UAS SAINS DATA GENOM

Disusun oleh:

Diki Wahyudi 2106709131

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS INDONESIA

DESEMBER 2023

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

I PENDAHULUAN

Kanker dimulai ketika sel-sel dalam tubuh mulai tumbuh di luar kendali. Sel di hampir semua bagian tubuh bisa menjadi sel kanker, dan kemudian menyebar ke area lain di tubuh. Kanker dapat tumbuh di bagian mana saja dalam tubuh manusia. Salah satu daerah tubuh yang dapat ditumbuhi olek kanker yaitu prostat. Kanker prostat dimulai ketika sel-sel di kelenjar prostat mulai tumbuh di luar kendali. Prostat adalah kelenjar yang hanya ditemukan pada pria. Prostat berada di bawah kandung kemih (organ berongga tempat penyimpanan urin) dan di depan rektum (bagian terakhir dari usus). Tepat di belakang prostat terdapat kelenjar yang disebut vesikula seminalis yang menghasilkan sebagian besar cairan untuk air mani. Uretra, yaitu saluran yang membawa urin dan air mani keluar tubuh melalui penis, melewati bagian tengah prostat [1].

Hampir semua kanker prostat adalah *adenocarcinomas*. Kanker ini berkembang dari sel kelenjar (sel yang membuat cairan prostat yang ditambahkan ke air mani). Jenis kanker lain yang dapat bermula di prostat, yaitu *carcinomas* sel kecil, tumor *neuroendocrine* (selain *carcinomas* sel kecil), *carcinomas* sel transisi, *sarcomas*, dan lain-lain. Jenis kanker prostat lainnya, yang telah disebutkan sebelumnya, jarang terjadi. Jika seseorang diberi tahu bahwa seseorang tersebut mengidap kanker prostat, hampir pasti kanker itu adalah *adenocarcinomas*. Beberapa kanker prostat tumbuh dan menyebar dengan cepat, namun sebagian besar tumbuh lambat. Faktanya, penelitian otopsi menunjukkan bahwa banyak pria lanjut usia (dan bahkan beberapa pria lebih muda), yang meninggal karena sebab lain, juga menderita kanker prostat yang tidak pernah menyerang mereka selama hidup mereka. Dalam banyak kasus, baik mereka maupun dokter tidak mengetahui bahwa mereka mengidap penyakit tersebut.

Kanker yang tumbuh di bagian prostat dapat menyebar ke bagian tubuh lainnya. Kanker prostat metastatik merupakan kanker yang telah menyebar dari prostat ke bagian tubuh lain. Kanker tersebut terkadang disebut sebagai kanker prostat stadium lanjut. Penyakit ini paling sering menyebar ke kelenjar getah bening di bagian lain tubuh atau ke tulang. Kanker tersebut juga bisa menyebar ke organ lain, seperti paru-paru [3].

Penelitian ini menggunakan data ekspresi gen GDS2546. Dataset ini mengandung data ekspresi gen dari empat jaringan berbeda, yaitu *normal prostate tissue*, *normal prostate adjacent to tumor*, *primary prostate tumor*, dan *metastatic prostate tumor*. Penelitian ini dilakukan untuk menganalisis *metastatic prostate tumor* dan *primary prostate tumor*, serta jaringan donor normal dan jaringan normal yang berdekatan dengan tumor tersebut. Analisis akan dilakukan dengan menggunakan metode *differentially expressed genes*, *clustering*, dan *classification*. Metode LIMMA akan digunakan untuk menganalisis perbedaan gen. Selain itu, berbagai macam metode pengelompokkan dan klasifikasi akan dicoba untuk menganalisis data gen tersebut. Hasil dari analisis ini diharapkan dapat memberikan wawasan tentang mekanisme molekuler yang mendasari proses metastasis.

II METODE ANALISIS

Linear Models for Microarray Data (LIMMA)

Saat melakukan uji t, kita mengesitimasi variansi setiap gen secara individu. Hal tersebut baik-baik saja jika kita memiliki cukup banyak replikasi. Namun, dengan jumlah replikasi yang sedikit (katakanlah 2–5 per kelompok), estimasi variansinya akan sangat tinggi variasinya. Dalam statistik t yang dimoderasi, estimasi varian spesifik gen s_g^2 digantikan oleh rata-rata berbobot s_g^2 dan s_0^2, yaitu estimasi variansi global yang diperoleh dari pengumpulan (*pooled*) semua gen. Hal tersebut memberikan interpolasi antara uji t dan kriteria *fold-change*. Salah satu contohnya yaitu `limma`. LIMMA pada dasarnya merupakan modifikasi dari model linier yang digunakan khusus untuk analisis ekspresi gen atau protein.

LIMMA adalah *library* untuk analisis data ekspresi gen microarray, khususnya digunakan pada model linier untuk menganalisis eksperimen yang dirancang dari *differential expression*. LIMMA memberikan kemampuan untuk menganalisis perbandingan antara banyak target RNA secara bersamaan dalam eksperimen yang dirancang rumit dan *random*. Metode empiris Bayesian digunakan untuk memberikan hasil yang stabil meskipun jumlah arraynya sedikit. Model linier dan fungsi ekspresi diferensial berlaku untuk semua teknologi ekspresi gen, termasuk microarray, RNA-seq, dan PCR kuantitatif [7].

Clustering

Analisis pengelompokkan (*clustering*) bertujuan untuk mengelompokkan observasi berdasarkan karakteristik tertentu sedemikian sehingga observasi dalam suatu kelompok lebih mirip (similar atau homogen), sedangkan antarklaster (kelompok) berbeda antara satu dengan lainnya berdasarkan karakteristik yang sama. Dengan kata lain, observasi dalam satu grup berbeda dengan observasi pada klaster atau kelompok lain.

Definisi similar atau homogen tergantung dari tujuan penelitian. Ukuran kesamaan/kemiripan yang biasa digunakan adalah fungsi jarak Euclid. Pendekatan *clustering* yang dapat dilakukan terdapat 3 cara yaitu melalui partisi, hierarki, atau *density-based*. Dalam partisi, dibuat partisi dan evaluasi berdasarkan kriteria tertentu, misalnya meminimalkan *sum of square errors*. Contoh metode yang menggunakan partisi, yaitu k-means, k-medoids, dan CLARANS.

Dalam teknik hierarki, dibuat struktur *hierarchical* menggunakan kriteria tertentu. Dalam metode hierarki, terdapat dua tipe dasar kalster yaitu *agglomerative* (pemusatan) dan *divisive* (penyebaran). Dalam metode *agglomerative*, setiap objek atau observasi dianggap sebagai sebuah klaster tersendiri. Dalam tahap selanjutnya, dua klaster yang mempunyai kemiripan digabungkan menjadi sebuah klaster baru, demikian seterusnya. Sebaliknya, dalam metode *divisive*, kita beranjak dari sebuah klaster besar yang terdiri dari semua objek atau observasi. Selanjutnya, objek atau observasi yang paling tinggi nilai ketidakmiripannya dipisahkan, demikian seterusnya. Ukuran *linkage* yang biasa digunakan dalam metode hierarki yaitu

1. *Single Linkage (Nearest Neighbor)*

$$D(A,B) = \min\{\boldsymbol{y}_i, \boldsymbol{y}_j, \text{ untuk } \boldsymbol{y}_i \text{ dalam } A \text{ dan } \boldsymbol{y}_j \text{ dalam } B\}$$

2. *Complete Linkage (Farthest Neighbor)*

$$D(A,B) = \max\{\boldsymbol{y}_i, \boldsymbol{y}_j, \text{ untuk } \boldsymbol{y}_i \text{ dalam } A \text{ dan } \boldsymbol{y}_j \text{ dalam } B\}$$

3. *Average Linkage*

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(\boldsymbol{y}_i, \boldsymbol{y}_j)$$

4. *Centroid*

$$D(A, B) = d(\bar{\boldsymbol{y}}_i, \bar{\boldsymbol{y}}_j)$$

Nonhierarki merupakan kebalikan dari metode hierarki. Metode nonhierarki tidak meliputi proses "*treelike construction*". Metode ini justru menempatkan objek-objek ke dalam klaster sekaligus sehingga terbentuk sejumlah klaster tertentu. Langkah pertama adalah memilih sebuah klaster sebagai pusat klaster awal, dan semua objek dalam jarak tertentu ditempatkan pada kalster yang terbentuk. Kemudian, dipilih kalster selanjutnya dan penempatan dilanjutkan sampai semua objek ditempatkan. Objek-objek bisa ditempatkan lagi jika jaraknya lebih dekat pada klaster lain daripada klaster asalnya. Dalam pendekatan *density-based*, dibuat berdasarkan *connectivity* dan *density functions*. Contoh metode yang menggunakan *density-based*, yaitu DBSACN, OPTICS, dan DenClue.

Biclustering

Seperti pada pendekatan analisis pengelompokkan (*clustering*), kita memulai algoritma dengan $n \times m$ matriks data $\boldsymbol{A}$:

	y_1	$\cdots$	y_i	$\cdots$	y_m
x_1	a_{11}	$\cdots$	a_{i1}	$\cdots$	a_{m1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
x_j	a_{1j}	$\cdots$	a_{ij}	$\cdots$	a_{mj}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
x_n	a_{1n}	$\cdots$	a_{in}	$\cdots$	a_{mn}

dengan objek X, variabel Y, dan entri a_{ij}. Tujuan dari analisis *biclustering* adalah untuk menemukan subgrup $\boldsymbol{A}_{IJ}$ dari objek $I = i_1, ..., i_k, k \leq n, I \subset X$ yang semirip mungkin satu sama lain pada subset variabel $J = j_1, ..., j_l, l \leq m, J \subset Y$ dan sebisa mungkin berbeda dengan objek dan variabel lainnya. *Biclsuter* z kemudian didefinsikan sebagai $BC_z = (I_z, J_z) = A_{I_z J_z}$.

Situasi umum untuk menghitung *bicluster* adalah ketika dataset berdimensi tinggi dengan banyak variabel sehingga algoritma *cluster* normal memberikan hasil yang tersebar karena banyak variabel yang tidak berkorelasi. *Biclustering* juga berguna jika terdapat asumsi koneksi objek dan beberapa variabel dalam dataset, misalnya beberapa objek memiliki pola "mirip" untuk sekumpulan variabel tertentu.

Sama seperti pengelompokan (*clustering*) tradisional, ada banyak kemungkinan untuk menghitung kesamaan dalam *bicluster*. Madeira dan Oliveira (2004) mengidentifikasi empat kelompok utama struktur di dalam submatriks:

1. *Bicluster* dengan nilai konstan:

$$a_{ij} = \mu$$

2. *Bicluster* dengan nilai konstan pada baris atau kolom:

$$(a_{ij} = \mu + \alpha_i \text{ atau } a_{ij} = \mu \times \alpha_i) \text{ dan } (a_{ij} = \mu + \beta_j \text{ atau } a_{ij} = \mu \times \beta_j)$$

3. *Bicluster* dengan nilai koheren:

$$a_{ij} = \mu + \alpha_i + \beta_j \text{ atau } a_{ij} = \mu \times \alpha_i \times \beta_j$$

4. *Bicluster* dengan evolusi koheren.

$$a_{ih} \leq a_{ir} \leq a_{it} \leq a_{id} \text{ atau } a_{hj} \leq a_{rj} \leq a_{tj} \leq a_{dj}$$

Dalam kasus yang paling sederhana, algoritma ini mampu menemukan subset dari baris dan kolom dengan nilai konstan. Metode yang sedikit ditingkatkan dapat mengidentifikasi *bicluster* dengan nilai konstan pada baris maupun nilai konstan pada kolom. Pendekatan lainnya mencari nilai yang koheren pada kolom atau baris ekspresi matriks. Artinya, setiap kolom atau baris dapat dihitung hanya dengan menjumlahkan atau mengalikan konstanta. Tipe selanjutnya bertujuan untuk menemukan *bicluster* dengan koheren evolusi. Dengan kata lain, nilai numerik pasti dari elemen matriks tidak menjadi masalah. Sebaliknya, algoritma mencari subset kolom dan baris dengan perilaku yang koheren. Jelas bahwa kasus ini disertai dengan hilangnya informasi karena matriks harus didiskritisasi karena nilai numerik pasti dari matriks tidak menjadi masalah [6].

Classification

Klasifikasi adalah teknik dalam data science yang digunakan oleh data scientist untuk mengkategorikan data ke dalam sejumlah kelas tertentu. Teknik ini dapat dilakukan pada data terstruktur atau tidak terstruktur dan tujuan utamanya adalah untuk mengidentifikasi kategori atau kelas tempat data baru akan berada. Teknik ini juga memiliki algoritma yang dapat digunakan untuk mengaktifkan perangkat lunak analisis teks untuk melakukan tugas seperti menganalisis sentimen berbasis aspek dan mengkategorikan teks yang tidak terstruktur berdasarkan topik dan polaritas opini. Dalam metode klasifikasi, terdapat berbagai jenis algoritma. Contoh alogiritma-algoritma dalam klasifikasi, yaitu sebagai berikut.

1. *Classifier* Linier

   Dalam machine learning, tugas utama klasifikasi statistik adalah menggunakan karakteristik objek untuk menemukan kelasnya. Tugas ini dicapai dengan membuat keputusan klasifikasi berdasarkan nilai kombinasi linier dari karakteristik. Contoh algoritma klasifikasi linier yaitu sebagai berikut.

   - Regresi Logistik
     Regresi logistik adalah proses pemodelan probabilitas hasil diskrit dengan diberikan variabel masukan (*input*). Regresi logistik yang paling umum memodelkan hasil biner, yaitusesuatu yang dapat berupa dua nilai, seperti benar/salah, ya/tidak, dan seterusnya.
   - *Classifier* Naive Bayes
     Naive Bayes adalah teknik klasifikasi berdasarkan teorema Bayes dengan asumsi independensi antarprediktor. Secara sederhana, pengklasifikasi (*classifier*) naive Bayes mengasumsikan bahwa keberadaan fitur tertentu dalam suatu kelas tidak terkait dengan keberadaan fitur lainnya. Hal tersebut memperbarui pengetahuan selangkah demi selangkah dengan menggunakan informasi baru.
   - Diskriminan Linier Fisher
     Diskriminan linier Fisher dapat digunakan sebagai pengklasifikasi dalam *supervised learning*. Dengan adanya data berlabel, pengklasifikasi dapat menemukan sekumpulan bobot untuk menarik batas keputusan, mengklasifikasikan data. Diskriminan linier Fisher berupaya menemukan vektor yang memaksimalkan pemisahan antarkelas data yang diproyeksikan.

2. *Support Vector Machine* (SVM)

   SVM adalah algoritma *supervised learning* yang digunakan untuk klasifikasi dan analisis regresi. Dalam SVM, setiap item data diplot sebagai titik dalam ruang n-dimensi dengan nilai atribut masing-masing, yaitu nilai koordinat tertentu.

3. *Classifier* Kuadrat

Algoritma klasifikasi kuadrat didasarkan pada teorema Bayes. Algoritma pengklasifikasi ini berbeda dalam pendekatannya untuk klasifikasi dari regresi logistik. Dalam regresi logistik, dimungkinkan untuk menurunkan probabilitas pengamatan secara langsung untuk suatu kelas ($Y = k$) untuk pengamatan tertentu ($X = x$). Namun, dalam klasifikasi kuadrat, pengamatan dilakukan dalam dua langkah berikut.

(a) Pada langkah pertama, identifikasi distribusi input X untuk setiap grup atau kelas.

(b) Setelah itu, distribusi dibalik dengan bantuan teorema Bayes untuk menghitung probabilitas.

4. Estimasi Kernel

Estimasi kernel adalah cara nonparametrik untuk memperkirakan *Probability Density Function* (PDF) dari variabel acak kontinu. Teknik tersebut merupakan teknik nonparametrik karena mengasumsikan tidak ada distribusi implisit untuk variabel. Pada dasarnya, pada setiap datum, sebuah fungsi kernel dibuat dengan datum sebagai pusatnya. Hal ini memastikan bahwa kernel simetris terhadap datum. PDF kemudian diestimasi dengan menambahkan semua fungsi kernel ini dan membaginya dengan jumlah data untuk memastikannya memenuhi dua properti PDF:

(a) setiap kemungkinan nilai PDF harus nonnegatif;

(b) integral tetap dari PDF pada daerah asalnya harus sama dengan 1.

5. *k-Nearest Neighbors*

k-NN (*k-Nearest Neighbors*) merupakan algoritma *classifier* yang pembelajarannya didasarkan pada kesamaan data (vektor) satu dengan lainnya. k-NN juga dapat digunakan untuk menyimpan semua kasus yang tersedia dan mengklasifikasikan kasus baru berdasarkan ukuran kesamaan (misalnya fungsi jarak).

6. *Decision Tree* (Pohon Keputusan)

Algoritma *decision tree* termasuk dalam algoritma *supervised learning*. Algoritma ini dapat digunakan untuk menyelesaikan regresi dan masalah klasifikasi lainnya. *Decision tree*membangun model klasifikasi atau regresi dalam bentuk struktur pohon. *Decision tree* memecah dataset menjadi subset yang lebih kecil dan lebih kecil sementara pada saat yang sama pohon keputusan terkait dikembangkan secara bertahap. Tujuan penggunaan algoritma *decision tree* adalah untuk memprediksi kelas atau nilai variabel target dengan mempelajari aturan keputusan sederhana yang disimpulkan dari data sebelumnya.

7. *Random Forest*

Random forest adalah metode pembelajaran ensemble untuk klasifikasi, regresi, dan tugas lain yang beroperasi dengan membangun beberapa pohon keputusan pada waktu pelatihan. Untuk tugas klasifikasi, *output* dari *random forest* adalah kelas yang dipilih oleh sebagian besar pohon. Untuk tugas regresi, prediksi mean atau mean dari setiap pohon merupakan *output*-nya. *Random forest* umumnya mengungguli *decision tree*, tetapi memiliki akurasi yang lebih rendah daripada pohon yang ditingkatkan dengan gradien. Namun, karakteristik data dapat mempengaruhi kinerjanya.

8. *Neural Network*

Neural network adalah sekumpulan algoritma yang berupaya mengidentifikasi hubungan mendasar dalam kumpulan data melalui proses yang meniru cara kerja otak manusia. Dalam data science, *neural network* membantu mengelompokkan dan mengklasifikasikan hubungan yang kompleks. Neural network dapat digunakan untuk mengelompokkan data yang tidak berlabel sesuai dengan kesamaan di antara input contoh dan mengklasifikasikan data ketika telah memiliki kumpulan data berlabel untuk dilatih.

III HASIL DAN PEMBAHASAN

Preprocessing

Import Library

```
library(Biobase)
library(GEOquery)
library(knitr)
library(kableExtra)
library(dplyr)
library(genefilter)
library(factoextra)
library(hgu95b.db)
library(ggplot2)
library(limma)
library(cluster)
library(caret)
library(MASS)
library(class)
library(party)
library(randomForest)
library(glmnet)
library(tidyverse)
library(biclust)
library(annotate)
library(GO.db)
library(clusterProfiler)
source("D:/Materi Kuliah UI/Sains Data Genom/CreateRadialPlot.R")
```

Import Data

Akan digunakan data dari NCBI dengan kode GDS2546. Dataset ini digunakan untuk menganalisis *metastatic prostate tumor* dan *primary prostate tumor*, serta jaringan donor normal dan jaringan normal yang berdekatan dengan tumor tersebut. Dataset tersebut terdiri dari 12620 gen (*features*) yang berasal dari 167 sampel. Chip menggunakan *platform* GPL92, dengan anotasi HG_U95B. Sampel dalam dataset tersebut adalah Homo sapiens (manusia).

```
dtgeo <- getGEO('GDS2546', destdir = ".")
```

Selanjutnya, dataset diubah ke *expression set* agar bisa diproses lebih lanjut lagi ke *phenotype set*.

Expression Set

Fungsi `GDS2eSet` digunakan untuk mengambil struktur data GDS (*Genomic Data Sharing*) dari `getGEO` dan menyesuaikannya ke dalam limma MALists atau Expression Sets. GDS ditransformasikan menjadi $\log_2$ sebelum dimasukkan ke dalam struktur data baru.

```
eset <- GDS2eSet(dtgeo, do.log2 = TRUE)
```

Phenotype Set

Akan diakses data fenotipik dan metadata yang terkait dengan eksperimen.

```
phdtgeo <- pData(eset)
```

Expression Data

Selanjutnya, akan diakses data ekspresi gen yang disimpan dalam objek yang berasal dari kelas `eSet`.

```
expdtgeo <- exprs(eset)
dim(expdtgeo)
```

```
[1] 12620   167
```

```
kable(expdtgeo[1:10, 1:6], format = "latex", booktabs = TRUE,
      align = rep("c", 6), caption = "\\textit{Head} dari expdtgeo") %>%
  kable_styling(position = "center", latex_options = c("HOLD_position"))
```

Tabel 1: *Head* dari expdtgeo

	GSM152822	GSM152823	GSM152824	GSM152825	GSM152826	GSM152827
41880_at	5.488644	6.951867	4.053111	6.704595	6.461070	6.504620
41881_at	4.847997	5.951867	4.129283	5.385431	6.221104	7.422065
41882_at	2.432959	2.378512	5.560715	2.378512	2.700440	2.944858
41883_at	3.137503	3.185866	4.169925	2.608809	3.954196	4.683696
41884_at	3.185866	5.514122	4.852998	4.112700	5.289097	5.048759
41885_at	5.686501	4.770829	6.033423	5.532940	3.887525	5.738768
41886_r_at	7.311067	8.715276	8.662846	8.406418	8.076281	7.366322
41887_at	3.017922	5.289097	6.102238	2.584963	4.683696	3.336283
41888_at	5.491853	7.066089	5.409391	6.773469	5.426265	6.427941
41889_at	4.877744	4.517276	5.307429	4.683696	4.852998	3.897240

```
annotation(eset) <- "hgu95b"
```

Gene Filtering

Selanjutnya, akan dilakukan *gene filtering*. Hal tersebut dilakukan untuk mengeluarkan gen-gen yang tidak banyak bervariasi antarsampel, memiliki ekspresi yang kecil di seluruh sampel, dan juga gen yang tidak memiliki cukup anotasi. Proses ini dilakukan agar dapat mengurangi terjadinya *false positif* (kesalahan tipe I yaitu $\alpha = \Pr(\text{Menolak } H_0 | H_0 \text{ benar})$) yang akan meningkatkan *power* (peluang hasil uji statistik untuk bebas dari kesalahan statistik tipe II) dari analisis data ini.

```
esetFilt <- nsFilter(eset)
# Hasi filtering
esetFilt
```

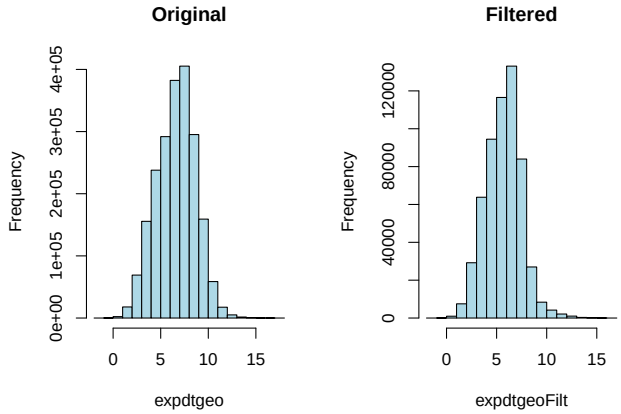
```
$eset
ExpressionSet (storageMode: lockedEnvironment)
assayData: 3430 features, 167 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM152822 GSM152823 ... GSM152905 (167 total)
  varLabels: sample tissue description
  varMetadata: labelDescription
featureData
  featureNames: 51020_at 47789_at ... 45662_at (3430 total)
  fvarLabels: ID Gene title ... GO:Component ID (21 total)
  fvarMetadata: Column labelDescription
experimentData: use 'experimentData(object)'
  pubMedIds: 17430594
Annotation: hgu95b

$filter.log
$filter.log$numDupsRemoved
[1] 2288

$filter.log$numLowVar
[1] 3430

$filter.log$numRemoved.ENTREZID
[1] 3453

$filter.log$feature.exclude
[1] 19
```

Akhirnya, ekspresi gen yang sudah di-*filter* bisa diekstraksi untuk diproses lebih lanjut.

```r
# Extract the expression of the filtered dataset
expdtgeoFilt <- exprs(esetFilt$eset)
```

Setelah tahap tersebut, bisa dilakukan *plotting* untuk melihat perbedaan sebelum dan sesudah *filtering*.

```r
par(mfrow = c(1, 2))
hist(expdtgeo, main = "Original", col = "lightblue")
hist(expdtgeoFilt, main = "Filtered", col = "lightblue")
```



Gambar 1: Grafik Perbandingan Ekspresi Gen Sebelum dan Sesudah *Filtering*

```
par(mfrow = c(1, 1))
```

Dari hasil di atas, didapatkan informasi bahwa sebelum dilakukan *filtering* terdapat 12620 *features/gene* dan setelah *filtering* tersisa 3430 gen. Gen yang disaring merupakan gen yang memiliki ekspresi rendah seperti terlihat pada histogram di atas.

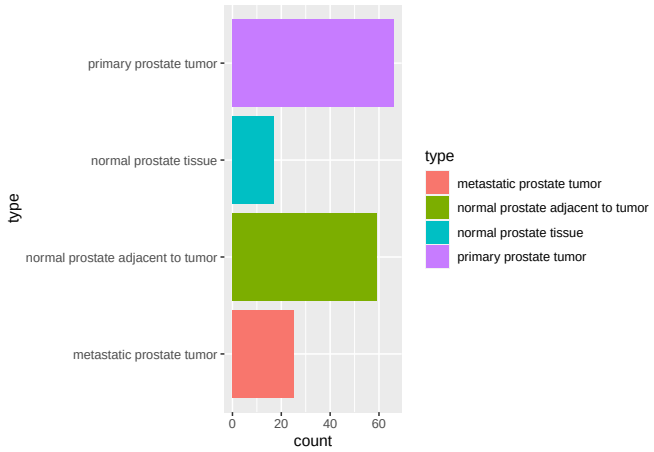
Exploratory Data Analysis (EDA)

Faktor dalam data ekspresi gen ini merupakan jaringan yang terdiri atas 4 jenis, yaitu *normal prostate tissue*, *normal prostate adjacent to tumor*, *primary prostate tumor*, dan *metastatic prostate tumor*.

```
vargrp <- phdtgeo[, 2]
table(vargrp)
```

```
vargrp
        metastatic prostate tumor normal prostate adjacent to tumor
                               25                                 59
           normal prostate tissue            primary prostate tumor
                               17                                 66
```

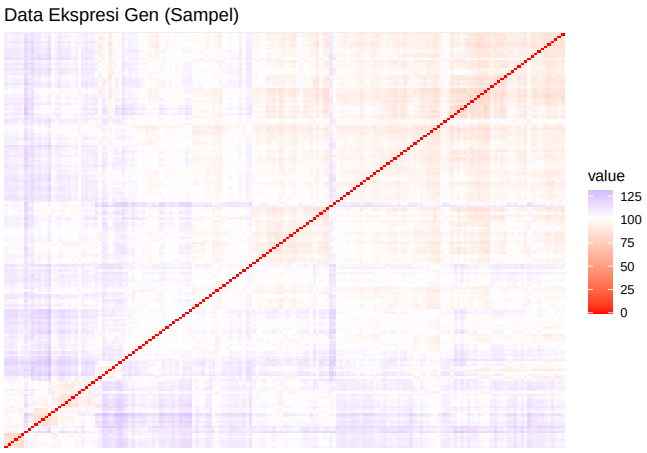
```
ggplot(data.frame(type = vargrp), aes(y = type, fill = type)) + geom_bar()
```



Gambar 2: *Bar Plot* Jumlah Kategori *Tissue*

Dalam data ini, terlihat bahwa *primary prostate tumor* merupakan jenis *tissue* (jaringan) terbanyak, dengan persentase sekitar 39.52%. Selanjutnya, akan dibuat *dissimilarity matrix plot* dari `t(expdtgeoFilt)`, yaitu *transpose* dari `expdtgeoFilt` agar ukuran matriks disimilaritasnya kecil.

```
dist.eucl <- dist(t(expdtgeoFilt), method = "euclidean")
fviz_dist(dist.eucl, show_labels = FALSE) + labs(title = "Data Ekspresi Gen (Sampel)")
```



Gambar 3: *Dissimilarity Matrix Plot* dari t(expdtgeoFilt)

Dari warna grafik tersebut, terlihat bahwa makin ke kanan atas, jarak antara ekspresi gen sampel makin dekat.

Linear Models for Microarray Data (LIMMA)

Pada bagian ini, akan dilakukan analisis LIMMA pada keempat jenis jaringan kanker prostat. Pertama, akan dibuat matriks model (*design matrix*) berdasarkan jenis jaringan. Dalam model matriks, akan digunakan `+ 0` dalam model yang menyetel intersep ke 0 sehingga efek jaringan menangkap ekspresi untuk grup tersebut, bukan perbedaan dari grup terhadap *base level*.

```
# Create the design matrix
des_mat <- model.matrix(~ vargrp + 0)
colnames(des_mat) <- c("mpt", "npat", "nptissue", "ppt")
head(des_mat, 15)
```

```
   mpt npat nptissue ppt
1    0    0        1   0
2    0    0        1   0
3    0    0        1   0
4    0    0        1   0
5    0    0        1   0
6    0    0        1   0
7    0    0        1   0
8    0    0        1   0
9    0    0        1   0
10   0    0        1   0
11   0    0        1   0
12   0    0        1   0
13   0    0        1   0
14   0    0        1   0
15   0    0        1   0
```

Selanjutnya, akan dilakukan *fitting* model *differential expression* pada data. Model LIMMA untuk data ini yaitu

$$\boldsymbol{Y}_j^T = (Y_{1j}, ..., Y_{N_j})$$
$$E(\boldsymbol{Y}_j) = \boldsymbol{X}\boldsymbol{\beta}_j$$

di mana $\boldsymbol{Y}_j$: ekspresi gen, $\boldsymbol{X}$: matriks model (*design matrix*) yang *full rank* (misalnya kondisi grup), dan $\boldsymbol{\beta}_j$: vektor efek dari kolom di matriks $\boldsymbol{X}$, $\boldsymbol{\beta}_j^T = (\beta_{j1}, \beta_{j2}, \beta_{j3}, \beta_{j4})$ dengan β_{jk} merupakan ekspektasi dari level ekspresi dari gen j dalam grup k. Keterangan level: *metastatic prostate tumor* (`mpt`) = 1, *normal prostate adjacent to tumor* (`npat`) = 2, *normal prostate tissue* (`nptissue`) = 3, dan *primary prostate tumor* (`ppt`) = 4.

```
# Apply linear model to data
fit <- lmFit(expdtgeoFilt, design = des_mat)
# Apply empirical Bayes to smooth standard errors
fit <- eBayes(fit)
```

Setelah *fitting* model, ingin diselidiki perbedaan di antara semua kelompok dengan menggunakan *contrast* sebagai berikut.

$$\boldsymbol{C} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & -1 & -1 \end{bmatrix}$$

$$
\begin{aligned}
\boldsymbol{T}_j &= \boldsymbol{C}^T \boldsymbol{\beta}_j \\
&= \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_{j1} \\ \beta_{j2} \\ \beta_{j3} \\ \beta_{j4} \end{bmatrix} \\
&= \begin{bmatrix} \beta_{j1} - \beta_{j2} \\ \beta_{j1} - \beta_{j3} \\ \beta_{j1} - \beta_{j4} \\ \beta_{j2} - \beta_{j3} \\ \beta_{j2} - \beta_{j4} \\ \beta_{j3} - \beta_{j4} \end{bmatrix}
\end{aligned} \tag{1}
$$

Keterangan: $\boldsymbol{T}_j$ merupakan ekspektasi dari perbedaan dalam level ekspresi dari gen j dengan perbandingan *metastatic prostate tumor* versus *normal prostate adjacent to tumor*, *metastatic prostate tumor* versus *normal prostate tissue*, *metastatic prostate tumor* versus *primary prostate tumor*, *normal prostate adjacent to tumor* versus *normal prostate tissue*, *normal prostate adjacent to tumor* versus *primary prostate tumor*, dan *normal prostate tissue* versus *primary prostate tumor*.

```
contrast_matrix <- makeContrasts(
  "mpt.vs.npat" = mpt - npat,
  "mpt.vs.nptissue" = mpt - nptissue,
  "mpt.vs.ppt" = mpt - ppt,
  "npat.vs.nptissue" = npat - nptissue,
  "npat.vs.ppt" = npat - ppt,
  "nptissue.vs.ppt" = nptissue - ppt,
  levels = colnames(des_mat)
)
contrast_matrix
```

```
          Contrasts
Levels     mpt.vs.npat mpt.vs.nptissue mpt.vs.ppt npat.vs.nptissue npat.vs.ppt
  mpt                1               1          1                 0           0
  npat              -1               0          0                 1           1
  nptissue           0              -1          0                -1           0
  ppt                0               0         -1                 0          -1
          Contrasts
Levels     nptissue.vs.ppt
  mpt                    0
  npat                   0
  nptissue               1
  ppt                   -1
```

```
fit <- contrasts.fit(fit, contrast_matrix)
```

Beberapa koreksi pengujian diperlukan setiap kali beberapa pengujian hipotesis (*multiple hypothesis tests*) dilakukan, untuk meminimalkan jumlah *false positif* yang diperoleh. Dalam analisis ini, akan

digunakan metode *False Discovery Rate* (FDR) untuk melakukan koreksi *multiple hypothesis tests*, dan menetapkan batas signifikansi pada 0.05. Ini berarti bahwa hanya gen dengan nilai $p - value$ yang disesuaikan dengan FDR < 0.05 dan perubahan $\log_2$ absolut sebesar 1 atau lebih yang akan dianggap *berbeda secara signifikan.*

```
# Identifying differentially expressed genes
results <- decideTests(fit, p.value = 0.05, adjust.method = "fdr")
summary(results)
```

```
       mpt.vs.npat mpt.vs.nptissue mpt.vs.ppt npat.vs.nptissue npat.vs.ppt
Down           635             414        539                5          48
NotSig        2273            2622       2507             3403        3262
Up             522             394        384               22         120
       nptissue.vs.ppt
Down               114
NotSig            3226
Up                  90
```

Didapatkan informasi bahwa jenis jaringan yang paling banyak berbeda secara signifikan yaitu *metastatic prostate tumor* versus *normal prostate adjacent to tumor*, dengan jumlah yang signifikan berbeda sebanyak $635 + 522 = 1157$. Selanjutnya, akan dibuat tabel hasil berdasarkan model yang dilengkapi kontras. Langkah ini akan menerapkan koreksi *multiple hypothesis tests* Benjamini-Hochberg. Default fungsi `topTable()` adalah menggunakan metode koreksi Benjamini-Hochberg.

```
# Re-smooth the Bayes
contrasts_fit <- eBayes(fit)
# Apply multiple testing correction and obtain stats
stats_df <- topTable(contrasts_fit, number = nrow(expdtgeoFilt)) %>%
  tibble::rownames_to_column("Gene")
kable(head(stats_df, 15), format = "latex", booktabs = TRUE,
      align = rep("c", 11), caption = "\\textit{Head} dari stats\\_df") %>%
  kable_styling(position = "center", latex_options = c("scale_down", "HOLD_position"))
```

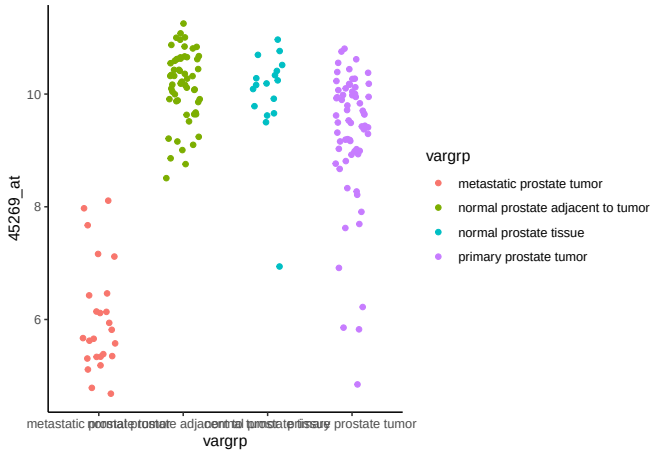
Tabel 2: *Head* dari stats_df

Gene	mpt.vs.npat	mpt.vs.nptissue	mpt.vs.ppt	npat.vs.nptissue	npat.vs.ppt	nptissue.vs.ppt	AveExpr	F	P.Value	adj.P.Val
45269_at	-4.157382	-4.000436	-3.264411	0.1569456	0.8929708	0.7360252	9.169726	109.90301	0	0
52946_at	-4.826910	-4.898510	-3.898026	-0.0715999	0.9288838	1.0004838	9.100647	100.99183	0	0
52140_at	-5.749362	-5.350365	-4.338072	0.3989968	1.4112897	1.0122929	10.268617	97.33389	0	0
50361_at	-4.929657	-4.908223	-3.706627	0.0214338	1.2230304	1.2015966	11.258848	92.32705	0	0
46276_at	-4.324847	-4.442108	-3.304251	-0.1172615	1.0205961	1.1378576	11.688339	91.38222	0	0
50298_at	-4.084435	-3.585757	-2.966931	0.4986784	1.1175042	0.6188259	7.263469	73.88825	0	0
45217_at	-3.012935	-2.736566	-2.266847	0.2763688	0.7460878	0.4697190	9.922762	70.96904	0	0
56409_at	-3.639961	-4.147528	-2.519044	-0.5075671	1.1209175	1.6284846	9.392204	69.92366	0	0
43014_at	3.576697	3.559412	2.998527	-0.0172848	-0.5781698	-0.5608850	7.297168	67.89217	0	0
51214_at	-3.925990	-2.589943	-3.259488	1.3360470	0.6665011	-0.6695459	8.018877	65.09797	0	0
54668_at	-3.069502	-2.949274	-2.309891	0.1202278	0.7596100	0.6393823	10.036330	64.78944	0	0
53766_at	-3.965050	-3.233337	-3.140086	0.7317135	0.8249641	0.0932506	6.695557	58.63878	0	0
58494_r_at	2.807184	3.051040	2.648321	0.2438566	-0.1588627	-0.4027194	9.443107	58.29595	0	0
46183_at	-2.731837	-3.142705	-2.228452	-0.4108684	0.5033846	0.9142530	8.066897	58.18608	0	0
43355_s_at	-2.179226	-2.858407	-1.412796	-0.6791804	0.7664305	1.4456109	9.977102	58.00408	0	0

Untuk menguji apakah hasil tersebut masuk akal, dapat dibuat plot dari salah satu gen teratas. Akan diekstrak data untuk gen `45269_at`, kemudian akan dibuat dataframe untuk tujuan visualisasi. Berdasarkan hasil di `stats_df`, diperkirakan keempat jenis jaringan tersebut berbeda secara signifikan.

```r
top_gene_df <- data.frame(X45269_at = expdtgeoFilt["45269_at", ], vargrp)
```

```r
ggplot(top_gene_df, aes(x = vargrp, y = X45269_at, color = vargrp)) +
  labs(y = "45269_at") +
  geom_jitter(width = 0.2, height = 0) + # Make this a jitter plot
  theme_classic() # This makes some aesthetic changes
```



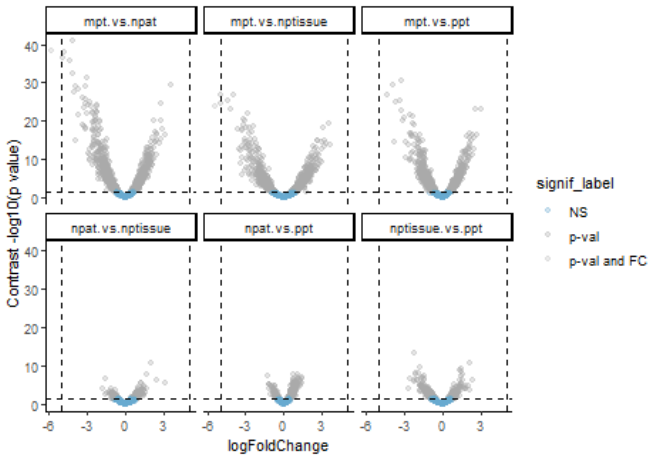
Gambar 4: *Scatter Plot* Ekspresi Gen 45269_at Berdasarkan Jenis Jaringan

Hasil visualisasi tersebut sejalan dengan `stat_df` sebelumnya di mana keempat jenis jaringan berbeda secara signifikan. Selanjutnya, akan dibuat *volcano plot* dari data tersebut.

```r
# Let's extract the contrast p-values for each and transform them with -log10()
contrast_p_vals_df <- -log10(contrasts_fit$p.value) %>%
  # Make this into a dataframe
  as.data.frame() %>%
  # Store genes as their own column
  tibble::rownames_to_column("Gene") %>%
  # Make this into long format
  tidyr::pivot_longer(dplyr::contains("vs"),
                      names_to = "contrast",
                      values_to = "neg_log10_p_val")
# Let's extract the fold changes from stats_df
log_fc_df <- stats_df %>%
  # Only want to keep the `Gene` column as well
  dplyr::select("Gene", dplyr::contains("vs")) %>%
  # Make this a longer format
  tidyr::pivot_longer(dplyr::contains("vs"),
                      names_to = "contrast",
                      values_to = "logFoldChange")
plot_df <- log_fc_df %>%
  dplyr::inner_join(contrast_p_vals_df,
                    by = c("Gene", "contrast"),
                    # This argument will add the given suffixes to the column names
                    # from the respective dataframes, helping us keep track of which columns
```

```
                    # hold which types of values
                    suffix = c("_log_fc", "_p_val"))
```

```
# Convert p-value cutoff to negative log 10 scale
p_val_cutoff <- -log10(0.05)
# Absolute value cutoff for fold changes
abs_fc_cutoff <- 5
plot_df <- plot_df %>%
  dplyr::mutate(signif_label = dplyr::case_when(
    abs(logFoldChange)>abs_fc_cutoff & neg_log10_p_val>p_val_cutoff ~ "p-val and FC",
    abs(logFoldChange)>abs_fc_cutoff ~ "FC",
    neg_log10_p_val>p_val_cutoff ~ "p-val",
    TRUE ~ "NS"))
volcanoes_plot <- ggplot(plot_df,
  aes(
    x = logFoldChange, # Fold change as x value
    y = neg_log10_p_val, # -log10(p-value) for the contrasts
    color = signif_label # Color code by significance cutoffs variable
  )) +
  # Make a scatter plot with points that are 30% opaque using `alpha`
  geom_point(alpha = 0.3) +
  # Draw our `p_val_cutoff` for line here
  geom_hline(yintercept = p_val_cutoff, linetype = "dashed") +
  # Using our `abs_fc_cutoff` for our lines here
  geom_vline(xintercept = c(-abs_fc_cutoff, abs_fc_cutoff), linetype = "dashed") +
  # Specify color
  scale_colour_manual(values = c("#67a9cf", "darkgray", "gray", "#a1d76a")) +
  # Let's be more specific about what this p-value is in our y axis label
  ylab("Contrast -log10(p value)") +
  # This makes separate plots for each contrast
  facet_wrap(~ contrast) +
  theme(text = element_text(size = 7)) + theme_classic()
# Print out the plot
volcanoes_plot
```



Gambar 5: *Volcano Plot* dari Masing-Masing *Contrast*

Akan dicari top 50 dari gen-gen yang berbeda antara keempat jenis jaringan tersebut menggunakan fungsi `topTable`, sama seperti yang telah dilakukan sebelumnya.

```
topResult <- topTable(contrasts_fit, number = 50)
```

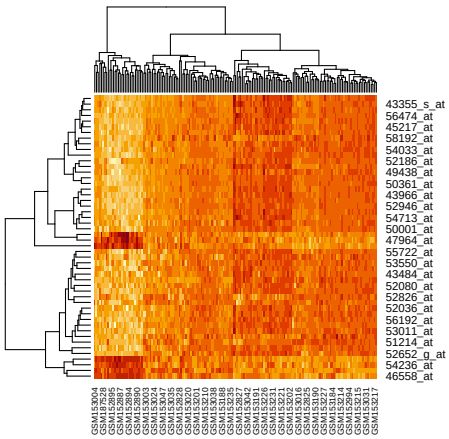
Selanjutnya, akan ditampilkan pola ekspresi dari 50 gen tersebut dengan menggunakan *heat map* dan *box plot*.

```
# Selected genes
rownames(topResult)
```

```
 [1] "45269_at"   "52946_at"   "52140_at"   "50361_at"   "46276_at"
 [6] "50298_at"   "45217_at"   "56409_at"   "43014_at"   "51214_at"
[11] "54668_at"   "53766_at"   "58494_r_at" "46183_at"   "43355_s_at"
[16] "54033_at"   "43966_at"   "56474_at"   "57194_at"   "49438_at"
[21] "50001_at"   "54713_at"   "56192_at"   "43506_at"   "50411_at"
[26] "52826_at"   "53550_at"   "53785_at"   "55722_at"   "53011_at"
[31] "52652_g_at" "48069_at"   "44119_at"   "58192_at"   "58617_at"
[36] "47964_at"   "48587_at"   "45680_at"   "43484_at"   "46558_at"
[41] "45260_at"   "54236_at"   "45199_at"   "50658_s_at" "52186_at"
[46] "58917_at"   "43076_at"   "45939_at"   "52080_at"   "52036_at"
```

```
# Extract selected genes names
selected <- rownames(expdtgeoFilt) %in% rownames(topResult)
# Extract the expression of the selected genes
exptop50 <- expdtgeoFilt[selected, ]
```

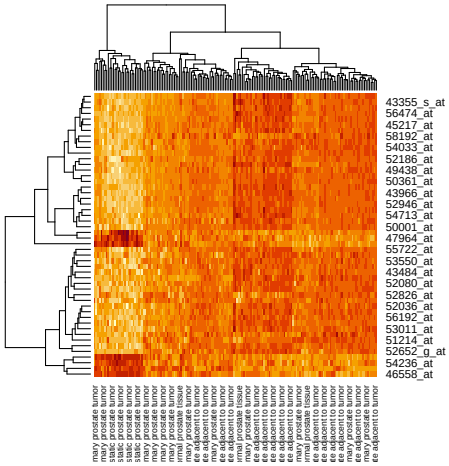
```
# Heat map of the top genes
heatmap(exptop50)
```



Gambar 6: *Heat Map* dari Ekspresi Gen Dataframe expdtgeoFilt

```
# Heat map dari kategori grup
exptop50_2 <- exptop50
colnames(exptop50_2) <- vargrp
```

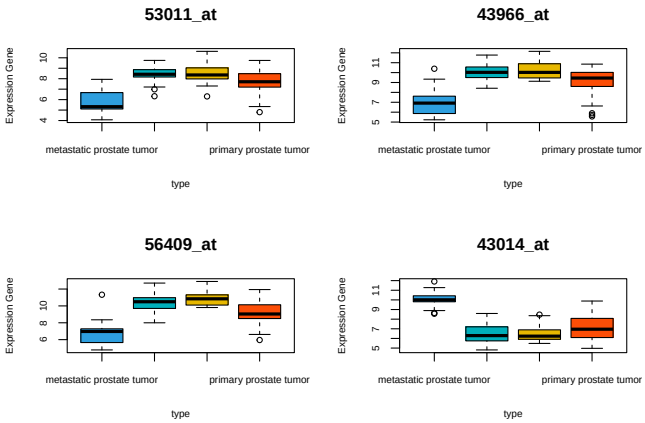
```
heatmap(exptop50_2)
```



Gambar 7: *Heat Map* dari Ekspresi Gen Dataframe expdtgeoFilt dengan Label Grup

Terlihat bahwa ekpresi gen dari jenis jaringan *metastatic prostate tumor* secara umum berbeda dibandingkan ketiga jaringan lainnya.

```
# Boxplot for the top 4 genes
par(mfrow = c(2, 2))
for(i in 1:4){
  df_bp <- data.frame(y = exptop50[i, ], vargrp)
  df_bp$vargrp <- factor(df_bp$vargrp)
  boxplot(df_bp$y ~ df_bp$vargrp, xlab = "type", ylab = "Expression Gene",
          col = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
          cex.axis = 0.7, cex.lab = 0.7, main = rownames(exptop50)[i])
}
```



Gambar 8: *Box Plot* dari Top 4 Gen *Differentialy Expressed* Dataframe expdtgeoFilt

```
par(mfrow = c(1, 1))
```

Dari *box plot* di atas, terlihat bahwa setiap gen memiliki karakteristik yang berbeda terhadap keempat jaringan.

Clustering of Samples

Akan dilakukan berbagai macam algoritma *clustering* berdasarkan sampel. Karena label dari sampel telah diketahui, *clustering* pada sampel akan dilakukan dengan prinsip klasifikasi untuk mengetahui model yang terbaik. Setelah mendapatkan hasil klaster, di mana klaster dibagi menjadi 4 sesuai dengan grup aslinya, selanjutnya dibuat tabel kontingensi dari hasil klaster dengan grup aslinya. Kemudian, dilakukan perhitungan akurasi pada tabel kontingensi tersebut, di mana model dengan akurasi yang paling tinggi akan digunakan sebagai model terbaik.

1. k-Means

Akan di-*set* jumlah klaster $k = 4$ berdasarkan grup aslinya. Kemudian, akan dilakukan pengelompokan k-means berdasarkan sampel.

```
cl <- vargrp
k <- 4
seed <- 123
set.seed(seed)
k_means <- kmeans(t(expdtgeoFilt), centers = k)
table(k_means$cluster, cl)
```

```
   cl
    metastatic prostate tumor normal prostate adjacent to tumor
  1                        24                                 0
  2                         0                                 4
  3                         0                                34
  4                         1                                21
   cl
    normal prostate tissue primary prostate tumor
  1                      0                      4
  2                      7                      0
  3                     10                     33
  4                      0                     29
```

```
# Akurasi
(ac1 <- sum(diag(table(k_means$cluster, cl)))/sum(table(k_means$cluster, cl)))
```

```
[1] 0.4011976
```

```
sprintf("Akurasi = %.2f%%", ac1*100)
```

```
[1] "Akurasi = 40.12%"
```

Dengan menggunakan k-means dengan $k = 4$ didapatkan akurasi sebesar 40.1197605%.

2. Partitioning Around Medoids (PAM)

Selanjutnya, akan dilakukan pengelompokkan dengan menggunakan metode PAM. Metode dengan mempartisi di sekitar medoid merupakan generalisasi dari k-means.

```
result <- pam(t(expdtgeoFilt), k)
groups <- result$clustering
table(groups, cl)
```

```
      cl
groups metastatic prostate tumor normal prostate adjacent to tumor
     1                          0                                37
     2                          0                                22
     3                         11                                 0
     4                         14                                 0
      cl
groups normal prostate tissue primary prostate tumor
     1                     13                     17
     2                      4                     45
     3                      0                      2
     4                      0                      2
```

```
# Akurasi
ac2 <- sum(diag(table(groups, cl)))/sum(table(groups, cl))
sprintf("Akurasi = %.2f%%", ac2*100)
```

```
[1] "Akurasi = 14.37%"
```

Dengan menggunakan PAM dengan $k = 4$ didapatkan akurasi sebesar 14.37126%.

3. Hierarchical Clustering

Selanjutnya, akan dilakukan pengelompokkan dengan metode hierarki dengan berbagai macam *linkage*.

```
d <- dist(t(expdtgeoFilt))
# Single Linkage
res1 <- hclust(d, method = "single" )
gr1 <- cutree(res1, k = k)
table(gr1, cl)
```

```
   cl
gr1 metastatic prostate tumor normal prostate adjacent to tumor
  1                        24                                59
  2                         0                                 0
  3                         0                                 0
  4                         1                                 0
   cl
gr1 normal prostate tissue primary prostate tumor
  1                     17                     64
  2                      0                      1
  3                      0                      1
  4                      0                      0
```

```r
# Akurasi
ac3 <- sum(diag(table(gr1, cl)))/sum(table(gr1, cl))
sprintf("Akurasi = %.2f%%", ac3*100)
```

```
[1] "Akurasi = 14.37%"
```

```r
# Complete Linkage
res2 <- hclust(d, method = "complete")
gr2 <- cutree(res2, k = k)
table(gr2, cl)
```

```
   cl
gr2 metastatic prostate tumor normal prostate adjacent to tumor
  1                         0                                37
  2                         0                                21
  3                         1                                 1
  4                        24                                 0
   cl
gr2 normal prostate tissue primary prostate tumor
  1                      10                     32
  2                       5                     29
  3                       2                      1
  4                       0                      4
```

```r
# Akurasi
ac4 <- sum(diag(table(gr2, cl)))/sum(table(gr2, cl))
sprintf("Akurasi = %.2f%%", ac4*100)
```

```
[1] "Akurasi = 16.17%"
```

```r
# Average Linkage
res3 <- hclust(d, method = "average")
gr3 <- cutree(res3, k = k)
table(gr3, cl)
```

```
   cl
gr3 metastatic prostate tumor normal prostate adjacent to tumor
  1                         0                                59
  2                        24                                 0
  3                         0                                 0
  4                         1                                 0
   cl
gr3 normal prostate tissue primary prostate tumor
  1                      17                     60
  2                       0                      5
  3                       0                      1
  4                       0                      0
```

```r
# Akurasi
ac5 <- sum(diag(table(gr3, cl)))/sum(table(gr3, cl))
sprintf("Akurasi = %.2f%%", ac5*100)
```

```
[1] "Akurasi = 0.00%"
```

```r
# Ward's Linkage
res4 <- hclust(d, method = "ward.D")
gr4 <- cutree(res4, k = k)
table(gr4, cl)
```

```
   cl
gr4 metastatic prostate tumor normal prostate adjacent to tumor
  1                         0                               38
  2                         0                               16
  3                         0                                5
  4                        25                                0
   cl
gr4 normal prostate tissue primary prostate tumor
  1                      10                     20
  2                       7                     24
  3                       0                     20
  4                       0                      2
```

```r
# Akurasi
ac6 <- sum(diag(table(gr4, cl)))/sum(table(gr4, cl))
sprintf("Akurasi = %.2f%%", ac6*100)
```

```
[1] "Akurasi = 10.78%"
```

Berdasarkan hasil di atas, didapatkan bahwa metode hierarki dengan *complete linkage* menghasilkan akurasi yang paling tinggi, yaitu 16.16766%.

Cluster Validation

Karena label pada sampel sudah diketahui, jadi tidak perlu lagi dilakukan validasi dari nilai $k = 4$ karena nilai tersebut didasarkan pada label aslinya pada data.

Model Akhir

Karena model dengan akurasi yang paling tinggi adalah k-means dengan akurasi 40.1197605%, maka model akhir yang akan dipakai adalah k-means dengan $k = 4$.

Cluster Profiling

Pada bagian ini, akan dibuat *cluster plot* dan *radar plot* dengan menggunakan metode k-means ($k = 4$).

```r
set.seed(seed)
kmres <- kmeans(t(expdtgeoFilt), centers = k)
table(kmres$cluster, cl)
```

```
   cl
    metastatic prostate tumor normal prostate adjacent to tumor
  1                        24                                  0
  2                         0                                  4
  3                         0                                 34
  4                         1                                 21
   cl
    normal prostate tissue primary prostate tumor
  1                      0                      4
  2                      7                      0
  3                     10                     33
  4                      0                     29
```

```
# Akurasi
(ac1 <- sum(diag(table(kmres$cluster, cl)))/sum(table(kmres$cluster, cl)))
```

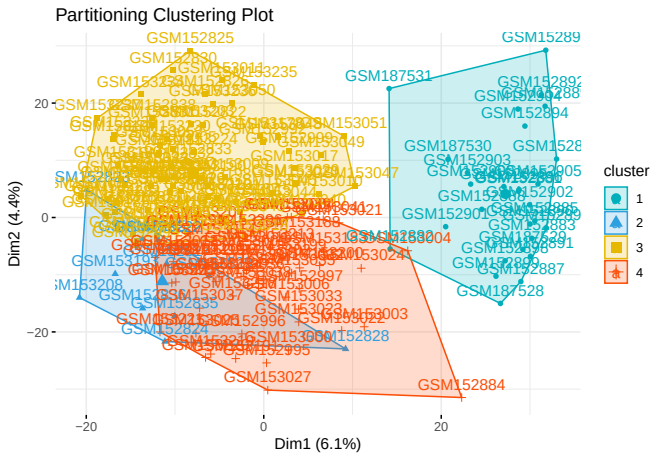
```
[1] 0.4011976
```

```
sprintf("Akurasi = %.2f%%", ac1*100)
```

```
[1] "Akurasi = 40.12%"
```

1. Cluster Plot

```
fviz_cluster(kmres, data = t(expdtgeoFilt),
             palette = c("#00AFBB", "#2E9FDF", "#E7B800", "#FC4E07"),
             ggtheme = theme_minimal(),
             main = "Partitioning Clustering Plot")
```



Gambar 9: Plot *Clustering* Berdasarkan Sampel dengan Metode k-Means

Dari hasil *cluster plot* di atas, dapat dilihat bahwa klaster 2, 3, dan 4 tidak memisah dengan sempurna karena terdapat sampel yang saling tumpang-tindih.

2. Radar Plot

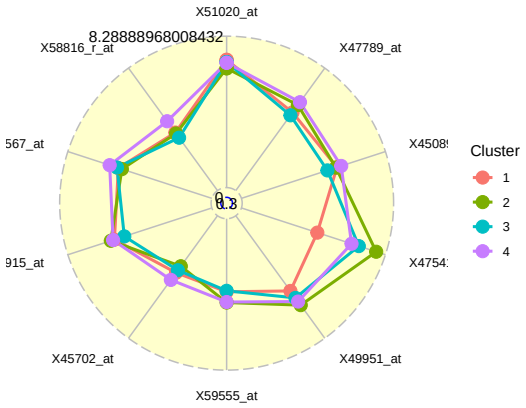
Karena jumlah variabelnya banyak, akan digunakan 10 variabel saja agar visualisasinya tidak menumpuk.

```
temp <- data.frame(group = c(1:4), kmres$centers)
kable(temp[, 1:11], format = "latex", booktabs = TRUE,
      align = rep("c", 11), caption = "Cuplikan Dataframe temp") %>%
  kable_styling(position = "center", latex_options = c("scale_down", "HOLD_position"))
```

Tabel 3: Cuplikan Dataframe temp

group	X51020_at	X47789_at	X45089_at	X47541_at	X49951_at	X59555_at	X45702_at	X44915_at	X44567_at	X58816_r_at
1	7.072935	5.466947	5.667474	4.605541	5.281584	4.257335	4.125952	5.770176	5.555602	4.188577
2	6.636065	5.935638	5.614971	7.788890	6.185372	4.812492	3.711021	5.963322	5.366931	4.129699
3	6.958624	5.284554	5.150762	6.847945	5.724530	4.209342	3.966833	5.234532	5.629535	3.861688
4	6.921604	6.119427	5.896725	6.455878	5.956617	4.785607	4.578938	5.842378	6.025908	4.908585

```
CreateRadialPlot(temp[, 1:11], grid.max = max(temp[, 1:11]) + 0.5,
                 grid.min = min(temp[, 1:11]) - 0.5,
                 centre.y = min(temp[, 1:11]) - 0.7,
                 label.centre.y = TRUE, label.gridline.min = FALSE)
```



Gambar 10: *Radar Plot Cluster* Berdasarkan Sampel

Hasil radar plot di atas menunjukkan bahwa keempat klaster memiliki variasi yang hampir mirip dengan hampir sejajarnya titik-titik dan garis antara keempat klaster. Selain itu, terlihat bahwa gen 4754 memiliki variabilitas yang tinggi terhadap empat klaster.

Clustering of Genes

Ukuran data pada data ekspresi gen yang telah di-*filter*, yaitu jumlah sampel = 167 dan jumlah gen = 3430. Kedua entitas tersebut tidak memiliki jumlah proporsional jika dilakukan pengelompokkan terhadap gen karena jumlah gen > jumlah sampel. Oleh karena itu, akan dilakukan *filtering* lagi terhadap jumlah gen agar jumlahnya proporsional. Selain itu, untuk meningkatkan hasil penelitian, kita harus mencoba menghindari gen yang hanya menyumbang *noise* dan tidak memberikan informasi. Pendekatan sederhana yang dapat dilakukan untuk mengatasi hal tersebut, yaitu dengan mengecualikan semua gen yang tidak menunjukkan varian di semua sampel.

Gene Selection Before Clustering Samples

Pada bagian ini, akan dilakukan seleksi gen berdasarkan variansinya kemudian dipilih 100 gen dengan variansi tertinggi.

```
genes.var <- apply(expdtgeoFilt, 1, var)
genes.var.select <- order(-genes.var)[1:100]
data.s <- expdtgeoFilt[genes.var.select, ]
```

k-Means

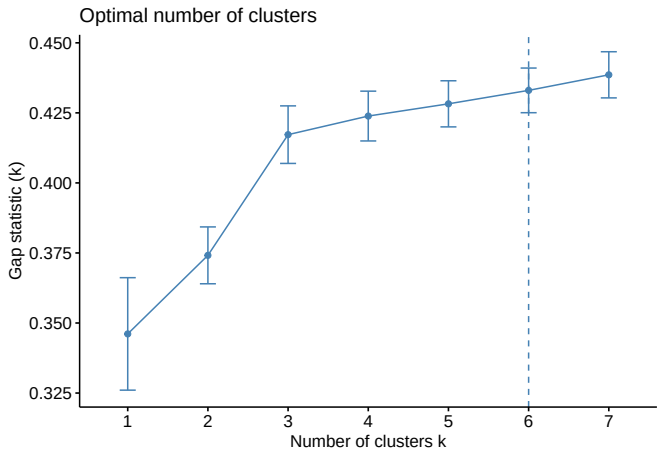
Akan dilakukan *clustering* berdasarkan gen dengan menggunakan metode k-means. Untuk menentukan nilai k, akan digunakan *cluster validation* dengan metode *gap statistic*. Didapatkan hasil sebagai berikut.

```
set.seed(seed)
gap_stat <- clusGap(data.s, FUN = kmeans, nstart = 25, K.max = 7, B = 50)
# Print the result
print(gap_stat, method = "firstmax")
```

```
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = data.s, FUNcluster = kmeans, K.max = 7, B = 50, nstart = 25)
B=50 simulated reference sets, k = 1..7; spaceH0="scaledPCA"
 --> Number of clusters (method 'firstmax'): 7
         logW   E.logW       gap      SE.sim
[1,] 7.021823 7.367928 0.3461053 0.020071783
[2,] 6.774555 7.148704 0.3741497 0.010145560
[3,] 6.664161 7.081387 0.4172256 0.010271631
[4,] 6.622570 7.046432 0.4238623 0.008884905
[5,] 6.589109 7.017335 0.4282260 0.008227817
[6,] 6.558135 6.991165 0.4330297 0.007970814
[7,] 6.528360 6.966943 0.4385826 0.008237159
```

Dari hasil di atas, terlihat bahwa kolom pertama dan kolom kedua masing-masing adalah nilai $\log W_k$ dan $\log W_{kb}$ yang dilanjutkan dengan *gap statistic* dan standard eror dari *gap statistic*. Untuk menentukan jumlah klaster yang optimal, dapat menggunakan plot sebagai berikut.

```
fviz_gap_stat(gap_stat)
```



Gambar 11: Grafik Jumlah Klaster Berdasarkan *Gap Statistic*

Dari plot tersebut, didapatkan bahwa klaster yang optimal adalah $k = 6$.

Model Akhir

Berdasarkan *cluster validation*, didapatkan bahwa nilai k yang paling optimal adalah $k = 6$. Oleh karena itu, model akhir yang akan dipakai adalah k-means dengan $k = 6$.

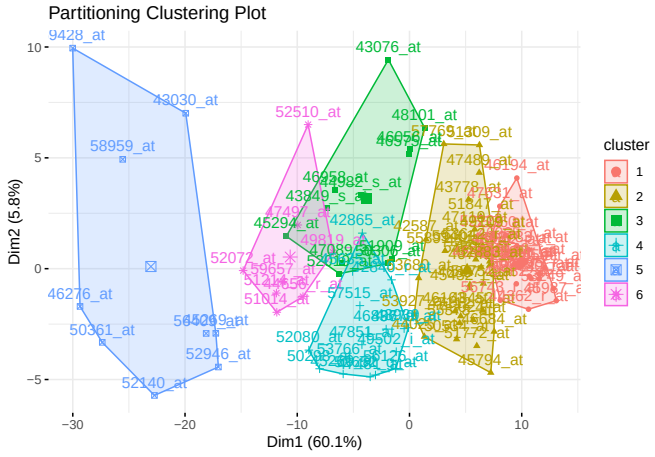
Cluster Profiling

Pada bagian ini, akan dibuat *cluster plot* dan *radar plot* dengan menggunakan metode k-means ($k = 6$).

```
set.seed(seed)
kmres2 <- kmeans(data.s, centers = 6)
```

1. Cluster Plot

```
fviz_cluster(kmres2, data = data.s,
             ggtheme = theme_minimal(),
             main = "Partitioning Clustering Plot")
```



Gambar 12: Plot *Clustering* Berdasarkan Gen dengan Metode k-Means

Dari hasil *cluster plot* di atas, dapat terlihat bahwa klaster 1 dan 2 tidak memisah dengan sempurna karena terdapat gen yang saling tumpang-tindih.

2. Radar Plot

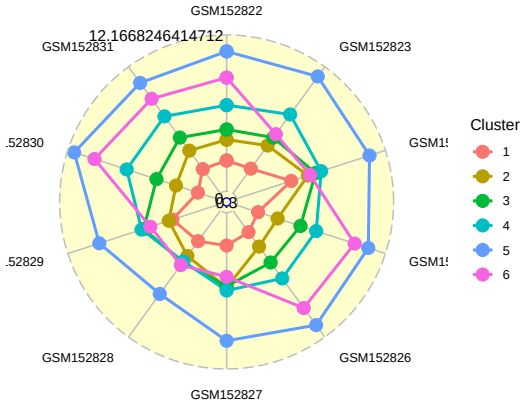
Karena jumlah variabelnya banyak, akan digunakan 10 variabel saja agar visualisasinya tidak menumpuk.

```
temp2 <- data.frame(group = c(1:6), kmres2$centers)
kable(temp2[, 1:11], format = "latex", booktabs = TRUE,
      align = rep("c", 11), caption = "Cuplikan Dataframe temp2") %>%
  kable_styling(position = "center", latex_options = c("scale_down", "HOLD_position"))
```

Tabel 4: Cuplikan Dataframe temp2

group	GSM152822	GSM152823	GSM152824	GSM152825	GSM152826	GSM152827	GSM152828	GSM152829	GSM152830	GSM152831
1	2.793693	2.773037	4.774478	2.170486	2.485995	2.956553	3.324746	3.956008	1.958140	2.731654
2	4.352213	4.905212	6.081315	3.705021	3.819003	6.094978	4.683738	4.243447	3.672782	4.434798
3	5.108980	5.640024	6.676555	5.502328	5.289033	5.996516	5.276103	6.211438	5.206518	5.641137
4	6.927162	7.771899	7.123564	6.733630	6.760236	6.315117	5.190664	6.376500	7.548423	7.595607
5	10.940444	11.278925	10.919263	10.823123	11.068468	10.067354	8.191939	9.694687	11.666825	10.687979
6	8.981323	5.951171	6.246275	9.752977	9.491505	5.292986	5.520632	5.686165	10.069631	9.223050

```
CreateRadialPlot(temp2[, 1:11], grid.max = max(temp2[, 1:11]) + 0.5,
                 grid.min = min(temp2[, 1:11]) - 0.5,
                 centre.y = min(temp2[, 1:11]) - 0.7,
                 label.centre.y = TRUE, label.gridline.min = FALSE)
```



Gambar 13: *Radar Plot Cluster* Berdasarkan Gen

Hasil *radar plot* di atas menunjukkan keenam klaster memiliki variasi yang sangat berbeda. Klaster 5 memiliki variasi ekspresi gen yang paling tinggi, sedangkan klaster 1 memiliki variasi ekspresi gen yang paling rendah.

Biclustering

Pada bagian ini, akan dilakukan pengelompokkan (*clustering*) berdasarkan baris dan kolom secara bersamaan. *Biclustering* akan dilakukan dengan menggunakan fungsi `biclust` dengan metode

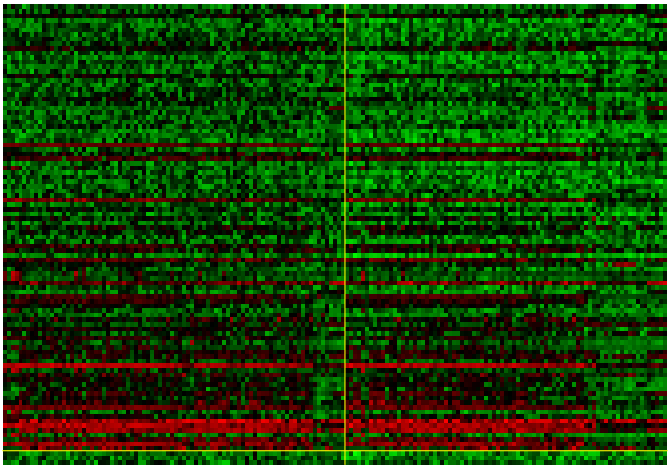
`BCBimax()`. Data yang digunakan pada bagian ini adalah `data.s`, yang merupakan data 100 ekspresi gen varian teratas.

```
biclust_m <- biclust(data.s, method = BCBimax())
biclust_m
```

```
An object of class Biclust

call:
    biclust(x = data.s, method = BCBimax())

Number of Clusters found:  100

First  5  Cluster sizes:
                  BC 1 BC 2 BC 3 BC 4 BC 5
Number of Rows:    100   99   99   98   98
Number of Columns:  81   82   82   83   83
```

Terlihat bahwa terdapat 100 klaster yang ditemukan, dengan klaster pertama terdiri dari 100 baris dan 81 kolom. Selanjutnya, hasil *biclustering* tersebut akan divisualisasikan.

```
drawHeatmap2(x = data.s, bicResult = biclust_m, number = 100)
```



Gambar 14: *Heat Map* dari *Biclsutering*

Dari *heat map* tersebut tidak terlihat pola yang jelas dari 100 klaster yang terbentuk.

Classification

Karena perintah dari soal adalah klasifikasi pada 2 kelompok, maka akan dibuat dataframe baru, yaitu `df2`, yang merupakan subset dari dataframe `expdtgeoFilt` dengan jenis jaringan *normal prostate tissue* dan *metastatic prostate tumor* saja. Akan dilakukan kalsifikasi untuk memprediksi berdasarkan data ekspresi gen yang diberikan apakah ekspresi gen tersebut merupakan jaringan *normal prostate tissue* atau *metastatic prostate tumor*.

```
tipe <- c("normal prostate tissue", "metastatic prostate tumor")
tipe_2 <- vargrp[which(vargrp %in% tipe)]
df2 <- expdtgeoFilt[, which(vargrp %in% tipe)]
df2 <- data.frame(t(df2), tipe = tipe_2)
df2$tipe <- ifelse(df2$tipe=="normal prostate tissue", "ntissue", "mpt")
dim(df2)
```

```
[1]   42 3431
```

```
kable(df2[1:7, 1:7], format = "latex", booktabs = TRUE,
      align = rep("c", 7), caption = "\\textit{Head} dari df2") %>%
  kable_styling(position = "center", latex_options = c("HOLD_position"))
```

Tabel 5: *Head* dari df2

	X51020_at	X47789_at	X45089_at	X47541_at	X49951_at	X59555_at	X45702_at
GSM152822	8.050393	5.770829	2.035624	6.785289	6.490249	0.7655347	3.944858
GSM152823	8.295998	4.727920	2.887525	9.324181	6.392317	5.3680699	3.446256
GSM152824	5.232661	6.539159	5.837943	7.319221	7.995484	3.1858665	6.411087
GSM152825	6.057450	4.620586	5.752213	7.220136	4.426265	3.0179219	3.807355
GSM152826	7.937815	5.465974	4.000000	6.925999	4.452859	4.2094534	3.837943
GSM152827	6.316508	5.765535	5.778734	7.077884	6.350497	6.6639138	3.232661
GSM152828	6.091700	6.955359	6.072535	7.085340	5.263034	4.5235620	3.523562

Kemudian, data tersebut akan dibagi menjadi data `train` dan `test` dengan perbandingan 70/30.

```
set.seed(seed)
train.index <- createDataPartition(df2$tipe, p = 0.7, list = FALSE)
train <- df2[train.index, ]
train$tipe <- factor(train$tipe)
test <- df2[-train.index, ]
test$tipe <- factor(test$tipe)
table(train$tipe)/sum(table(train$tipe))
```

```
    mpt ntissue
    0.6     0.4
```

```
table(test$tipe)/sum(table(test$tipe))
```

```
      mpt   ntissue
0.5833333 0.4166667
```

Selanjutnya, akan dilakukan berbagai macam algoritma klasifikasi dengan menggunakan *k-fold cross validation* dan *hyperparameter tuning*. Nilai k yang digunakan adalah $k = 5$. Didapatkan hasil sebagai berikut.

```r
# k-fold (k = 5)
ctrl <- trainControl(method = "cv", number = 5)
```

1. Linear Discriminant Analysis (LDA)

Akan dilakukan klasifikasi pada sampel berdasarkan grupnya dengan menggunakan metode LDA. Klasifikasi dilakukan pada dataframe `df2`. Didapatkan hasil sebagai berikut.

```r
lda.fit <- train(tipe ~ ., data = train, trControl = ctrl, method = "lda")
kable(lda.fit$results, format = "latex", booktabs = TRUE,
      align = rep("c", 5), caption = "Hasil Akurasi dengan Metode LDA") %>%
  kable_styling(position = "center", latex_options = c("HOLD_position"))
```

Tabel 6: Hasil Akurasi dengan Metode LDA

parameter	Accuracy	Kappa	AccuracySD	KappaSD
none	0.9047619	0.7391304	0.1467718	0.4336943

2. Regresi Logistik

Akan dilakukan klasifikasi pada sampel berdasarkan grupnya dengan menggunakan metode regresi logistik. Klasifikasi dilakukan pada dataframe `df2`. Didapatkan hasil sebagai berikut.

```r
logit.fit <- train(tipe ~ ., data = train, trControl = ctrl,
                   method = "glm", family = "binomial")
kable(logit.fit$results, format = "latex", booktabs = TRUE,
      align = rep("c", 5), caption = "Hasil Akurasi dengan Metode Regresi Logistik") %>%
  kable_styling(position = "center", latex_options = c("HOLD_position"))
```

Tabel 7: Hasil Akurasi dengan Metode Regresi Logistik

parameter	Accuracy	Kappa	AccuracySD	KappaSD
none	0.4514286	-0.1402635	0.2286309	0.4792977

3. k-Nearest Neighbour (k-NN)

Akan dilakukan klasifikasi pada sampel berdasarkan grupnya dengan menggunakan metode k-NN. Klasifikasi dilakukan pada dataframe `df2`. Didapatkan hasil sebagai berikut.

```r
knn.fit <- train(tipe ~ ., data = train, trControl = ctrl, method = "knn",
                 preProcess = c("center", "scale"),
                 tuneGrid = expand.grid(k = seq(1, 10, by = 2)))
kable(knn.fit$results, format = "latex", booktabs = TRUE,
      align = rep("c", 5), caption = "Hasil Akurasi dengan Metode k-NN") %>%
  kable_styling(position = "center", latex_options = c("HOLD_position"))
```

Tabel 8: Hasil Akurasi dengan Metode k-NN

k	Accuracy	Kappa	AccuracySD	KappaSD
1	1.0000000	1.0000000	0.0000000	0.000000
3	0.9666667	0.9142857	0.0745356	0.191663
5	1.0000000	1.0000000	0.0000000	0.000000
7	0.9666667	0.9142857	0.0745356	0.191663
9	0.9666667	0.9142857	0.0745356	0.191663

4. Decision Tree

Akan dilakukan klasifikasi pada sampel berdasarkan grupnya dengan menggunakan metode *decision tree*. Klasifikasi dilakukan pada dataframe `df2`. Didapatkan hasil sebagai berikut.

```
dt.fit <- train(tipe ~ ., data = train, trControl = ctrl, method = "ctree")
kable(dt.fit$results, format = "latex", booktabs = TRUE,
              align = rep("c", 5), caption = "Hasil Akurasi dengan Metode Decision Tree") %>%
  kable_styling(position = "center", latex_options = c("HOLD_position"))
```

Tabel 9: Hasil Akurasi dengan Metode Decision Tree

mincriterion	Accuracy	Kappa	AccuracySD	KappaSD
0.01	0.8714286	0.7249524	0.0726093	0.1628187
0.50	0.8714286	0.7249524	0.0726093	0.1628187
0.99	0.6009524	0.0000000	0.0701796	0.0000000

5. Random Forest

Akan dilakukan klasifikasi pada sampel berdasarkan grupnya dengan menggunakan metode *random forest*. Klasifikasi dilakukan pada dataframe `df2`. Didapatkan hasil sebagai berikut.

```
rf.fit <- train(tipe ~ ., data = train, trControl = ctrl, method = "rf")
kable(rf.fit$results, format = "latex", booktabs = TRUE,
              align = rep("c", 5), caption = "Hasil Akurasi dengan Metode Random Forest") %>%
  kable_styling(position = "center", latex_options = c("HOLD_position"))
```

Tabel 10: Hasil Akurasi dengan Metode Random Forest

mtry	Accuracy	Kappa	AccuracySD	KappaSD
2	0.96	0.9090909	0.0894427	0.2032789
82	1.00	1.0000000	0.0000000	0.0000000
3430	1.00	1.0000000	0.0000000	0.0000000

6. Penalized Logistic Regression

Akan dilakukan klasifikasi pada sampel berdasarkan grupnya dengan menggunakan metode *penalized logistic regression*. Klasifikasi dilakukan pada dataframe `df2`. Didapatkan hasil sebagai berikut.

```
plogit.fit <- train(train[, -3431], train[, 3431], trControl = ctrl,
                    method = "glmnet", family = "binomial")
kable(plogit.fit$results, format = "latex", booktabs = TRUE,
      align = rep("c", 6),
      caption = "Hasil Akurasi dengan Metode Penalized Logistic Regression") %>%
  kable_styling(position = "center", latex_options = c("HOLD_position"))
```

Tabel 11: Hasil Akurasi dengan Metode Penalized Logistic Regression

alpha	lambda	Accuracy	Kappa	AccuracySD	KappaSD
0.10	0.0290469	1.0000000	1.0000000	0.0000000	0.0000000
0.10	0.0918544	1.0000000	1.0000000	0.0000000	0.0000000
0.10	0.2904692	1.0000000	1.0000000	0.0000000	0.0000000
0.55	0.0290469	0.9666667	0.9142857	0.0745356	0.1916630
0.55	0.0918544	0.9666667	0.9142857	0.0745356	0.1916630
0.55	0.2904692	0.9666667	0.9142857	0.0745356	0.1916630
1.00	0.0290469	0.9666667	0.9142857	0.0745356	0.1916630
1.00	0.0918544	0.9666667	0.9142857	0.0745356	0.1916630
1.00	0.2904692	0.8380952	0.6534161	0.2044017	0.4110046

Model Akhir

Dari *k-fold cross validation* dengan $k = 5$ di atas, didapatkan bahwa secara umum akurasi prediksi model cukup tinggi. Hanya model regresi logistik yang memiliki rata-rata akurasi yang relatif rendah dibandingkan model yang lainnya, yaitu 45.14286%. Model k-NN dengan $k = 5$ menghasilkan rata-rata akurasi yang tinggi, yaitu 100% dengan standar deviasi 0. Model ini dipilih sebagai model akhir karena rata-rata akurasi dari model k-NN lebih stabil dibandingkan dengan model lainnya. Jadi, model klasifikasi terbaik untuk data ekspresi gen ini adalah k-NN dengan $k = 5$.

```
knn.fit$finalModel
```

```
5-nearest neighbor model
Training set outcome distribution:

    mpt ntissue
     18      12
```

Prediksi

Akan dilakukan prediksi dari data `test` dengan menggunakan model terbaik k-NN ($k = 5$) yang telah didapatkan sebelumnya.

```
pred <- predict(knn.fit, test)
(cm <- table(pred, test$tipe))
```

```
pred      mpt ntissue
  mpt       7       0
  ntissue   0       5
```

```r
ac <- sum(diag(cm))/sum(cm)
sprintf("Akurasi = %.2f%%", ac*100)
```

```
[1] "Akurasi = 100.00%"
```

Terlihat bahwa akurasi prediksi data `test` dengan menggunakan model ini adalah 100%.

IV KESIMPULAN

Dari hasil analisis sebelumnya, didapatkan tiga *insight*, yaitu sebagai berikut.

Insight 1

Analisis *differentially expressed genes* telah dilakukan dengan menggunakan metode LIMMA. Dengan menggunakan *contrast* pada persamaan (1), didapatkan gen-gen dengan ekpresi gen yang berbeda secara signifikan. Pada bagian ini, akan dicari nama dan deskripsi dari gen-gen tersebut.

```
# See gene name and description
ids1 <- rownames(topResult)
GeneSelected1 <- AnnotationDbi::select(hgu95b.db, ids1,
                                       c("SYMBOL", "ENTREZID", "GENENAME", "GO"))
kable(head(GeneSelected1), format = "latex", booktabs = TRUE,
      align = rep("c", 7), caption = "\\textit{Head} dari GeneSelected1") %>%
  kable_styling(position = "center", latex_options = c("scale_down", "HOLD_position"))
```

Tabel 12: *Head* dari GeneSelected1

PROBEID	SYMBOL	ENTREZID	GENENAME	GO	EVIDENCE	ONTOLOGY
45269_at	RAB34	83871	RAB34, member RAS oncogene family	GO:0003924	IEA	MF
45269_at	RAB34	83871	RAB34, member RAS oncogene family	GO:0005515	IPI	MF
45269_at	RAB34	83871	RAB34, member RAS oncogene family	GO:0005525	IEA	MF
45269_at	RAB34	83871	RAB34, member RAS oncogene family	GO:0005794	IDA	CC
45269_at	RAB34	83871	RAB34, member RAS oncogene family	GO:0005795	IDA	CC
45269_at	RAB34	83871	RAB34, member RAS oncogene family	GO:0005929	IEA	CC

Setelah mengetahui nama gen, selanjutnya ingin diketahui model dan fungsi dari gen tersebut dengan menggunakan Gene Ontology. Terdapat tiga aspek fungsi gen yaitu

1. *molecular function* (aktivitas molekuler dari hasil gen);
2. *cellular component*, di mana hasil gen tersebut aktif;
3. *biological process*, informasi dimana proses dan *pathways* biologi dari gen tersebut ikut serta.

Untuk melakukan hal tersebut, diperlukan *package* `GO.db` yang akan menghubungkan nama-nama gen dengan Gene Ontology.

```
# Gene ontology for the top genes
GOselected1 <- AnnotationDbi::select(GO.db, GeneSelected1$GO, c("TERM", "GOID"))
# Combine the result
finalres1 <- cbind(GeneSelected1, GOselected1)
kable(head(finalres1), format = "latex", booktabs = TRUE,
      align = rep("c", 9), caption = "\\textit{Head} dari finalres1") %>%
  kable_styling(position = "center", latex_options = c("scale_down", "HOLD_position"))
```

Tabel 13: *Head* dari finalres1

PROBEID	SYMBOL	ENTREZID	GENENAME	GO	EVIDENCE	ONTOLOGY	GOID	TERM
45269_at	RAB34	83871	RAB34, member RAS oncogene family	GO:0003924	IEA	MF	GO:0003924	GTPase activity
45269_at	RAB34	83871	RAB34, member RAS oncogene family	GO:0005515	IPI	MF	GO:0005515	protein binding
45269_at	RAB34	83871	RAB34, member RAS oncogene family	GO:0005525	IEA	MF	GO:0005525	GTP binding
45269_at	RAB34	83871	RAB34, member RAS oncogene family	GO:0005794	IDA	CC	GO:0005794	Golgi apparatus
45269_at	RAB34	83871	RAB34, member RAS oncogene family	GO:0005795	IDA	CC	GO:0005795	Golgi stack
45269_at	RAB34	83871	RAB34, member RAS oncogene family	GO:0005929	IEA	CC	GO:0005929	cilium

Untuk versi lengkapnya, dapat dilihat pada *link* berikut. gende_uas.csv

Dari top 50 gen tersebut, akan dipilih tiga gen untuk dicari deskripsi dan kaitannya dengan kanker prostat. Didapatkan hasil sebagai berikut.

1. RAB34

   RAB34 mengkodekan protein yang termasuk dalam keluarga protein RAB, yang merupakan GTPase kecil yang penting dalam mengatur transduksi sinyal dan proses seluler. Peran potensial RAB34 pada kanker telah dikemukakan, namun hanya sedikit penelitian yang melaporkan fungsinya pada kanker epitel manusia. Dilaporkan bahwa pasien kanker prostat (PCa) lebih rentan terhadap kekambuhan biokimia ketika RAB34 diturunkan regulasinya dalam jaringan PCa dibandingkan dengan jaringan prostat jinak (BPT), yang dikaitkan secara negatif dengan miR-148, miRNA yang meningkatkan pertumbuhan garis sel dan diatur dalam sampel PCa. RAB34 adalah gen target yang diduga untuk miR-9 [2].

2. Phosphoglucomutase 5 (PGM5)

   Kanker prostat (PCa) merupakan keganasan paling umum pada pria di negara maju. Antigen spesifik prostat (PSA) tetap menjadi penanda serum yang paling banyak digunakan untuk kanker prostat. Dari hasil penelitian, didapatkan bahwa ekspresi protein mirip fosfoglukomutase 5 (PGM5) secara signifikan lebih rendah pada jaringan kanker prostat. Rendahnya ekspresi PGM5 dan tanda gen terkait ditemukan terkait dengan hasil klinis yang buruk dan skor Gleason yang tinggi. Uji in vitro menunjukkan bahwa ekspresi berlebih PGM5 secara signifikan menekan proliferasi dan migrasi sel kanker prostat. Analisis GO dan jalur menunjukkan pengayaan gen dalam regulasi pertumbuhan dan migrasi sel, serta jalur yang terkait dengan kanker. Penurunan regulasi PGM5 berkaitan erat dengan metilasi DNA. Secara keseluruhan, ekspresi PGM5 dikaitkan dengan perkembangan kanker prostat. Hasil ini juga menyoroti alasan praklinis bahwa PGM5 mewakili penanda prognostik dan target yang menjanjikan untuk strategi terapi baru pada kanker prostat [10].

3. Myosin Light Chain Linase (MYLK)

   Di negara maju, kanker prostat (PCa) merupakan kanker yang sering didiagnosis dengan tingkat kematian tertinggi kedua. Hasil dalam penelitian ini, tingkat ekspresi circMYLK secara signifikan lebih tinggi pada sampel PCa dan sel PCa dibandingkan pada jaringan normal dan sel prostat normal. CirRNA-MYLK yang diregulasi mendorong proliferasi, invasi, dan migrasi sel PCa. Namun, si-circRNA-MYLK secara signifikan mempercepat apoptosis sel PCa. Selain itu, didapatkan bahwa fungsi circRNA-MYLK pada sel PCa dipengaruhi melalui penargetan miR-29a. CircRNA-MYLK adalah onkogen pada PCa dan mengungkapkan mekanisme baru yang mendasari circRNA-MYLK dalam perkembangan PC (*prostate cancer*) [4].

Dengan menggunakan *contrast* pada persamaan (1), jumlah gen-gen yang signifikan berbeda, yaitu sebagai berikut.

Tabel 14: *Differentially Expressed Genes* dari Dataframe expdtgeoFilt

	mpt vs npat	mpt vs nptissue	mpt vs ppt	npat vs nptissue	npat vs ppt	nptissue vs ppt
Down	635	414	539	5	48	114
Not Significance	2273	2622	2507	3403	3262	3226
Up	522	394	384	22	120	90

Dari tabel tersebut, didapatkan informasi bahwa jenis jaringan paling banyak berbeda secara signifikan, yaitu *metastatic prostate tumor* versus *normal prostate adjacent to tumor* dengan jumlah yang signifikan berbeda sebanyak 635 + 522 = 1157.

Berikut ini merupakan top 50 gen yang berbeda secara signifikan berdasarkan analisis LIMMA sebelumnya.

```
rownames(topResult)
```

```
 [1] "45269_at"   "52946_at"   "52140_at"   "50361_at"   "46276_at"
 [6] "50298_at"   "45217_at"   "56409_at"   "43014_at"   "51214_at"
[11] "54668_at"   "53766_at"   "58494_r_at" "46183_at"   "43355_s_at"
[16] "54033_at"   "43966_at"   "56474_at"   "57194_at"   "49438_at"
[21] "50001_at"   "54713_at"   "56192_at"   "43506_at"   "50411_at"
[26] "52826_at"   "53550_at"   "53785_at"   "55722_at"   "53011_at"
[31] "52652_g_at" "48069_at"   "44119_at"   "58192_at"   "58617_at"
[36] "47964_at"   "48587_at"   "45680_at"   "43484_at"   "46558_at"
[41] "45260_at"   "54236_at"   "45199_at"   "50658_s_at" "52186_at"
[46] "58917_at"   "43076_at"   "45939_at"   "52080_at"   "52036_at"
```

Gen tersebut mempunyai nama sebagai berikut.

```
unique(finalres1$GENENAME)
```

```
 [1] "RAB34, member RAS oncogene family"
 [2] "MAS related GPR family member F"
 [3] "phosphoglucomutase 5"
 [4] "synaptopodin 2"
 [5] "myosin light chain kinase"
 [6] "formin homology 2 domain containing 3"
 [7] "platelet derived growth factor C"
 [8] "sorbin and SH3 domain containing 1"
 [9] "endosome associated trafficking regulator 1"
[10] "plakophilin 1"
[11] "phosphodiesterase 5A"
[12] "period circadian regulator 3"
[13] "metastasis associated lung adenocarcinoma transcript 1"
[14] "RAB23, member RAS oncogene family"
[15] "tubulin beta 6 class V"
[16] "prostate androgen-regulated mucin-like protein 1"
[17] "inositol 1,4,5-triphosphate receptor associated 1"
[18] "heat shock protein family B (small) member 8"
[19] "prickle planar cell polarity protein 2"
```

```
[20] "cysteine rich secretory protein LCCL domain containing 2"
[21] "cadherin like and PC-esterase domain containing 1"
[22] "mitogen-activated protein kinase kinase kinase 20"
[23] "protocadherin 7"
[24] "osteoglycin"
[25] "sorbin and SH3 domain containing 2"
[26] "GH3 domain containing"
[27] "ring finger protein 150"
[28] "LIM zinc finger domain containing 2"
[29] "mir-100-let-7a-2-mir-125b-1 cluster host gene"
[30] "myocardial zonula adherens protein"
[31] "collagen type VIII alpha 2 chain"
[32] "nuclear paraspeckle assembly transcript 1"
[33] "RAS like family 12"
[34] "family with sequence similarity 135 member A"
[35] "ANTXR cell adhesion molecule 2"
[36] "formin binding protein 4"
[37] "KLF transcription factor 4"
[38] "teashirt zinc finger homeobox 3"
[39] "sodium voltage-gated channel alpha subunit 7"
[40] "baculoviral IAP repeat containing 6"
[41] "family with sequence similarity 107 member A"
[42] "musashi RNA binding protein 2"
[43] "discoidin domain receptor tyrosine kinase 2"
[44] "eva-1 homolog C"
[45] "niban apoptosis regulator 1"
[46] "dishevelled binding antagonist of beta catenin 3"
[47] "hepsin"
[48] "contactin 3"
[49] "ECRG4 augurin precursor"
[50] "myocardin"
```

Gen-gen tersebut signifikan berbeda pada analisis LIMMA sebelumnya. Gen tersebut direkomendasikan sebagai gen yang berekpresi berbeda di antara empat jenis jaringan, yaitu *normal prostate tissue*, *normal prostate adjacent to tumor*, *primary prostate tumor*, dan *metastatic prostate tumor*.

Insight 2

Analisis *clustering* berdasarkan gen telah dilakukan pada dataframe `data.s` yang merupakan data 100 ekspresi gen dengan varian teratas. Melalui *cluster validation*, didapatkan jumlah klaster yang optimal adalah $k = 6$. Dengan menggunakan k-means ($k = 6$), didapatkan klaster dari 100 gen dengan varian yang tinggi tersebut. Pada bagian ini, akan dicari nama dan deskripsi gen pada masing-masing klaster yang terbentuk.

```
# See gene name and description
ids2 <- rownames(data.s)
GeneSelected2 <- AnnotationDbi::select(hgu95b.db, ids2,
                                       c("SYMBOL", "ENTREZID", "GENENAME", "GO"))
kable(head(GeneSelected2), format = "latex", booktabs = TRUE,
      align = rep("c", 7), caption = "\\textit{Head} dari GeneSelected2") %>%
  kable_styling(position = "center", latex_options = c("scale_down", "HOLD_position"))
```

Tabel 15: *Head* dari GeneSelected2

PROBEID	SYMBOL	ENTREZID	GENENAME	GO	EVIDENCE	ONTOLOGY
45294_at	PIGR	5284	polymeric immunoglobulin receptor	GO:0001580	IDA	BP
45294_at	PIGR	5284	polymeric immunoglobulin receptor	GO:0001792	IDA	MF
45294_at	PIGR	5284	polymeric immunoglobulin receptor	GO:0001895	HEP	BP
45294_at	PIGR	5284	polymeric immunoglobulin receptor	GO:0002415	IBA	BP
45294_at	PIGR	5284	polymeric immunoglobulin receptor	GO:0002415	IDA	BP
45294_at	PIGR	5284	polymeric immunoglobulin receptor	GO:0004888	IBA	MF

Setelah mengetahui nama gen, selanjutnya ingin diketahui model dan fungsi dari gen tersebut dengan menggunakan Gene Ontology.

```
GOselected2 <- AnnotationDbi::select(GO.db, GeneSelected2$GO, c("TERM", "GOID"))
# Combine the result
finalres2 <- cbind(GeneSelected2, GOselected2)
clust <- data.frame(CLUSTER = kmres2$cluster, PROBEID = names(kmres2$cluster))
finalres2 <- merge(finalres2, clust, by = "PROBEID")
finalres2 <- finalres2[order(finalres2$CLUSTER), ]
kable(head(finalres2), format = "latex", booktabs = TRUE,
      align = rep("c", 10), caption = "\\textit{Head} dari finalres2") %>%
  kable_styling(position = "center", latex_options = c("scale_down", "HOLD_position"))
```

Tabel 16: *Head* dari finalres2

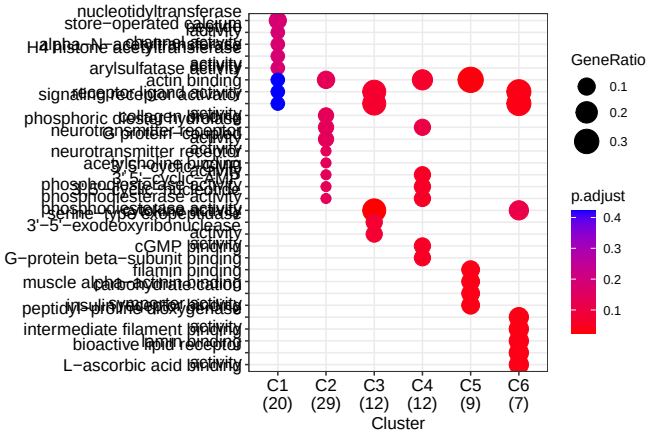
PROBEID	SYMBOL	ENTREZID	GENENAME	GO	EVIDENCE	ONTOLOGY	GOID	TERM	CLUSTER
42251_r_at	CPNE8	144402	copine 8	GO:0003674	ND	MF	GO:0003674	molecular_function	1
42251_r_at	CPNE8	144402	copine 8	GO:0005515	IPI	MF	GO:0005515	protein binding	1
42251_r_at	CPNE8	144402	copine 8	GO:0005544	IBA	MF	GO:0005544	calcium-dependent phospholipid binding	1
42251_r_at	CPNE8	144402	copine 8	GO:0005886	IBA	CC	GO:0005886	plasma membrane	1
42251_r_at	CPNE8	144402	copine 8	GO:0008150	ND	BP	GO:0008150	biological_process	1
42251_r_at	CPNE8	144402	copine 8	GO:0046872	IEA	MF	GO:0046872	metal ion binding	1

Untuk versi lengkapnya, dapat dilihat pada *link* berikut. gencl_uas.csv

Selanjutnya, akan dilakukan *enrichment analysis* pada setiap klaster gen yang telah dibentuk. Didapatkan hasil sebagai berikt.

```
clust2 <- merge(clust, finalres2[, c(1, 3)], by = "PROBEID")
clust2 <- clust2[!duplicated(clust2), ]
clust3 <- list(C1 = clust2[clust2$CLUSTER==1, ]$ENTREZID,
               C2 = clust2[clust2$CLUSTER==2, ]$ENTREZID,
               C3 = clust2[clust2$CLUSTER==3, ]$ENTREZID,
               C4 = clust2[clust2$CLUSTER==4, ]$ENTREZID,
               C5 = clust2[clust2$CLUSTER==5, ]$ENTREZID,
               C6 = clust2[clust2$CLUSTER==6, ]$ENTREZID)
test_enrich <- compareCluster(geneCluster = clust3,
                              OrgDb = hgu95b.db,
                              pAdjustMethod = "BH",
                              pvalueCutoff = 0.5,
                              qvalueCutoff = 0.5,
                              fun = enrichGO)
```

```
dotplot(test_enrich, showCategory = 5)
```



Gambar 15: *Dot Plot* dari *Enrichment Analysis* pada Setiap Klaster

Dari hasil *enriched terms* tersebut, didapatkan bahwa klaster 6 (C6) berkaitan erat dengan *receptor ligand activity* dan *signaling receptor activator activity*, sedangakan klaster 1 (C1) tidak berkaitan erat dengan kedua *term* tersebut.

Insight 3

Pada analisis ini, prediksi dilakukan dengan menggunakan metode penggolompokkan (*clustering*) menggunakan label sampel dari data asli dan menggunakan metode klasifikasi (*classification*). Melalui pendekatan *clustering* dengan prinsip klasifikasi, akurasi prediksi label sampel asli tidak cukup tinggi. Algoritma *clustering* yang dipakai hanya mampu memprediksi dengan akurasi maksimum sebesar 40.1197605%. Sebaliknya, prediksi dengan algoritma klasifikasi menghasilkan rata-rata akurasi yang tinggi secara umum. Akurasi terbaik dengan *k-fold cross validation* dengan $k = 5$ dimiliki oleh model k-NN dengan $k = 5$ karena hasil akurasi prediksinya tinggi, standar deviasinya rendah, dan hasil akurasinya relatif stabil. Model akhir klasifikasi dengan menggunakan k-NN ($k = 5$) menghasilkan akurasi prediksi 100% pada data `test`.

DAFTAR PUSTAKA

[1] American Cancer Society. (2019). *What Is Prostate Cancer?*. https://www.cancer.org/cancer/types/prostate-cancer/about/what-is-prostate-cancer.html

[2] Benedetti, I., Barrios, L., & Rebollo, J. (2022). Rab34 is downregulated in human prostate cancer tissue. *Cancer Research*, *82*(12_Supplement), 5718-5718.

[3] Cancer Research UK. (n.d.). *What is metastatic prostate cancer?*. https://www.cancerresearchuk.org/about-cancer/prostate-cancer/metastatic-cancer/what-is-metastatic-prostate-cancer

[4] Dai, Y., Li, D., Chen, X., Tan, X., Gu, J., Chen, M., & Zhang, X. (2018). Circular RNA Myosin Light Chain Kinase (MYLK) Promotes Prostate Cancer Progression through Modulating Mir-29a Expression. *Medical science monitor: international medical journal of experimental and clinical research*, *24*, 3462–3471. https://doi.org/10.12659/MSM.908009

[5] Dalpiaz , D. (2020).*R for Statistical Learning*. https://daviddalpiaz.github.io/r4sl/the-caret-package.html

[6] Kaiser, S. (2011). *Biclustering: Methods, Software and Application*.

[7] MIT. (n.d.). *Introduction to the LIMMA Package*. https://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/limma/html/01Introduction.html#

[8] Multimatics Insight. (n.d.). *Five Types of Classification Algorithms in Data Science*. https://multimatics.co.id/blog/jun/5-types-of-classification-algorithms-in-data-science.aspx

[9] Rencher, A. C., & Christensen, W. F. (2020). *Methods of Multivariate Analysis* (3rd ed). John Wiley & Sons.

[10] Sun, J., Wang, F., Zhou, H., Zhao, C., Li, K., Fan, C., & Wang, J. (2022). Downregulation of PGM5 expression correlates with tumor progression and poor prognosis in human prostate cancer. *Discover. Oncology*, *13*(1), 63. https://doi.org/10.1007/s12672-022-00525-x