Tecniche per risolvere il problema di Data Leakage

Mattia Manna

2024-12-13

Indice

1 Importazione dati, pulizia e filtraggio **1**

1.1 Importazione dati . . . 1

1.2 Pulizia dati . . . 1

1.3 Filtrare . . . 2

1.4 Prendere count GBM e SANI e filtrare di nuovo . . . 3

2 Calcolo dei DEGS sul TRAIN set **4**

2.0.1 DEGS su tutto il dataframe senza TMM . . . 4

2.1 Differential co-expressed network . . . 5

3 Metriche di centralità **6**

3.1 Istogrammi . . . 7

3.1.1 Degree . . . 7

3.1.2 Betweenness centrality . . . 8

3.1.3 Closeness centrality . . . 9

3.1.4 Eigenvector centrality . . . 10

3.1.5 Local clustering coefficient . . . 11

3.2 Betweenness centrality . . . 12

3.3 Closeness centrality . . . 13

3.3.1 Ultimo 5% (0.95) . . . 13

3.3.2 Primo 5 % (0.05) . . . 13

3.4 Eigenvector centrality . . . 14

3.5 Local Clustering coefficient . . . 14

```r
dataframe.formato.classificazione <- function(dataframe){

  # Trasporre il dataframe delle conte di test, in questo modo si hanno le conte
  # disposte nel modo giusto per la classificazione: pazienti sulle righe e geni sulle colonne
  dataframe <- as.data.frame(t(dataframe))

  # Aggiungere le label y (cancer sano) alle conte
  dataframe <- merge(dataframe,patients.summary,by.x="row.names",
                                    by.y="sample_name",all = FALSE)

  # Sistemare il dataframe dopo l'operazione di merge, rimettere i rownames al posto giusto
  rownames(dataframe) <- dataframe$Row.names

  # Eliminare la colonna rownames
  dataframe <- dataframe %>% select(-"Row.names")

  return(dataframe)
}
```

1 Importazione dati, pulizia e filtraggio

1.1 Importazione dati

Importare i dati salvati localmente.

```r
tep.expr <- read.delim("~/Desktop/Università/Magistrale/Tesi/R/GSE68086/GSE68086_TEP_data_matrix.txt", r
dim(tep.expr)
## [1] 57736   285

patients <- read.csv("~/Desktop/Università/Magistrale/Tesi/R/GSE68086/patients.csv", row.names=1, sep="
dim(patients)
## [1] 285  51

gene.info <- read.delim("~/Desktop/Università/Magistrale/Tesi/R/GSE68086/Human.GRCh38.p13.annot.tsv", ro
dim(gene.info)
## [1] 39376    17
```

1.2 Pulizia dati

Siccome ci sono delle incongruenze tra il nome dei sample nei dataframe:

- patients, dataset contenente le informazioni per ogni sample
- tep.expr, dataset contenente la conta per i geni

sistemare queste incongruenze.

```r
patients$sample_name <- sapply( patients$source_name_ch1,    # Vettore al quale applicare la funzione

                                # Funzione da applicare ad ogni valore nel vettore
```

```
                                function(x) gsub("-", ".", x)
                                )
colnames(tep.expr)<- sapply(colnames(tep.expr), function(x) sub("X", "", x))
```

1.3 Filtrare

Filtrare per la condizione: almeno un sample con count > 5

```
tep.expr.filtr <- tep.expr %>% filter_all(any_vars(. >= 5))
dim(tep.expr.filtr)
```

```
## [1] 16347   285
```

1.4 Prendere count GBM e SANI e filtrare di nuovo

```
# Specificare la condizione da analizzare
condizione.da.analizzare = "GBM"

# Estrarre MALATI con la condizione specificata
condition.samples <- patients[patients$cancer.type.ch1==condizione.da.analizzare,]$sample_name

# Dimensioni del samples (controllo per la correttezza dei risultati)
n.condition.samples <- length(condition.samples)
print(n.condition.samples)
## [1] 40



# Estrarre il nome/codice dei  sample SANI
healhty.samples <- patients[patients$cancer.type.ch1=="HC",]$sample_name

## Dimensioni del samples (controllo per la correttezza dei risultati)
length(healhty.samples)
## [1] 55


# Unire SANI e MALATI
tutti.samples <- c(condition.samples,healhty.samples)
length(tutti.samples)
## [1] 95


# Filtrare la matrice delle conte di nuovo perchè sono state selezionate solo alcune patologie
counts <- tep.expr.filtr[, tutti.samples]
counts <- counts %>% filter_all(any_vars(. >= 5))
dim(counts)
## [1] 14411    95
```

Estrarsi in sommario dei pazienti presi in considerazione.

```
patients.summary <- patients[patients$sample_name %in% colnames(counts),c("sample_name","cancer.type.ch
patients.summary[1:5,]
```

```
##               sample_name cancer.type.ch1
## GSM1662606      VU280.GBM             GBM
## GSM1662607 VU284.GBM.vIII             GBM
## GSM1662608 VU306.GBM.vIII             GBM
## GSM1662609   VU360.GBM.WT             GBM
## GSM1662610 VU372.GBM.vIII             GBM
```

2 Calcolo dei DEGS sul TRAIN set

2.0.1 DEGS su tutto il dataframe senza TMM

Numero degs su tutto il dataframe senza TMM (no divisione train e test).

```
nomiDEGS <- rownames(DEGs.metriche)
dim(DEGs.metriche)
```

```
## [1] 740   5
```

```
tep.expr.filtr.normDEGS <- normalized_counts[rownames(normalized_counts) %in%
                                                nomiDEGS,]

tep.expr.filtr.normDEGS <- dataframe.formato.classificazione(tep.expr.filtr.normDEGS)

# Esportare il dataframe
write.csv(tep.expr.filtr.normDEGS,
          file = "/Users/mattia/Desktop/Università/Magistrale/Tesi/Python Data Leakage GSE183635/Data/GE
          row.names =T )
```

2.1 Differential co-expressed network

```
DEGS.tep.expr.filtr.normalized.gbmVShc <- normalized_counts[rownames(normalized_counts) %in% rownames(DE
```

```
matrix.condition1 <- DEGS.tep.expr.filtr.normalized.gbmVShc[, condition.samples]
matrix.condition2 <- DEGS.tep.expr.filtr.normalized.gbmVShc[, healhty.samples]
Zscores.gbmVShc <- evaluate_zscores(matrix.condition1,matrix.condition2)

# Soglia
z_threshold <- 3

output.gbmVShc <- differential_coexpression_network(Zscores.gbmVShc,z_threshold)
```

```
## [1] 34  2
```

```
##                            gene degree
## ENSG00000013016 ENSG00000013016    159
## ENSG00000054267 ENSG00000054267     41
## ENSG00000070614 ENSG00000070614     39
## ENSG00000076924 ENSG00000076924     78
## ENSG00000089289 ENSG00000089289     49
## ENSG00000096060 ENSG00000096060     55
## ENSG00000099622 ENSG00000099622     51
## ENSG00000105287 ENSG00000105287     49
## ENSG00000111707 ENSG00000111707    105
## ENSG00000114439 ENSG00000114439     78
```

```
tep.expr.filtr.normHUBS <- normalized_counts[rownames(normalized_counts) %in%
                                                output.gbmVShc$hubs$gene,]

tep.expr.filtr.normHUBS <- dataframe.formato.classificazione(tep.expr.filtr.normHUBS)

# Esportare il dataframe
write.csv(tep.expr.filtr.normHUBS,
          file = "/Users/mattia/Desktop/Università/Magistrale/Tesi/Python Data Leakage GSE183635/Data/GE
          row.names =T )
```

3 Metriche di centralità

```
matrice.adiacenza <- output.gbmVShc$adjacency.matrix
matrice.adiacenza <- abs(matrice.adiacenza)
```

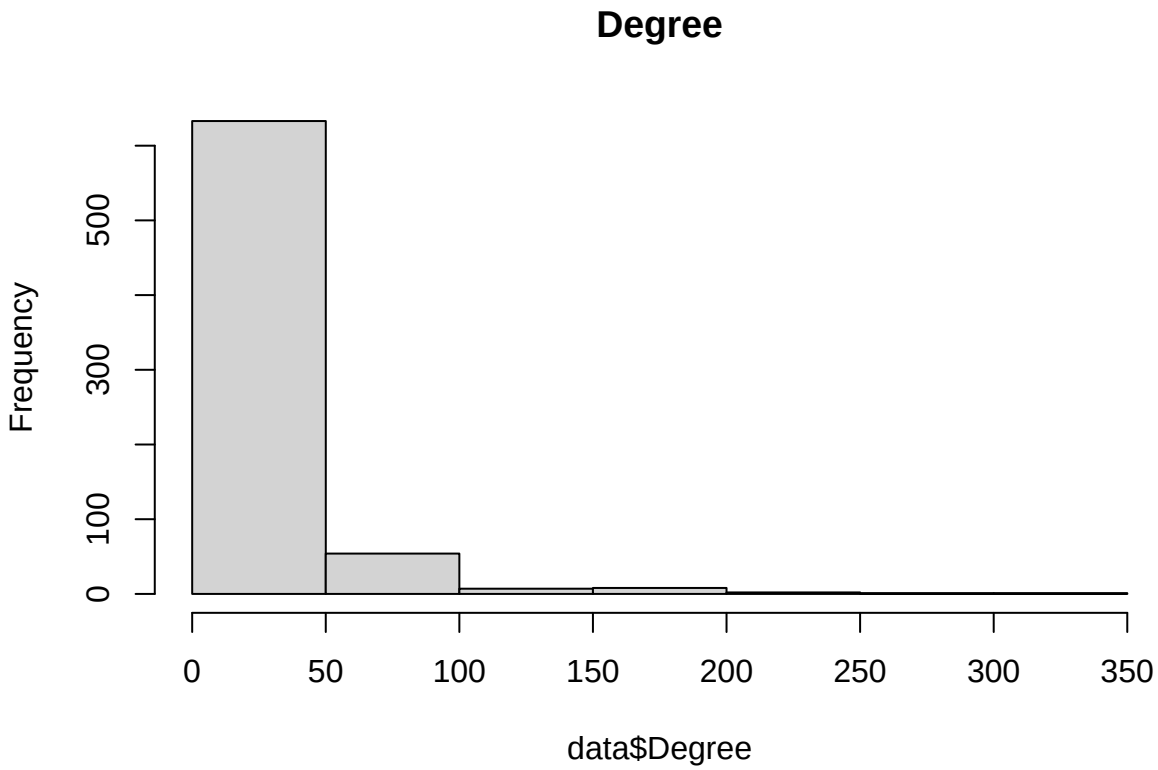
```
# Creare il network igraph dalla matrice di adiacenza
network <- graph_from_adjacency_matrix(matrice.adiacenza)

data <- data.frame(
  "Gene" = V(network)$name,
  "Degree" = degree(network),
  "betweenness_centrality"= betweenness(network),
  "closeness_centrality"=   closeness(network),
  "eigenvector_centrality" = eigen_centrality(network)$vector,
  "local_clustering_coefficient" =transitivity(network, type="local")
)
data[is.na(data)] <- 0
dim(data)
```

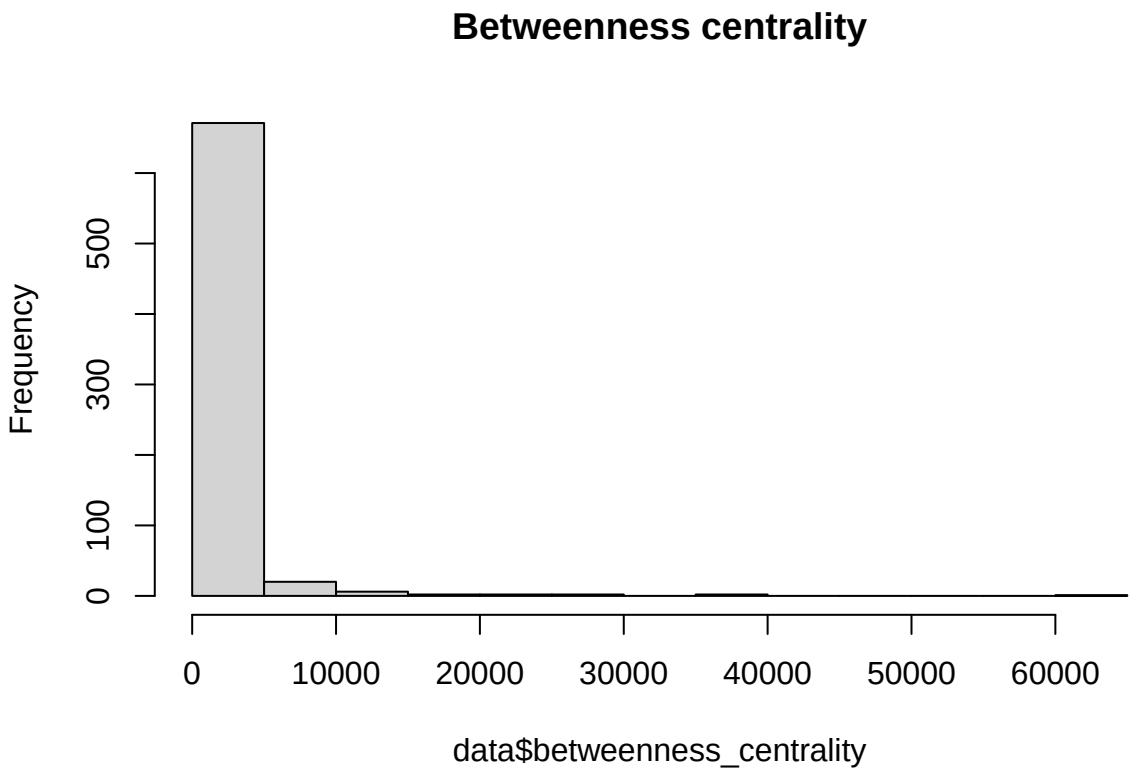
```
## [1] 706   6
```

3.1 Istogrammi

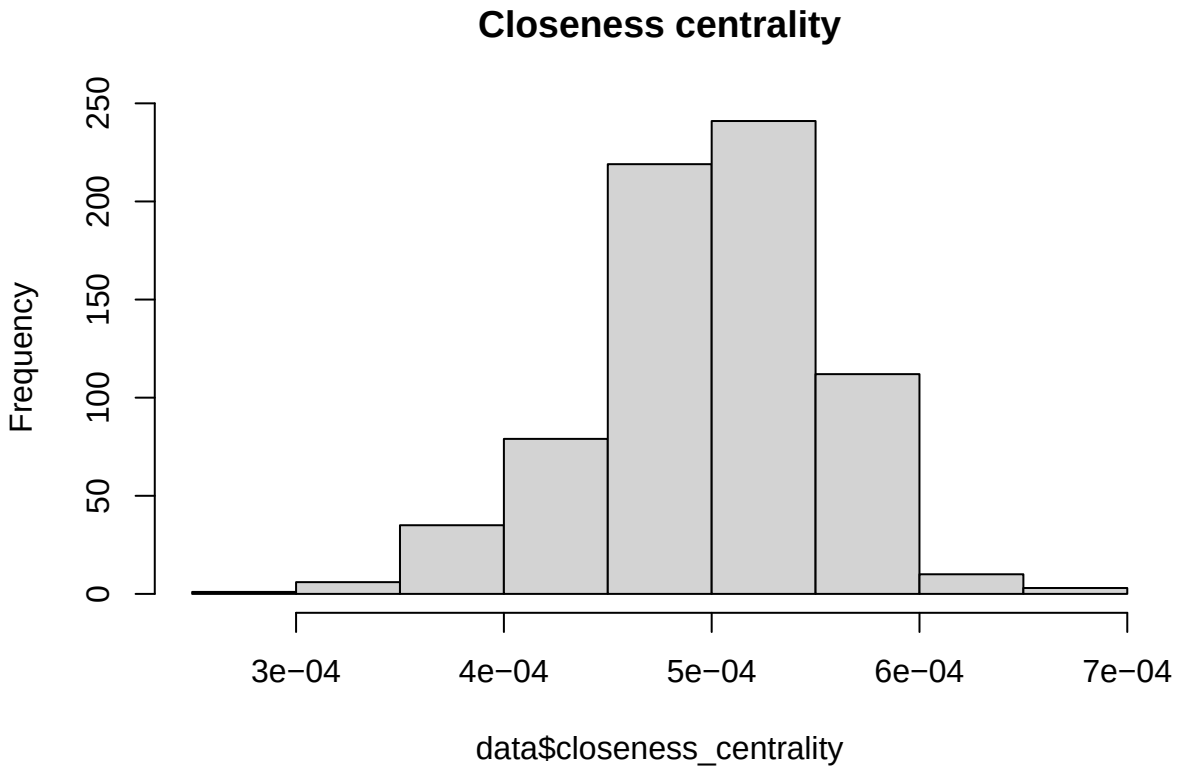
3.1.1 Degree



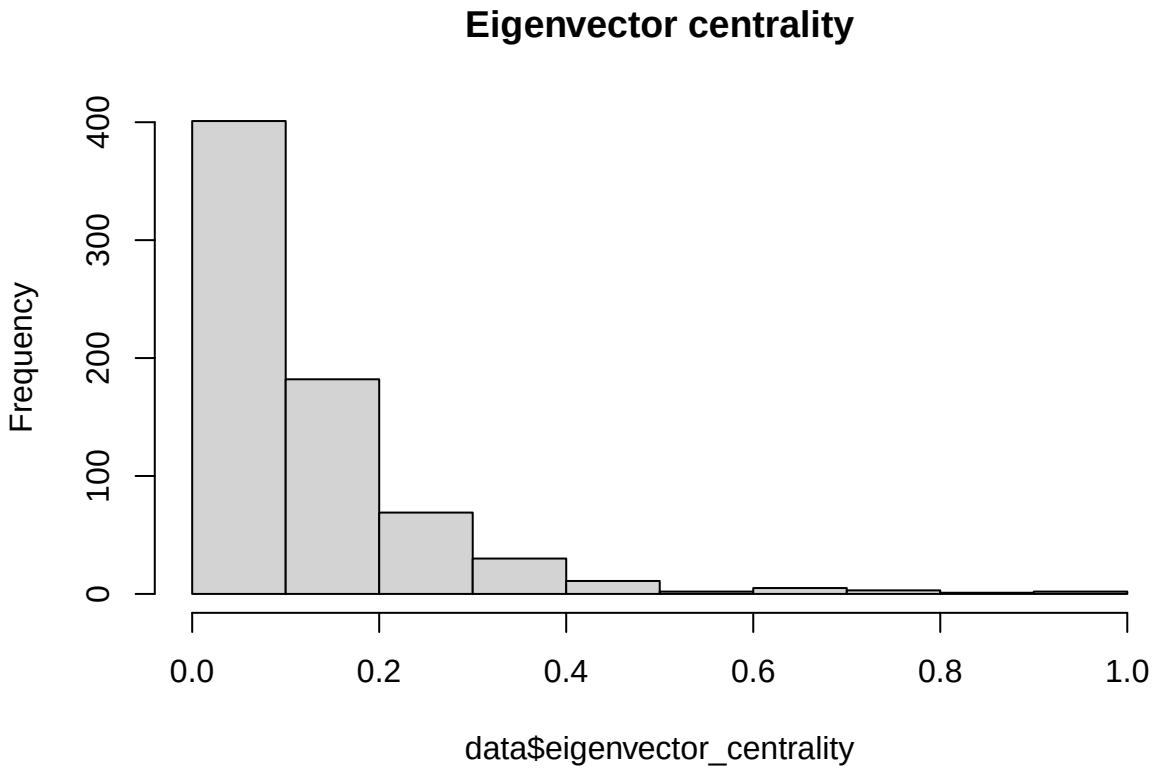
3.1.2 Betweenness centrality



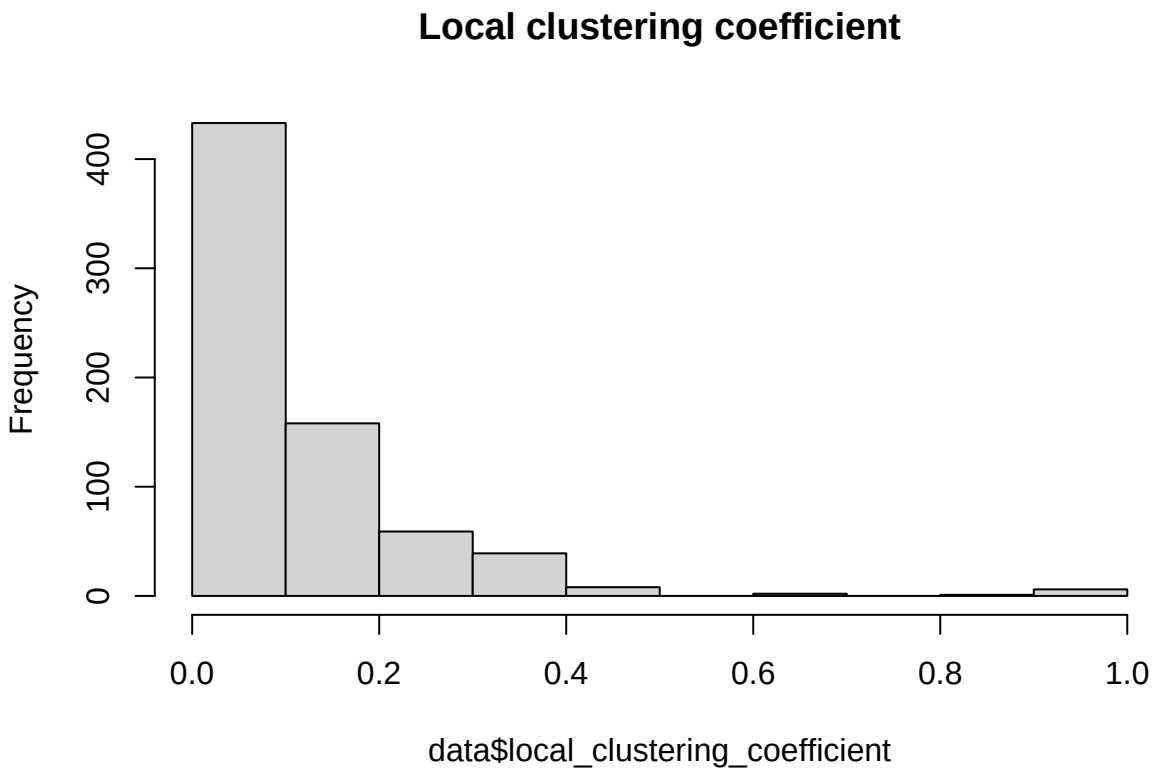
3.1.3 Closeness centrality



3.1.4 Eigenvector centrality



3.1.5 Local clustering coefficient



3.2 Betweenness centrality

```
soglia <- 0.95

metrica <- data$betweenness_centrality
q <- quantile(metrica, probs = c(soglia))

# Selezionare solo i geni con eigenvector_centrality superiore al 95% percentile
genes <- data[metrica > q,]$Gene

# Estrarre solo i geni con eigenvector_centrality superiore al 95% percentile
# dalla matrice delle conte filtrata e normalizzata
tep.expr.filtr.normBETWEENNESS <- normalized_counts[rownames(normalized_counts) %in%
                                                      genes,]
dim(tep.expr.filtr.normBETWEENNESS)
## [1] 36 95
nrow(tep.expr.filtr.normBETWEENNESS) # Numero geni selezionati tramite metrica
## [1] 36

tep.expr.filtr.normBETWEENNESS <- dataframe.formato.classificazione(tep.expr.filtr.normBETWEENNESS)
dim(tep.expr.filtr.normBETWEENNESS)
## [1] 95 37

# Esportare il dataframe
write.csv(tep.expr.filtr.normBETWEENNESS,
          file = "/Users/mattia/Desktop/Università/Magistrale/Tesi/Python Data Leakage GSE183635/Data/G
          row.names =T )
```

3.3 Closeness centrality

3.3.1 Ultimo 5% (0.95)

```
soglia <- 0.95

metrica <- data$closeness_centrality
q <- quantile(metrica, probs = c(soglia))

# Selezionare solo i geni con eigenvector_centrality superiore al 95% percentile
genes <- data[metrica > q,]$Gene

# Estrarre solo i geni con eigenvector_centrality superiore al 95% percentile
# dalla matrice delle conte filtrata e normalizzata
tep.expr.filtr.normCLOSENESS <- normalized_counts[rownames(normalized_counts) %in%
                                                    genes,]
dim(tep.expr.filtr.normCLOSENESS)
## [1] 36 95
nrow(tep.expr.filtr.normCLOSENESS) # Numero geni selezionati tramite metrica
## [1] 36

tep.expr.filtr.normCLOSENESS <- dataframe.formato.classificazione(tep.expr.filtr.normCLOSENESS)
dim(tep.expr.filtr.normCLOSENESS)
## [1] 95 37

# Esportare il dataframe
write.csv(tep.expr.filtr.normCLOSENESS,
          file = "/Users/mattia/Desktop/Università/Magistrale/Tesi/Python Data Leakage GSE183635/Data/GE
          row.names =T )
```

3.3.2 Primo 5 % (0.05)

Ha due code quindi considerare anche il primo 5%.

```
soglia <- 0.05
metrica <- data$closeness_centrality
q <- quantile(metrica, probs = c(soglia))
genes <- data[metrica < q,]$Gene
tep.expr.filtr.normCLOSENESS_last5 <- normalized_counts[rownames(normalized_counts) %in%
                                                    genes,]
dim(tep.expr.filtr.normCLOSENESS_last5)
## [1] 36 95
nrow(tep.expr.filtr.normCLOSENESS_last5) # Numero geni selezionati tramite metrica
## [1] 36
tep.expr.filtr.normCLOSENESS_last5 <- dataframe.formato.classificazione(tep.expr.filtr.normCLOSENESS_las
dim(tep.expr.filtr.normCLOSENESS_last5)
## [1] 95 37
# Esportare il dataframe
write.csv(tep.expr.filtr.normCLOSENESS_last5,
          file = "/Users/mattia/Desktop/Università/Magistrale/Tesi/Python Data Leakage GSE183635/Data/GE
          row.names =T )
```

3.4 Eigenvector centrality

```
soglia <- 0.95

metrica <- data$eigenvector_centrality
q <- quantile(metrica, probs = c(soglia))

# Selezionare solo i geni con eigenvector_centrality superiore al 95% percentile
genes <- data[metrica > q,]$Gene

tep.expr.filtr.normEIGEN <- normalized_counts[rownames(normalized_counts) %in%
                                                genes,]
dim(tep.expr.filtr.normEIGEN)
## [1] 36 95
nrow(tep.expr.filtr.normEIGEN) # Numero geni selezionati tramite metrica
## [1] 36

tep.expr.filtr.normEIGEN <- dataframe.formato.classificazione(tep.expr.filtr.normEIGEN)
dim(tep.expr.filtr.normEIGEN)
## [1] 95 37

# Esportare il dataframe
write.csv(tep.expr.filtr.normEIGEN,
          file = "/Users/mattia/Desktop/Università/Magistrale/Tesi/Python Data Leakage GSE183635/Data/G
          row.names =T )
```

3.5 Local Clustering coefficient

```
soglia <- 0.95

metrica <- data$local_clustering_coefficient
q <- quantile(metrica, probs = c(soglia))

# Selezionare solo i geni con eigenvector_centrality superiore al 95% percentile
genes <- data[metrica > q,]$Gene

tep.expr.filtr.normCLUSTER <- normalized_counts[rownames(normalized_counts) %in%
                                                genes,]
dim(tep.expr.filtr.normCLUSTER)
## [1] 27 95
nrow(tep.expr.filtr.normCLUSTER) # Numero geni selezionati tramite metrica
## [1] 27

tep.expr.filtr.normCLUSTER <- dataframe.formato.classificazione(tep.expr.filtr.normCLUSTER)
dim(tep.expr.filtr.normCLUSTER)
## [1] 95 28

# Esportare il dataframe
write.csv(tep.expr.filtr.normCLUSTER,
          file = "/Users/mattia/Desktop/Università/Magistrale/Tesi/Python Data Leakage GSE183635/Data/G
          row.names =T )
```