**

---

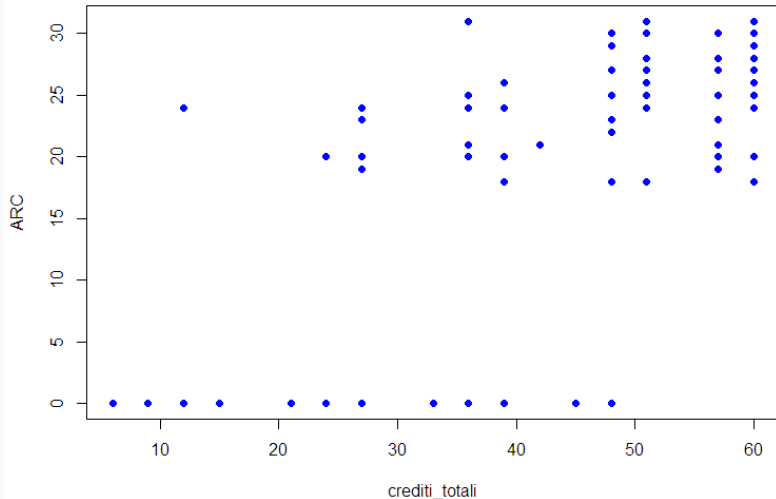
## Tabella di correlazione

	coorte	crediti totali	crediti con voto	voto medio	ASD	ARC	PRG	ANI	MDL	ING	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
ING	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

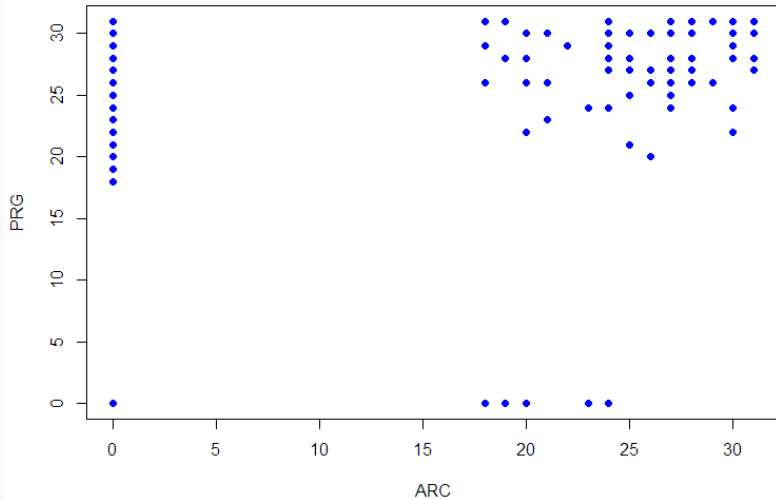
# Scatterplot Crediti totale e ASD



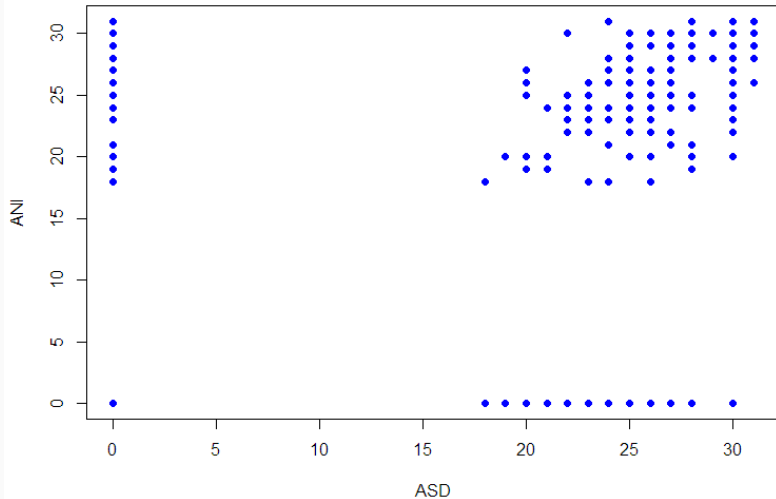
## Scatterplot Crediti totale e ARC



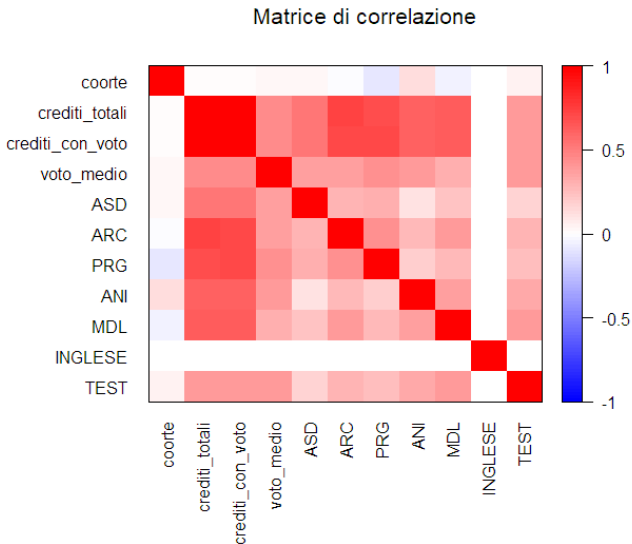
## Scatterplot ARC e PRG



## Scatterplot ASD e ANI



# Matrice di correlazione





# Clustering

---

- crediti totali, architetture, programmazione;
- algoritmi e strutture dati, architetture, programmazione, analisi 1 e matematica discreta e logica;
- voto medio e test.

- l'analisi effettuata con tecniche di clustering gerarchico è stata effettuata su un sottoinsieme dei dati a disposizione selezionato in base alla coorte dello studente (anno 2010);
- nel caso dell'algoritmo di Kmeans viene stabilito preventivamente il numero dei cluster possibili utilizzando valori ritenuti sensati di volta in volta;
- l'algoritmo DBSCAN è stato utilizzato per l'analisi relativa ai voti dei diversi esami scegliendo preventivamente i valori di MinPts e eps ritenuti sensati di volta in volta.

## Cluster ARC e PRG $k = 2$

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

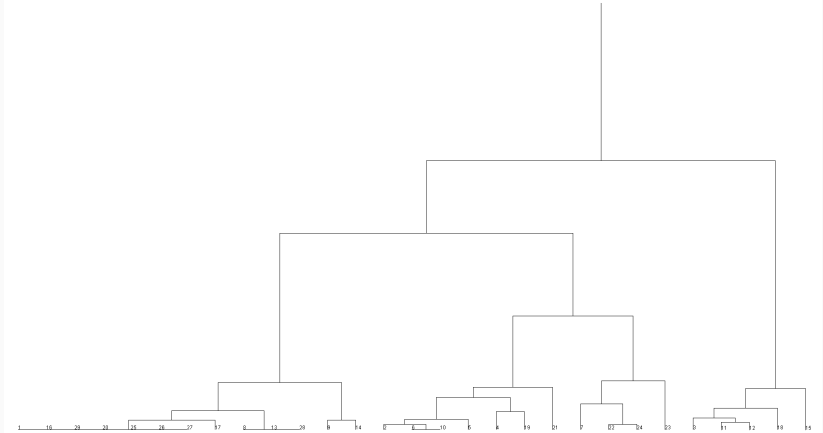
**Tabella 1:** Cluster con ARC e PRG con  $k = 2$  SSE 51.35

## Cluster ARC e PRG $k = 3$

	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%)

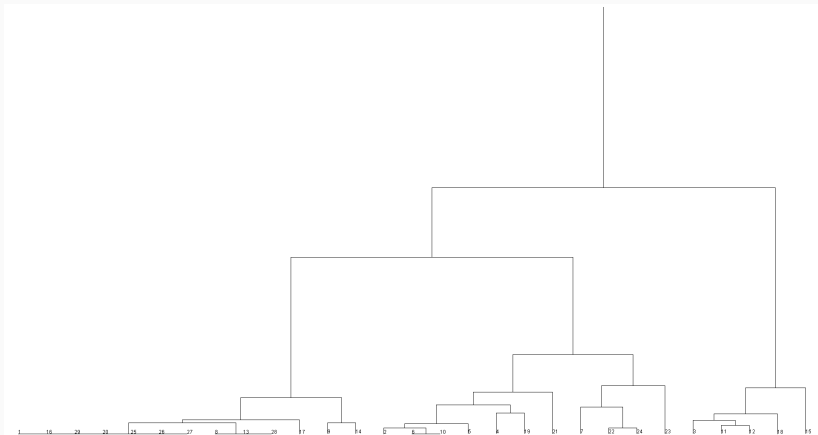
**Tabella 2:** Cluster con ARC e PRG con  $k = 3$  SSE 14.85

# Dendrogramma



**Figura 1:** Dendrogramma relativo al clustering gerarchico con metodo completo.

# Dendrogramma



**Figura 2:** Dendrogramma relativo al clustering gerarchico con metodo average.

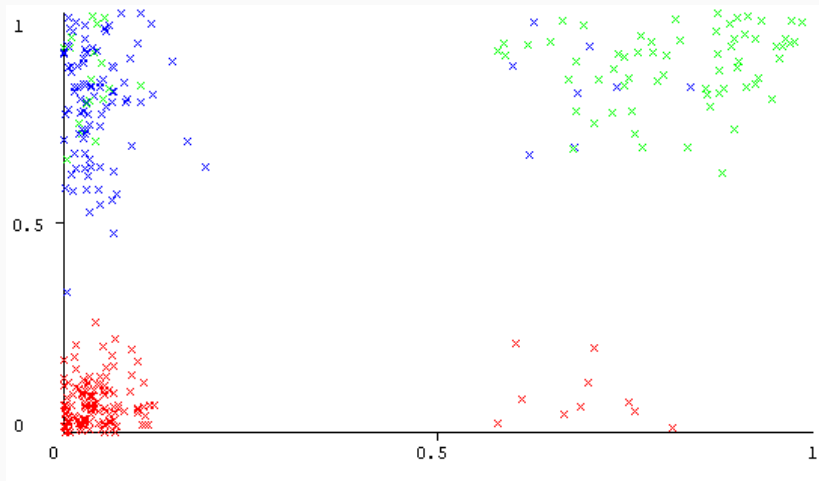
	ASD	ARC	PRG	ANI	MDL	Istanze
0	0.73	0.05	0.81	0.43	0.02	100 ( 32%)
1	0.65	0.05	0	0.50	0.12	133 ( 42%)
2	0.91	0.65	0.89	0.88	0.60	83 ( 26%)

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



- 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.

## Scatter plot dei cluster



**Figura 3:** Scatter plot relativo ai cluster dei voti di Architetture degli Elaboratori e Programmazione

	Istanze
0	40 (14%)
1	46 (17%)
2	41 (15%)
3	13 (5%)
4	45 (16%)
5	33 (12%)
6	6 (2%)
7	18 (6%)
8	22 (8%)
9	14 (5%)

**Tabella 4:** Cluster ottenuti con DBSCAN eseguito con MinPts=6 e  $\text{eps}=0.5$ .

	Istanze
0	40 (15%)
1	46 (17%)
2	41 (15%)
3	13 (5%)
4	45 (17%)
5	33 (12%)
6	18 (7%)
7	22 (8%)
8	14 (5%)

**Tabella 5:** Cluster ottenuti con DBSCAN eseguito con MinPts=10 e  $\text{eps}=0.4$ .

	Istanze
0	40 (18%)
1	46 (20%)
2	41 (18%)
3	45 (20%)
4	33 (15%)
5	22 (10%)

**Tabella 6:** Cluster ottenuti con DBSCAN eseguito con MinPts=20 e  $\text{eps}=0.4$ .

	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 7:** Cluster con Voto\_medio e Test con  $k = 3$  SSE 9.6

## **Valutazione del clustering e model selection**

---

- Selezione del numero "ottimale" di cluster per il K-means
- Valutazione del K-means
- Valutazione DBSCAN



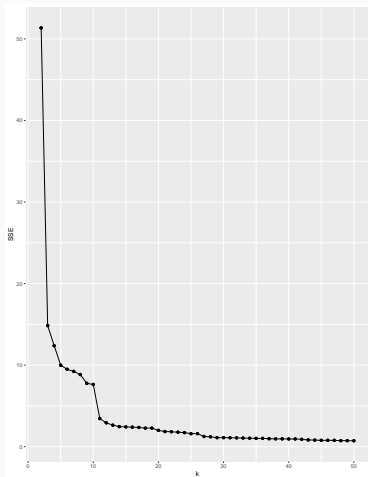
## Selezione numero di cluster nel K-means

Viene effettuata tramite la seguente procedura

- Determinazione SSE in funzione di  $k$
- Selezione del valore ottimale di  $k_{opt}$

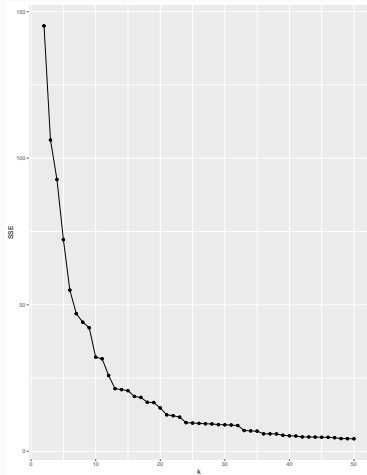
successivamente è possibile valutare e confrontare i risultati ottenuti dall'algoritmo con i diversi valori di  $k$ .

## Example



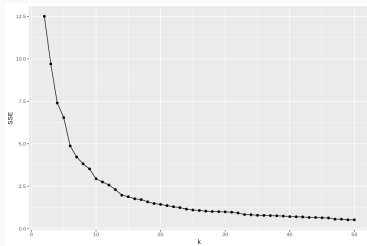
**Figure 4:** Dependency update

## Example



**Figure 5:** Dependency update

# Example



**Figura 6:** Dependency update

## Selezione Eps fissato MinPts in DBSCAN

Viene effettuata tramite la seguente procedura

- Ordino i punti rispetto alla loro distanza dal loro  $k$ -esimo punto più vicino item pongo  $\text{MinPts}=k$
- Determino un grafico con indici punti ordinati e distanze dal  $k$ -esimo più vicino
- Selezione come valore di Eps quello per cui c'è un picco.

La valutazione dei clustering ottenuti con K-means e DBSCAN è stata fatta con la seguente procedura

- Calcolo matrice distanze tra i punti
- Calcolo matrice di incidenza dei cluster
- "Serializzazione" e calcolo della correlazione

successivamente è possibile valutare e confrontare i risultati ottenuti dai clustering ottenuti con il K-means con i diversi valori di  $k$  e con il DBSCAN.

```
# Matrice di incidenza
matriceIncidenza <- function(data){
  nr = nrow(data)
  nc = ncol(data)
  C = matrix(nrow = nr, ncol = nr)
  for(i in 1:nr){
    for(j in 1:nr){
      if(data[i,nc] == data[j,nc])
        C[i,j] = 1
      else
        C[i,j] = 0
    }
  }
  return(C)
```

```
# matrice distanza
matriceDistanza <- function(data){
  return(as.matrix(dist(data[,1:(ncol(data)-1)],method =
    'euclidean',diag = TRUE,upper = TRUE)))
}

calcoloCorrelazione <- function(data){
  MI <- matriceIncidenza(data)
  D <- matriceDistanza(data)
  mi = as.vector(t(MI))
  d = as.vector(t(D))

  return(cor(mi,d,method="pearson"))
}

calcoloCorrelazione(crediti_totali_prg_arc_clustered)
```



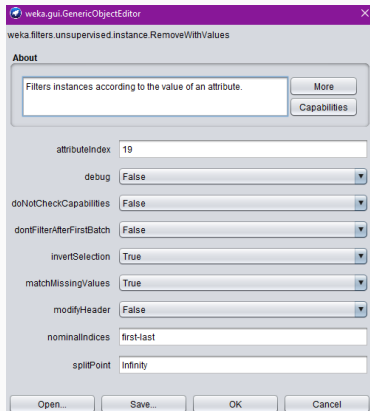
# Valori Correlazione K-means

	coorte	crediti totali	crediti con voto	voto medio	ASD	ARC	PRG	ANI	MDL	ING	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
ING	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

- Procedura model selection non efficace
- Valori scelti inizialmente sono migliori

- E' possibile calcolare lo stesso valore di correlazione anche per il DBSCAN
- Necessaria preventiva rimozione di rumore

# Rimozione rumore con Weka



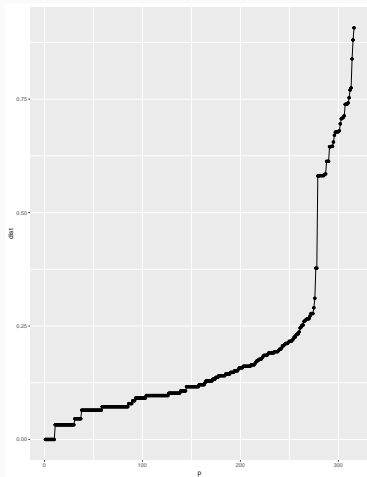
**Figura 7:** Dependency update

# Correlazione e rumore DBSCAN

	coorte	crediti totali	crediti con voto	voto medio	ASD	ARC	PRG	ANI	MDL	ING	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
ING	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

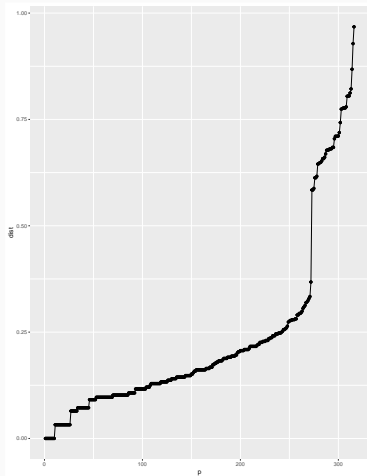
```
Hello world
```

# Rimozione rumore con Weka



**Figura 8:** Dependency update

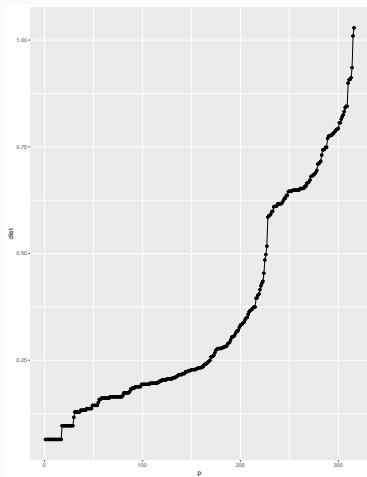
# Rimozione rumore con Weka



**Figura 9:** Dependency update



# Rimozione rumore con Weka



**Figura 10:** Dependency update

# Correlazione e rumore DBSCAN

	coorte	crediti totali	crediti con voto	voto medio	ASD	ARC	PRG	ANI	MDL	ING	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
ING	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

## Conclusioni

---

- Architetture degli elaboratori esame più difficile
- La media alla fine del primo anno non sempre conferma i risultati ottenuti al test di ingresso
- Non tutti gli esami sono generalmente sostenuti al primo anno

Grazie per l'attenzione