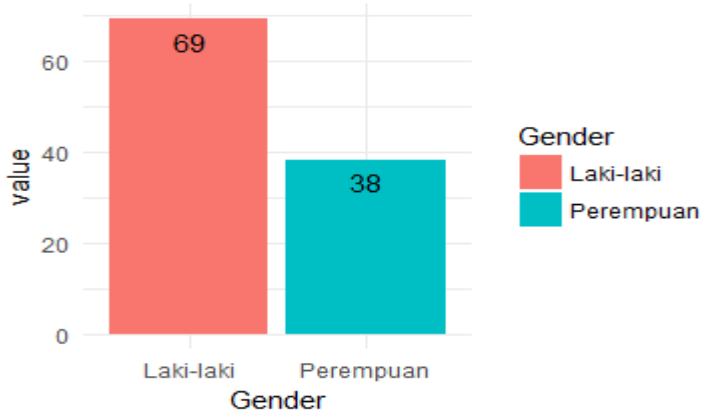
BAB V
PEMBAHASAN

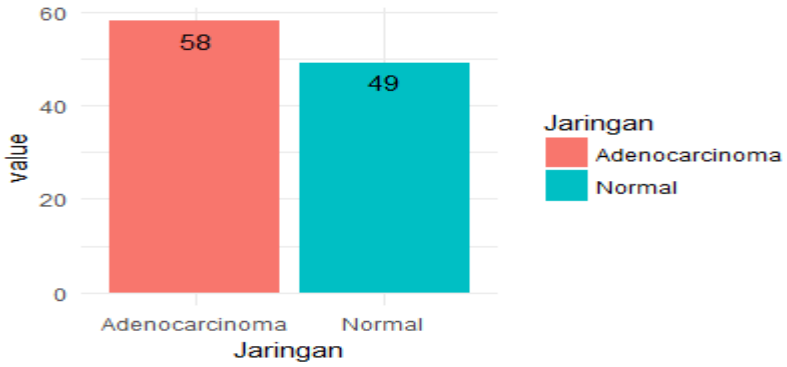
5.1 Deskripsi Data

Berdasarkan *gene expression data* pasien kanker paru-paru dengan jenis kelamin laki-laki lebih banyak dibandingkan jumlah perempuan yaitu 69 pasien laki-laki dan 38 pasien perempuan. Jumlah pasien laki-laki lebih besar bisa saja dikarenakan adanya kebiasan merokok pasien yang dapat meningkatkan resiko terkena kanker paru-paru.



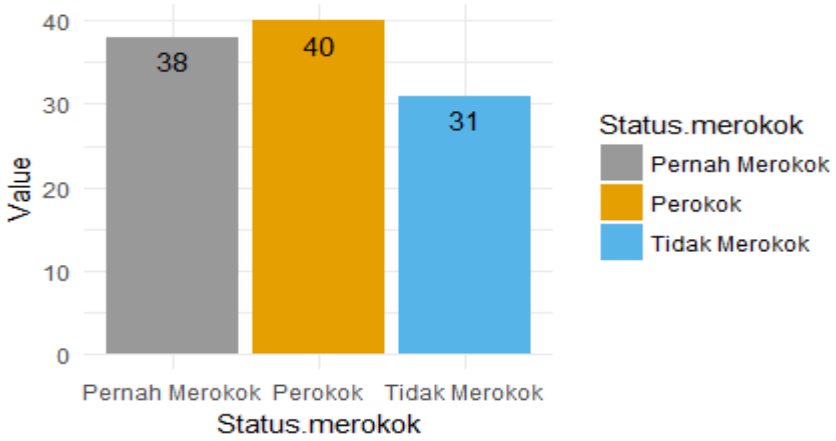
Gambar 5.1 *Bar plot jenis kelamin pasien kanker paru-paru*

Berdasarkan jaringan tempat kanker berada terdapat 58 pasien menderita kanker pada jaringan *Adenocarsinoma* sedangkan 49 pasien menderita kanker pada jaringan paru-paru normal.



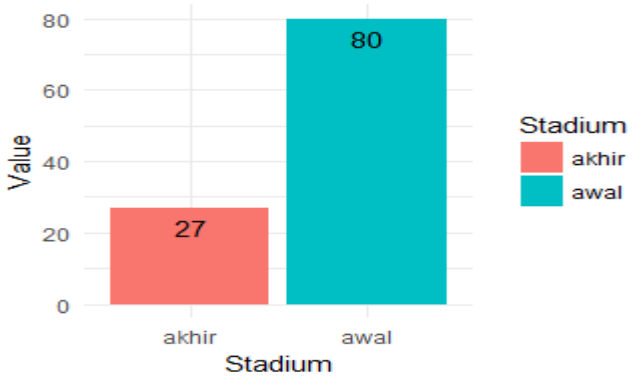
Gambar 5.2 *Bar plot jaringan pasien kanker paru-paru*

Merokok adalah salah satu faktor resiko yang paling besar mempengaruhi tingkat resiko seseorang terkena kanker paru-paru. Berdasarkan *bar plot* pasien kanker paru-parupada gambar 5.3dapat dilihat bahwa pasien perkok dan pernah merokok lebih banyak dibandingkan tidak merokok. Pasien dengan status perokok sebanyak 40 orang, yang pernah merokok 38 orang, dan tidak merokok adalah 31 .



Gambar 5.3 *Bar plot status merokok pasien kanker paru-paru*

Stadium kanker paru-paru di kelompokan menjadi dua yaitu stadium awal dan akhir. Dimana, stadium awal adalah gabungan dari stadium IA,IB,IIA, dan IIB, sedangkan stadium akhir adalah gabungan stadium IIIA,IIIB dan IV. pada kasus ini jumlah pasien dengan stadium awl lebih banyak dibandingkan stadium akhir yaitu 80 dan 27.



Gambar 5.4 *Bar plot stadium pasien kanker paru-paru*

5.2 *Proses Data Microarray*

*Gene expression microarray data*akandi proses dengan tiga tahap yaitu *preprocessing*, *filtering*, dan*feture selection*.*Preprocessing*adalah suatu proses yang berfungsi untuk menghilangkan efek non biologi pada data. Proses kedua adalah *filtering*yaitu proses yang dilakukan utuk meningakan kekuatan analisis, *filtering* berfungsi dengan cara membuang probe yang tidak informatif berdasarkan variansi, atau dari rata-rata sinyal yang dipancarkan.Tahap terakhir adalah *feature selection*adalah peilihan variabel yang relvan untuk digunakan dala proses klasifikasi.

Data yang digunakan dalam penelitian adalah *gene expression data* yang dengan jumlah gen sebanyak 22.238 dan jumlah sample 107 yang akan melewati tahap proses *microarray*.

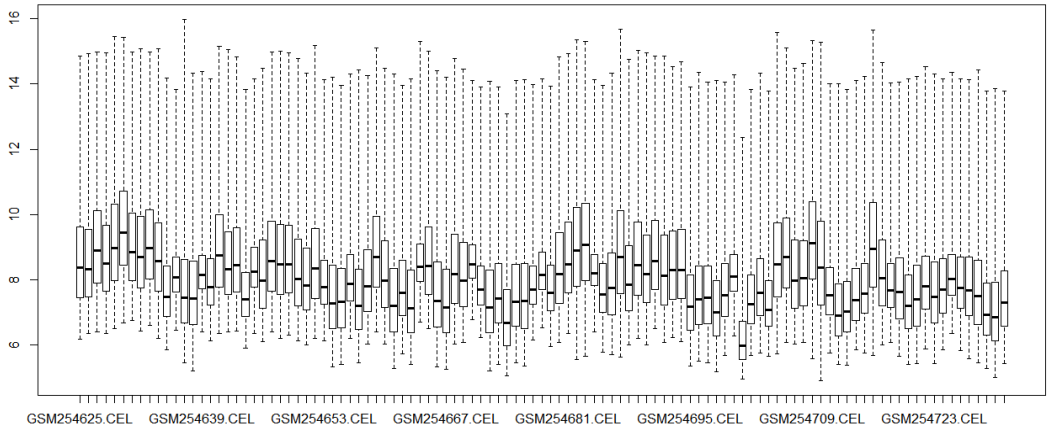
Tabel. 5.1*Gene expression data*

	X219513_at	X203445_s_at	X205381_at	.....	X204400_at
GSM254625	7.12511	1.254579	7.147354	.....	7.489513
GSM254626	6.064396	1.176146	6.634613	.....	7.815476
GSM254627	6.136757	1.048556	6.073493	.....	7.363241
GSM254628	6.068789	1.169542	6.492673	.....	7.814594
:	:	:	:	:	:
:	:	:	:	:	:
GSM254629	6.417902	1.116274	6.245343	..	8.374207

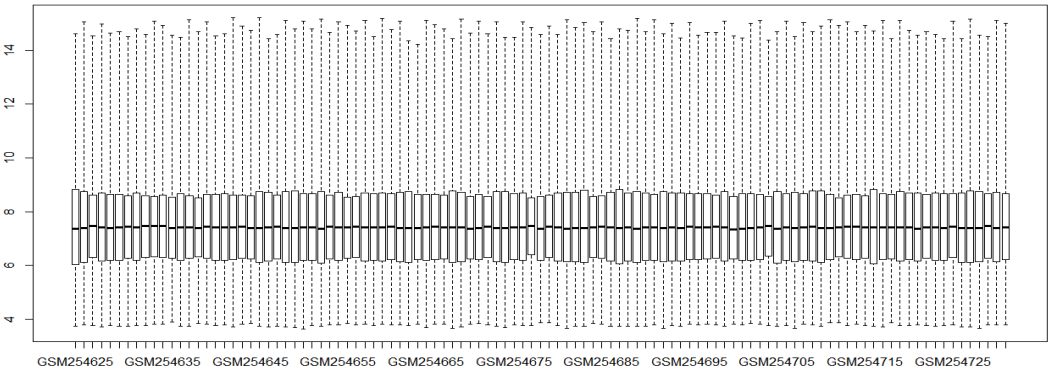
Proses *preprocessing* memiliki tiga tahap yaitu *backgroundcorrection*, *normalization*, dan *summerization*. Tahapbackground correction akan mengkoreksi kesalahan inensitas yang terdetesi keluar dari *chip microarray*ataupun kesalah yang dapat terjadi karena adanya noda yang melekat ada *chip*, tahap *normalization* akan dilakukan untu menghlangkan variansi dari efek non biologis, dan *summarization*

akan menghasikan *gene expression* dengan cara menggabungkan beberapaprobe menjadiprobe set.

Gambar 5.2memperlihatkandata yang di *preprocessing* dengandata sebelum dan setelah proses *preprocessing* dapat dilihat dengan menggunakan *box plot*. *Boxplot*dari *raw data* tidak beraturan dan rentang yang berbeda menyatakan data belum normal..



Gambar 5.5*Box plot raw data GSE10072*

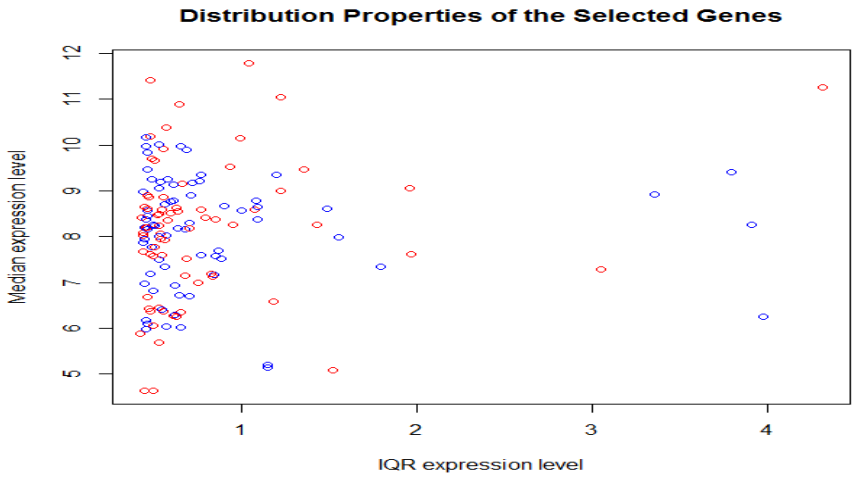


Gambar 5.6*Box plot preprocessing data GSE10072*

Setelah dilakukanya proses *preprocessing*maka 22.238 gen melakuakan proses *filtering*dari 22.238 varibel gen didapatkan 6217 gene, proses *fltering* yang digunakan adalah *non specific filtering* yaitu filtering dengan membuang probe yang duplikat atau sama, gen yang tidak terdeteksi, dan *probe control*, kemudian dilakukan proses*feature selection* dan didapatkan 141 gen dimana proses ini

membuang 97.7% gene yang tidak relevan. adapun *feature selection* yang digunakan adalah metode filter dengan menggunakan *t-test*.*Feature selection t-test* dapat membedakan variabel gen secara signifikan berasal dari jenis stadium.

Hasil *feature selection* pada *gambar 5.6*memperlihatkan bahwa sebagian besar data telah berkumpul mendekati nol pada *IQR* dan berpusat pada nilaimedian delapan, kemudian hanya sedikit data yang menjauh dari nilai nol hal ini menjelaskan bahwa data yang dipilih sudah homogen dan data telah siap untuk dilakukan analisis lanjutan.



Gambar 5.7*Gen yang terpilih*

5.3 Pemilihan nilai *k*

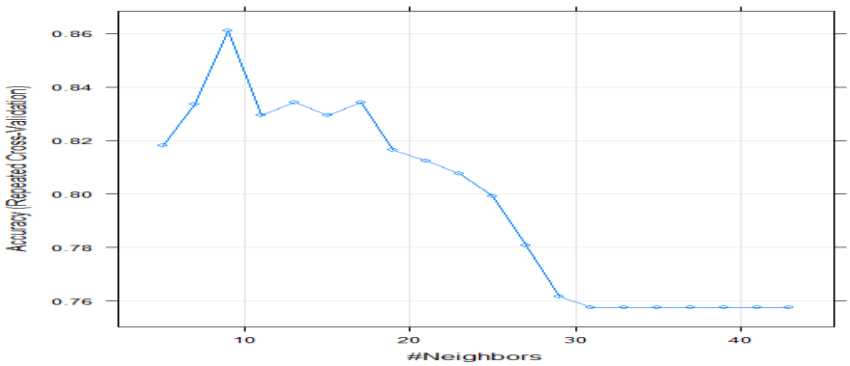
Setelah dilakukan *feature selection* dari gen kemudian akan ditambahkan dengan data fenotip yang berupa umur, jaringan, status merokok, dan stadium kanker.

Klasifikasi dengan menggunakan *gene expression data* pasien dengan jumlah variabel yang terpilih adalah 144 dan 107 sample yang akan di klasifikasikan berdasarkan *stadium*dari jaringan tumor kanker paru-paru NSCLC (*non- small cell lung cancer)*, kemudian dibagi menjadi data *testing* dan data *training* dengan rasio0.7 untuk data *training* dan 0.3 untuk data *testing* maka, didapatkanlah data *training* sebesar 74 dan 33 data *testing*.

Tabel 5.2 *Tabel nilai akurasi dan kappa*

k	Accuracy	Kappa
5	0,82	0,320
7	0,83	0,361
9	0,86	0,438
11	0,83	0,304
13	0,83	0,324
15	0,83	0,310
17	0,83	0,324
19	0,82	0,257
21	0,81	0,237
23	0,81	0,217
25	0,80	0,177
27	0,78	0,098
29	0,76	0,020
31	0,76	0,000

Pemilihan nilai *k* terbaik dengan menggunakan *10-fold cross validation* dengan tabel 5.2 menampilkan nilai akurasi dan nilai *kappa* dari parameter K-NN didapakan nilai akurasi tertinggi ketika *k = 9* dengan nilai 0.86 dan nilai *kappa* sebesar 0.438 yang menyatakan bahwa nilai *k* yang tepilih bagus.



Gambar 5.8 *Pemilihan nilai k terbaik*

5.4 Hasil Klasifikasi

Pemilihan nilai *k* terbaik berdasarkan plot akurasi dapat dilihat bahwa akurasi yang tertinggi berada pada *k=9* dan terpilih sebagai *k* optimal. Adapun hasil prediksi klasifikasi K-NNdengan menggunakan nilai *k* optimal adalah sebagai brikut :

Tabel 5.3 *Hasil prediksi test data*

No	Data *Testing*	Hasil Prediksi	No	Data T*esting*	Hasil Prediksi
1	Stadium akhir	Stadium awal	18	Stadium awal	Stadium awal
2	Stadium akhir	Stadium awal	19	Stadium awal	Stadium awal
3	Stadium akhir	Stadium awal	20	Stadium awal	Stadium awal
4	Stadium akhir	Stadium awal	21	Stadium awal	Stadium awal
5	Stadium akhir	Stadium awal	22	Stadium awal	Stadium awal
6	Stadium akhir	Stadium akhir	23	Stadium awal	Stadium awal
7	Stadium akhir	Stadium akhir	24	Stadium awal	Stadium awal
8	Stadium akhir	Stadium akhir	25	Stadium awal	Stadium awal
9	Stadium akhir	Stadium akhir	26	Stadium awal	Stadium awal
10	Stadium awal	Stadium awal	27	Stadium awal	Stadium awal
11	Stadium awal	Stadium awal	28	Stadium awal	Stadium awal
12	Stadium awal	Stadium awal	29	Stadium awal	Stadium awal
13	Stadium awal	Stadium awal	30	Stadium awal	Stadium awal
14	Stadium awal	Stadium awal	31	Stadium awal	Stadium awal
15	Stadium awal	Stadium awal	32	Stadium awal	Stadium awal
16	Stadium awal	Stadium awal	33	Stadium awal	Stadium awal
17	Stadium awal	Stadium awal			

Hasil klasifikasi pada stadium kanker paru-paruyang terbagi menjadi dua kelompok yaitu jaringan tumor stadium awal dan jaringan tumor stadium akhir dengan masing-masing pada stadium awal terdapat 24 dan 9 pada stadium akhirdata *testing*. Kemudian dilakukan klasifikasi dengan mengunakan*gene expressiondata*dan didapatkan bahwa pengelompokan pada jaringan tumor stadium akhir terdapat kesalahan klasifikasi, dimana terdapat 5 jaringan tumor stadium akhir dikelompokan dalam jaringan tumor stadium awal.

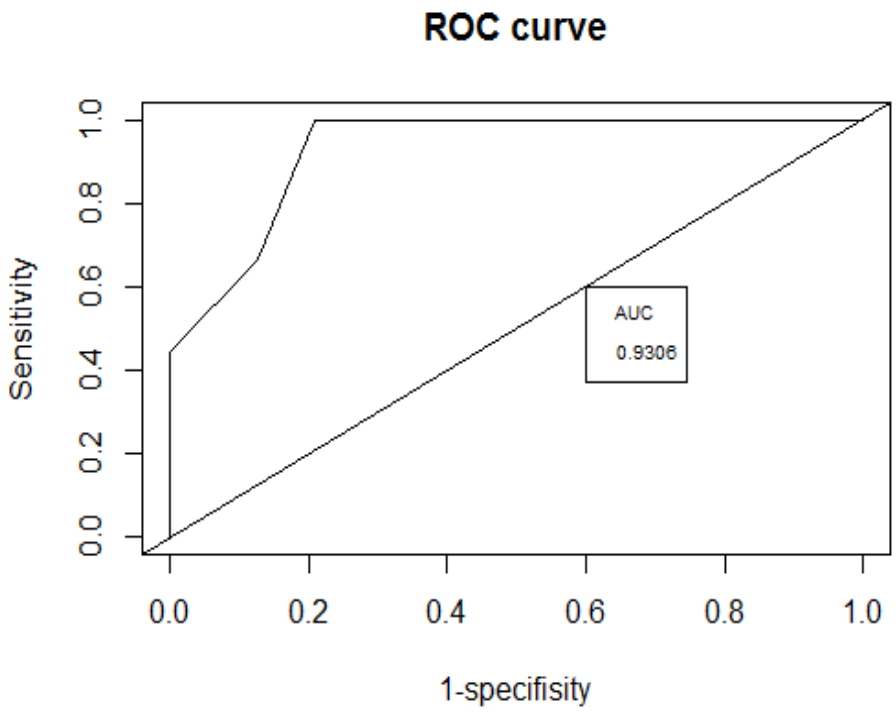
Tabel 5.4 *Confusion matrix*

	Actual value	
Predict value	Jaringan tumor stadium awal	Jaringan tumor stadium akhir
Jaringan tumor stadium awal	24	5
Jaringan tumor stadium akhir	0	4

Validasi klasifikasi pada penelitian klasifikasi dengan menggunakan tabel *confusion matrix* didapatkan bahwa hasil klasifikasi jaringan tumor stadium awal dan diklasifikasi dengan benar adalah sebanyak 24 dan terdapat sebesar 0 kesalahan. Sedangan dari 9 hasil data *test* prediksi jaringan tumor stadium akhir hanya 4 yang diprediksi dengan benar.

Berdasarkn tabel *confusion matrix* dari klasifikasi K-NN dengan $k = 9$ dapat dihitung nilai akurasi, *recall*, dan presisi. Untuk nilai akurasi dari klasifikasi yang didapatkan adalah 0.848 yang dapat diartikan seberapa sering klasifikasi dilakukan secara benar. Untuk nilai *recall* pada analisis ini adalah 1 dan dapat diartikan bahwa hasil klasifiksi jaringan tumor stadium awal yang di klasifikasikan sebagai jaringan tumor stadium awal dengan benar sebesar 1 atau tidak terjadi kesalahan dalam klasifikasi. Sedangkan nilai presisi yang didapatkan adalah 0.827 yang menjelaskan bahwa prediksi klasifikasi jaringan tumor stadium awal yang di klasifikasikan dengan benar.

Menentukan suatu klasifikasi dikatakan baik atau tidak kita tidak hanya bisa mengacu pada hasil akurasi, *recall*, dan *precision*. Hasil klasifikasi dapat dientukan dengan nilai kappa yaitu 0.5378 menjelaskan hasil klasifikasi adalah klasifikasi yang baik.

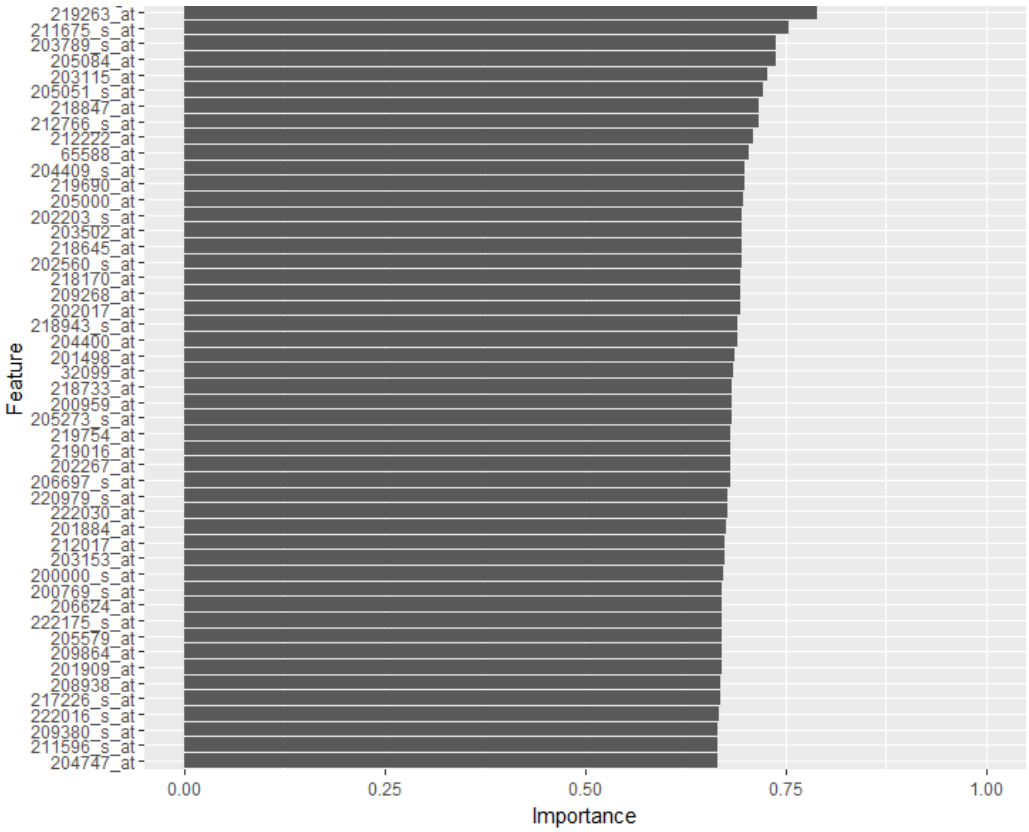


Gambar 5.9*Kurva ROC*

Hasil kurva ROC yang didapatkan untuk dalam klasifikasi untuk membedakan jaringan tumor stadium awal dan jaringan tumor stadium akhir dijelasan bahwa apabila kurva mendekati garis *y(0,1)* maka akan semakin baik klasifikasi dalam membedakan kedua jaringan stadium tumor pada gambar 5.9 didapatkan bahwa garis kurva mendekati nilai y.

Pada kurva ROC digunakan nilai AUC (area dibawah kurva) untuk menghitung keakuratan prediksi. Biasanya nilai AUC berada pada rentang nilai 0.5 hingga 1. Jika nilai AUC mendekati 1 (satu) maka prediksi klasifiksi semakin akurat, dan akan tidak akurat ketika mendekati nilai 0.5 . Berasarkan skala yang di jelaskan oleh Gorunescu (2011) tingkat keakuratan klasifikasi model adalah baik dengan nilai 0.9306.

Penelitian ini menambahkan variabel apa yang berpangaruh dalam dalam menentukan hasil klasifikasi dari K-NN pemilihan variabel berpengaruh menggunakan nilai ROC yang pada setiap variabel gambar 5.10 memperlihatkan bahwa 50 variabel yang berpengaruh adalah variabel gen.



Gambar 5.10 Variabel yang paling berpengaruh

Variabel yang mempengaruhi klasifikasi tidak memiliki rentang perbedaan yang sangat besar pada setiap variabelnya. Berikut adalah 5 (lima) variabel gen yang paling berpengaruh adalah :

PROBEID	SYMBOL	ENTERID	GENENAME
219263_at	RNF128	79589	ring finger protein 128, E3 ubiquitin protein ligase
211675_s_at	MDFIC	29969	MyoD family inhibitor domain containing
205084_at	BCAP29	55973	B-cell receptor associated protein 29
203789_s_at	SEMA3C	10512	semaphorin 3C
203115_at	FECH	2235	Ferrochelatase

Tabel 5.5 *Annotation variabel yang paling berpengaruh*

Adapun variabel yang paling berpengaruh dari klasifikasi K-NN yang pertama biasanya dikenal sebagai GRAIL yang merupakan protein yang dikodekan oleh gen yang merupakan protein trans membran tipe 1 yang terlokasi pada jalur endositik, *gene expression* ini menghambat produksi IL2 dan IL4. Yang kedua adalah MyoD family inhibitor domain containing produk gen ini adalah anggota keluarga protein yang ditandai dengan domain C-terminal cysteine yang kaya, yang terlibat dalam regulasi transkripsi ekspresi genom virus. Inisiasi translasi alternatif dari hulu non-AUG (GUG), dan kodon AUG dalam-frame,menghasilkan dua isoform, p40 dan p32, masing-masing memiliki lokalisasi subselular yang berbeda. semaphorin class 3 family isyarat bimbingan neuronal. Protein yang dikodekan mengandung domain sema N-terminal, integrin dan domain mirip imunoglobulin, dan domain dasar C-terminal. Homodimerisasi dan pembelahan proteolitik propeptida C-terminal diperlukan untuk fungsi protein yang dikodekan. Ini mengikat reseptor neuropilin sebelum membentuk kompleks heterotrimer dengan pleksus terkait. Peningkatan ekspresi gen ini berkorelasi dengan peningkatan invasi sel kanker dan adhesi. Mutasi alami pada gen ini terkait dengan penyakit Hirschsprung, dan ferrochelatase yang merupakan protein yang dikodekan oleh gen ini dilokalisasi ke mitokondria, di mana ia mengkatalisis penyisipan bentuk besi besi menjadi protoporfirin IX pada jalur sintesis heme. Mutasi pada gen ini terkait dengan eritropoietik protoporphyria. Dua varian transkrip yang mengkodekan isoform yang berbeda telah ditemukan untuk gen ini. Sebuah pseudogene dari gen ini ditemukan pada kromosom 3.