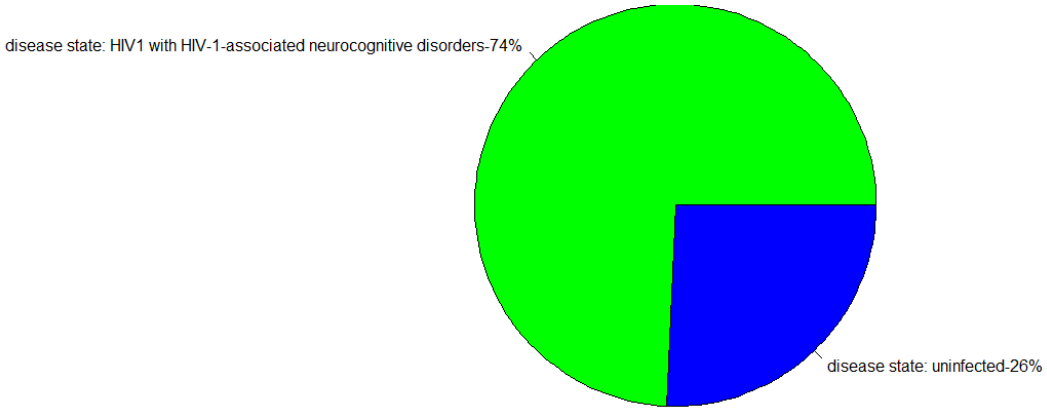
BAB V

PEMBAHASAN

5.1 Analisi Deskriptif

Gambaran umum pada penderita HIV-1 disajikan menggunakan analisis deskriptif suatu data untuk menjadi gambaran suatu data.



Gambar 5.1 *Pie chart* status penyakit

(Sumber : https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE28160)

Berdasarkan gambar diatas data berasal dari NCBI dengan GSE28160 yang berisi 35 sampel data HIV-1 yang terbagi menjadi 9 orang atau 26% yang tidak terinfeksi dan 26 orang atau 74% yang terinfeksi HIV-1. Banyaknya jumlah gen bisa didapat dari hasil gen yang ada di 35 sampel pasien baik yang terkena maupun tidak terkena HIV-1 yakni sebanyak 54675 gen.

5.2 Pengelolaan Data Bioinformatika

Dalam pengolahan data bioinformatika dibutuhkan *package* dari biokonduktor dari aplikasi R. Data yang diolah ialah data *matrix*, *vector*, *frame*, dll dengan aplikasi R. Dengan menggunakan packages `affy` dalam membaca data bioinformatika pada data *AffyBatch*. Dengan informasi mengenai pasien yang tersimpan dalam *pheno* data yang berguna untuk menyimpan berbagai macam informasi pasien.

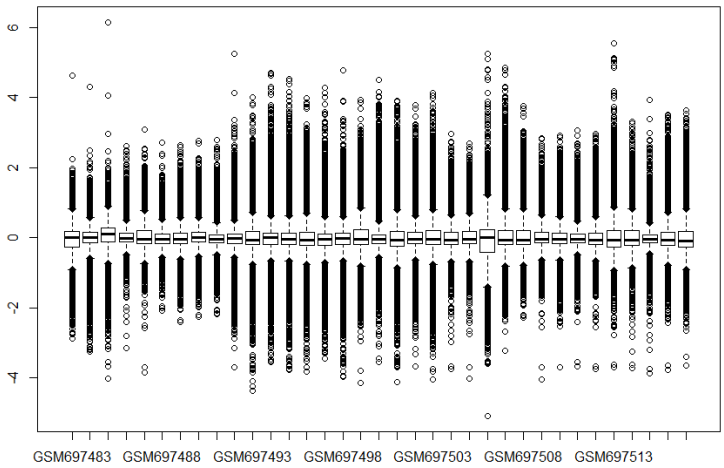
```
 [1] "title"                    "geo_accession"            "status"
 [4] "submission_date"          "last_update_date"         "type"
 [7] "channel_count"            "source_name_ch1"          "organism_ch1"
[10] "characteristics_ch1"      "characteristics_ch1.1"    "characteristics_ch1.2"
[13] "molecule_ch1"             "extract_protocol_ch1"     "label_ch1"
[16] "label_protocol_ch1"       "taxid_ch1"                "hyb_protocol"
[19] "scan_protocol"            "description"              "data_processing"
[22] "platform_id"              "contact_name"             "contact_email"
[25] "contact_phone"            "contact_laboratory"       "contact_department"
[28] "contact_institute"        "contact_address"          "contact_city"
[31] "contact_state"            "contact_zip/postal_code"  "contact_country"
[34] "supplementary_file"       "data_row_count"           "disease state:ch1"
[37] "tissue:ch1"               "treatment:ch1"
```

Gambar 5.2 *Pheno* Data

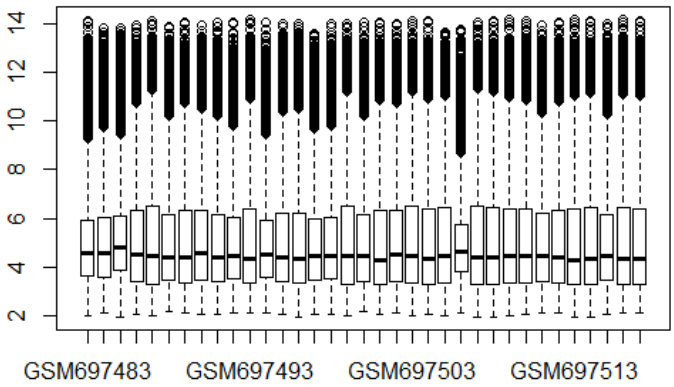
Berdasarkan gambar diatas bisa diperoleh informasi tentang pasien yang dijadikan sampel penelitian, dimana pasien diberikan kode GSM yang mencakup *geo_accession*, jaringan, status, identitas pasien, hingga keadaan penyakit pasien. Dari gambar *pheno* data diatas mempunyai keseluruhan 38 variabel dengan jumlah sampel sebanyak 35

5.3 *Pre-Processing*

Pada tahap *pre-processing* diperlukan packacge yakni `affyPLM` dengan sebuah fungsi `threestep()` yang dapat digunakan menghilangkan *non* biologis dan *noise* untuk data. Pada tahapannya ekspresi data gen untuk menghilangkan nilai *non* biologis, sehingga *boxplot* yang sudah di *pre-processing* cuma ada data bersifat biologis. Hasil dari proses *pre-processing* dapat berupa *boxplot*. Pada bentuk tabung yang vertikal terdapat tiga bagian yakni kuartil pertama (Q1) bagian paling atas, lalu kuartil kedua (Q2) yang merupakan median berada ditengah, dan kuartil ketiga (Q3) bagian paling bawah. Terdapat garis patah-patah yang berada diatas dan bawah *boxplot* yang dinamakan dengan *whiskers*, *whiskers* sendiri diartikan nilai yang lebih rendah atau lebih tinggi dari data yang berada pada IQR. Sehingga jika ada nilai yang berada paling atas atau bawah dari *whiskers* itu sendiri dinamakan *outlier* atau nilai ekstrim.



Gambar 5.3 Boxplot sebelum *pre-processing*



Gambar 5.4 Boxplot sesudah *pre-processing*

Berdasarkan gambar 5.3 banyak sekali faktor *non* biologis yang terdapat dalam *microarray*, selain itu terdapat gen yang nilai rata-ratanya mendekati 0 dan juga ditakutkan variansi yang ditimbulkan sangat rendah. Sehingga setelah dilakukan *pre-processing* seperti gambar 5.4 terlihat tidak ada gen yang rata-ratanya 0 yang juga mempengaruhi dari variansi, karena dalam ilmu statistika jika variansi rendah maka sulit atau tidak dapat dibedakan ekspresi gen baik gen dari klasifikasi itu sendiri ataupun antar klasifikasi yakni klasifikasi 1 dan klasifikasi 2. Oleh karena itu perlu dilakukan penghapusan agar pengerjaan lebih efisien dan tidak mempengaruhi memori dan lamanya pengerjaan dalam komputer.

5.4 *Filtering*

Pada proses *filtering* untuk data *microarray* ialah sebuah proses yang memilih subset dari suatu *probe* yang tersedia, namun terdapat pengecualian didalam analisis dalam program R dibutuhkan *package* `genefilter` dalam

melakukan sebuah penyaringan dan juga dibutuhkan *function* `nsFilter` dalam melakukan *filtering*.

Tabel 5.1 *Filtering*

	Gen	Sampel
Sebelum *Filtering*	54675	35
Sesudah *Filtering*	10093	35

Selepas melakukan *filtering* data, dalam penghapusan data dengan nilai IQR tertentu dalam menghilangkan ID gen entrez yang sama dan menyaring sebuah *probe control* yang mana dalam data ekspresi gen digunakan *probe control* `AFFX` yang tidak disertakan dalam suatu analisis. Hasil dari *filtering* data didapat 10093 gen data dari 54675 gen dalam 35 sampel.

5.5 *Feature Selection*

Tahap *feature selection* bertujuan menseleksi variabel yang paling berpengaruh dari hasil filtering antara data yang terkena dan tidak terkena HIV. Pada tahap ini digunakan *package* `multtest` yang bisa diolah apabila data dalam bentuk matriks atau data frame. Untuk menghasilkan data frame atau matriks dapat dilakukan dengan merubah data yang dalam bentuk *AffyBatch* menjadi matriks dengan bantuan *function* `exprs`, lalu melakukan *multiple testing* dengan *function* `mt.teststat` yang dapat memberikan cara menghitung nilai uji statistik dengan menggunakan uji *t-test* dikarenakan data memiliki dua kelas antara pasien yang terkena HIV-1 dan yang tidak terkena dengan memakai nilai signifikansi 0,0000001. Penggunaan uji T tersebut juga untuk menseleksi variabel yang paling berpengaruh dari hasil filtering data. Hasil dari proses *filtering* yakni berdimensi 70 × 35 yakni data gen menjadi 70 data dari 35 sampel.

Tabel 5.2 *Feature Selection*

Gen	Sampel
70	35

5.6 Klasifikasi SVM

Pada tahapan sesudah pengolahan data dan didapatkan data berdimensi 70×35, tahap selanjutnya ialah analisis d ata dengan melakukan klasifikasi untuk mempelajari sebuah pola dengan menggunakan data *training* dan untuk hasil pembelajaran digunakan data *testing*. Penggunaan data *training* didalamnya dibagi dua yakni terkena penyakit HIV-1 dan tidak terkena HIV-1 yang mempelajari pola data didasarkan pada masing-masing kelas. Untuk hasil pembelajaran SVM selanjutnya diuji menggunakan *testing* sehingga tingkat akurasi diukur untuuk menguji data baru, proses tersebut dinamai *Machine Learning*.

Pada klasifikasi menggunakan dua data dengan sebuah perbandingan data *training* dengan *testing* yakni 80%:20% atau sebanyak 28 sampel data train dan 7 sampel data test menggunakan perintah *ratio* untuk penentuan data yang diambil secara random.

Tabel 5.3 Pembagian Data

	Data Training	Data Testing
Ratio	80%	20%
Jumlah	28	7

Dari perbandingan 4 *kernel* yakni Linear, Polynomial, RBF, dan Sigmoid yang dilakukan didapatkan hasil akurasi terbaik yakni *kernel* Linear, Polynomial, dan Sigmoid yang didasarkan percobaan memakai data HIV-1. Karena hasil dari ketiga *kernel* sama maka peneliti memilih menggunakan *kernel* Linear, berikut tabel hasil yang didapat.

Tabel 5.4 Hasil *Kernel*

Kernel	Akurasi Data Test	Best Parameter		Best Performance
		Cost	Gamma	
Linear	85,71%	1	-	0,33333333
Polynomial	85,71%	100	-	-
RBF	71,42%	0,1	1	0,23333333
Sigmoid	85,71%	0,1	1	-

Berdasarkan penggunaan *kernel* Linear pada metode SVM data HIV-1 yang terkena HIV-1 dan tidak terkena HIV-1, dari data *training* yang digunakan diperoleh 7 pasien yang masuk ke dalam kelas tidak terkena HIV-1 dan tidak ada kesalahan, lalu ada 21 pasien yang masuk ke dalam yang terkena HIV-1 dan tidak ada kesalahan dengan keseluruhan 28 pasien dari data training. Selanjutnya dari data *testing* diperoleh 2 pasien yang tidak terkena HIV-1 dengan satu nilai kesalahan, lalu terdapat 4 pasien yang terkena HIV-1 dan tidak ada kesalahan dengan keseluruhan 7 pasien dari data testing. Hasil dapat dilihat pada **Tabel 5.5** dibawah.

Tabel 5.5 *Confusion Matrix Linear*

Data	Prediksi	Linear		Jumlah	TOTAL
		NO HIV	HIV		
Training	NO HIV	7	0	7	28
	HIV	0	21	21	
Testing	NO HIV	2	1	3	7
	HIV	0	4	4	

Setelah melakukan *confusion matrix* maka akan dapat dicari nilai *accuracy*, *recall*, dan presisi dengan perhitungan sebagai berikut :

Recall Training $= \frac{\text{TP}}{\text{TP+FN}} \times 100\%$

$= \frac{7}{7} \times 100\% = 100\%$

Precission Training $= \frac{\text{TP}}{\text{FP+TP}} \times 100\%$

$= \frac{7}{7} \times 100\% = 100\%$

Accuracy Training $= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \times 100\%$

$= \frac{28}{28} \times 100\% = 100\%$

Berdasarkan perhitungan data *training* didapat hasil *recall training* sebesar 100%, *precission training* sebesar 100%, dan *accuracy training* sebesar 100%.

Dari semua hasil yang sama maka dapat ditarik kesimpulan nilai *recall*, *precission*, dan *accuracy* dari klasifikasi yakni tepat atau bisa dikatakan klasifikasi data *training* sudah tepat.

Recall Testing = $\frac{\text{TP}}{\text{TP+FN}} \times 100\%$

= $\frac{2}{2} \times 100\% = 100\%$

Precission Testing = $\frac{\text{TP}}{\text{FP+TP}} \times 100\%$

= $\frac{2}{3} \times 100\% = 66{,}67\%$

Accuracy Testing = $\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \times 100\%$

= $\frac{6}{7} \times 100\% = 85{,}72\%$

Specificity = $\frac{\text{TN}}{\text{TN+FP}} \times 100\%$

= $\frac{4}{5} \times 100\% = 80\%$

FPR = 1 − Spesificity

= 1 – 0,8 = 0,2

AUC = $\frac{1+\text{Recall}-\text{FPR}}{2}$

= $\frac{1+1-0{,}2}{2} = 0{,}9$

Berdasarkan hasil perhitungan dengan data *testing* diperoleh hasil nilai *recall testing* 100%, *precission testing* 66,67%, *accuracy testing* 85,72%, *specificity* 80%, FPR 20%, dan AUC 90%. Dari hasil nilai *recall*, *precission*, dan *accuracy testing* yang tinggi maka bisa dikatakan klasifikasi data *testing* sudah tepat. Dari pengukuran lainnya yakni *specificity*, dan AUC juga cukup besar dan bisa dikatakan pula bahwa klasifikasi cukup baik. Dari nilai *confusion matrix* juga menghasilkan perhitungan nilai kesalahan dalam klasifikasi dengan perhitungan yakni.

$$Error = \frac{\text{jumlah prediksi salah}}{\text{jumlah prediksi keseluruhan}}$$

$$= \frac{1}{7} = 0{,}143 = 14{,}3\%$$

Dari hasil *error* yang diperoleh maka bisa disimpulkan tingkat kesalahan yang dimiliki kecil yakni 14,3% atau 0,143.

5.7 Model Klasifikasi

Dalam melakukan model klasifikasi dengan SVM didapatkan sebuah model *hyperlane* yang dapat memisahkan dua kelas bahkan lebih dan juga memiliki bobot yang dapat memberi jarak satu kelas dengan yang lainnya. Untuk membentuk sebuah model peneliti cukup menggunakan 10 model sebagai acuan dari nilai tertinggi dan didapatkan hasil bahwa probe yang paling berperan dalam menyebabkan HIV-1.

Tabel 5.6 Bobot gen 10 terbesar

PROBE	W
X.210193_at.	8.7866025
X.226702_at.	8.6508198
X.205269_at.	8.4661205
X.203153_at.	8.2444033
X.218559_s_at.	8.0574257
X.202086_at.	7.8296829
X.200838_at.	7.7826144
X.202510_s_at.	7.7209012
X.204037_at.	7.7198682
X.202430_s_at.	7.5236031

Berdasarkan tabel 5.6 diatas seriap probe memiliki peran dan probe X.210193_at ialah probe tertinggi atau paling berpengaruh menyebabkan HIV dengan bobot (w) sebesar 8,7866025, kemudian disusul dengan *probe* X.226702_at dengan bobot 8,6508198, lalu probe X.205269_at dengan bobot 8,4661205, dan selanjutnya dapat dilihat pada Gambar 5.6 dibawah.