Clustering studenti informatica

Tommaso Ceccarini, Filippo Mameli

3 ottobre 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

Il dataset

- Valutazione conseguita all'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

La 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

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'
```

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

Analisi dei dati

Tabella di correlazione

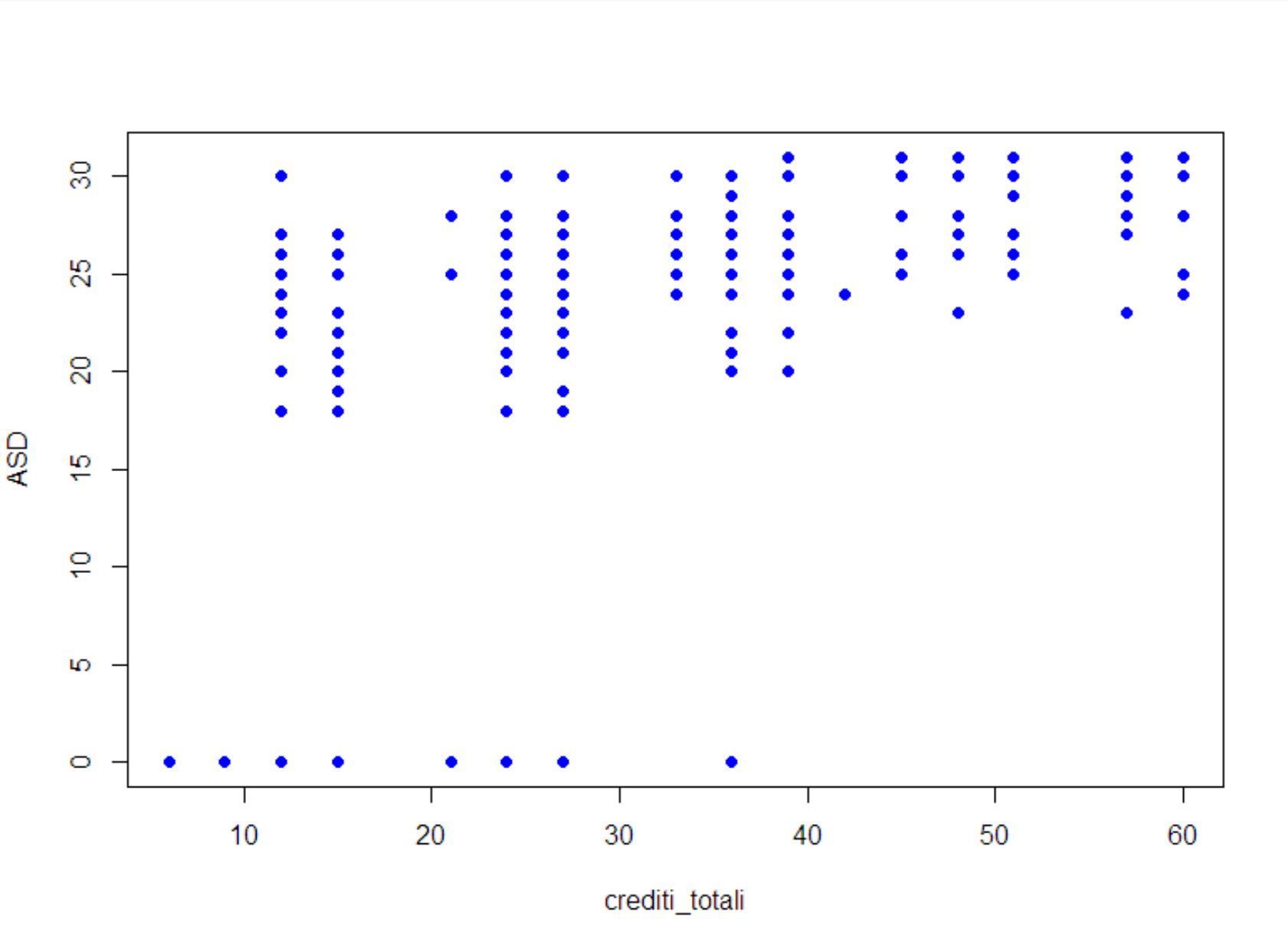
	coorte	crediti totali	crediti con voto	voto medio	ASD	ARC	PRG	ANI	MDL	ING	TEST
coorte	1	0.013343	0.01821	0.03655	0.03581	-0.01609	-0.0822	0.13386	-0.04033	NA	0.04126
crediti_totali	0.01334	1	0.99522	0.44571	0.52984	0.72508	0.69882	0.61015	0.62789	NA	0.38433
crediti_con_voto	0.01821	0.99522	1	0.44838	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

Correlazioni

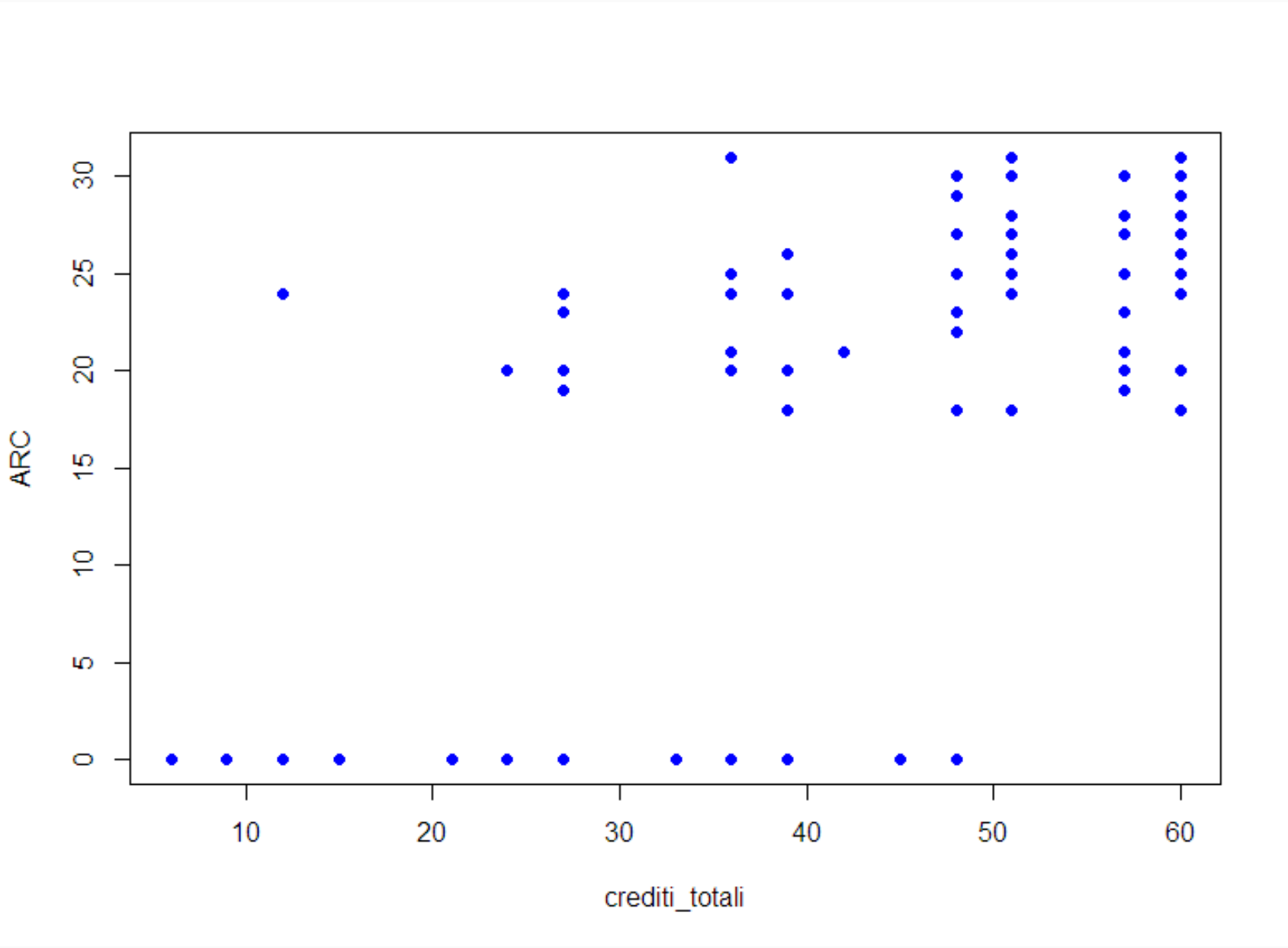
Osservazioni:

- Ovvia correlazione tra Crediti con voto e Crediti totali
- Coorte non è correlato con nessun attributo
- Correlazione alta tra Architetture degli elaboratori e Crediti totali
- Bassa correlazione tra Algoritmi e Crediti totali
- L'attributo Test è maggiormente correlato con Voto medio
- Architetture e Programmazione hanno il valore di correlazione più alto tra gli esami, Analisi I e Algoritmi il più basso

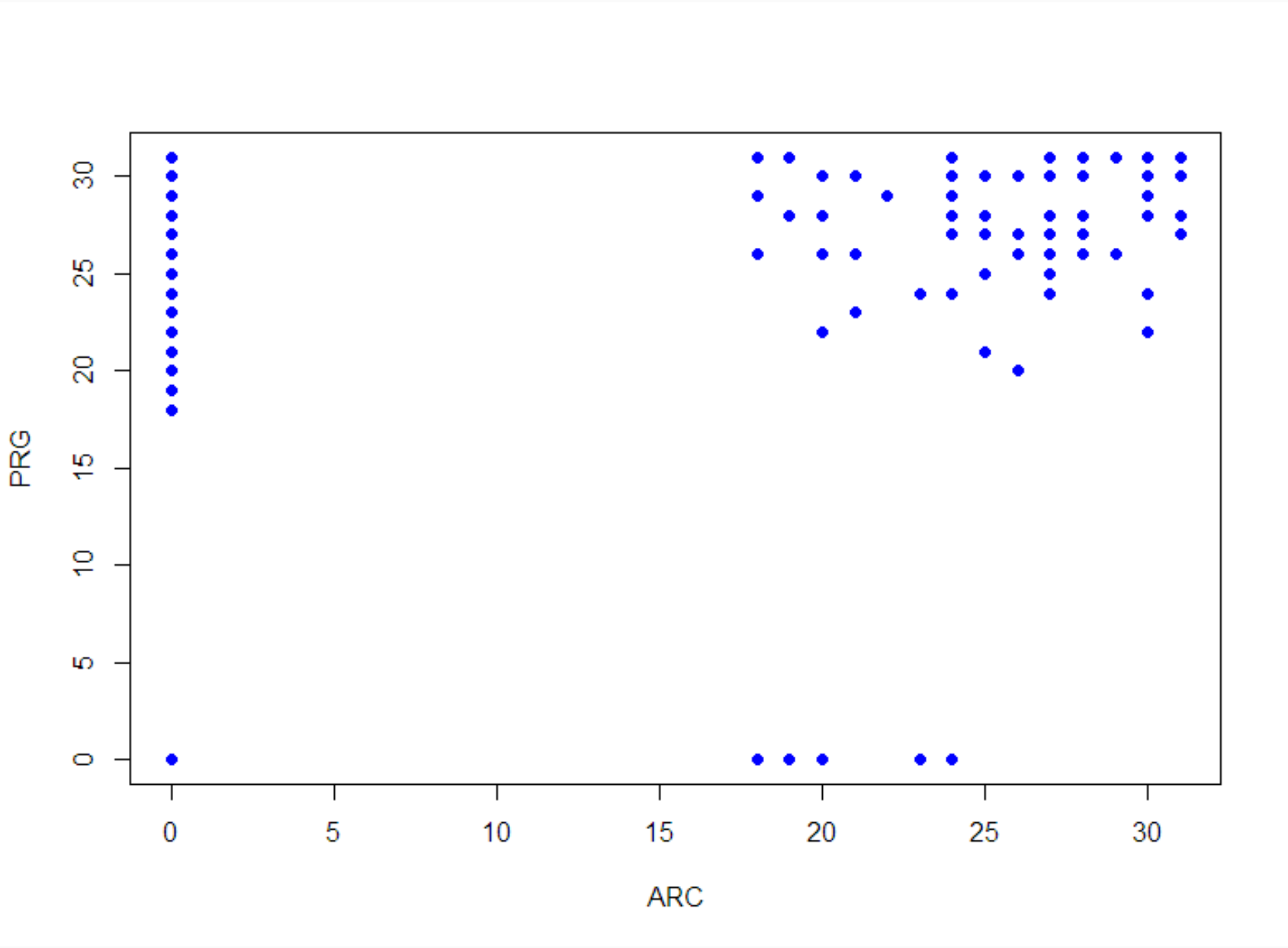
Scatterplot Crediti totale e ASD



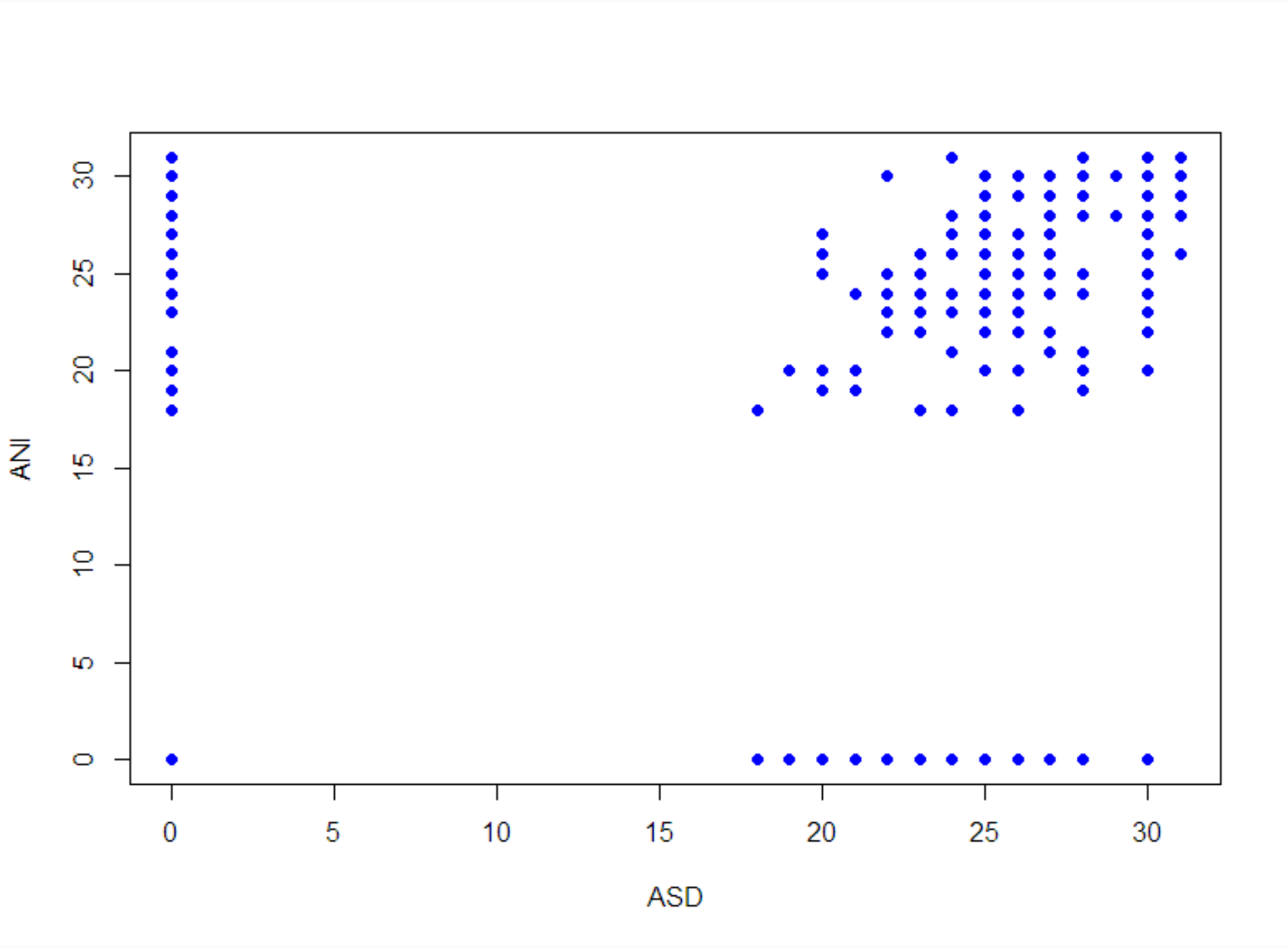
Scatterplot Crediti totale e ARC



Scatterplot ARC e PRG



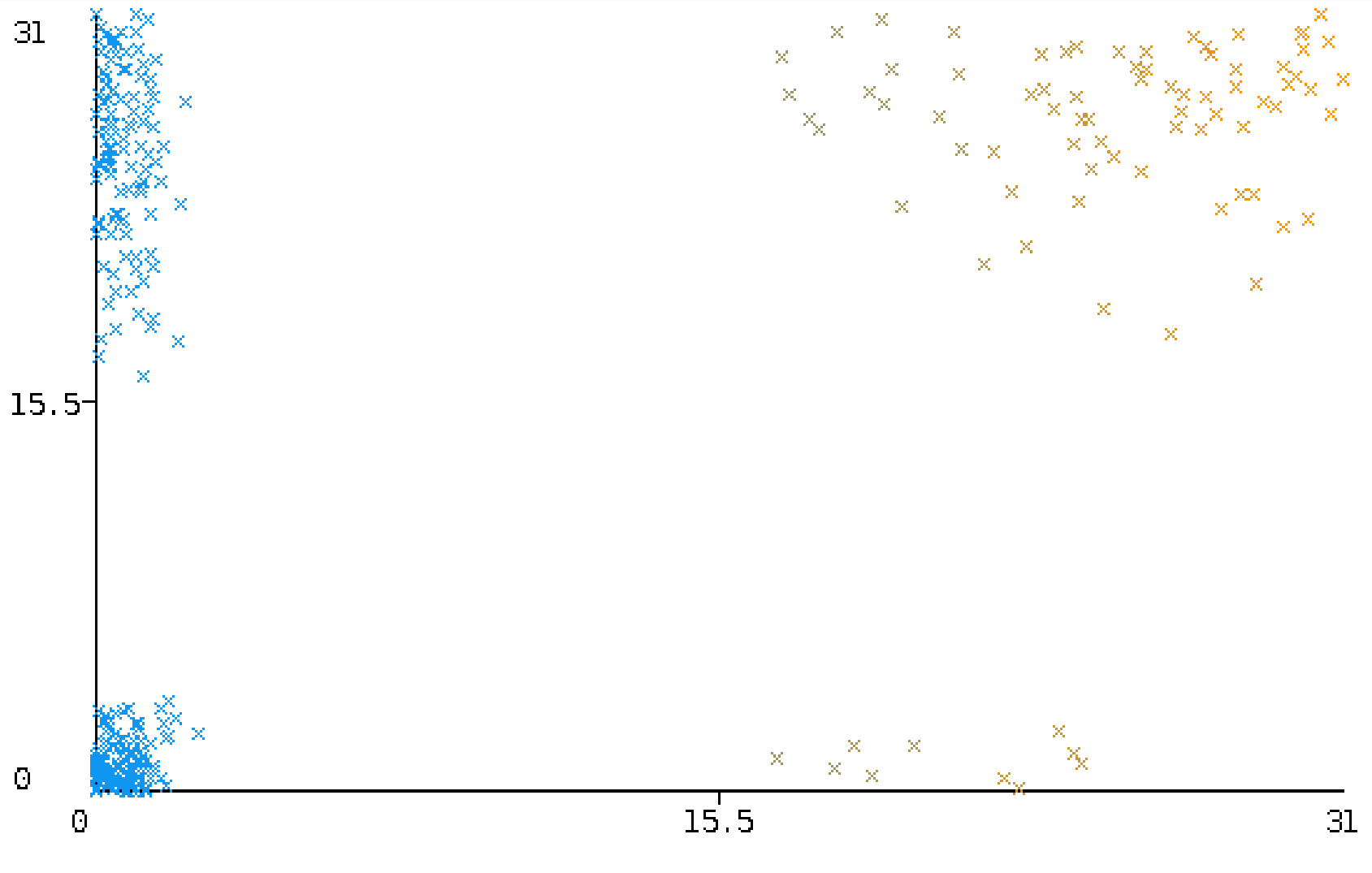
Scatterplot ASD e ANI



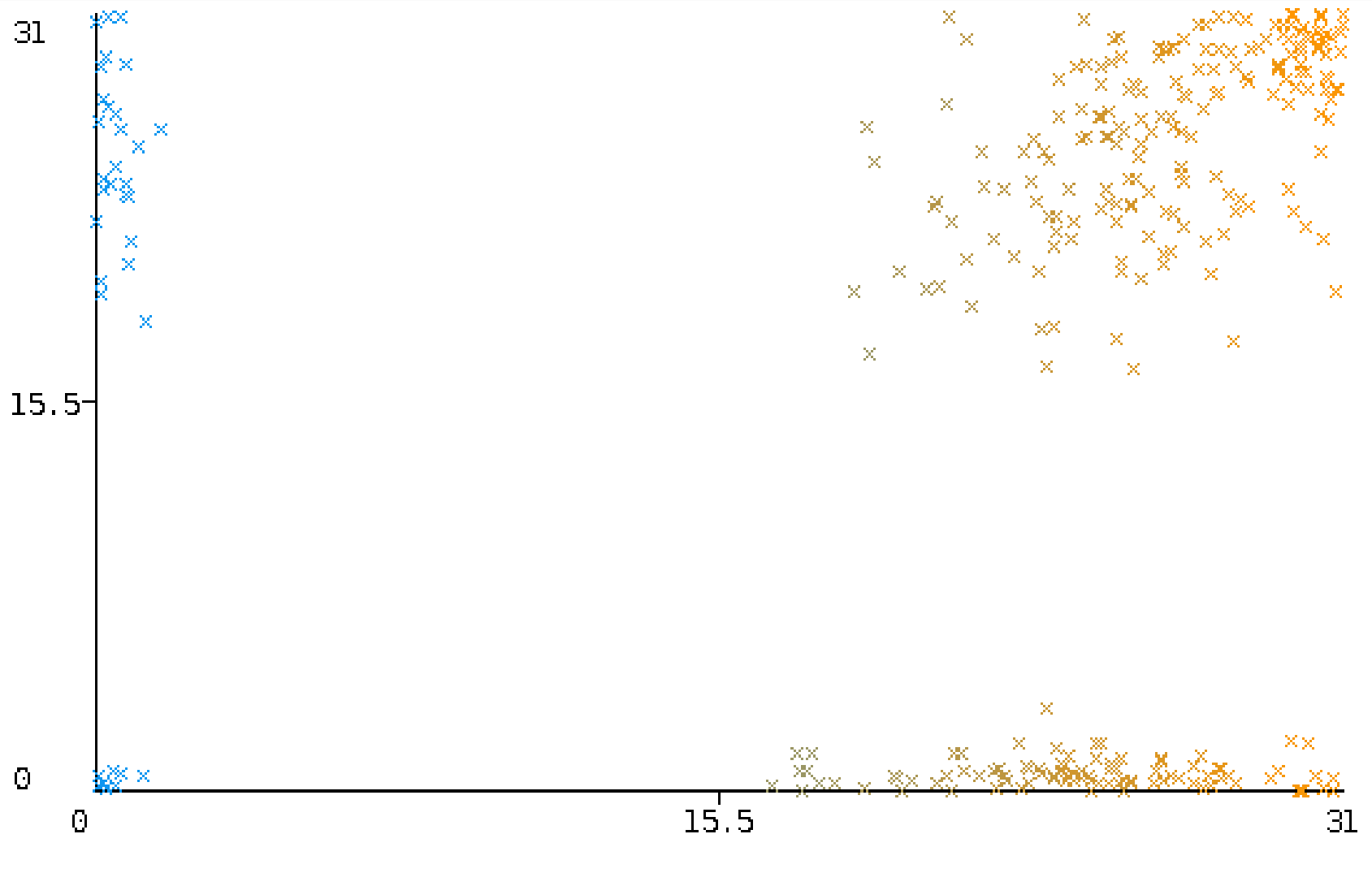
Osservazione

In uno scatterplot non è possibile distinguere oggetti con gli stessi valori per gli attributi considerati e quindi i dati che influenzano maggiormente la correlazione tra Architetture e Programmazione non vengono essenzialmente mostrati

Scatterplot ARC e PRG con Jitter



Scatterplot ASD e ANI con Jitter



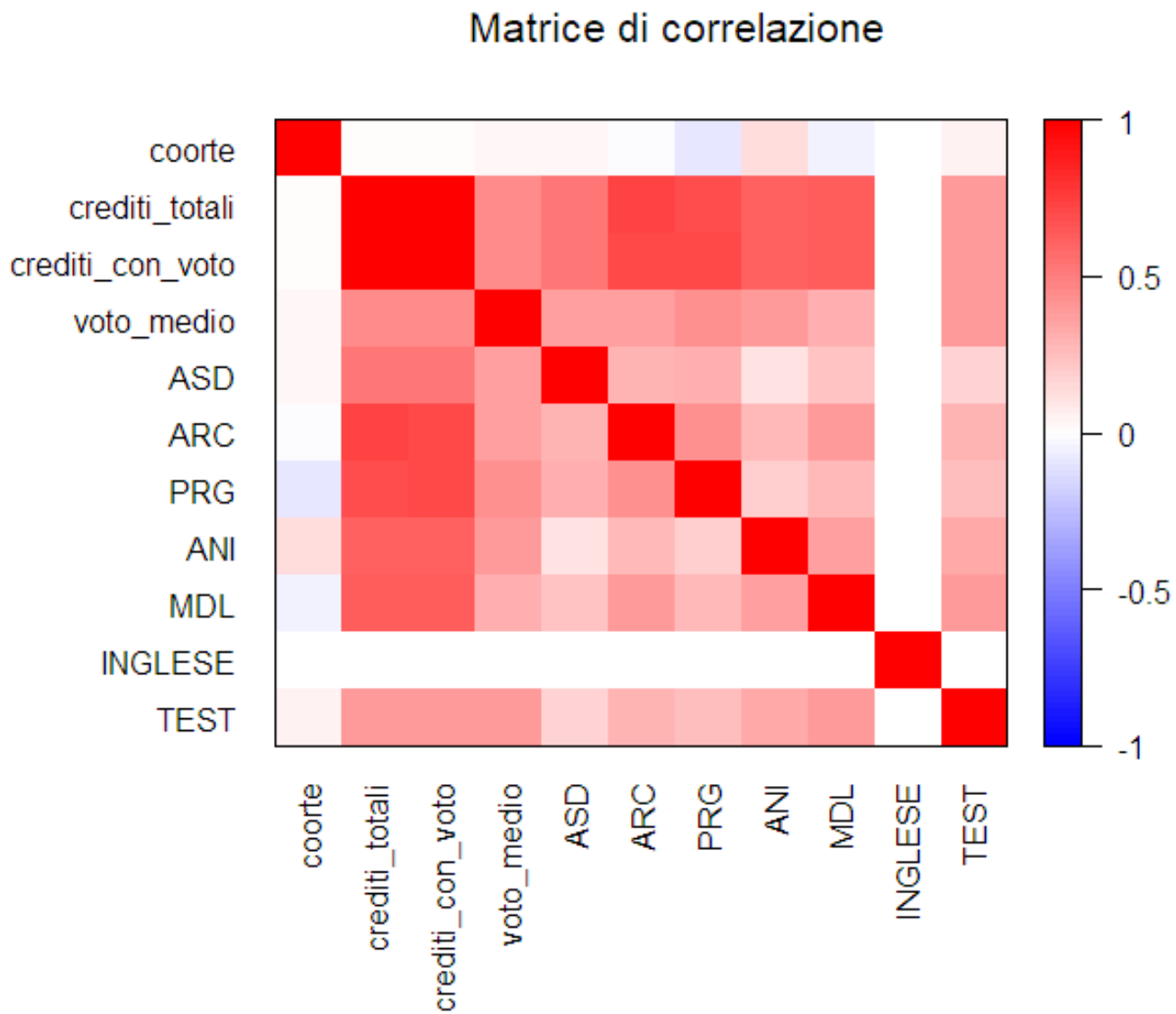
Analisi scatterplot con Jitter

- Il numero di studenti che non ha sostenuto ne Architteture ne Programmazione è notevolmente superiore al numero di studenti che non hanno sostenuto ne algoritmi ne analisi 1. Ciò influenza la correlazione complessiva.
- Gli studenti che hanno sostenuto Programmazione ma non hanno sostenuto Architetture sono molti di più rispetto a quelli che hanno sostenuto Analisi 1 ma non hanno sostenuto Algoritmi. Viceversa, gli studenti che hanno sostenuto Architetture ma non Programmazione sono molti meno di quelli che hanno sostenuto Algoritmi ma non Analisi 1. Essendo in quantità paragonabili influenzeranno le correlazione complessive tra le due diverse coppie di attributi circa in egual misura;

Analisi scatterplot con Jitter

- Confrontando gli studenti che hanno sostenuto sia Architetture che Programmazione con quelli che hanno sostenuto sia Algoritmi che Analisi 1 si nota una migliore correlazione per la seconda coppia di attributi su tali sottoinsiemi del dataset. Tuttavia, essendo i due sottoinsieme (studenti che hanno sostenuto sia Architetture che Programmazione e studenti che hanno sostenuto sia Algoritmi che Analisi 1) di dimensioni paragonabili influenzeranno in modo minore la correlazione complessiva rispetto a quanto lo fanno gli studenti discussi nel primo punto.

Matrice di correlazione



Clustering

Clustering

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

Clustering

- 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

Osservazioni

- 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.

Osservazioni

- Non esiste la categoria di studenti che ha sostenuto con profitto l'esame di architetture, ma non ha sostenuto l'esame di programmazione

Dendogramma

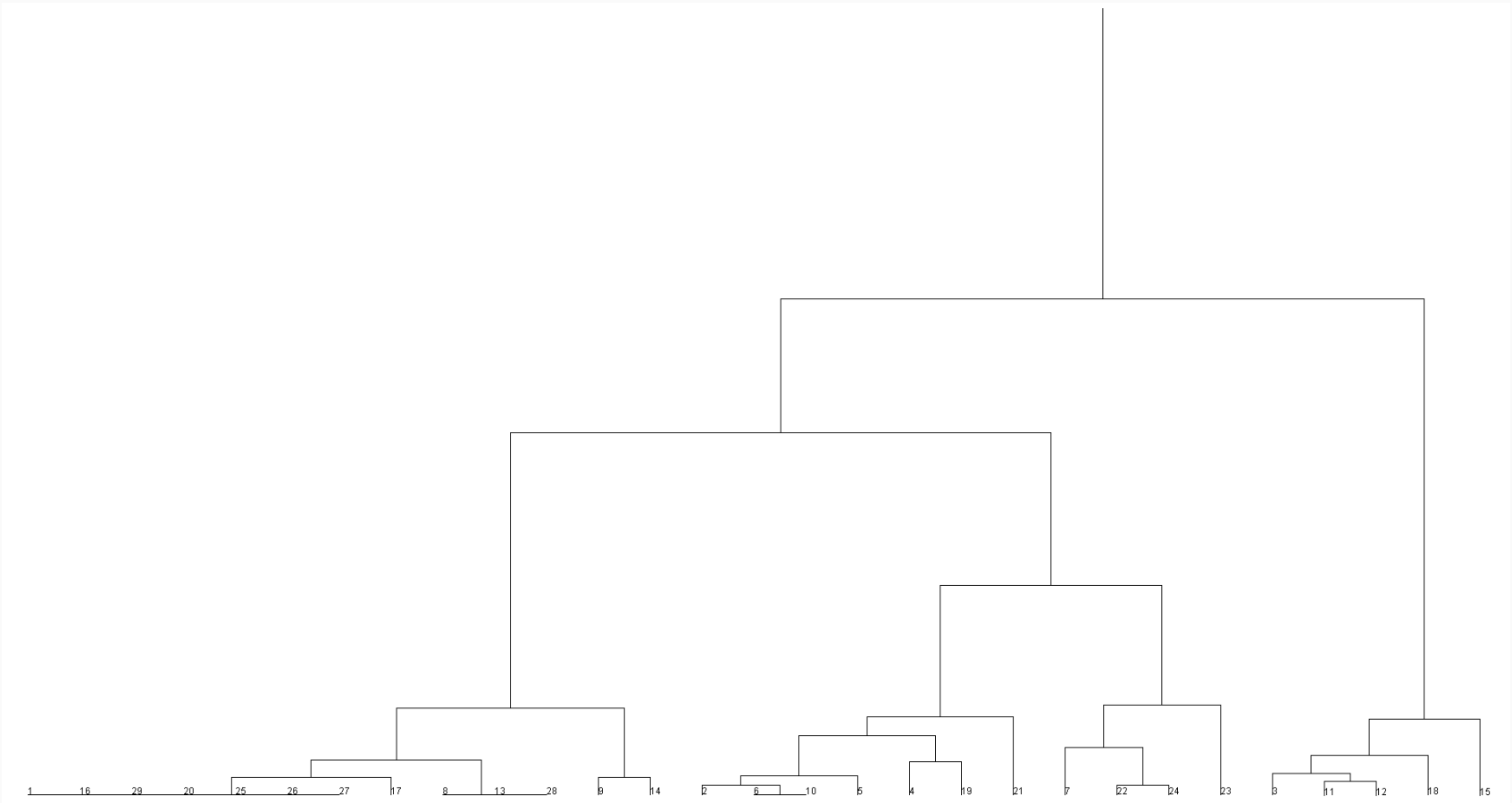


Figura 1: Dendogramma relativo al clustering gerarchico con metodo complete.

Dendogramma

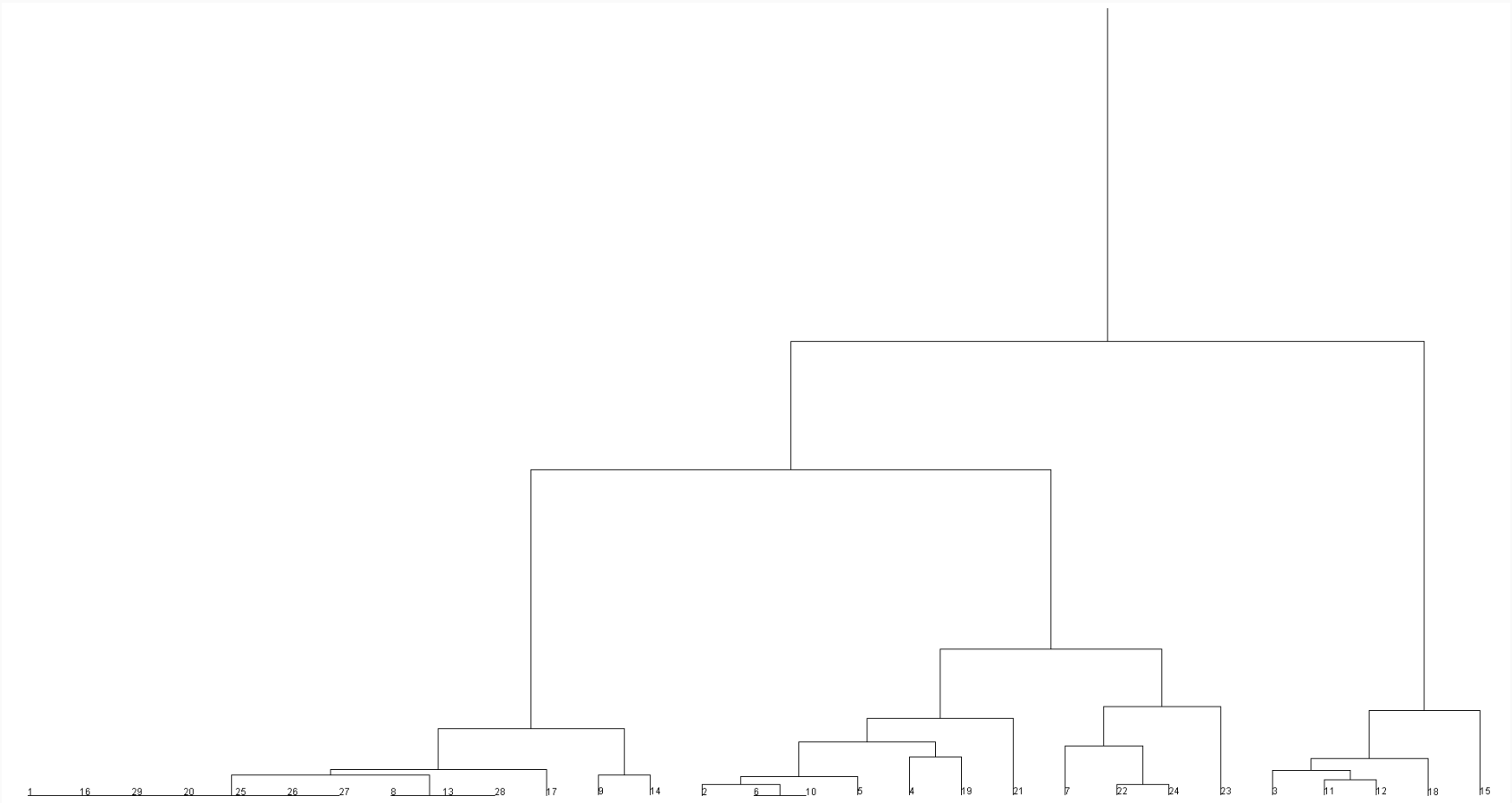


Figura 2: Dendogramma relativo al clustering gerarchico con metodo average.

Cluster di tutti i voti

	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

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.

Scatter plot dei cluster

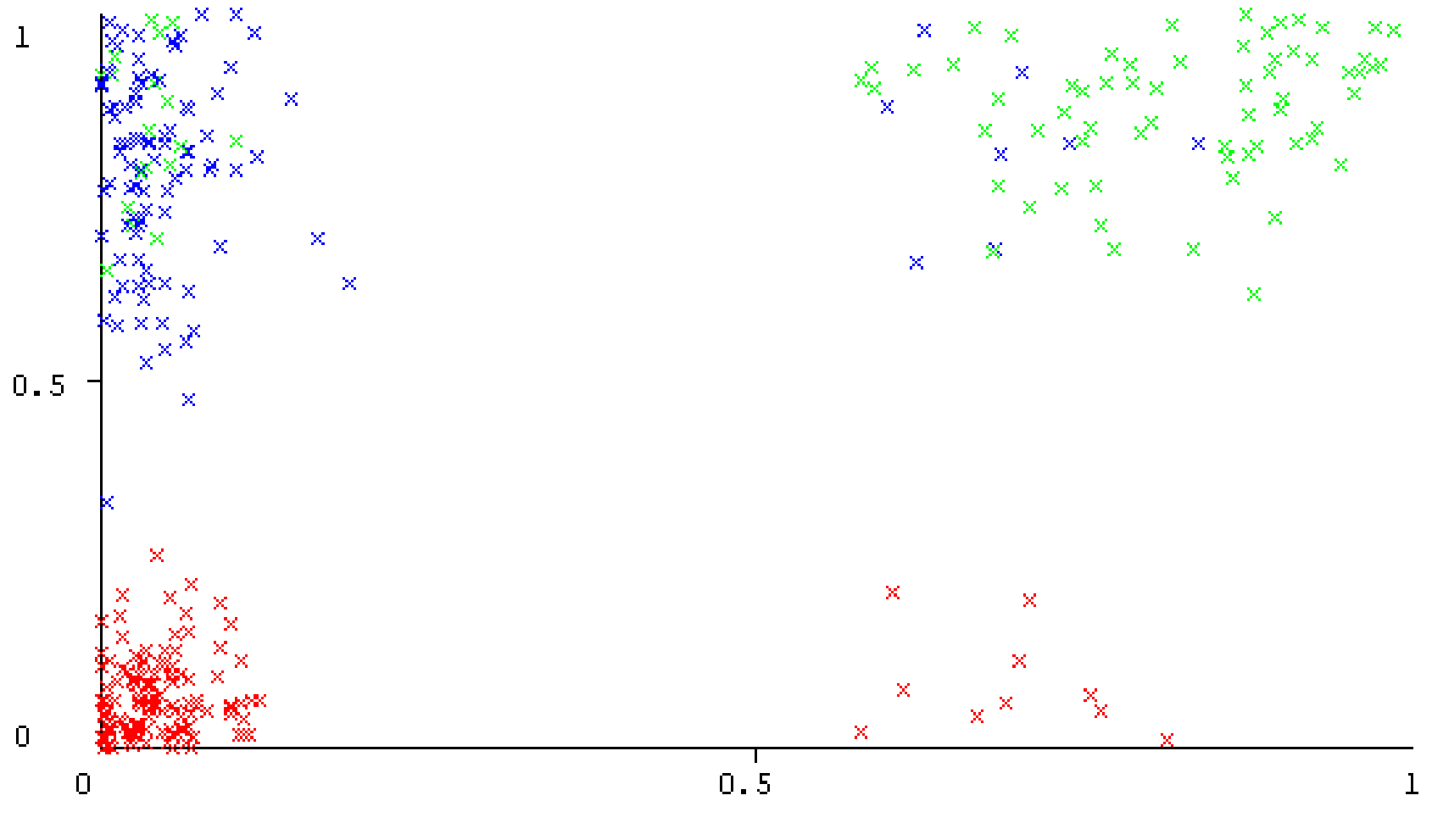


Figura 3: Scatter plot relativo ai cluster dei voti di Architetture degli Elaboratori e Programmazione

Cluster DBSCAN

	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 `eps=0.5`.

Osservazioni

- 38 record dei 316 totali sono stati marcati come rumore dall'algoritmo
- Cluster di piccole dimensioni
- `MinPts` troppo basso

Cluster DBSCAN

	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 `eps=0.4`.

Osservazioni

- 44 record dei 316 totali sono stati marcati come rumore dall'algoritmo
- Il numero dei cluster diminuisce e passa a 8
- I cluster di piccole dimensioni diventano solo due

Cluster DBSCAN

	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 `eps=0.4`.

Osservazioni

- 89 record dei 316 totali sono stati marcati come rumore dall'algoritmo
- Il numero dei cluster è pari a 6
- La distribuzione dei record nei cluster risulta decisamente più uniforme
- I cluster di piccole dimensioni non sono presenti

Cluster Voto medio e Test

	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

Osservazioni

- Si determinano tre cluster ben distinti
- I primi due cluster identificano gli studenti "migliori" e quelli "peggiori"
- Nel terzo cluster gli studenti hanno conseguito un punteggio al test d'ingresso decisamente positivo, ma non hanno mantenuto una media dei voti altrettanto buona

Valutazione del clustering e model selection

Valutazione del clustering e model selection

- Selezione parametri "ottimali" per K-means e DBSCAN
- 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.

Esecuzione automica K-means in Java (1)

```
public class SseWeka {
public static void main(String[] args) throws Exception {
DataSource source = new
    DataSource("./voti-con-data.arff");
Instances data = source.getDataSet();
data.setClassIndex(-1);
Normalize normalize = new Normalize();
normalize.setInputFormat(data);
Instances nData = Filter.useFilter(data, normalize);
Remove remove = new Remove();
int[] attributeIndexesToRemove = {1,6,8};
remove.setInvertSelection(true);
remove.setAttributeIndicesArray(attributeIndexesToRemove);
remove.setInputFormat(nData);
```

Esecuzione automica K-means in Java (2)

```
Instances creditiArcPrg = Filter.useFilter(nData, remove);
SimpleKMeans kMeans = new SimpleKMeans();
kMeans.setPreserveInstancesOrder(true);
int maxK = 50;
int minK = 2;
ArrayList<double[]> resultSet = new
    ArrayList<double[]>(maxK);
double[] a;
for (int i = minK; i < maxK; i++) {
kMeans.setNumClusters(i);
kMeans.buildClusterer(creditiArcPrg);
a = new double[2];
a[0] = i; a[1] = kMeans.getSquaredError();
resultSet.add(a);
}
```

Esecuzione automica K-means in Java (3)

```
for (int i = 0; i < maxK - minK; i++) {
 System.out.println(resultSet.get(i)[1]);
}
try {
 FileWriter writer = new FileWriter("creditiArcSse.txt",
 false);
 for (int i = minK; i < maxK; i++)
  writer.write(i + "," + resultSet.get(i - minK)[1] +
  "\r\n");
 writer.close();
} catch (IOException e) {
 e.printStackTrace();
 }
}
}
```

Grafico K-SSE crediti totali, architetture e programmazione

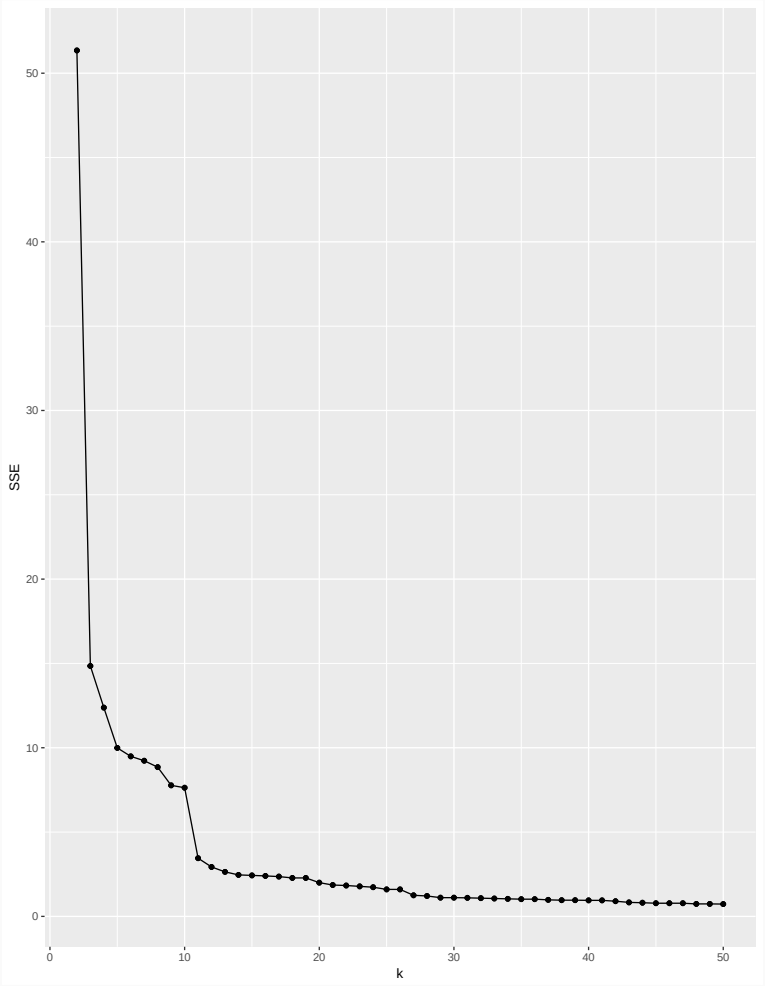


Figura 4: Andamento del valore del SSE in funzione del numero di cluster

Grafico K-SSE relativo ai voti

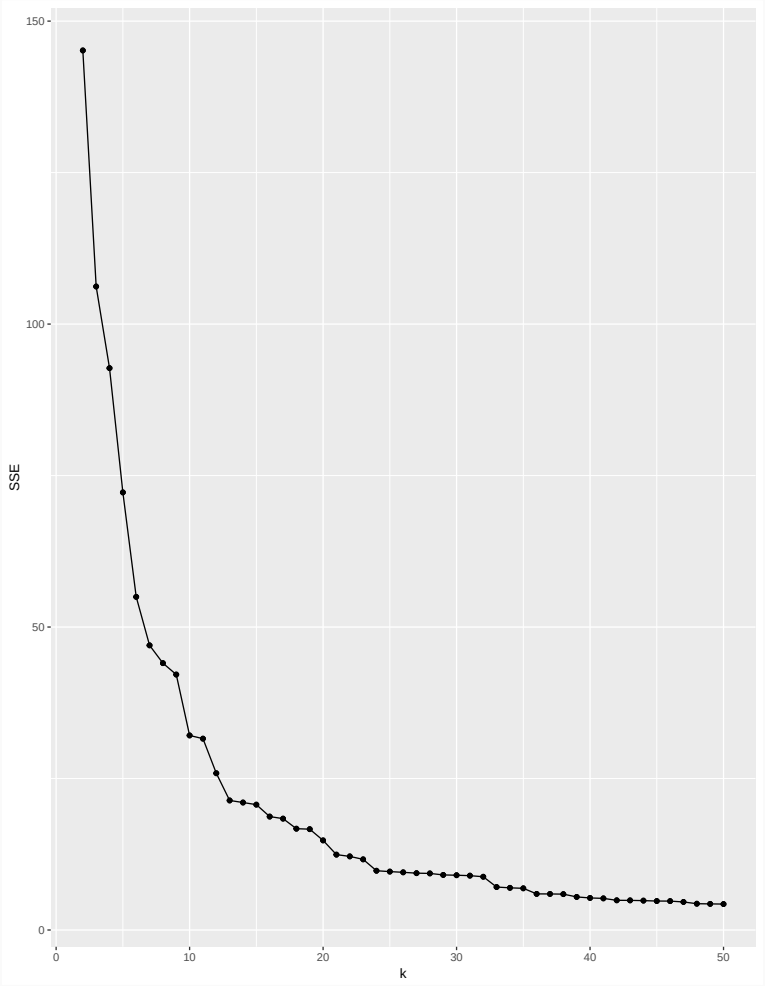


Figura 5: Andamento del valore del SSE in funzione del numero di cluster

Grafico K-SSE test, voto-medio

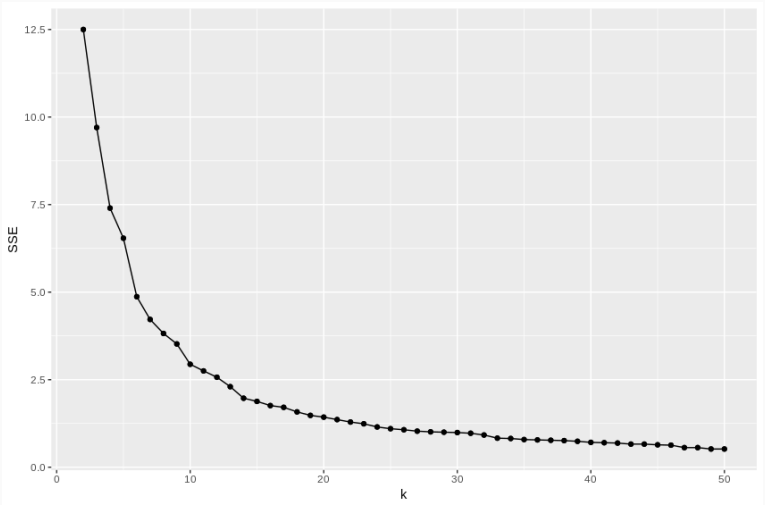


Figura 6: Andamento del valore del SSE in funzione del numero di cluster

Selezione valori ottimali

- viene determinato il primo valore di k per cui $SSE_{k+1} - SSE_k \leq \varepsilon$;
- risultati ottenuti con $\varepsilon = 0.01$.

Attributi	k_{opt}
crediti-totali, ARC, PRG	18
ARC,ASD,AN1,PRG,MDL	36
voto-medio,test	43

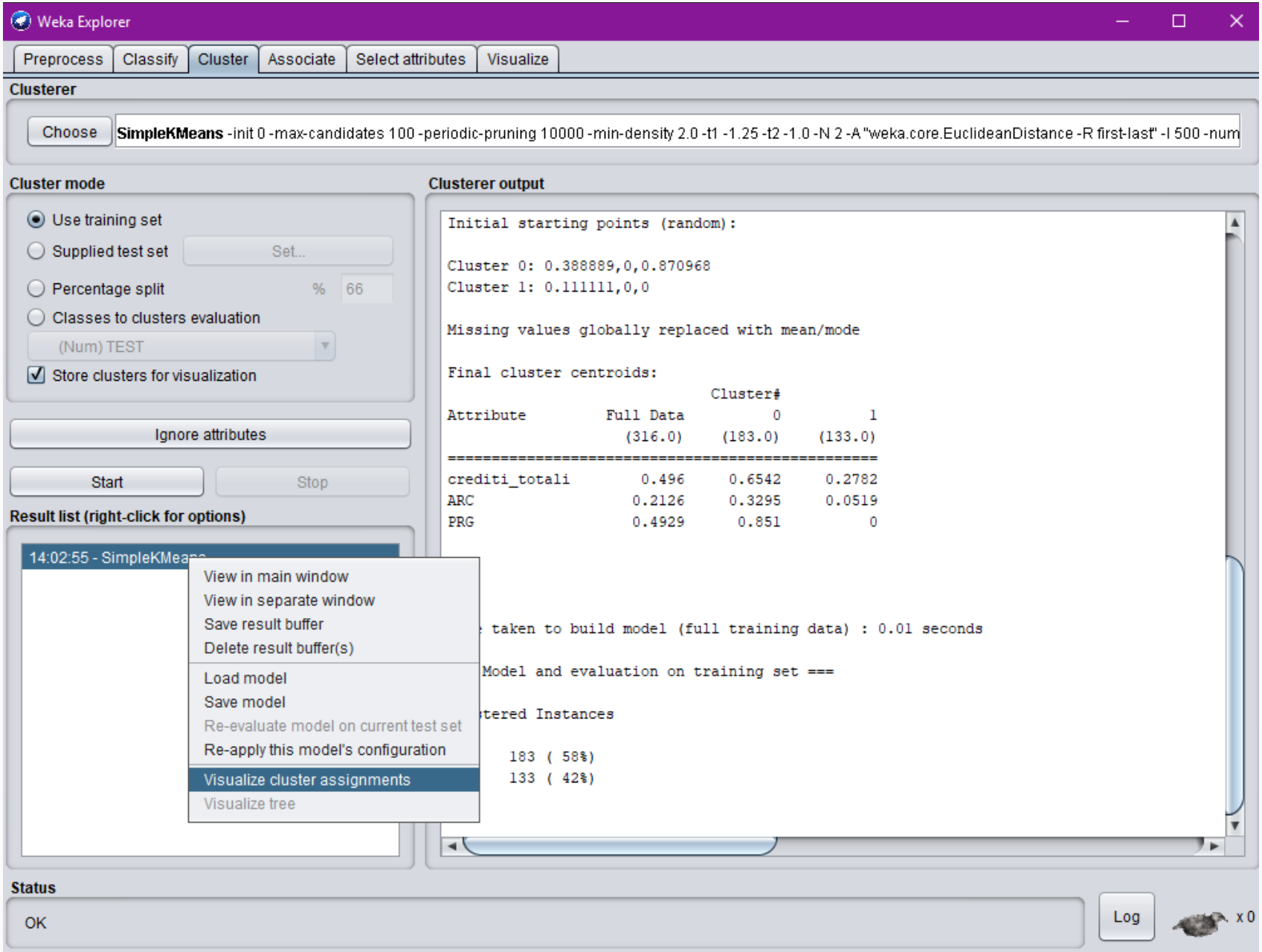
Valutazione

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

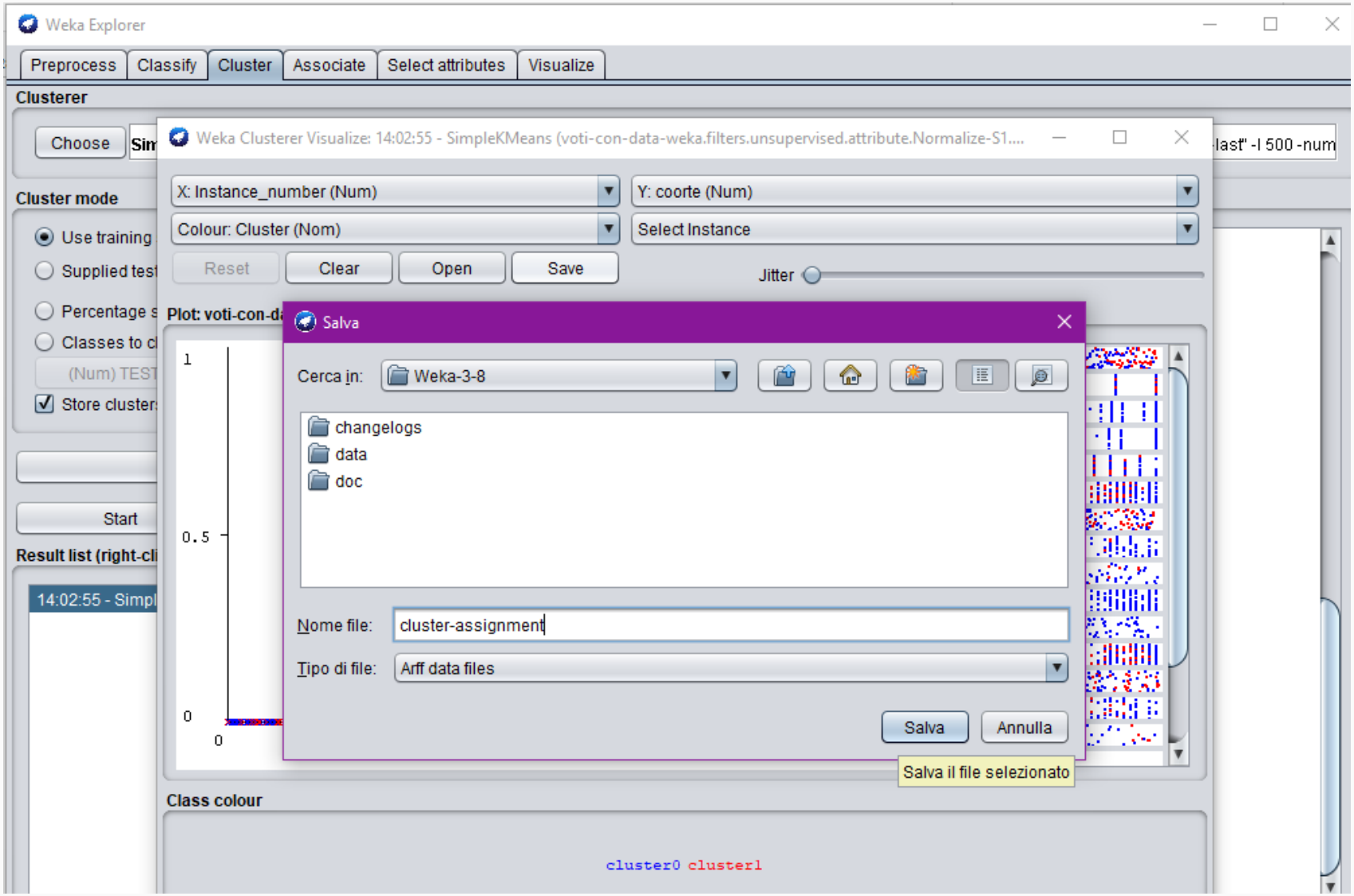
- Calcolo matrice distanze tra i punti
- Calcolo matrice di incidenza dei cluster
- "Serializzazione" e calcolo della correlazione

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.

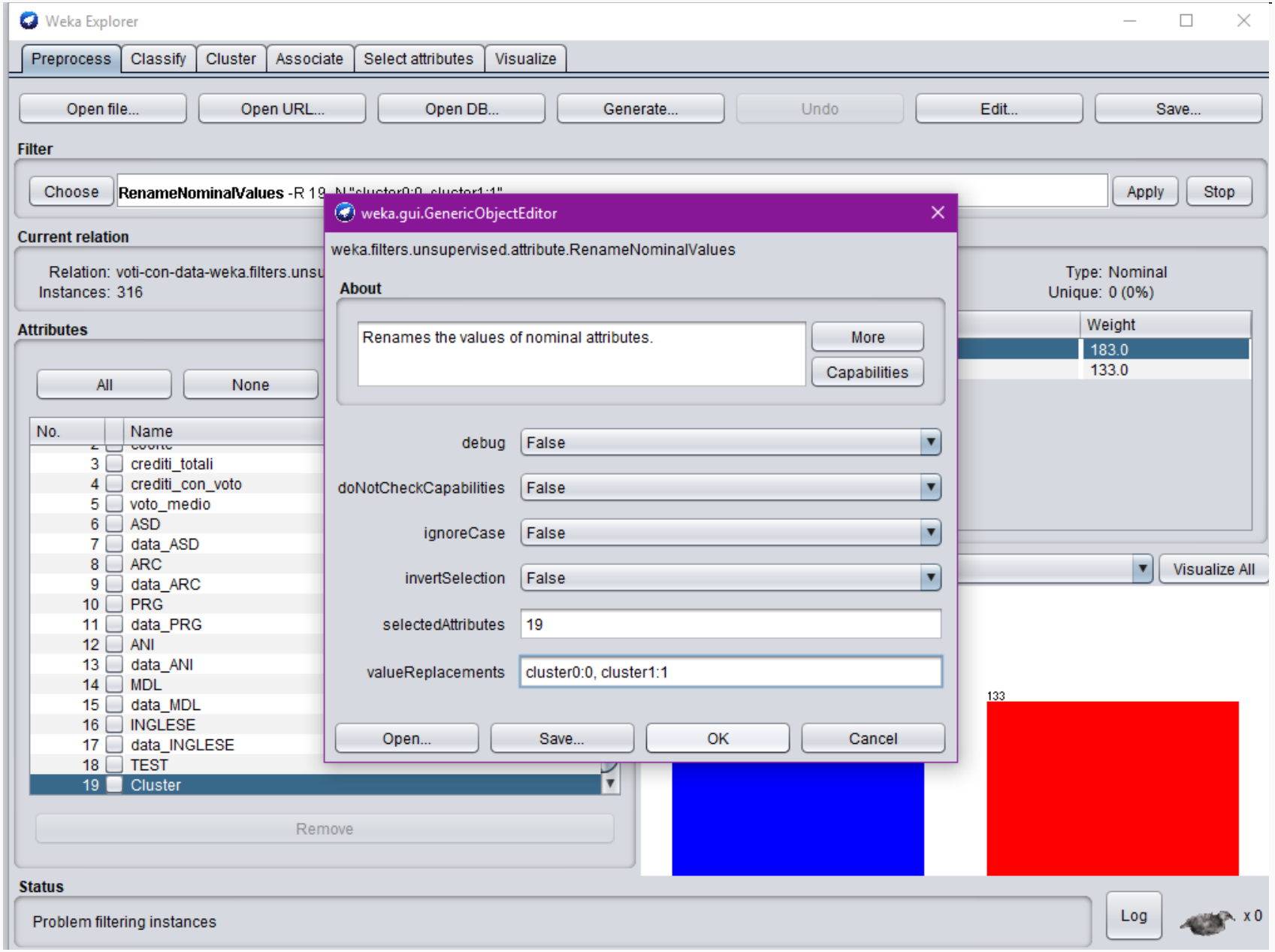
Salvataggio etichette dei cluster (1)



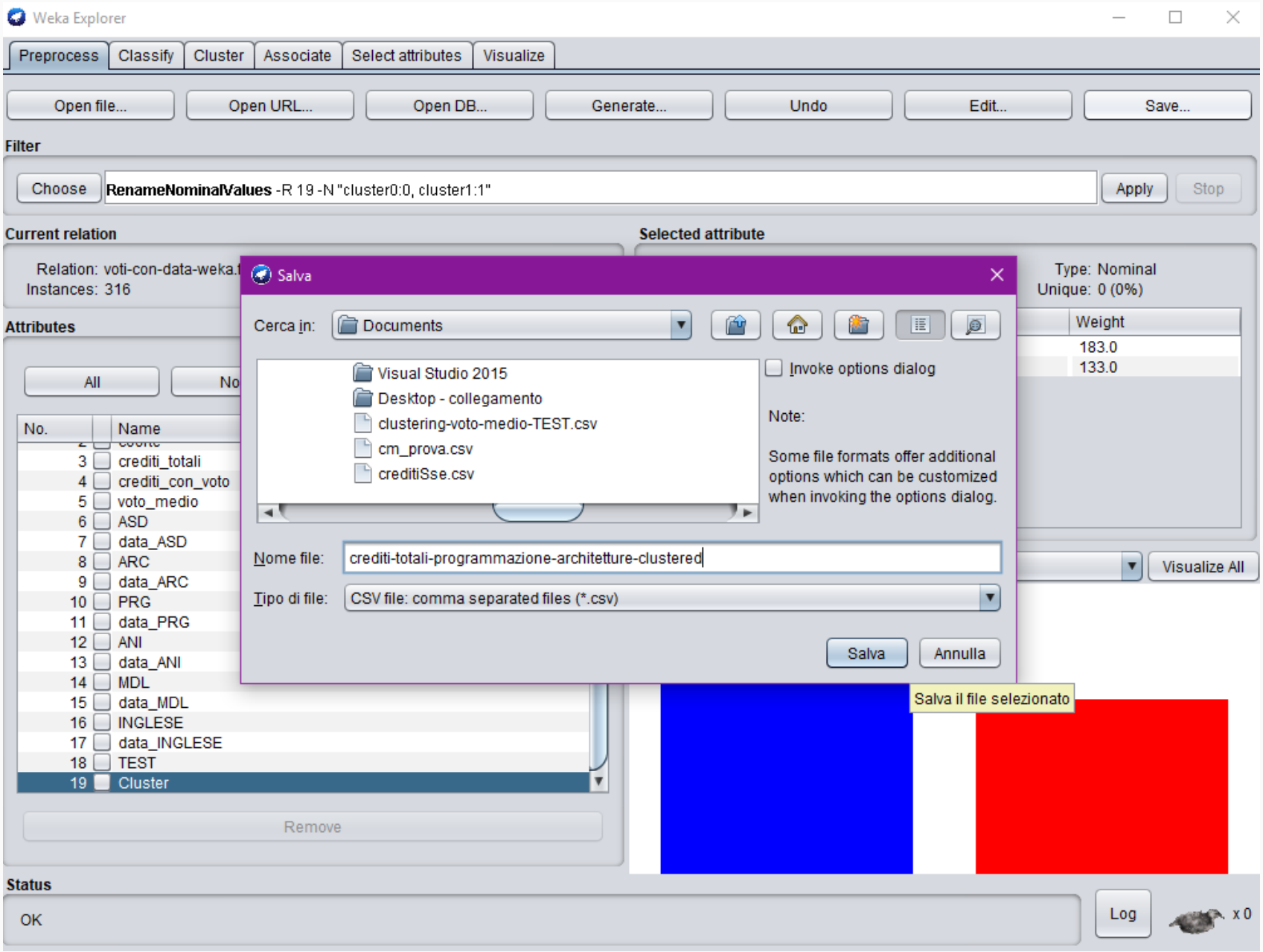
Salvataggio etichette dei cluster (2)



Salvataggio etichette dei cluster (3)



Salvataggio etichette dei cluster (4)



Calcolo matrice di incidenza con R

```
# 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)
```

Calcolo matrice delle distanze con R

```
# 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

Attributi analizzati	k	Correlazione
crediti totali, ARC, PRG	2	-0.687
	3	-0.854
	18	-0.489
ARC, ASD, PRG, MDL, AN1	2	-0.520
	3	-0.618
	36	-0.424
voto medio, test	2	-0.476
	3	-0.465
	43	-0.273

Tabella 8: Valori correlazione tra la Matrice di incidenza dei cluster e la matrice delle distanze in funzione di k.

Valutazione k-Means

- Procedura model selection non efficace (anche con valori maggiori di ε);
- Valori scelti inizialmente sono migliori.

Valutazione DBSCAN

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

Rimozione rumore con Weka

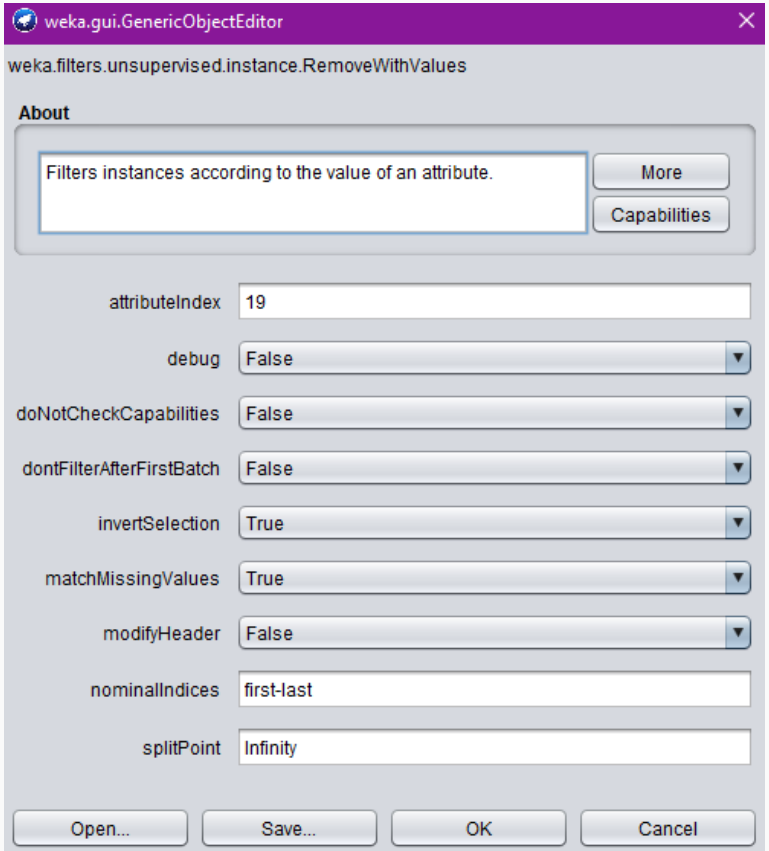


Figura 7: Filtro Weka impiegato per rimozione rumore

Correlazione e rumore DBSCAN

`MinPts, eps`	Correlazione	Rumore
6,0.5	−0.725	38
10,0.4	−0.728	44
20,0.4	−0.758	89

Tabella 9: Valori riepilogativi di correlazione e rumore per DBSCAN

Selezione `Eps` fissato `MinPts` in DBSCAN

Viene effettuata la seguente procedura

- Ordinamento dei punti rispetto alle distanze dal loro k-esimo punto più vicino;
- pongo `MinPts=k`;
- Determinazione del grafico che mostra l'andamento crescente delle k-distanze dei punti;
- Scelta del valore di `Eps` come la prima k-distanza per cui verifica una sostanziale crescita.

Codice model selection DBSCAN con R

```
kDBScan <- function(data,k){
  library(ggplot2)
  D = as.matrix(dist(data[,1:ncol(data)-1],method =
      'euclidean',diag = TRUE,upper = TRUE))
  D_1 = D
  for(i in 1:nrow(data)){
    D_1[i,] = sort(D[i,])
  }
  p = 1:nrow(data)
  dist = sort(D_1[, k])
  data = data.frame(p,dist)
  ggplot(data, aes(x=p, y=dist)) +geom_point(shape=1) +
      geom_line() + geom_point(color = 'black')
}
kDBScan(crediti_totali_prg_arc_clustered, 6)
```

Grafico k-distanze con $k = 6$

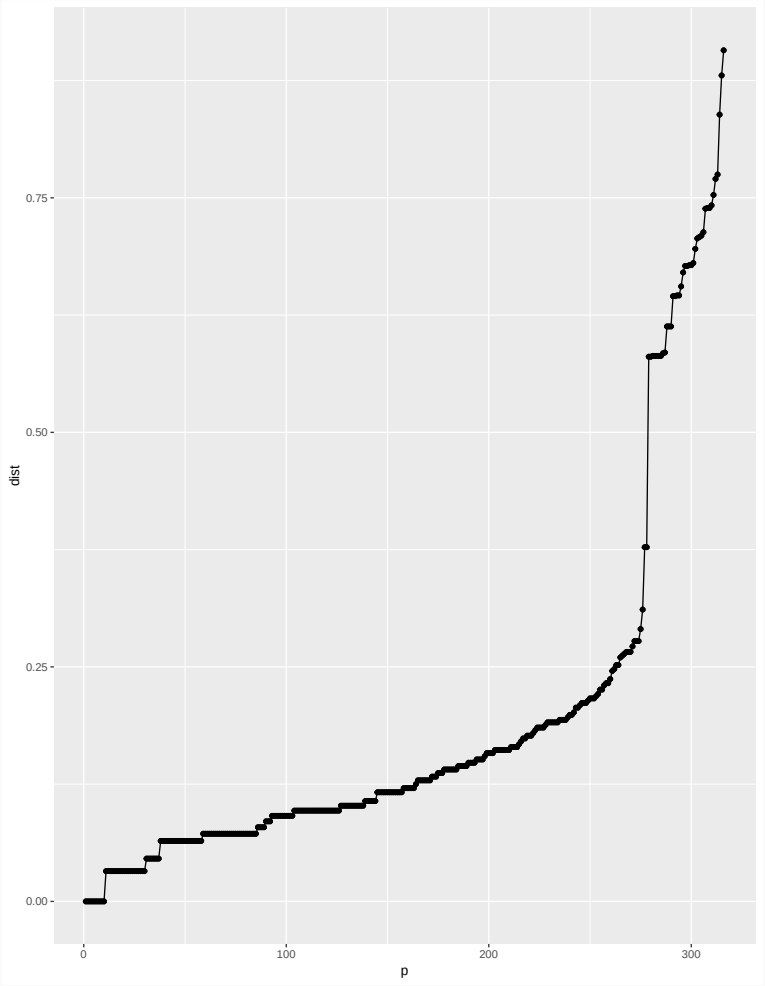


Figura 8: Andamento k-distanze con $k = 6$

Grafico k-distanze con $k = 10$

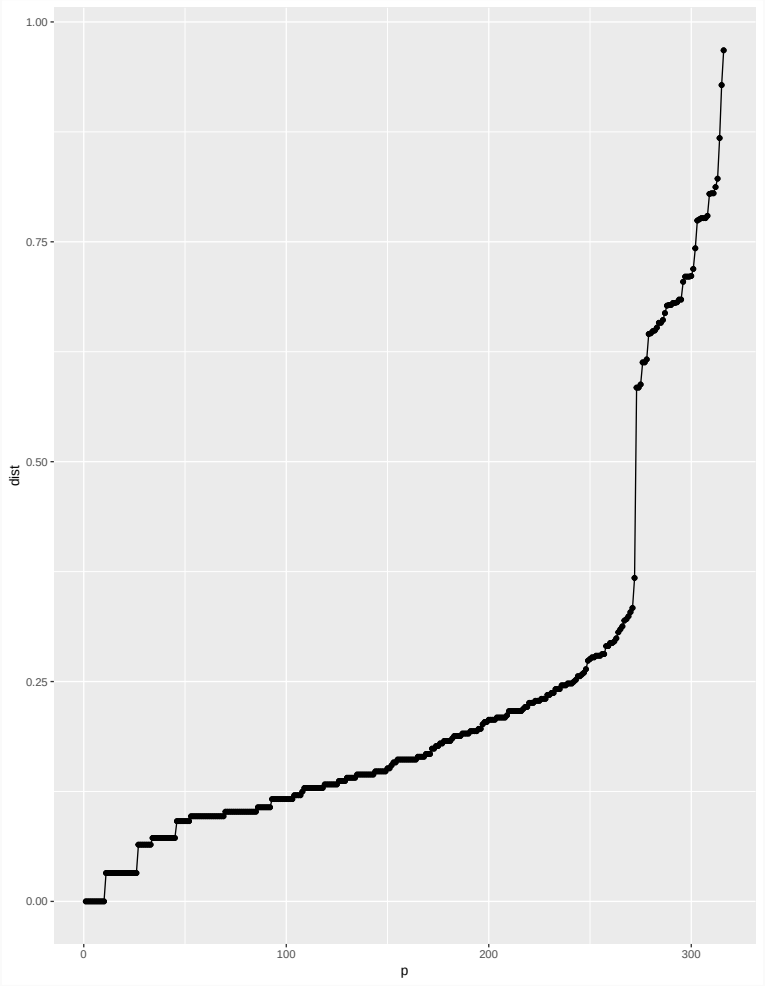


Figura 9: Andamento k-distanze con $k = 10$

Grafico k-distanze con $k = 20$

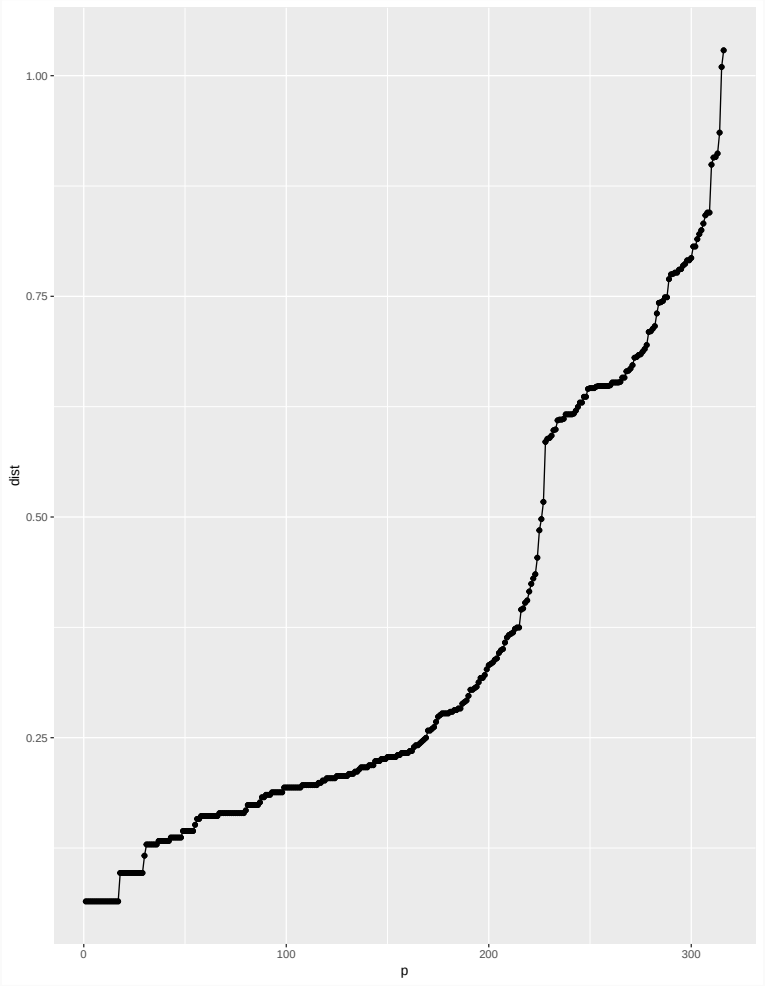


Figura 10: Andamento k-distanze con $k = 20$

Valori scelti dalla procedura

Tale procedura suggerisce quindi di scegliere i valori di eps nei seguenti intervalli

K	eps
6	[0.37,0.56]
10	[0.36,0.58]
20	[0.54,0.58]

Confronto tra clustering

Attributi analizzati	Algoritmo Migliore	Parametri ottimali	Correlazione
crediti totali, ARC, PRG	K-means	$k = 3$	-0.854
ARC, ASD, PRG, MDL, AN1	DBSCAN	`MinPts=20,eps=0.4`	-0.758
voto medio, test	K-means	$k = 2$	-0.476

Tabella 10: Confronto tra modelli.

Conclusioni

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;
- DBSCAN stabilisce un numero elevato di cluster;
- Analisi clustering gerarchico non svolta.

Grazie per l'attenzione