Cognome: ..... Nome: ..... Matricola: .....

Corso di laurea: ☐ SGI ☐ SSE Anno: ☐ II ☐ III o più

Prova scritta di ASM - Modulo Analisi Esplorativa del 10.02.2017

La durata della prova è di 60 minuti.
Si svolgano gli esercizi 1 e 2 riportando il risultato dove indicato.

Esercizio 1. Punteggio: 7

Alla matrice $\underset{n\times p}{X}$ sono associati i seguenti autovalori $\lambda_1 = 6, \lambda_2 = 4$ e autovettori normalizzati $\underset{2\times 1}{v_1} = \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$, $\underset{2\times 1}{v_2} = \begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix}$.

a. Quante sono le colonne di $\underset{n\times p}{X}$? $p = \ldots$

b. Determinare la matrice di varianze/covarianze $\underset{p\times p}{S} = \begin{bmatrix} \end{bmatrix}$

c. Riportare
   - varianza totale = ... e generalizzata = ...
   - l'indice di variabilità relativo (arrontondare al secondo decimale) = ...

d. Determinare $\underset{p\times p}{S^2} = \begin{bmatrix} \end{bmatrix}$

e. Calcolare la proporzione di varianza spiegata dalla prima componente principale

$= \ldots\ldots\ldots$

f. Calcolare la correlazione tra la prima colonna $\underset{n\times 1}{\tilde{x}_1}$ di $\underset{n\times p}{\tilde{X}}$ e i punteggi $\underset{n\times 1}{y_1}$ della prima componente principale, arrotondando al secondo decimale

$= \ldots\ldots\ldots$

g. Determinare (arrotondando al secondo decimale) la matrice di correlazione $\underset{p\times p}{R} = \begin{bmatrix} \end{bmatrix}$

h. Riportare (arrotondando al secondo decimale) gli autovalori e gli autovettori normalizzati di $\underset{p\times p}{R}$ calcolata al punto precedente.

```
##      [,1] [,2]
## [1,]  4.4  0.8
## [2,]  0.8  5.6
```

```
## [1] 24
```

```
## [1] 10
```

```
## [1] 0.97
```

```
##      [,1] [,2]
## [1,]   20    8
## [2,]    8   32
```

```
## [1] 0.6
```

```
## [1] 0.52
```

```
##      [,1] [,2]
## [1,] 1.00 0.16
## [2,] 0.16 1.00
```

```
## [1] 1.16 0.84
```

```
##      [,1]  [,2]
## [1,] 0.71 -0.71
## [2,] 0.71  0.71
```

Esercizio 2. Punteggio: 6.5

Si consideri il dataset `quakes` presente nella libreria `datasets`, che contiene $n = 1000$ osservazioni (eventi sismici) su cui sono state misurate le seguenti 5 variabili:

- *lat* latitudine dell'evento sismico
- *long* longitudine dell'evento sismico
- *depth* profondità (in km) dell'evento sismico
- *mag* magnitudo (scala Richter)
- *stations* Numero di stazioni che hanno riportato l'evento sismico

a. Si consideri la matrice $\underset{1000\times5}{X}$ che contiene le seguenti variabili: *lat*, *long*, *depth*, *mag* e *stations*. Si costruisca il diagramma a scatola con baffi (*boxplot*) per ciascuna delle variabili presenti in $\underset{1000\times5}{X}$ e si riporti il numero di valori anomali (*outliers*).

	lat	long	depth	mag	stations
numero di valori anomali					

```
##      lat     long    depth      mag stations
##       32      204        0        7       54
```

b. Per la matrice $\underset{1000\times5}{X}$ calcolata al punto a., si calcoli il quadrato della distanza di Mahalanobis di ciascuna unità statistica u_i' dal baricentro $\bar{x}'$ e si riporti il valore minimo e il valore massimo, arrotondando i calcoli al secondo decimale.

```
## [1] 0.57
```

```
## [1] 25.9
```

$\min_{i=1,\ldots,1000}\{d_M^2(u_i,\bar{x})\} = \ldots\ldots\ldots\ldots$ $\max_{i=1,\ldots,1000}\{d_M^2(u_i,\bar{x})\} = \ldots\ldots\ldots\ldots$

c. Utilizzare l'algoritmo delle K-medie (specificando `algorithm = "Lloyd"`) per formare $K = 4$ gruppi sulla base della matrice dei dati standardizzati $\underset{1000\times5}{Z}$ ottenuta a partire da $\underset{1000\times5}{X}$, inizializzando i centroidi con le osservazioni di riga 200, 400, 600 e 800, ed eseguendo l'algoritmo una sola volta. Riportare i valori dei centroidi dei 4 gruppi ottenuti, arrotondando alla seconda cifra decimale.

```
##     lat  long depth   mag stations
## 1 -0.15  0.23  0.02  1.61     1.82
## 2  0.95 -1.89 -0.76  0.22    -0.10
## 3 -0.51  0.70 -0.79 -0.26    -0.34
## 4  0.01  0.25  1.03 -0.59    -0.46
```

	Centroidi				
	lat	long	depth	mag	stations
Gruppo 1					
Gruppo 2					
Gruppo 3					
Gruppo 4					

d. Calcolare, arrotondando al secondo decimale, l'indice di Calinski and Harabasz per i quattro gruppi individuati al punto c.

```
## [1] 474.8148
```

Indice di Calinski and Harabasz =