

**IMPLEMENTASI METODE SUPPORT VECTOR MACHINE
DAN RANDOM FOREST PADA DATA EKSPRESI GEN
MICROARRAY**

(Studi Kasus : *Novel Mutation Target Distinct Subgroups of Medulloblastoma in America*)

TUGAS AKHIR



Revika Dwi Merianti

16611042

**PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS ISLAM INDONESIA
YOGYAKARTA
2020**

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul : Implementasi Metode Support Vector Machine Dan
Random Forest Pada Data Ekspresi Gen Microarray.
(Studi Kasus : *Novel Mutation Target Distinct Subgroups
of Medulloblastoma in America*)

Nama Mahasiswa : Revika Dwi Merianti

NIM : 16611042

**TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK
DIUJIKAN**

Yogyakarta, 16 Juli 2020

Pembimbing


(Dr. techn. Rohmatul Fajriyah, S.Si., M.Si)

HALAMAN PENGESAHAN

TUGAS AKHIR

IMPLEMENTASI METODE SUPPORT VECTOR MACHINE DAN RANDOM FOREST PADA DATA EKSPRESI GEN MICROARRAY

(Studi Kasus : *Novel Mutation Target Distinct Subgroups of Medulloblastoma in America*)

Nama Mahasiswa : Revika Dwi Merianti

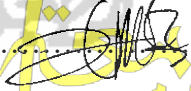
NIM : 16611042

TUGAS AKHIR INI TELAH DIUJIKAN

PADA TANGGAL: 24 Juli 2020

Nama Penguji:

Tanda Tangan

1. Dr. techn. Rohmatul Fajriyah, S.Si., M.Si
2. Ayundyah Kesumawati, S.Si., M.Si
3. Dina Tri Utari, S.Si., M.Sc

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam



(Prof. Riyanto, S.Pd., M.Si., Ph.D.)

KATA PENGANTAR

Assalamualaikum Wr.Wb

Alhamdulillah rabbil 'alamin, puji dan syukur penulis panjatkan kehadirat Allah SWT yang telah melimpahkan rahmat dan hidayah-Nya, serta memberi kemudahan dan kesehatan fisik serta akal pikiran, sehingga penulis dapat menyelesaikan Tugas Akhir ini dengan tepat waktu, guna untuk memenuhi syarat menyelesaikan studi Strata 1 dan guna untuk mencapai gelar Sarjana Statistika di prodi Statistika Universitas Islam Indonesia Yogyakarta. Solawat serta salam tidak lupa penulis haturkan kepada junjungan umat Islam yaitu baginda Nabi Muhammad SAW agar umat Islam memperoleh syafatnya di akhirat nanti.

Tugas akhir ini merupakan tugas yang harus diselesaikan oleh penulis sebagai hasil akhir dari kegiatan belajar selama berada di prodi Statistika untuk mendapatkan gelar S1 Statistika. Tugas akhir yang ditulis berjudul **“Implementasi Metode Support Vector Machine Dan Random Forest Pada Data Ekspresi Gen Microarray”** dengan studi kasus yang diangkat yaitu **“Novel Mutation Target Distinct Subgrups of Medulloblastoma in America”**. Selain itu, tugas akhir ini ditulis juga untuk memberikan informasi terkait medulloblastoma dan pengklasifikasian terhadap pasien medulloblastoma.

Penulis juga tidak lupa mengucapkan banyak terima kasih kepada pihak yang telah membantu, membimbing, dan mendukung dalam proses penulisan tugas akhir ini. Oleh karena itu, pada kesempatan kali ini, penulis ingin mengucapkan rasa terima kasih tersebut kepada:

1. Kedua orang tua, kakak dan adik saya yang selalu memberikan doa, semangat, serta dukungan materi dari awal kuliah sampai penyusunan tugas akhir ini selesai.

2. Bapak Prof. Fathul Wahid, S.T., M.Sc., Ph.D. selaku Rektor Universitas Islam Indonesia Yogyakarta.
3. Bapak Prof. Riyanto, S.Pd., M.Si., Ph.D. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia Yogyakarta.
4. Bapak Dr. Edy Widodo, S.Si., M.Si. selaku Ketua Prodi Statistika Universitas Islam Indonesia Yogyakarta.
5. Ibu Dr. techn. Rohmatul Fajriyah, S.Si., M.Si. selaku Dosen Pembimbing Skripsi yang telah bersedia meluangkan waktunya untuk memberikan arahan, bimbingan, dan bantuan selama proses penyusunan tugas akhir.
6. Ibu Kariyam, S.Si., M.Si., selaku Dosen Pembimbing Akademik penulis yang sangat membantu mengarahkan selama masa perkuliahan hingga sampai menyelesaikan Tugas Akhir ini.
7. Semua dosen yang selama ini telah memberikan ilmunya kepada penulis dengan sabar dan ikhlas dan *staff* prodi Statistika yang selalu membantu penulis dalam urusan administrasi.
8. Hanum Salsabiella, Salsabilla Novianingrum, Dianawati Khaerunnisa, Ella Anggraini, Nadia Ihya Muthi, Yesih Nurmallasari, dan Anggit Novietasari sebagai sahabat yang selalu membantu, memberi semangat dan doa kepada penulis selama kuliah sampai penyusunan tugas akhir ini selesai.
9. Teman-teman sebimbingan skripsi yang selalu saling mengingatkan dan menyemangati satu sama lain dan tentunya saling membantu.
10. Semua pihak yang tidak dapat penulis sebutkan satu per satu.

Tentunya dalam penulisan tugas akhir ini, penulis menyadari bahwa masih belum sempurna dan tentunya terdapat banyak kekurangan karena penulis juga masih dalam tahap belajar. Dengan ini, penulis sangat mengharapkan kritik dan saran dari semua pembaca agar bisa menjadi catatan bagi penulis untuk penulisan yang akan datang dan tidak terjadi kesalahan yang sama. Atas kekurangan dan kesalahan yang ada dalam penulisan tugas akhir ini penulis memohon maaf yang sebesar-besarnya. Dan juga penulis sangat berharap semoga tugas akhir ini dapat bermanfaat dan dapat memberi informasi bagi yang membaca, dan tentunya

dapat memberikan informasi akan penyakit medulloblastoma yang penulis angkat dalam tugas akhir ini.

Wassalamualaikum Wr.Wb

Yogyakarta, 16 Juli 2020



Revika Dwi Merianti



DAFTAR ISI

HALAMAN SAMPUL	i
HALAMAN PERSETUJUAN PEMBEMBIING	ii
HALAMAN PENGESAHAN	iii
KATA PENGANTAR	iv
DAFTAR ISI	vii
DAFTAR TABEL	x
DAFTAR GAMBAR	xi
DAFTAR LAMIRAN	xii
PERNYATAAN	xiii
ABSTRAK	xiv
ABSTRACT	xv
BAB I	1
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah.....	3
1.4 Tujuan Penelitian.....	3
1.5 Manfaat Penelitian.....	4
BAB II	5
TINJAUAN PUSTAKA	5
BAB III	13
LANDASAN TEORI	13
3.1 Medulloblastoma	13
3.2 Bioinformatika.....	14
3.3 Microarray	15
3.4 <i>Preprocessing</i>	15
3.5 <i>Filtering</i>	16
3.6 <i>Feature Selection</i>	17
3.7 <i>Machine Learning</i>	17
3.8 <i>Data Mining</i>	18
3.9 Klasifikasi.....	18
3.10 <i>Support Vector Machine (SVM)</i>	19
3.10.1 <i>Support Vector Machine (SVM) Binary</i>	19

3.10.2	Kernel <i>support vector machine</i> (SVM).....	21
3.10.3	<i>Support Vector Machine</i> (SVM) Multiclass	23
3.11	<i>Random Forest</i>	25
3.12	<i>Confusion Matrix</i>	26
3.13	ROC (<i>Receiver Operating Characteristic</i>) Curve	27
BAB IV		29
METODOLOGI PENELITIAN		29
4.1	Jenis dan Sumber Data	29
4.2	Tempat dan waktu Penelitian	29
4.3	Variabel Penelitian	29
4.4	Metode Analisis data	29
4.5	Tahapan Penelitian	30
BAB V		32
HASIL DAN PEMBAHASAN		32
5.1	Deskripsi Data	32
5.2	<i>Preprocessing</i>	35
5.3	<i>Filtering</i>	36
5.4	Analisis <i>support vector machine</i> (SVM).....	38
5.4.1	<i>Confussion Matrix support vector machine</i> (SVM).....	38
5.4.2	ROC Curve SVM	41
5.4.3	Model Klasifikasi.....	41
5.5	Analisis <i>Random Forest</i>	45
5.5.1	Hasil <i>Mtry</i> dan <i>Ntree</i>	45
5.5.2	<i>Confussion Matrix Random Forest</i>	46
5.5.3	ROC Curve <i>Random Forest</i>	47
5.5.4	<i>Importance Variable</i>	47
5.6	Perbandingan Metode <i>Support Vector Machine</i> (SVM) Dan <i>Random Forest</i>	51
BAB VI		53

PENUTUP.....	53
6.1 Kesimpulan.....	53
6.2 Saran.....	53
DAFTAR PUSTAKA	54
LAMPIRAN.....	58

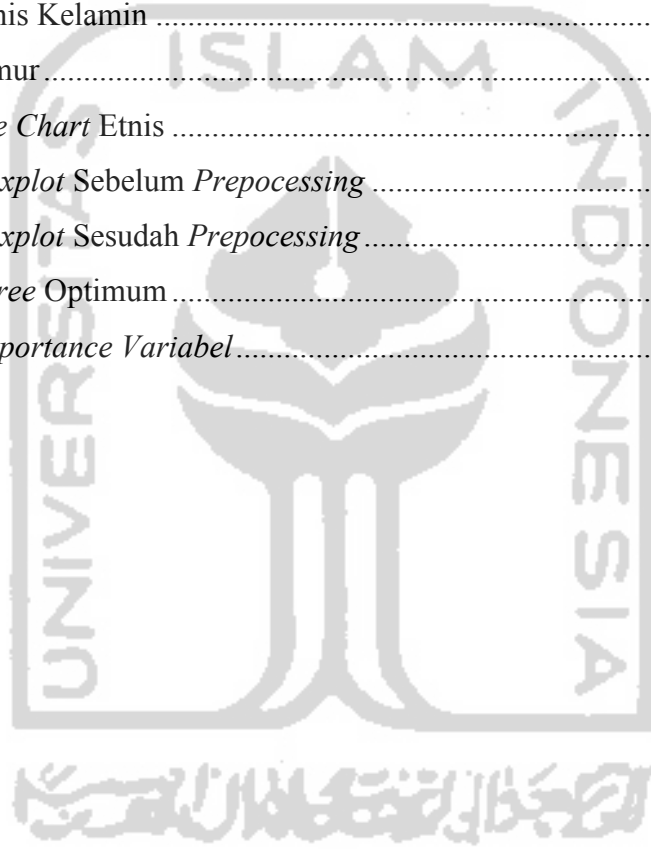


DAFTAR TABEL

Tabel 2.1 Jurnal Penelitian	5
Tabel 3.1 Ringkasan Metode <i>Preprocessing</i> Data <i>Microarray</i>	16
Tabel 3.2 Contoh SVM Pendekatan <i>One-Against-All</i>	24
Tabel 3.3 Contoh SVM Pendekatan <i>One-Against-One</i>	24
Tabel 5.1 Filtering nsFilter	37
Tabel 5.2 Filtering Multttest	37
Tabel 5.3 Pembagian Data Latih (<i>Train</i>) dan Data Uji (<i>Test</i>)	38
Tabel 5.4 <i>Confussion Matrix</i> (SVM) Data Latih (<i>Train</i>)	38
Tabel 5.5 <i>Confussion Matrix</i> (SVM) Data Uji (<i>Test</i>) Kernel Linear	39
Tabel 5.6 <i>Confussion Matrix</i> (SVM) Data Uji (<i>Test</i>) Kernel <i>Polynomial</i>	40
Tabel 5.7 <i>Confussion Matrix</i> (SVM) Data Uji (<i>Test</i>) Kernel Sigmoid	40
Tabel 5.8 Bobot <i>Probe</i> ID	42
Tabel 5.9 Keterangan <i>Probe</i> ID dari Hasil Analisis SVM Kernel Linear	43
Tabel 5.10 Nilai dar Masing-masing <i>Mtry</i>	45
Tabel 5.11 <i>Confussion Matrix</i> <i>Random Forest</i> Data Uji (<i>Test</i>)	46
Tabel 5.12 Keterangan <i>Probe</i> ID dari Hasil Analisis <i>Random Forest</i>	48
Tabel 5.13 Perbandingan Akurasi SVM dan <i>Random Forest</i>	51

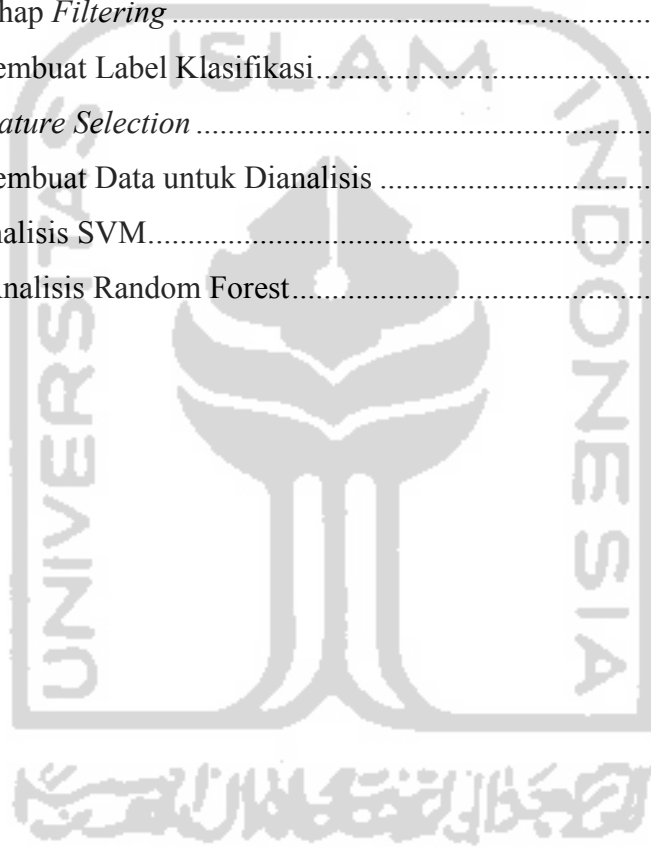
DAFTAR GAMBAR

Gambar 3.1 Langkah-Langkah dalam Pengolahan <i>Gen Expression</i>	15
Gambar 3.2 Penentuan <i>Hyperplane</i> Terbaik	19
Gambar 3.3 Data <i>Nonlinear</i>	22
Gambar 4.1 <i>Flowchart</i> Penelitian.....	30
Gambar 5.1 <i>Pie Chart</i> Subgrup <i>Medulloblastoma</i>	32
Gambar 5.2 Jenis Kelamin	33
Gambar 5.3 Umur	34
Gambar 5.4 <i>Pie Chart</i> Etnis	34
Gambar 5.5 <i>Boxplot</i> Sebelum <i>Preprocessing</i>	36
Gambar 5.6 <i>Boxplot</i> Sesudah <i>Preprocessing</i>	36
Gambar 5.7 <i>Ntree</i> Optimum	45
Gambar 5.8 <i>Importance Variabel</i>	47



DAFTAR LAMPIRAN

Lampiran 1 Membaca Data <i>Microarray</i>	59
Lampiran 2 Melihat dan Menyimpan Pheno Data	59
Lampiran 3 visualisasi Data	60
Lampiran 4 Tahap <i>Preprocessing</i>	60
Lampiran 5 Tahap <i>Filtering</i>	61
Lampiran 6 Membuat Label Klasifikasi	61
Lampiran 7 <i>Feature Selection</i>	61
Lampiran 8 Membuat Data untuk Dianalisis	62
Lampiran 9 Analisis SVM	62
Lampiran 10 Analisis Random Forest	63



PERNYATAAN

Dengan ini saya menyatakan bahwa dalam tugas akhir ini tidak terdapat karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi dan sepanjang sepengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan orang lain, kecuali yang diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 24 Juli 2020



Revika Dwi Merianti

ABSTRAK

IMPLEMENTASI METODE SUPPORT VECTOR MACHINE DAN RANDOM FOREST PADA DATA EKSPRESI GEN MICROARRAY

(studi kasus : *Novel Mutation Target Distinct Subgroups of Medulloblastoma in America*)

Revika Dwi Merianti

Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Islam Indonesia

Medulloblastoma adalah kanker otak kecil yang tumbuh cepat pada bagian belakang otak. Antara 250 hingga 500 anak-anak didiagnosis dengan medulloblastoma setiap tahunnya. Lebih dari 70% terjadi pada anak-anak umur dibawah 10 tahun. Kemungkinan perkembangan medulloblastoma berkurang seiring bertambahnya usia. Medulloblastoma diklasifikasikan sebagai tumor Grade IV yang mana tumor ini ganas dan tumbuh cepat. Ada empat subgrup yang telah diidentifikasi, yaitu WNT, SHH, *Group 3* dan *Group 4*. Subgrup WNT ini sering menempati *fourth ventricle* (ruang yang berisi cairan di tengah fossa posterior), subgrup SHH terutama terletak di tubuh otak kecil dan sering terlihat lateralisasi ke sisi organ, jika terdapat sinyal SHH tidak terkendali maka akan menyebabkan kanker, untuk *group 3* terletak di *fourth ventricle* dan saat didiagnosa sudah menyebar ke area lain dari otak (metastasis), dan *group 4* ini mirip dengan *group 3* dan WNT letaknya tetapi tidak metastasis dan lebih agresif. Untuk mengetahui subgrup ini dapat menggunakan informasi biologis dari pasien yaitu dengan menggunakan data ekspresi gen *microarray*, dimana datanya berukuran sangat besar, menggunakan teknik klasifikasi data *mining*. Metode klasifikasi yang akan digunakan untuk pemecahan masalah medulloblastoma ini adalah *Support Vector Machine* (SVM) dan *Random Forest*. Berdasarkan penelitian diperoleh nilai akurasi klasifikasi dari metode SVM dengan kernel linear, polynomial dan sigmoid masing-masing adalah 90.92%, 77.27%, dan 86.36%. akurasi dengan menggunakan metode *Random Forest* adalah sebesar 90.92%. Nilai AUC SVM dan *Random Forest* yaitu 0.9754 atau 97.54%, dan dikategorikan sebagai excellent classification. Oleh karena itu, dengan melihat hasil dari kedua metode sama dan dengan faktor penyebab medulloblastoma masih sangat minim maka peneliti merekomendasikan untuk menggunakan kedua metode yaitu SVM dan *Random Forest* dalam klasifikasi data subgrup medulloblastoma.

Kata kunci : *Support Vector Machine*, *Random Forest*, Klasifikasi, Medulloblastoma, Kanker Otak.

ABSTRACT

IMPLEMENTATION OF SUPPORT VECTOR MACHINE AND RANDOM FOREST METHODS IN MICROARRAY GEN EXPRESSION DATA

(Case Study : Novel Mutation Target Distinct Subgroups of Medulloblastoma in America)

Revika Dwi Merianti

Department of Statistics, Faculty of Mathematics and Natural Science

Islamic University of Indonesia

Medulloblastoma is a cerebellum cancer that grows rapidly in the back of the brain. Between 250 and 500 children are diagnosed with medulloblastoma every year. More than 70% occurs in children under 10 years old. The possibility of developing medulloblastoma reduced as we get older. Medulloblastoma is all classified as grade IV tumors which is tumor is malignant and grow fast. There are four subgroup that have been identified, namely WNT, SHH, Group 3, and Group 4. The WNT subgroup is often occupy the fourth ventricle (the fluid filled space in the middle of the posterior fossa), the SHH subgroup is mainly located in the cerebellum and is often seen lateralization to the sides. organs, if there is an uncontrolled SHH signal it will cause cancer, for group 3 it is located in the fourth ventricle and when diagnosed has spread to other areas of the brain (metastatic), and group 4 is similar to group 3 and WNT, but not metastatic and more aggressive. To find out this subgroup can use microarray gene expression data, where the data is very large, using the data mining classification technique. The classification method that will be used for the problem solving this medulloblastoma is Support Vector Machine (SVM) and Random Forest. Based on the study obtained the value of the classification of the the classification of the SVM method with linear, polynomial, and sigmoid kernels each of which is, 90.92%, 77.27% and 86.36%. Accuracy using the Random Forest method is equal to 90.92%. the value of AUC SVM and Random Forest is 0.9754 or 97.54%, and is categorized as excellent classification. Therefore, by looking at the results of the two methods there and with the factors causing medulloblastoma are still very minimal, the researchers recommend using both methods, namely SVM and Random Forest in calculating the medulloblastoma subgroup data.

Keywords : *support vector machine, random forest, classification, medulloblastoma, brain cancer.*

BAB I

PENDAHULUAN

1.1 Latar Belakang

Tingkat insiden kanker otak menyumbang sekitar 5% dari kanker seluruh tubuh, menduduki tingkat 70% dari kanker anak (Modern Cancer Hospital, 2020). Kanker otak dapat bersifat primer (kanker mulai tumbuh di otak) atau sekunder (lebih sering disebut otak atau metastasis otak), yang terjadi ketika sel-sel kanker dari bagian lain dari tubuh menyebar ke otak. Ada dua jenis utama kanker otak pada anak salah satunya yaitu Medulloblastoma. (Cancer Council, 2019)

Medulloblastoma adalah kanker otak kecil yang tumbuh cepat pada bagian belakang otak. Antara 250 dan 500 anak-anak didiagnosis dengan medulloblastoma setiap tahun. Lebih dari 70% terjadi pada anak-anak umur dibawah 10 tahun. Kemungkinan perkembangan medulloblastoma berkurang seiring bertambahnya usia. Medulloblastoma dapat terjadi pada orang dewasa, tetapi lebih jarang terjadi. (American Society of Clinical Oncology ® Cancer.Net, Doctor-Approved Patient Information, 2019)

Medulloblastoma semuanya diklasifikasikan sebagai tumor Grade IV. Ini berarti mereka ganas dan tumbuh cepat. Ada empat subgrup yang telah diidentifikasi, yaitu WNT, SHH, Group 3 dan Group 4 (Institute, 2019). Untuk mengetahui subgrup ini dapat menggunakan informasi biologis dari pasien yang dimana dapat ditemukan dalam aplikasi dari alat komputasi dan analisa dalam menangkap dan menginterpretasi data-data biologi yaitu Bioinformatika.

Bioinformatika merupakan salah satu cabang baru ilmu biologi yang dimana merupakan perpaduan antara ilmu biologi dan teknologi informasi. Menurut Durso (1997) bioinformatika adalah manajemen dan analisis dalam informasi biologis yang disimpan dalam database. Dimana informasi biologis itu berupa genomik, proteomik (protein), ekspresi gen dan struktur RNA. Profil transkripsi

yang luas dari medulloblastomas ini akan lebih tepat mengklasifikasikannya sesuai dengan profil ekspresi mRNA.

Pada setiap pasien pasti memiliki DNA dan dari DNA dapat diperoleh sifat dan informasi dari setiap individu termasuk informasi mengenai penyakit kanker otak anak. Dan terdapat Teknologi DNA microarray yang dimana digunakan untuk menentukan tingkat ekspresi ribuan gen yang dilakukan dalam sekali percobaan, dan secara simultan memantau proses biologi yang sedang berlangsung (Siang, et al., 2015). Teknologi ini mampu menghasilkan data yang dibutuhkan untuk proses prediksi dan klasifikasi gen yang diambil dari beberapa jaringan tertentu pada manusia untuk digolongkan ke dalam kanker atau bukan.

Seperti yang telah diketahui bahwa kendala dari suatu data *microarray* adalah besarnya dimensi, sehingga menyebabkan beban komputasi menjadi tidak stabil. Besarnya dimensi ini berkaitan dengan banyaknya informasi yang dapat digali dengan tepat. Proses penggalian informasi yang sebelumnya tidak diketahui dan memproses informasi dapat dilakukan dengan basis data besar kemudian digunakan untuk membuat keputusan yang penting yang disebut dengan *Data Mining* (K, S, & K, 2016). Semakin banyaknya informasi maka kebutuhan akan metode yang akan digunakan untuk mengatasi masalah semakin dirasakan. Dalam hal ini Data Mining menawarkan solusi untuk mengelolah ledakan data dengan skala besar yang terjadi dalam berbagai bidang termasuk dibidang Bioinformatika.

Berdasarkan, klasifikasi dan identifikasi penyakit kanker otak pada anak atau Medulloblatoma dengan menggunakan data ekspresi gen *microarray*, dengan jumlah dan bentuk data yang besar maka dapat dilakukan prediksi menggunakan suatu teknik klasifikasi data mining. Metode klasifikasi yang akan digunakan untuk pemecahan masalah medulloblastoma ini adalah *Support Vector Machine* (SVM) dan *Random Forest*. Peneliti memilih menggunakan metode *Support Vector Machine* (SVM) dan *Random Forest* karena mengacu pada penelitian yang telah dilakukan oleh (Cahyo, 2018) yang membahas tentang Implementasi Metode *Support Vector Machine* untuk Melakukan Klasifikasi pada Data

Bioinformatika dan penelitian yang telah dilakukan oleh (Klassen, 2010) membahas tentang *Learning Microarray Cancer Dataset By Random Forest And Support Vector Mechines* pada data *colon cancer, lung cancer, lymphoma*, dan *prostate cancer*.

1.2 Rumusan Masalah

Berdasarkan pada latar belakang yang telah dipaparkan oleh peneliti, maka terdapat beberapa rumusan masalah yaitu.

1. Berapa nilai optimal yang dihasilkan saat diberlakukan metode *Support Vector Machine* (SVM) dan *Random Forest* ?
2. Bagaimana hasil perbandingan klasifikasi *Support Vector Machine* (SVM) dan *Random Forest* ?

1.3 Batasan Masalah

Agar penelitian ini tidak menyimpang dari pembahasan maka perlu diberikan suatu pembatasan masalah yaitu sebagai berikut.

1. Data yang digunakan adalah data sekunder yang diambil dari *website* NCBI yaitu *microarray gene expression* data GSE37418.
2. Alat bantu yang digunakan adalah *Rstudio* versi 3.6.2 dengan beberaa *package* bioinformatika dari *bioconductor*.
3. Metode statistik yang digunakan adalah *Support Vector Machine* (SVM) dan *Random Forest*.

1.4 Tujuan Penelitian

Terdapat beberapa tujuan peneliti berdasarkan rumusan masalah yang diambil yaitu sebagai berikut :

1. Mengetahui nilai optimal yang dihasilkan saat diberlakukan metode *Support Vector Machine* (SVM) dan *Random Forest*.

2. Menjelaskan hasil perbandingan klasifikasi *Support Vector Machine* (SVM) dan *Random Forest*.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat menjadi salah satu acuan untuk menganalisis dan menjadi referensi pembelajaran tentang bioinformatika. Memberikan suatu informasi tentang pengklasifikasian pasien yang terdiagnosa medulloblastoma berdasarkan *gene expression* dengan nomor series GSE37418. Semoga hasil dari penelitian ini dapat dikembangkan dan dianalisis lebih mendalam dengan menambahkan beberapa faktor lainnya yang mempengaruhi penyakit medulloblastoma kanker otak.



BAB II

TINJAUAN PUSTAKA

Penelitian ini merujuk pada penelitian sebelumnya atau jurnal yang berhubungan dengan penelitian yang dilakukan oleh penulis. Penelitian atau jurnal yang dirujuk penulis adalah jurnal yang berkaitan dengan data bioinformatik dan metode klasifikasi seperti Support Vector Machine (SVM) dan Random Forest.

Tabel 2.1 Jurnal Penelitian

No	Nama dan Judul	Metode	Variabel	Hasil
1.	Klassen (<i>Learning microarray cancer dataset by random forest and support vector machines</i> , 2010)	<ul style="list-style-type: none"> • <i>Support Vector Machine (SVM)</i> - <i>Karnel Linear</i> - <i>Karnel Polynomial</i> - <i>Karnel RBF</i> - <i>Karnel Sigmoidal</i> 	<i>Classes</i> <ul style="list-style-type: none"> • <i>Colon cancer</i> <ul style="list-style-type: none"> - 44 Tumor - 22 Normal • <i>Lung cancer</i> <ul style="list-style-type: none"> - 31 MPM - 150 ADCA • <i>Lymphoma</i> <ul style="list-style-type: none"> - 11 DLBCL 	<ul style="list-style-type: none"> • <i>Colon cancer</i> besar menunjukkan tingkat terendah 0,8125 dan 0,85 masing-masing dengan hutan acak dan SVM. • Dalam Penelitian kali ini <i>Support Vector Machine (SVM)</i> menunjukkan kinerja yang lebih baik dari pada kinerja Random Forest dengan tingkat klasifikasi yang benar sebesar 85%.

No	Nama dan Judul	Metode	Variabel	Hasil
		<ul style="list-style-type: none"> <i>Random Forest</i> 	<ul style="list-style-type: none"> - 42 FL - 9 CLL • <i>Prostate cancer</i> - 52 Normal - 50 Tumor 	
2.	Aliady,Tuasikal,&Widodo (Implementasi <i>Support Vector Machine</i> (SVM) dan <i>Random Forest</i> pada Diagnosis Kanker Payudara, 2018)	<i>Support vector machine (SVM)</i> dan <i>Random Forest</i>	<ul style="list-style-type: none"> • Variabe dependen (Y) terdiri dari 2 yaitu <i>meligant</i> dan <i>benign</i>. • Variabel independen Radius (X1),Tekstur(X2),Primeter(X3),Ar ea(X4),<i>Smoothness</i>(X5),<i>Compactness</i>(X6),<i>Concavit</i> 	<ul style="list-style-type: none"> • Nilai akurasi yang didapatkan menggunakan metode SVM sebesar 93.1%, • Nilai akurasi yang didapatkan menggunakan metode <i>random forest</i> sebesar 94.5%.

No	Nama dan Judul	Metode	Variabel	Hasil
			$y(X7)$, <i>Cocave Point</i> (X8), dan <i>Simetry</i> (X9)	
3.	Huljanah, Rustam, Utama, & Siswantining (<i>Feature Selection Using Random Forest Classifier for Predicting Prostate Cancer</i> , 2019)	<i>Random Forest</i>	<ul style="list-style-type: none"> - <i>States Age</i> - <i>PSA (Prostate-Specific Antigen)</i> - <i>Haemoglobin</i> - <i>Leukocytes</i> - <i>Haematocrit</i> - <i>Thrombocytes</i> - <i>Diagnose</i> 	<ul style="list-style-type: none"> • <i>Random Forest</i> dapat memprediksi kanker prostat dengan baik dan cepat. • Fitur yang paling berpengaruh dalam memprediksi kanker prostat adalah tingkat PSA, leukosit, dan hematokrit. Ketiga, Seleksi Fitur meningkatkan akurasi model hingga 100%.
4.	Octaviani & Rustam (<i>Random Forest for Breast Cancer Prediction</i> , 2018)	<i>Random Forest</i>	<ul style="list-style-type: none"> - <i>Clump Thickness</i> - <i>Unidormity of Cell Size</i> - <i>Uniformity of Cell Shape</i> 	<ul style="list-style-type: none"> • Hasil akurasi yang didapat yaitu lebih dari 99%. • ketika menggunakan data <i>train</i> 80% dan 90% akurasi yang di dapatkan adalah 100%, sehingga

No	Nama dan Judul	Metode	Variabel	Hasil
			<ul style="list-style-type: none"> - <i>Marginal Adhesion</i> - <i>Single Epithelial Cell Size</i> - <i>Bare Nuclei</i> - <i>Bland Chormatin</i> - <i>Normal Necleoli</i> - <i>Mitoses</i> 	dapat disimpulkan akurasi keseluruhan terbaik untuk prediksi kanker payudara
5.	Diani, Wisesty, & Aditsania (Analisis Pengaruh Kernel <i>Support Vector Mechine</i> (SVM) pada Klasifikasi Data, 2017)	<i>Support Vector Machine</i> (SVM)	<p>Jumlah Class</p> <ul style="list-style-type: none"> • Leukimia <ul style="list-style-type: none"> - 47 ALL - 25 AML • <i>Colon Tumor</i> <ul style="list-style-type: none"> - 40 Negatif - 22 Positif • <i>Lung Cancer</i> <ul style="list-style-type: none"> - 31 Mesothelioma - 150 ADCA 	<ul style="list-style-type: none"> • Akurasi yang dihasilkan skenario dua lebih besar dibandingkan skenario satu, seperti pada dataset leukimia, lung cancer dan colon tumor. Akurasi skenario dua meningkat masing – masing sebesar 0.95%, 0,33%, 0.67%. Peningkatan akurasi tersebut membuktikan bahwa pentingnya melakukan perhitungan korelasi untuk mengetahui seberapa kuat hubungan antara atribut (gen) dengan classnya (kanker atau non kanker).

No	Nama dan Judul	Metode	Variabel	Hasil
			<ul style="list-style-type: none"> • <i>Ovarian Cancer</i> <ul style="list-style-type: none"> - 91 Negatif - 162 Positif 	<ul style="list-style-type: none"> • Peningkatan akurasi yang dihasilkan oleh skenario tiga pada dataset leukimia, lung cancer dan colon tumor sebesar 2.45%, 1.01% dan 3.83%. Akurasi tersebut meningkat karena setelah melakukan perhitungan korelasi dilakukan proses seleksi fitur oleh ANOVA, sehingga pasangan gen informatif yang dihasilkan dari ANOVA menjadi pasangan terbaik untuk masukan kedalam SVM dan menghasilkan akurasi terbaik. • Karakteristik dataset kanker yang digunakan di dalam pengujian mempengaruhi akurasi setiap kernel yang diujikan. Sehingga tidak semua dataset akan memiliki akurasi tertinggi dari satu kernel yang sama. • Nilai parameter C, d dan σ yang diubah – ubah berdampak signifikan terhadap akurasi yang

No	Nama dan Judul	Metode	Variabel	Hasil
				<p>dihasilkan seperti pada dataset leukimia dan lung cancer yang mendapatkan akurasi sebesar 100% pada skenario tiga.</p> <ul style="list-style-type: none"> Akurasi yang dihasilkan data colon tumor merupakan akurasi terkecil jika dibandingkan dengan dataset lainnya, sehingga diperlukan penelitian kembali dengan menggunakan metode seleksi fitur atau metode klasifikasi yang berbeda untuk mendapatkan akurasi yang lebih baik. Kemudian penelitian dengan metode yang berbeda dapat dilakukan dengan menambah dataset kanker lain, selain empat dataset yang telah digunakan pada penelitian ini seperti dataset breast cancer dan dataset lainnya.
6.	Mase, Furqon & Rahayudi (Klasifikasi Penyakit Gigi dan Mulut Menggunakan	<i>Support Vector Machine</i> (SVM)	Penyakit kucing : <ul style="list-style-type: none"> - <i>Scabies</i> - <i>Gastritis</i> - <i>Helminthiasis</i> 	Nilai akurasi yang didapatkan dalam penelitian ini dihasilkan dengan menggunakan nilai parameter iterasi = 500, $\lambda = 500$, $\sigma = 1$, $C = 10$, $\gamma = 0.001$, $\varepsilon = 1.10^{-5}$,

No	Nama dan Judul	Metode	Variabel	Hasil
	Metode Support Vector Mechine, 2018)		<ul style="list-style-type: none"> - <i>Rhinitis</i> - <i>Dermatitis</i> - <i>Dermaphytosis</i> - <i>Otitis</i> - <i>Enteritis</i> 	sehingga nilai akurasi yang didapatkan yaitu sebesar 85% dan rerata dari nilai akurasinya sebesar 80.2%.
7.	Cahyo (Implementasi Metode Support Vector Machine untuk Melakukan Klasifikasi Pada Data Bioinformatika, 2018)	<i>Support Vector Machine (SVM)</i>	<ul style="list-style-type: none"> - Etnis - Gender - Subgrup 	<ul style="list-style-type: none"> • metode SVM mampu memprediksi kelas penderita dengan akurasi 95% dengan nilai AUC 98%. • Terdapat probe yang memiliki bobot cukup tinggi yang merupakan gen yang aktif pada berbagai tumor seperti insulinomas, kanker kerongkongan, dan kanker usus besar.
8.	Morfouace, et al. (ABCG2 <i>Transporter Expression Impacts Group 3 Medulloblastoma</i>)	ABCG2 <i>Transporter Expression</i>	<ul style="list-style-type: none"> - <i>Mouse Affymetrix chip (HT430PM)</i> - GSE33200 - <i>Human Affymetrix chip</i> 	Identifikasi ABCG2 sebagai faktor potensial kemoresisten pada G3 Medulloblastoma menunjukkan bahwa inhibitor ABCG2 ampuh dan selektif dapat digunakan untuk meningkatkan respon terhadap obat kemoterapi sebagai substrat dari ABCG2, termasuk topocetan.

No	Nama dan Judul	Metode	Variabel	Hasil
	Response to Chemotherapy, 2015)		(U133v2), GSE37418	
9.	Robinson, et al. (<i>Novel mutations target distinct subgroups of medulloblastoma</i> , 2015)	<i>Somatic SNVs (single-nucleotide variants)</i>	- array Affymetrix U133v2 - SNP 6. • Phs000409 • SRP008292	<i>Modelling of mutations in mouse lower rhombic lip progenitors that generate WNT-subgrup tumours.</i>

Penelitian ini dilakukan oleh penulis merujuk pada **Tabel 2.1.** penulis melakukan penelitian tentang implementasi metode klasifikasi pada data ekspresi gen microarray. Hal yang membedakan penelitian penulis dengan penelitian-penelitian sebelumnya seperti yang telah dijelaskan diatas adalah data yang digunakan (objek penelitian). Penulis menggunakan data *Novel mutations target distinct subgrups of medulloblastoma* dengan *update-an* terakhir tahun 2019. Selain itu dalam penelitian yang dilakukan oleh peneliti, ada 4 (empat) subgrup medulloblastoma yaitu SHH, WNT, Group 3 dan Group 4 yang telah tersedia dalam website resmi NCBI (NCBI, 2019).

BAB III

LANDASAN TEORI

3.1 Medulloblastoma

Medulloblastoma merupakan tumor sel biru bulat ganas (WHO Grade IV) yang mana merupakan tumor ganas yang paling umum terjadi pada anak-anak. Penyakit ini usia puncak awal yaitu pada usia 7-8 tahun, tetapi juga didiagnosis pada bayi dan memasuki usia dewasa, dan sering terjadi pada laki-laki dari pada perempuan dengan rasio perbandingan yaitu 1.5 : 1. Pada akhir tahun 2010, pertemuan konsensus dihadiri oleh kelompok-kelompok terkemuka di komunitas medulloblastoma yang mana diusulkan bahwa medulloblastoma harus dianggap terdiri dari empat subgrup yaitu : WNT, SHH, Group 3 dan Group 4. Dari hasil pertemuan konsensus terkait subgrup yang telah ditetapkan berikut penjabaran dari masing-masing subgrup (Northcott, et al., 2011).

1. WNT

Mereka yang termasuk kedalam subgrup WNT ini terhitung 10-15% dari kasus. Sebagian besar pasien yang didiagnosis pada masa-masa kanak-kanak dan remaja, hampir tidak pernah ditemukan pada bayi. Dan lebih banyak terjadi pada perempuan. Subgrup ini sering terjadi di bagian ventrikel empat dan pada tingkat seluler, subgrup ini menunjukkan akumulasi protein beta-katenin didalam nukleus yang dapat menyebabkan delesi kopi dari kromosom 6. Akumulasi beta-katenin ini juga yang menyebabkan aktivasi jalur WNT. Pasien-pasien yang menderita medulloblastoma dengan subgrup WNT akan menunjukkan keluaran yang baik dengan operasi, radiasi, dan kemoterapi.

2. SHH

Mereka yang termasuk ke dalam subgrup SHH terhitung sebanyak 25% dari kasus. Subgrup ini terjadi pada umur <5 dan >16 tahun. Subgrup ini biasanya dari serebrum dan umumnya terletak dibagian sebelah lateral. Jalur SHH penting dalam pertumbuhan normal serebrum, namun abnormalitas pada

jalur ini menyebabkan pertumbuhan dari tumor. Keluaran yang terjadi pada SHH ini bervariasi dan bergantung pada ada tidaknya metastase, gambaran histologi, dan umur tergantung pada saat pertamakali didiagnosis.

3. Group 3

Untuk Group 3 ini kira-kira terjadi sebanyak 25% dari keseluruhan kasus dan terjadi paling sering pada anak berumur 1-10 tahun. Subgrup ini hampir tidak pernah ditemukan pada usia dewasa. Dalam subgrup ini ketika terdiagnosis seringkali ditemukan metastase. Subgrup ini seperti subgrup WNT umumnya terdapat dibagian ventrikel empat dan berasal dari vermis. Secara mikroskopis, gambaran histologi yang sering terdapat pada group 3 ini adalah besar/anaplastik, namun dapat juga dijumpai tipe klasik. Keluaran dari *group* 3 ini yang paling buruk diantara subgrup yang lainnya.

4. Group 4

Subgrup ini yang paling sering dijumpai, sekitar 35-45% dari keseluruhan kasus. Subgrup ini dapat ditemui di semua kelompok umur, namun sering ditemui pada kelompok usia sekolah. Subgrup ini sama halnya dengan subgrup WNT dan Group 3 yaitu sering dijumpai di ventrikel empat. Ciri khas dari subgrup ini abnormalitas pada kromosom 17, tapi sifatnya tidak eksklusif. Keluaran dari subgrup ini cukup baik sekitar 80% *5-years survival rate* jika tanpa metastase.

3.2 Bioinformatika

Bioinformatika dapat digambarkan sebagai ilmu dalam pengumpulan, permodelan, penyimpanan, pencarian, anotasi dan analisis informasi biologis. Sehingga melibatkan serangkaian kegiatan mulai dari penanganan data, publikasi, hingga data mining dan analisis. Bagian penting dalam bioinformatika adalah membuat algoritma baru agar dapat menganalisis kumpulan data yang kompleks dan/atau besar. Database bioinformatika merupakan beberapa yang terbesar di dunia, yang dimana berurusan dengan masalah yang diciptakan oleh sejumlah besar jenis data baru yang diperoleh melalui eksperimen biologis novel. Bukan hanya masalah dengan datanya yang besar, tetapi ketika di terapkan pada genomik dan

proteomik diperlukan untuk membandingkan dan menganalisis banyak basis data besar yang dilakukan bersamaan, yang dimana harus dilakukan oleh beberapa sistem informasi sebelumnya. (S.K.Agarwal, 2008)

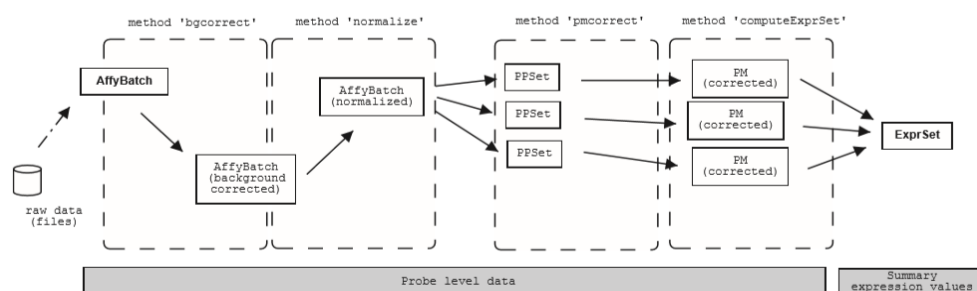
3.3 *Microarray*

Teknologi *microarray*, yang memungkinkan pengumpulan data ekspresi gen dalam skala besar, dan telah matang secara signifikan dalam beberapa tahun terakhir. *Microarray* menawarkan metode dalam mengumpulkan data yang efisien yang dapat digunakan untuk menentukan ekspresi ribuan gen hanya dalam beberapa jam. Metode ini memungkinkan peneliti untuk memeriksa mRNA dan jaringan yang berbeda dalam keadaan normal dan juga penyakit untuk menentukan gen serta kondisi lingkungan yang dapat menyebabkan penyakit. Metode *microarray* juga dapat digunakan untuk menentukan gen mana yang akan diekspresikan di bagian jaringan mana dan pada saat mana selama embrionik (S.K.Agarwal, 2008).

Data *microarray platform affymetrix*, yang berbentuk *affybatch* objek. Yang setelah melalui proses *preprocessing* akan berbentuk *dataframe* atau *matrix*, dimana baris merupakan informasi gen dan kolom merupakan informasi sampel penelitian. Sel-sel dari *matrix* atau *dataframe* merupakan nilai intensitas. Intensitas gen pada sampel tertentu.

3.4 *Preprocessing*

Preprocessing merupakan suatu proses upaya yang dilakukan untuk membuang efek *non* biologis pada data yang berguna untuk memberikan hasil yang lebih baik (Serin, 2011)



Gambar 3. 1 Langkah-Langkah dalam Pengolahan *Gen Expression*

Sumber : (Gautier, Cope, Bolstad, and, & Irizarry, 2004)

Pada kasus umumnya dimana dari data mentah berupa *affybatch* berahli ke bentuk *expression set* memiliki beberapa tahapan *preprocessing* yaitu : *background correction*, *normalization* dan *summarization* (Bolstad, 2004).

1. *Background Correction*

Background correction berfungsi untuk menghilangkan *background noise*, dimana menyesuaikan *cros hibdrization* yang merupakan pengikat dari DNA non spesifik yang melekat pada *array*.

2. *Normalization*

Suatu proses yang memiliki peran untuk menghilangkan variansi non biologis yang tidak diinginkan dan mungkin terdapat pada *microarray*.

3. *Summarization*

Summarization digunakan untuk mengukur *gen expression* yang melalui proses penggabungan beberapa intensitas *probe* menjadi *probe set*.

Untuk melakukan *preprocessing* ini menggunakan bantuan *packages* dari AffyPLM, dimana dapat menggunakan fungsi perintah *threestep* pada data *Affybatch* dengan menggunakan tiga tahapan dan *threestep* merupakan alternatif dalam perhitungan ukuran ekspresi. Berikut beberapa metode yang digunakan dalam *preprocessing* data *microarray* yang dapat digunakan :

Tabel 3. 1 Ringkasan Metode *Preprocessing* Data *Microarray*

	<i>Background correction</i>	<i>Normalization</i>	<i>Summarization</i>
<i>Preprocessing</i>	<i>Robust multichip average (RMA)</i>	<i>Scaling(constant)</i>	<i>Median polish</i>
	<i>MAS</i>	<i>Quantile</i>	<i>Tukey biweight</i>
	<i>Germa</i>	<i>loess</i>	<i>Li-Wong MBEI Expression Index</i>
		<i>Invariant set</i>	<i>Avgdiff</i>
		<i>Variance stabillization mrthode (VSM)</i>	

3.5 *Filtering*

Menurut Gentleman (2003), *filtering* merupakan proses pemilihan subset dari *probe* yang tersedia untuk dikecualikan atau dimasukkan ke dalam analisis dalam artian lain juga *filtering* berguna untuk menghapus sebagian *probe* yang tidak

dibutuhkan, yang dapat berguna untuk analisis. *Filtering* dapat membantu meningkatkan kualitas model, juga dapat membuat proses pemodelan lebih efisien. Fungsi `Nsfilter` (*non specific filter*) menyediakan *one-stop shop* (serba ada) untuk berbagai pilihan *filtering* fitur dari *expression set*. Dalam fungsi `Nsfilter` terdapat beberapa perintah yang dilakukan antara lain, `var.cutoff` berfungsi untuk memfilter gen dengan nilai tertentu untuk menghapus data dengan nilai IQR dibawah quartil. `Require.entrez TRUE` memiliki fungsi untuk memfilter anotasi *Entrez Gene ID*, dimana sistem pengenalan selain ID pusat, maka ID tersebut yang akan diperlukan. Selanjutnya ada `Remove.DupEntrez TRUE` yang akan perintah ini akan menghapus ID *gen Entrez* yang sama atau terduplikat. `Feature.exclude` berfungsi untuk menyaring *probe control* yaitu AFFX supaya tidak dimasukkan kedalam analisis (Gentleman, Carey, Huber, & Hahne, 2019).

3.6 Feature Selection

Feature selection berfungsi melakukan seleksi dari sekian banyak *feature* yang tersedia pada data set dan hanya akan digunakan sebagian saja (Bedy Purnama, 2019). Dan adalah proses untuk membantu membuat pengklasifikasian lebih efektif dengan cara mengurangi data-data yang tidak relevan, sehingga membuat lebih ringkas waktu untuk melakukan pengklasifikasian dan meningkatkan akurasi (Karabulut, Ozel, & Ibrikci, 2011). Dalam melakukan *feature selection* menggunakan bantuan packages `multtest` dengan perintah `mt.teststat` yang mampu menghitung statistik uji diantaranya yaitu uji t-test, wicoxon, uji F, untuk setiap baris kerangka data. Pada penelitian ini digunakan uji F karena memiliki lebih dari dua sampel. Uji F menurut Kuncoro (2009) digunakan untuk menguji signifikan tidaknya pengaruh variabel independen secara simultan terhadap variabel dependen.

3.7 Machine Learning

Machine learning merupakan bidang interdisipliner luas yang dibangun diatas konsep dari ilmu komputer, ilmu kognitif, statistik, teknik, teori optimasi dan banyak lagi disiplin ilmu matematika dan sains lainnya (Ghahramani, 2004).

Machine learning atau pembelajaran mesin banyak digunakan untuk menggantikan atau menirukan perilaku manusia untuk menyelesaikan masalah atau melakukan optimasi. *Machine learning* juga mencoba menirukan bagaimana proses manusia untuk belajar dan menggeneralisasi. *Machine learning* juga memiliki ciri khas yaitu adanya proses pelatihan dan pembelajaran, oleh karena itu data sangat dibutuhkan, data ini akan dipelajari atau biasa yang disebut dengan data *training* dan data untuk diuji atau biasa disebut sebagai data *testing* (Ahmad, 2017).

Data *training* adalah data set yang digunakan oleh komputer selama proses pembelajaran (*training*) dilakukan, sedangkan data *testing* adalah data set yang digunakan oleh komputer untuk menguji model. Proporsi untuk memilih dan menentukan data latih dan data uji ini tidak ada aturan baku ada yang menganjurkan pembagiannya sebanyak 75%:25% atau 80%:20% (Bedy Purnama, 2019).

3.8 Data Mining

Pada penerapan *machine learning* ke basis data besar disebut dengan *Data Mining* (Alpaydin, 2010). *Data mining* merupakan proses menganalisis data dalam berbagai sudut dan meringkas hasilnya menjadi informasi yang bermanfaat. Sebelum menjadi informasi yang bermanfaat maka data akan diproses dengan menggali informasi yang sebelumnya tidak diketahui dan memproses informasi tersebut dengan prangangkat lunak (Sumathi, Kannan, & Nagarajan, 2016). Proses penggalian informasi yang sebelumnya tidak diketahui dan memproses informasi dapat dilakukan dengan basis data besar kemudian digunakan untuk membuat keputusan yang penting yang disebut dengan *Data Mining*.

3.9 Klasifikasi

Klasifikasi pertama kali dilakukan oleh Carolus Linnaeus yang mana diterapkan untuk mengklasifikasikan spesies tanaman berdasarkan karakteristik disik yang dikenal sebagai bapak klasifikasi (Mardi, 2014). Klasifikasi merupakan teknik data minig yang menetapkan item ke dalam koleksi ke kategori target dengan tujuan untuk memprediksi secara akurat kategori yang tidak diketahui untuk setiap kasus dalam data. Membandingkan nilai prediktor dan nilai teget akan memberikan nilai akuraasi model. Akurasi model ini mengacu pada persentase dari contoh yang

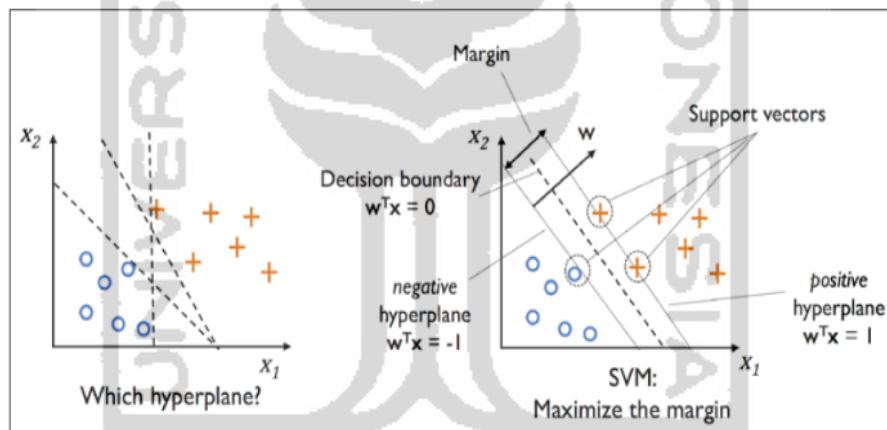
diklasifikasikan dengan benar yang dibuat oleh model jika dibandingkan dengan klasifikasi yang sebenarnya dalam data testing (Sumathi, Kannan, & Nagarajan, 2016)

3.10 Support Vector Machine (SVM)

Support vector machine (SVM) pertama kali di usulkan oleh Vapnik sekitar tahun 1992. Pada awalnya SVM digunakan untuk klasifikasi data ke dalam dua kelas dengan konsep secara matematis telah matang. Seiring berkembangnya, SVM dapat diperluas untuk klasifikasi multiclass dengan sejumlah skema.

3.10.1 Support Vector Machine (SVM) Binary

Konsep *support vector machine* (SVM) pada dasarnya yaitu menemukan *hyperplane* terbaik yang berfungsi memisahkan dua buah *class* pada *input space*.



Sumber : <https://medium.com/@samsudiney/penjelasan-sederhana-tentang-aa-itu-svm-149fec72bd02>

Gambar 3. 2 Penentuan *Hyperplane* Terbaik

Hyperplane pemisah terbaik antara kedua kelas bisa ditemukan dengan mengukur *margin hyperplane* dan mencari titik maksimalnya. *Margin* adalah jarak antara *hyperplane* dengan pattern terdekat dari masing-masing class. Pattern terdekat ini disebut sebagai *support vector*. Dimana *support vector* merupakan data point yang terdekat dengan batas (*boundary*). Sedangkan *boundary line* sendiri adalah dua buah garis yang berada diantara *hyperplane* yang membentuk *margin*. Pada gambar 3.2-b menunjukkan *hyperplane* terbaik, yaitu dimana *hyperplane* yang terbentuk optimum atau terbaik karena menghasilkan *margin* m yang lebar

(maksimum), sedangkan tanda berwarna orange dan biru pada gambar 3.2 menunjukkan *support vector*. (Tsuda, 2000)

Misal data yang tersedia dinotasikan sebagai $x_i \in R^d$ sedangkan label kelas dinyatakan sebagai $y_i \in \{-1, +1\}$ untuk $i = 1, 2, \dots, n$, dimana n merupakan jumlah data. Misalkan kedua kelas -1 dan +1 diasumsikan dapat dipisahkan secara sempurna dengan *hyperplane* berdimensi d , yang didefinisikan sebagai berikut :

$$w \cdot x + b = 0 \quad (3.1)$$

Postive hyperplane merupakan *hyperplane* dengan persamaan $w \cdot x + b = 1$ dan *negative hyperplane* merupakan *hyperplane* dengan persamaan $w \cdot x + b = -1$.

Margin terbesar dapat ditemukan dengan cara mengoptimumkan nilai jarak antara *hyperplane* dan titik terdekatnya, yaitu $\frac{1}{\|w\|}$. Pencarian *hyperplane* optimum ini

dengan cara memaksimalkan *margin* dapat dipandang sebagai sebuah masalah *Quadratic Programming* (QP) yaitu mencari titik minimal dari

$$\min_w \tau(w) = \frac{1}{2} \|w\|^2 \quad (3.2)$$

Dengan batasan

$$y_i (x_i \cdot w + b) - 1 \geq 0, \forall i \quad (3.3)$$

Permasalahan ini dapat dipecahkan menggunakan sejumlah cara, salah satu cara yang dapat digunakan yaitu *Lagrange Multiplier*.

$$L(w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^l a_i (y_i (x_i \cdot w + b) - 1), \text{ dimana } i = 1, 2, \dots, l \quad (3.4)$$

Dimana $a_i \geq 0$ merupakan nilai koefisien lagrange. Nilai optimum dari persamaan 3.4 dapat dihitung dengan meminimalkan L terhadap w dan b dan sekaligus memaksimalkan L terhadap a_i . Mengingat titik optimal gradien $L = 0$, maka persamaan 3.4 dapat dimodifikasi dengan memaksimalkan

$$\sum_{i=1}^l a_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l a_i a_j y_i y_j x_i x_j \quad (3.5)$$

Dengan batasan

$$a_i \geq 0 (i=1,2,\dots,l) \sum_{i=1}^l a_i y_i = 0 \quad (3.6)$$

Maksimasi ini menghasilkan sejumlah a_i yang bernilai positif. Data-data yang nanti berhubungan dengan a_i positif ini yang disebut sebagai *support vector*, yaitu data terluar pada kedua kelas.

Setelah mendapatkan support vector dengan a_i positif, maka selanjutnya dapat menghitung w dengan formula persamaan berikut :

$$w = \sum_{i=1}^n a_i y_i x_i \quad (3.7)$$

3.10.2 Kernel support vector machine (SVM)

Penjelasan diatas berdasarkan asumsi bahwa kedua kelas terpisah secara sempurna oleh *hyperplane*. Lalu bagaimana jika kedua kelas tersebut terpisah tidak sempurna oleh *hyperplane*. Maka untuk mengatasi masalah ini dapat menggunakan konsep *soft margin* (Kumar, 2009), yaitu dengan mengubah masalah *Quadratic Programing* (QP) dan batasannya dengan cara menambahkan sebuah variabel *slack* ξ_i yang mana bernilai lebih dari 0 sehingga persamaannya menjadi

$$\min_w \tau(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i \quad (3.8)$$

Dengan batasan

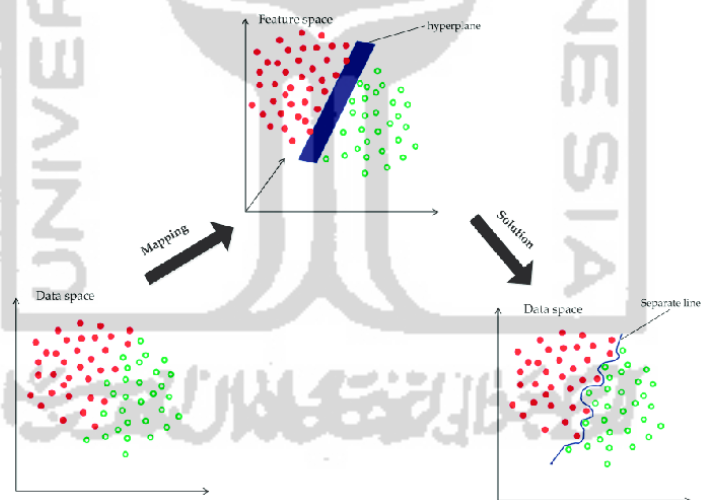
$$y_i (x_i \cdot w + b) \geq 1 - \xi_i, \forall i \quad (3.9)$$

Parameter C disini berguna untuk mengontrol optimasi antara *margin* dan kesalahan klasifikasi ξ . Jika semakin besar parameter C yang digunakan, maka semakin besar pula penalti terhadap kesalahan klasifikasi. Artinya jika menggunakan parameter C yang besar maka nantinya hanya sedikit data yang boleh melanggar batas *hyperplane*. (Suyanto, 2018)

Kelemahan *soft margin* ini yaitu tidak dapat mengeneralisasikan dengan baik dan tidak dapat menemukan pemisah dalam *hyperplane* sehingga tidak dapat memiliki akurasi yang besar. Oleh karena itu, kernel menjadi salah satu alternatif

yang dapat digunakan untuk mentransformasikan data ke dalam ruang dimensi yang lebih tinggi yang disebut ruang kernel yang berguna untuk memisahkan data secara linear.

Telah banyak teknik yang dikembangkan dengan asumsi linear pada *machine learning* dan *data mining*, sehingga sangat terbatas sekali algoritma yang dihasilkan untuk kasus-kasus *non linear*. Tak dapat dihindari bahwa permasalahan yang sering muncul dalam dunia nyata yaitu kasus yang tidak linear. Sebagai contoh dapat perhatikan **Gambar 3.3**. Pada **Gambar 3.3** data tidak dapat dipisahkan secara linear sehingga dalam penyelesaiannya diperlukan adanya bantuan metode kernel untuk menyelesaikannya. Metode kernel dimana suatu data x diinput *space* dimaping ke *feature space* yang lebih tinggi dengan menggunakan *mapping function* ϕ sebagai berikut $\phi: x \rightarrow \phi(x)$. Oleh karena itu data x pada input *space* menjadi $\phi(x)$ di *feature space*, dan nanti pada *feature space* akan membuat suatu fungsi pemisah yang linear yang dapat mewakili fungsi *nonlinear* pada *input space*.



Sumber : https://www.researchgate.net/figure/Illustration-of-nonlinear-data-separation-by-mapping-to-a-higher-dimensional-feature_fig1_318737266

Gambar 3. 3 Data *Nonlinear*

Pada **Gambar 3.3** dapat dilihat bahwa dari dua ruang dimensi *feature* ke *feature space* dua dimensi. Maka dari itu dengan memetakan ke *feature space* menjadikan tugas klasifikasi akan menjadi lebih mudah atau sederhananya fungsi kernel untuk mengkonversi himpunan data pada ruang masukkan (*input space*) ke

dalam ruang fitur (*feature space*) yang berdimensi lebih tinggi (Suyanto, 2018). Ada beberapa fungsi kernel yang digunakan yaitu antara lain kernel linear, polynomial dan sigmoid, dengan penjelasan sebagai berikut :

1. Kernel Linier

Kernel linier ini merupakan kernel yang paling sederhana, kernel ini digunakan ketika data yang diklasifikasikan telah dapat dipisahkan dengan *hyperplane*. (Honakan, 2018)

$$K(x, x_k) = x_k^T x \quad (3.10)$$

2. Kernel *Polynomial*

Sedangkan kernel *polynomial* ini merupakan sebuah metode yang populer dalam model *nonlinear*.

$$K(x, x_k) = (x_k^T x + 1)^d \quad (3.11)$$

3. Kernel Sigmoid

$$K(x, x_k) = \tanh[\kappa x_k^T x + \theta] \quad (3.12)$$

3.10.3 Support Vector Machine (SVM) Multiclass

Support vector machine (SVM) multiclass pada dasarnya dapat diimplementasikan dengan cara menggabungkan beberapa *support vector machine (SVM) biner*. Ada beberapa metode yang terdapat dalam *Support vector machine (SVM) multiclass* yaitu antara lain :

1. *One –Against-All*

Metode *One-Against-All* ini membandingkan satu kelas dengan semua kelas lainnya, sehingga metode ini juga dikenal sebagai *one-Against-the-rest*. Yang mana untuk mengklasifikasikan data kedalam k -kelas, harus membangun sejumlah k model SVM biner. Untuk k model SVM biner yang telah dibangun selanjutnya SVM biner ke- i dilatih menggunakan keseluruhan data *training* (latih), guna menjawab apakah sebuah data diklasifikasikan sebagai kelas ke- i atau tidak. (Suyanto, 2018)

Pada **Tabel 3.1** akan dijelaskan contoh permasalahan klasifikasi dengan menggunakan empat kelas.

Tabel 3. 2 Contoh SVM Pendekatan *One-Against-All*

$y_i = 1$	$y_i = -1$	Hipotesis
Kelas A	Bukan Kelas A	$f^1(x) = (w^1)x + b^1$
Kelas B	Bukan Kelas B	$f^2(x) = (w^2)x + b^2$
Kelas C	Bukan Kelas C	$f^3(x) = (w^3)x + b^3$
Kelas D	Bukan Kelas D	$f^4(x) = (w^4)x + b^4$

2. *One-Against-One*

Pada metode *one-again-one* ini dengan cara membangun sejumlah model SVM biner yang nanti akan dibandingkan satu kelas dengan kelas lainnya. Untuk mengklasifikasikan data ke k -kelas, maka harus membangun sejumlah $\frac{k(k-1)}{2}$ model SVM biner.

Pada **Tabel 3.2** akan dijelaskan contoh permasalahan klasifikasi dengan menggunakan empat kelas.

Tabel 3. 3 Contoh SVM Pendekatan *One-Against-One*

$y_i = 1$	$y_i = -1$	Hipotesis
Kelas A	Kelas B	$f^{12}(x) = (w^{12})x + b^{12}$
Kelas A	Kelas C	$f^{13}(x) = (w^{13})x + b^{13}$
Kelas A	Kelas D	$f^{14}(x) = (w^{14})x + b^{14}$
Kelas B	Kelas C	$f^{23}(x) = (w^{23})x + b^{23}$
Kelas B	Kelas D	$f^{24}(x) = (w^{24})x + b^{24}$
Kelas C	Kelas D	$f^{34}(x) = (w^{34})x + b^{34}$

3.11 *Random Forest*

Random forest atau *random decision forest* merupakan salah satu varian dari *bagging (bootstrap aggregating)* dan juga merupakan pengembangan dari metode CART (Breiman, 2001).

Random forest adalah metode klasifikasi yang lain berdasarkan pohon keputusan. Yang mana merupakan suatu kumpulan dari kelompok pohon prediktor. Sehingga setiap pohon yang terbentuk bergantung pada nilai-nilai vektor acak yang disampling secara independen dan dengan distribusi yang sama untuk semua pohon yang berada dalam hutan tersebut (Suyanto, 2018). Pada gugus data yang terdiri dari n amatan dan p merupakan peubah penjelas, *random forest* dilakukan dengan cara (Breiman & Cutler, 2003):

1. Tahapan *bootstrap* dimana dilakukannya penarikan contoh acak berukuran n dengan pemulihan pada gugus data.
2. Dengan menggunakan contoh *bootstrap*, pohon yang dibangun akan mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan dilakukan dengan memilih m peubah penjelas secara acak, dimana $m < p$ peubah penjelas tersebut. Tahapan ini dinamakan tahapan *random feature selection*.
3. Ulangi langkah 1 dan 2 sebanyak k kali, sehingga terbentuk sebuah hutan yang terdiri atas k pohon.

Metode *random forest* harus menentukan m jumlah variabel prediktor yang diambil secara acak dan k pohon yang akan dibentuk agar mendapatkan hasil yang optimal. Adapun nilai k yang disarankan untuk digunakan menurut Breiman (1984), nilai k yang disarankan untuk digunakan pada metode *bagging* yaitu $k = 50$ sudah cukup memberikan hasil yang memuaskan untuk masalah klasifikasi sedangkan menurut Sutton (2005) $k \geq 100$ cenderung menghasilkan tingkat misklasifikasi yang rendah.

Saat menggunakan metode *random forest* ukuran contoh peubah penjelas (m) sangat mempengaruhi koralasi dan kekuatan masing-masing pohon. Untuk menentukan m yaitu jumlah variabel prediktor yang diambil secara acak dengan

nilai p (banyak variabel *independent*/bebas), seperti berikut (Hastie, Tibshirani, & Friedman, 2008) :

1. Untuk klasifikasi, dalam menentukan nilai m dengan cara $\lfloor \sqrt{p} \rfloor$ dengan nilai $node$ (simpul terkecil) adalah 1.
2. Untuk regresi, dalam menentukan nilai m dengan cara $\lfloor \frac{p}{3} \rfloor$ dengan nilai $node$ (simpul terkecil) adalah 5.

Sedangkan menurut (Breiman & Cutler, 2003), terdapat tiga cara untuk mendapatkan nilai m untuk mengamati *error* :

$$m = \frac{1}{2} \lfloor \sqrt{p} \rfloor \quad (3.13)$$

$$m = \lfloor \sqrt{p} \rfloor \quad (3.14)$$

$$m = 2 \times \lfloor \sqrt{p} \rfloor \quad (3.15)$$

Dimana : p = total variabel

Menurut Breiman (2001), penggunaan m yang tepat dapat menghasilkan *random forest* dengan korelasi antar pohon cukup kecil. Namun, kekuatan setiap pohon cukup besar yang ditunjukkan dengan perolehan *error* bernilai kecil. Respon suatu amatan akan diprediksi dengan cara menggabungkan (*aggregating*) hasil prediksi k pohon. Untuk masalah klasifikasi dilakukan berdasarkan suara terbanyak (*majority vote*).

3.12 Confusion Matrix

Confusion matrix merupakan salah satu metode yang digunakan untuk mengukur kinerja metode lain seperti metode klasifikasi (*performance of a classification model or classifier*). (Bedy Purnama, 2019)

Confusion matrix mengandung informasi yang digunakan untuk membandingkan hasil klasifikasi yang telah dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya. *Confusion matrix* dinyatakan melalui matrix berikut :

		Actual Class	
		Positive (1)	Negative (0)
Predicted Class	Positive (1)	TP	FP
	Negative (0)	FN	TN

Ada beberapa singkatan atau istilah yang digunakan dalam *confusion matrix* yang tertera pada matrix diatas :

- TP (*True Positive*) : Jumlah anggota dari class 1 yang berhasil diprediksi dengan benar .
- TN (*True Negative*) : Jumlah anggota dari class 0 yang berhasil diprediksi dengan benar.
- FP (*False Positive*) : Jumlah anggota dari kelas 1 yang gagal diprediksi dengan benar .
- FN (*False Negative*) : Jumlah anggota dari kelas 0 yang gagal diprediksi dengan benar.

Dari tabel *confusion matrix*, kemudian dapat dihitung nilai akurasi, *recall*, dan nilai presisi. Akurasi adalah nilai akurasi dari suatu model, *recall* adalah seberapa banyak pelebelan yang dikenali oleh sistem, dan presisi yaitu merupakan hasil persentase dari pelebelan yang benar dari suatu label yang dikenali. berikut formulasi untuk menghitung nilai akurasi, *recall* dan nilai presisi pada pembentukan model klasifikasi.

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3.16)$$

$$Recall(sensitivity) = \frac{TP}{(TP + FN)} \quad (3.17)$$

$$Presisi = \frac{TP}{(TP + FP)} \quad (3.18)$$

3.13 ROC (*Receiver Operating Characteristic*) Curve

Receiver oprating characteristic (ROC) merupakan metode yang efektif untuk mengevaluasi kinerja diagnostik. Ukuran dasar kerja tes diagnostik ini adalah spesifisitas dan sensitivitas. Spesifisitas merupakan jumlah keputusan negatif yang

benar atau dengan kata lain yaitu jumlah kasus yang benar-benar negatif. Sedangkan sensitivitas adalah jumlah keputusan positif yang benar atau jumlah kasus yang benar-benar positif (Park, Goo, & Jo, 2004).

Dari kurva ROC ini akan mendapatkan nilai AUC (*Area Under Curve*). AUC berguna untuk mengukur kinerja keseluruhan tes diagnostik dan ditafsirkan sebagai nilai rata-rata sensitivitas untuk semua nilai yang mungkin dari spesifisitas. Secara teoritis nilai AUC memiliki rentang antara 50% - 100% yang berarti bahwa ketika nilainya 50 % adalah nilai terburuk dan 100% adalah nilai terbaik



BAB IV

METODOLOGI PENELITIAN

4.1 Jenis dan Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang berisi kumpulan data ekspresi gen yang diperoleh dari website NCBI (<https://www.ncbi.nlm.nih.gov/>) . Data yang digunakan memiliki kode akses GSE37418 dimana data merupakan data “*Novel mutations target distinct subgroups of medulloblastoma*” yang merupakan data ekspresi gen dari pasien medulloblastomas.

4.2 Tempat dan waktu Penelitian

Penelitian ini dilakukan di kampus terpadu Universitas Islam Indonesia Yogyakarta. Waktu pengambilan data dilakukan penulis pada bulan januari 2020 dan waktu penelitian yaitu pada bulan januari hingga Maret 2020.

4.3 Variabel Penelitian

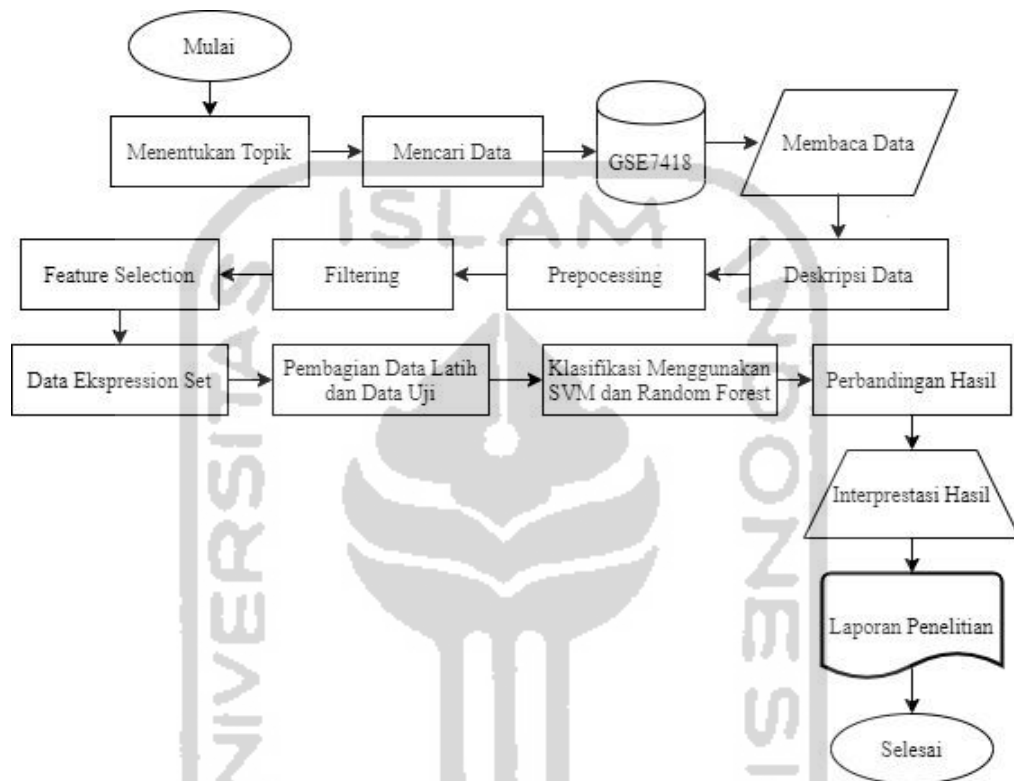
Jumlah sampel yang akan digunakan didalam penelitian ini adalah sebanyak 73 sampel medulloblastoma yang mewakili 4 kelas ekspresi (*subgroup* diantara ya WNT,SHH, *Group* 3, dan *Group* 4), dengan jumlah gen sebanyak 54675.

4.4 Metode Analisis data

Metode yang akan digunakan dalam penelitian ini yaitu klasifikasi menggunakan 2 metode yaitu metode SVM (*support vector machine*) dan Random Forest. Software yang digunakan dalam peelitian ini ialah R Studio dengan R *version* 3.6.2.

4.5 Tahapan Penelitian

Tahapan analisis yang dilakukan pada penelitian ini dapat digambarkan dengan *flowchart* pada Gambar 4.1 :



Gambar 4.1 *Flowchart* Penelitian

1. Tahapan pertama dalam penelitian ini adalah menentukan topik yang akan di angkat, topik yang dipilih adalah tentang bioinformatika khususnya akan mengangkat topik medulloblastoma.
2. Tahapan kedua peneliti mencari data medulloblastoma di website NCBI (<https://www.ncbi.nlm.nih.gov/>) dan menggunakan dataset GSE37418.
3. Selanjutnya peneliti membaca data dengan fungsi *list.celfiles* karena data yang akan digunakan merupakan *file.CEL*.
4. Kemudian setelah data dapat terbaca maka didapatkan informasi mengenai apa saja yang terdapat didalam data, sehingga peneliti dapat mendeskripsikan data sebagai gambaran umum dari data yang dimiliki.
5. Sebelum masuk ke tahap analisis ada beberapa tahapan yang perlu dilakukan oleh peneliti. Yang pertama adalah tahapan *preprocessing*. Tahapan *preprocessing*

ini adalah proses berahli data mentah yang berupa *affybatch* ke bentuk *expression set*.

6. Lanjut ketahap selanjutnya yaitu tahap *filtering*, tahapan ini bertujuan untuk menghapus fitur dari *expression set*. Termasuk penghapusan *probe* dengan nama AFFX.
7. Setelah tahap *filtering* lanjut ketahap *feature selection*, dimana *feature selection* ini merupakan pendekatan *filter* menggunakan anova karena data yang akan dipakai dalam penelitian memiliki lebih dari dua kelas.
8. Ketika semua tahapan selesai dan data telah berbentuk *expression set*. Maka, data siap untuk dianalisis. Akan tetapi sebelumnya dilakukan pembagian data *training* dan *testing* terlebih dahulu. Data *training* yang akan digunakan pada penelitian ini adalah 70 % dari total data.
9. Setelah itu analisis berupa klasifikasi data akan dilakukan yang pertama menggunakan metode *Support vector Machine* (SVM), dengan karnel yang digunakan adalah *karnel Linear*, selanjutnya melakukan optimasi untuk nilai parameter dengan proses *tuning* untuk menentukan nilai optimum, dan terakhir melakukan klasifikasi *Support vector Machine* (SVM) terhadap data *training* menggunakan nilai *cost* yang optimum agar menghasilkan nilai akurasi yang terbaik.
10. Dilanjutkan ke analisis selanjutnya yaitu *Random Forest*, dalam klasifikasi *random forest* ini juga menggunakan data *training* dan menentukan banyak pohon yang terbentuk (k atau *ntree*) dan menentukan banyak variabel prediktor (m) yang besarnya ditentukan oleh peneliti berdasarkan penelitian sebelumnya, tahapan selanjutnya adalah melakukan optimasi untuk nilai parameter dengan proses *tuning* untuk menentukan nilai parameter optimum. Langkah selanjutnya di *random forest* ini adalah melakukan klasifikasi terhadap data *training* menggunakan nilai k dan m yang optimum agar menghasilkan model yang terbaik.
11. Tahapan selanjutnya yaitu membandingkan hasil akurasi dari masing-masing metode yang digunakan.

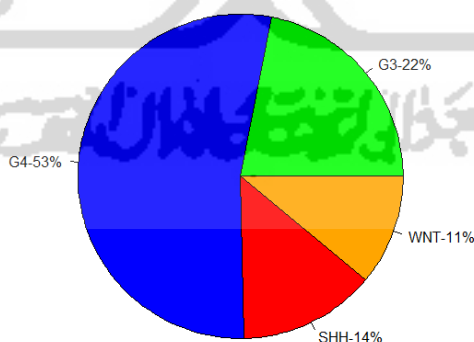
BAB V

HASIL DAN PEMBAHASAN

Medulloblastoma merupakan tumor otak ganas paling umum pada anak-anak dan merupakan tumor sistem saraf pusat primer (SSP). Ini berarti tumor bermula di otak atau sumsum tulang belakang, tumor yang bermula di otak terjadi di bagian otak kecil, yang dimana bagian dari otak yang terletak dipangkal tengkorak tepat diatas batang otak, sedangkan tumor yang bermula pada sumsum tulang belakang akan bisa menyebar ke otak dan ada penyebab yang memang terjadi mutasi gen keluarga tetapi secara garis besar penyebab dari medulloblastoma ini masih sangat minim diketahui.

5.1 Deskripsi Data

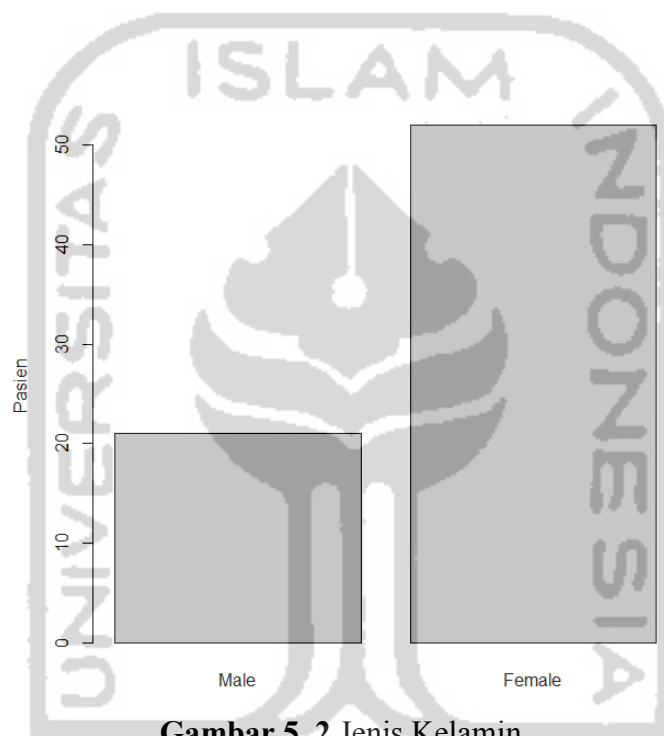
Gambaran secara umum dalam penelitian ini dari penderita medulloblastoma, akan disajikan dengan menggunakan analisis deskriptif. Dimana data GSE37418 memiliki 73 pasien dengan 4 subgrup. Oleh karena itu data yang digunakan adalah data pasien yang tergolong ke dalam 4 subgrup yang dimana subgrup tersebut yaitu SHH, WNT, Group 3, dan Group 4.



Gambar 5.1 Pie Chart Subgrup Medulloblastoma

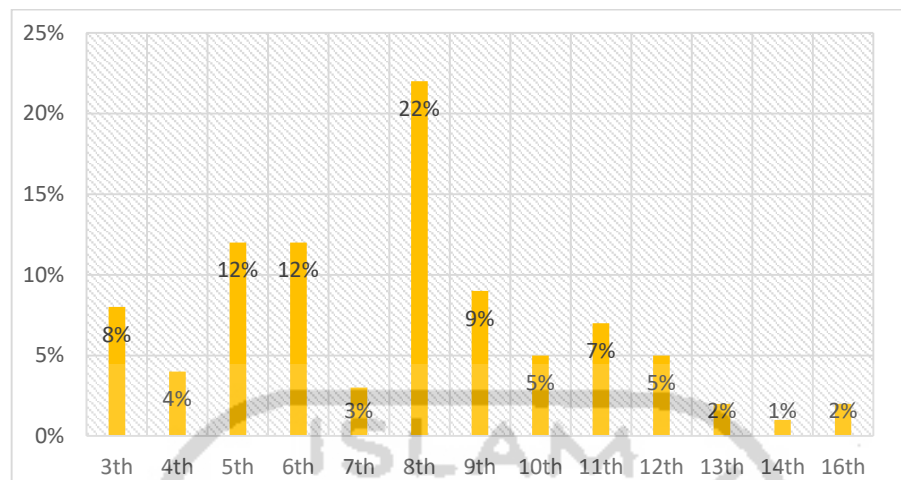
Pada Gambar 5.1 dapat dilihat bahwa Subgrup dari medulloblastoma ini sendiri terbagi menjadi 4 Subgrup, dimana Subgrup pertama yaitu group 4 (G4) ini memang banyak dan sering terjadi walaupun penyebab pastinya tidak diketahui yaitu sebanyak 53% hampir dari setengah pasien tergolong kedalam group 4 (G4)

ini. Untuk Subgrup yang kedua yaitu group 3(G3) sering terjadi pada pria dari pada wanita dan pusat tumor dibagian *fourth ventricle* dan ketika didiagnosis telah menyebar ke sumsum tulang belakang dalam group 3 (G3) ini terjadi sebanyak 22%. Untuk subgrup ketiga yaitu SHH sebanyak 14%, dimana subgrup ini seringnya terjadi karena disebabkan oleh mutasi gen . Dan untuk subgrup yang terakhir yaitu WNT sebanyak 11%. Dan tentunya perlakuan penanganan penyembuhan akan berbeda untuk setiap golongan.



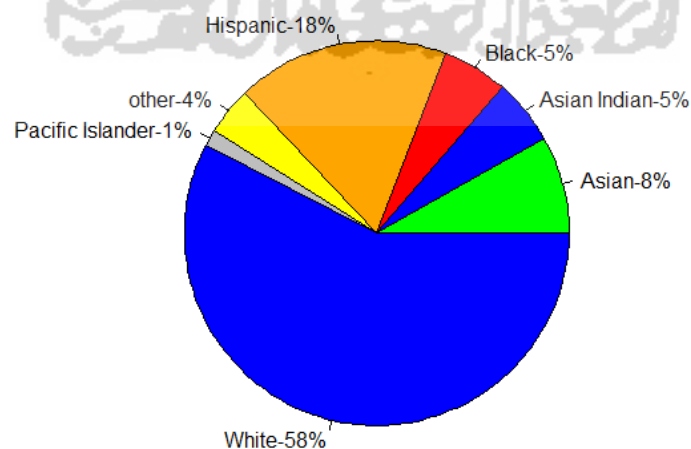
Gambar 5. 2 Jenis Kelamin

Pasien penderita medulloblastoma lebih banyak ditemukan pada pasien yang berjenis kelamin perempuan dibandingkan dengan pasien berjenis kelamin laki-laki, seperti yang ditunjukkan pada **Gambar 5.2**. Untuk dalam hal umur pasien yang menderita medulloblastoma seringnya terjadi pada anak-anak. Untuk melihat lebih jelas dapat dilihat dalam **Gambar 5.3** dibawah ini.



Gambar 5. 3 Umur

Dapat dilihat bahwa sampel pasien yang menderita medulloblastoma di rentang umur 3 tahun hingga 16 tahun. Untuk umur 3 tahun sebanyak 8 %, untuk umur 4 tahun 4%, untuk umur 5 tahun sebanyak 12 %, umur 6 tahun sebanyak 12%, umur 7 sebanyak 3 %, umur 8 tahun sebanyak 22%, untuk umur 9 tahun sebanyak 9%, umur 10 tahun sebanyak 5%, untuk umur 11 tahun sebanyak 7%, untuk umur 12 tahun sebanyak 5%, untuk umur 13 tahun sebanyak 2%, untuk umur 14 tahun sebanyak 1%, dan untuk umur 16 tahun sebanyak 2%. Sehingga jika dilihat dari persentase dari setiap umur pasien penderita medulloblastoma dari sampel yang ada kebanyakan di umur 8 tahun, dan cenderung banyak terjadi pada anak dibawah umur 10 tahun.



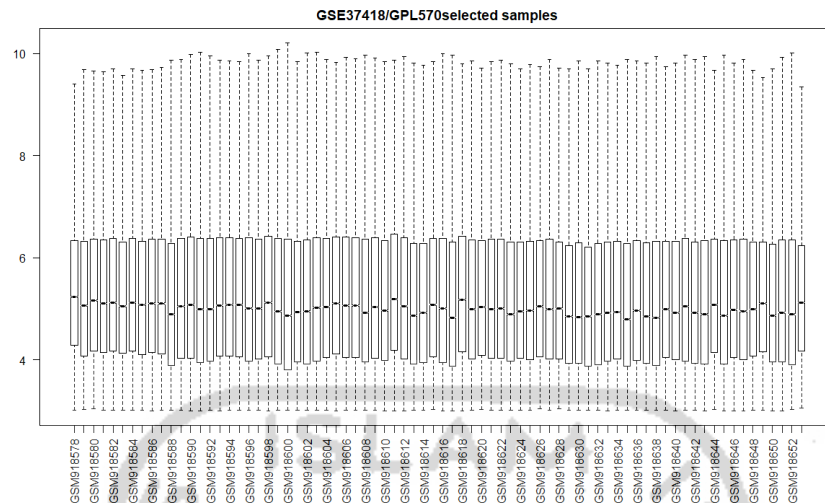
Gambar 5. 4 Pie Chart Etnis

Gambar 5.4 diatas menunjukkan jumlah masing-masing etnis dari pasien. Etnis kulit putih paling banyak yaitu sebesar 58%, hal tersebut dikarenakan sampel yang diambil dilakukan di *America* yang kebanyakan penduduk merupakan kulit putih. 18% penderita berasal dari etnis Hispanik yang merupakan etnis terbesar di *America*, yang dimana etnis ini memiliki perkembangan yang cukup pesat di beberapa bagian di Amerika Serikat sehingga dalam kasus ini etnis hispanik pasien terbanyak yang menderita medulloblastoma. Dapat dilihat pada **Gambar 5.4** dimana etnis yang lainnya seperti Pacific Islander sebanyak 1%, etnis kulit hitam sebanyak 5%, etnis Asia India sebanyak 5%, etnis Asia sebanyak 8% dan etnis lainnya sebanyak 4%.

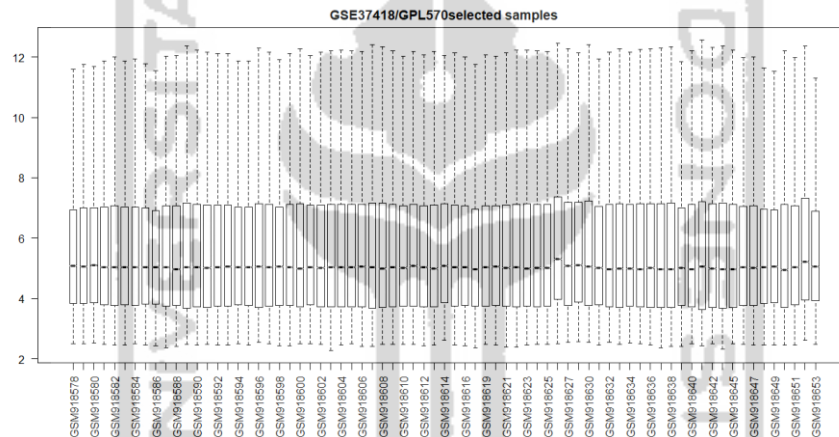
5.2 Preprocessing

Langkah *preprocessing* ini dapat dilakukan dengan menggunakan bantuan *packages AffyPLM* dari *Bioconductor* dengan perintah *Threestep*. Dalam penelitian ini peneliti menggunakan *default* dari R Studio untuk pemilihan metode dari tiap langkah *preprocessing*, karena pemilihan metode terbaik dalam tahap *preprocessing*, untuk *background correction* menggunakan *RMA2*, *Normalization* menggunakan *quatile*, dan untuk *summarization* menggunakan *median polish*.

Pada tahap *preprocessing* ini dapat dilihat dalam bentuk *boxplot* seperti pada **Gambar 5.5** dan **Gambar 5.6**. Dapat dilihat bahwa terdapat perbedaan nilai pada data sebelum dilakukannya *preprocessing* dan setelah dilakukannya proses *preprocessing*. Perbedaan yang sangat kontras terlihat perbedaannya adalah nilai median pada data sebelum dan setelah *preprocessing*. Setelah melalui tiga tahapan *preprocessing* mempunyai nilai yang sudah diperbaiki karena faktor non biologis, *background noise* dan variansi non biologis yang tidak diinginkan telah dihilangkan, penyesuaian *cross hibridization*, dan penggabungan intensitas *probe* menjadi *probe set* yang terdapat dalam *chip microarrays* menghasilkan nilai ekspresi gen.



Gambar 5.5 Boxplot Sebelum *Preprocessing*



Gambar 5.6 Boxplot Sesudah *Preprocessing*

5.3 Filtering

Tahap *filtering* merupakan proses pemilihan subset dari *probe* yang tersedia untuk dikecualikan atau dimasukkan ke dalam analisis dalam artian lain juga *filtering* berguna untuk menghapus sebagian *probe* yang tidak dibutuhkan, yang dapat berguna untuk analisis dan membantu meningkatkan kualitas model, juga dapat membuat proses pemodelan lebih efisien. Dalam proses *filtering* terdapat dua tahapan yaitu proses *filtering* data dan *feature selection*. Untuk tahapan pertama menggunakan bantuan *packages* khusus dari *bioinformatics* yang digunakan untuk memfilter data yang berbentuk *expression set* yaitu dengan menggunakan *library* *genefilter* dengan fungsi yang akan digunakan yaitu *nsFilter* (*Non specific*

filtering) dimana fungsi ini menyediakan suatu opsi *one-stop shop* (serba ada) untuk berbagai pilihan penghapusan fitur dari *expression set*. Tahapan ini berguna untuk mengeluarkan variabel yang memiliki nilai IQR tertinggi. Yang jelaskan kembali dalam **Tabel 5.1**.

Dalam perintah `nsFilter` terdapat beberapa perintah yang dilakukan yaitu seperti `require.entrez` untuk memfilter tanpa anotasi *Entrez Gene ID*, `remove.dupEntrez` untuk menghapus ID yang sama atau terduplikat, `var.cutoff` untuk memfilter gen dengan nilai tertentu, dan `feature.exclude` untuk menyaring *probe control* untuk tidak dimasukkan kedalam analisis (Gentlemen.R, Carey.V, Huber.W, & Hahne.F, 2020).

Tabel 5. 1 Filtering nsFilter

nsFilter	
Sampel	Gen
73	8070

Hasil dari data yang didapat dari tahapan ini berdimensi 8070 x 73 yang masih berbentuk *expressien set* dan kemudian dirubah menjadi data *frame* atau data *matrix* dengan menggunakan perintah `exprs`. Karena data yang dihasilkan masih memiliki dimensi yang besar, maka data akan dikurangi lagi pada tahap *feature selection*.

Selanjutnya masuk ke tahap *feature selection*. Tahapan ini menggunakan fungsi `multtest`, dimana `multtest` dengan bantuan uji F digunakan untuk membandingkan 4 subgrup yang diambil dari `characteristics_ch1` yang mempunyai variansi sama. Pada klasifikasi di penelitian ini menggunakan fungsi `multtest` yang digunakan untuk memfilter gen, sehingga dimensi akhir yang di dapatkan menjadi 73 x 4696. dijelaskan kembali pada **Tabel 5.2**.

Tabel 5.2 Filtering Multtest

Multtest	
Sampel	Gen
73	4696

5.4 Analisis *support vector machine* (SVM)

Dalam klasifikasi umumnya menggunakan data latih (*train*) dan data uji (*test*). Dalam penelitian ini data dibagi dengan porsi 70% untuk data latih (*train*) dan 30% untuk data uji (*test*). Untuk data latih (*train*) dan data uji (*test*) menggunakan data yang diambil secara *random* dengan mengutamakan semua *subgroup* masuk ke dalam data latih (*train*) dan data uji (*test*). Seperti telah dijelaskan pada **Gambar 5.1** dalam data medulloblastoma ini terdapat 4 *subgroup* yaitu SHH, WNT, *Group* 4 dan *Group* 3, dimana terdiri dari 73 sampel yang digunakan sehingga mendapatkan hasil seperti dijelaskan pada **Tabel 5.3**.

Tabel 5.3 Pembagian Data Latih (*Train*) dan Data Uji (*Test*)

	Data Latih (<i>Train</i>)	Data Uji (<i>Test</i>)
Ratio	70%	30%
Jumlah	51	22

5.4.1 *Confussion Matrix support vector machine* (SVM)

Analisis *support vector machine* (SVM) dalam penelitian ini menggunakan karnel *linear* . Proses klasifikasi pada analisis *support vector machine* (SVM) dilakukan dengan cara mempelajari pola data yang dengan menggunakan data latih (*train*) yang didalamnya terdapat empat *subgroup* yaitu SHH, WNT, *Group* 4 dan *Group* 3. Pada penelitian kali ini nilai *cost* yang terbaik adalah 1. Hasil dari *machine learning* dapat dilihat pada **Tabel 5.4** dan **Tabel 5.5**.

Tabel 5.4 *Confussion Matrix* (SVM) Data Latih (*Train*)

Prediksi	Data Latih (<i>Train</i>)			
	Group 3	Group 4	WNT	SHH
Group 3	11	0	0	0
Group 4	0	27	0	0
WNT	0	0	7	0

SHH	0	0	0	6
Akurasi : 100%				

Berdasarkan model pada data latih (*train*) prediksi klasifikasi yang dihasilkan sesuai dengan subgroup nya masing-masing, setiap *subgroup* tidak ada yang masuk ke klasifikasi yang salah. Hasil ini didapatkan karena data yang dipakai untuk membangun model *support vector machine* (SVM) sama dengan data yang diujikan. Berdasarkan hasil tersebut data uji digunakan untuk mengetahui model yang dibangun mampu untuk memprediksi data yang baru. Data uji diasumsikan sebagai data baru dan data uji tidak digunakan dalam membangun model *support vector machine* (SVM). Hasil prediksi yang didapat menggunakan data uji (*Test*) disajikan dalam **Tabel 5.5**.

Tabel 5.5 *Confussion Matrix* (SVM) Data Uji (*Test*) Kernel Linear

Prediksi	Data Uji (<i>Test</i>)			
	Kelas 0 (Group 3)	Kelas 1 (Group 4)	Kelas 2 (SHH)	Kelas 3 (WNT)
Kelas 0 (Group 3)	4	1	0	0
Kelas 1 (Group 4)	1	11	0	0
Kelas 2 (SHH)	0	0	3	0
Kelas 3 (WNT)	0	0	0	2
Akurasi : 90.92%				

Pada **Tabel 5.5** dapat dilihat bahwa pada proses klasifikasi dengan menggunakan data *testing* yang mana menggunakan sebanyak 22 sampel, menunjukkan bahwa terdapat 2 kesalahan yaitu pada kelas 0 (Group 3) dan kelas 1 (Group 4). Kesalahan yang terdapat pada kelas 0 menunjukkan bahwa dari 5 sampel terdapat 1 sampel yang dikenal sebagai kelas 1. Sedangkan, pada kelas 1 menunjukkan bahwa dari 12 sampel terdapat 1 sampel yang dikenali sebagai kelas 0. Untuk kelas 2 dan 3 sudah menunjukkan bahwa klasifikasi sudah benar.

Tabel 5.6 *Confussion Matrix (SVM) Data Uji (Test) Kernel Polynomial*

Prediksi	Data Uji (<i>Test</i>)			
	Kelas 0 (Group 3)	Kelas 1 (Group 4)	Kelas 2 (SHH)	Kelas 3 (WNT)
Kelas 0 (Group 3)	1	0	0	0
Kelas 1 (Group 4)	4	12	1	0
Kelas 2 (SHH)	0	0	2	0
Kelas 3 (WNT)	0	0	0	2
Akurasi : 77.27 %				

Pada **Tabel 5.6** dapat dilihat bahwa pada proses klasifikasi dengan menggunakan data *testing* yang mana menggunakan sebanyak 22 sampel, menunjukkan bahwa terdapat 2 kesalahan yaitu pada kelas 0 (Group 3) dan kelas 2 (SHH). Kesalahan yang terjadi pada kelas 0 menunjukkan bahwa dari 5 sampel terdapat 4 sampel yang dikenali sebagai kelas 1 sedangkan pada kelas 2 dari 3 sampel terdapat 1 sampel yang dikenali sebagai kelas 1. Untuk kelas 1 dan kelas 3 sudah menunjukkan bahwa klasifikasi sudah benar.

Tabel 5.7 *Confussion Matrix (SVM) data uji (Test) Kernel Sigmoid*

Prediksi	Data Uji (<i>Test</i>)			
	Kelas 0 (Group 3)	Kelas 1 (Group 4)	Kelas 2 (SHH)	Kelas 3 (WNT)
Kelas 0 (Group 3)	5	1	0	1
Kelas 1 (Group 4)	0	11	0	0
Kelas 2 (SHH)	0	0	3	0
Kelas 3 (WNT)	0	0	0	1

Akurasi : 86.36%

Pada **Tabel 5.7** dapat dilihat bahwa pada proses klasifikasi dengan menggunakan data testing yang mana menggunakan 22 sampel, menunjukkan bahwa terdapat 2 kesalahan yaitu pada kelas 1 (Group 4) dan kelas 0 (Group 3). Kesalahan yang terdapat pada kelas 1 yaitu menunjukkan bahwa dari 12 sampel terdapat 1 sampel yang dikenali sebagai kelas 0. Sedangkan, pada kelas 3 dari 2 sampel terdapat 1 sampel yang dikenali sebagai kelas 0. Untuk kelas 0 dan kelas 2 sudah menunjukkan bahwa klasifikasi sudah benar.

Berdasarkan *Confussion Matrix* hasil model yang didapatkan menggunakan 3 kernel yang telah dijelaskan seperti **Tabel 5.5, 5.6** dan **5.7** di peroleh nilai akurasi yang tertinggi yaitu menggunakan kernel linear dimana menghasilkan nilai akurasi sebesar 0.909191 atau 90.92%, dimana akurasi ini bisa dikatakan baik, akan tetapi untuk menguji validasi lebih lanjut digunakan ROC.

5.4.2 ROC Curve SVM

Nilai akurasi yang didapatkan dari model sebelumnya menggunakan data uji yaitu sebesar 90.91% dan hasil ini akan ditinjau kembali dengan melihat *area under ROC curve* (AUC).

Nilai *multi-class* AUC yang di dapatkan dalam penelitian ini yaitu sebesar 0.9764, nilai ini dapat dikatakan tinggi karena pada dasarnya nilai maksimal AUC sendiri yaitu 1. Jadi dalam penelitian ini didapatkan nilai akurasi yaitu 90.91% dan nilai *multi-class* AUC sebesar 0.9764 atau 97.64% maka dapat dikatakan jika model yang diperoleh baik karena nilai AUC menunjukkan akurasi prediksi dari model.

5.4.3 Model Klasifikasi

Hasil klasifikasi pasien medulloblastoma dengan kode akses GSE37418 memberikan bentuk dengan nilai AUC sebesar 0.9764 atau 97.64%. Ditahap sebelumnya yang sudah dilakukan oleh peneliti yaitu tahapan *filtering* di dapatkan sebanyak 4696 *probe* yang menjadi variabel pada penelitian dan ditambahkan dengan 4 variabel karakteristi tambahan dari sampel yaitu etnis, jenis kelamin, umur

dan subgroup pasien. Berdasarkan model yang didapatkan, berikut bobot 10 variabel dari 4696 variabel yang ditunjukkan ($\vec{w}$).

Tabel 5.8 Bobot *Probe ID*

<i>Probe ID</i>	w
X225784_s_at	0.003659362
X230237_at	0.00315555
X212182_at	0.003121758
X233720_at	0.002894109
X225589_at	0.002893663
X232059_at	0.002798406
X226092_at	0.00279201
X227461_at	0.002790291
X218951_s_at	0.002779925
X220559_at	0.002745847

Tabel 5.8 menunjukkan bobot dari masing-masing *probe*, dan bobot tertinggi adalah *probe* dengan ID X214438_at. *Probe ID* ini dapat menunjukkan gen name dan ontology dari gen tersebut. Pada penelitian ini khusus nya pada **Tabel 5.5** diambil 10 *probe* tersusun dari nilai bobot yang paling tinggi hingga nilai bobot yang paling rendah.

Dari *Probe ID* ini dapat diketahui seperti tabel **Tabel 5.9** menunjukkan bahwa *probe* berasal dari gen protein yang mengandung *domain zinc finger*. Merujuk pada situs NCBI bahwa setiap *probe* memiliki pengaruh yang berbeda-beda. *Probe* X225784_s_at Kode gen ini merupakan anggota dari keluarga protein yang mengandung *domain zinc finger*. Yang mana memiliki *domain zinc finger C-terminal* yang ditandai dengan empat residusistein dan dua residu histidin, dan juga termasuk daerah *coiled-coil*. Protein ini telah terdeteksi sebagai autoantigen pada

pasien karsinoma hepatoseluler. Gen ini telah teridentifikasi sebagai kandidat potensial untuk disabilitas kognitif terkait-X (NCBI, 2020).

Tabel 5.9 Keterangan *Probe ID* dari Hasil Analisis SVM Kernel Linear

No	<i>Probe ID</i>	Simbol	Keterangan
1.	X225784_s_at	ZC4H2	Kode gen ini merupakan anggota dari keluarga protein yang mengandung domain zinc finger. Yang mana memiliki domain zinc finger C-terminal yang ditandai dengan empat residusistein dan dua residu histidin, dan juga termasuk daerah coiled-coil. Protein ini telah terdeteksi sebagai autoantigen pada pasien karsinoma hepatoseluler. Gen ini telah teridentifikasi sebagai kandidat potensial untuk disabilitas kognitif terkait-X.
2.	X230237_at	ADCYAP1	Gen ini mengkodekan proprotein yang disekresikan dan selanjutnya akan di proses menjadi beberapa peptida matang. Peptida-peptida ini yang nantinya akan menstimulasi adenilat siklase dan meningkatkan tingkat siklik adenosin monofosfat (cAMP), dan menghasilkan aktiivasi gen target secara transkripsi. Produk yang dihasilkan dari gen ini merupakan mediator kunci dari respon stres neuroendokrin.
3.	X212182_at	NUDT4	Protein yang terdaat dalam gen ini mengatur pergantian polifosfat difosfoinositol. Omset polifosfat difosfoinositol berenergi tinggi ini mewakili aktivitaas perpindahan molekuler dengan konsekuensi pengaturan yang penting. Peralihan molekuler oleh polifosfat difosfoinositol dapat berkontribusi untuk jalannya intraseluler.
4.	X233720_at	SORBS2	Arg dan c-Abl merupakan representasi anggota mamalia dari keluarga Abelson dari protein-tirosin kinase non-reseptor. Mereka berinteraksi dengan perotein pengikat arg/abl melalui domain SH3 yang ada diujung karboksi dari kelompok protein terakhir. Gen ini sorbin dan SH3 merupakan domain yang mengandung 2 protein. Ia memiliki tiga domain SH3 terminal-C dan domain homologi sorbin N-terminal (SoHo) yang berinteraksi dengan protein rakit lipid. Lokalisasi subseluler dari protein ini dalam sel-sel epitel dan otot jantung menunjukkan bahwa ia berfungsi sebagai protein adaptor untuk merakit kompleks-kompleks pensinyalan dalam serat-serat stres.
5.	X225589_at	SH3RF1	Gen ini mengkodekan protein yang mengandung N-terminus RING-finger, 4 domain SH3, dan wilayah yang terlibat dalam pengikatan Rho GTPase Rac. Melalui RING-finger, protein yag dikodekan telah terbukti

No	Probe ID	Simbol	Keterangan
			berfungsi sebagai ligase protein-ubiquitin yang terlibat dalam penyortiran protein di jaringan trans-Golgi. Dan juga dapat bertindak sebagai perancah untuk pensinyalan c-Jun N-terminal kinase, memfasilitasi pembentukan modul pensinyalan fungsional.
6.	X232059_at	DSCAML1	Protein yang dikodekan oleh gen ini merupakan anggota superfamili Ig dari molekul adhesi yang terlibat dalam diferensiasi neuron. Protein terikat membran yang dikodekan melokalisasi ke permukaan sel, dimana ia membentuk agregat yang mengusir proses neuron dari jenis sel yang sama.
7.	X226092_at	MPP5	Gen ini mengkodekan anggota subfamili p55 seperti dari superfamili gen guanylate kinase (MAGUK) yang terkait membran. Protein yang dikodekan berpartisipasi dalam polarisasi sel-sel yang berdiferensiasi telah mengatur sel Schwann myelinating (PMID:20237282), dan juga merupakan salah satu komponen kompleks remah retina. Tikus yang mengekspresikan kadar protein ortologis lebih rendah mengalami degenerasi retina dan gangguan penglihatan (PMID:22114289). Beberapa varian transkrip yang menyandingkan isoform ditemukan untuk gen ini.
8.	X227461_at	STON2	Gen ini mengkodekan protein yang mana merupakan protein membran yang terlibat dalam mengatur kompleks endositosis. Produk protein digambarkan sebagai salah satu protein pemilahan yang terkait clatrin, molekul adaptor yang memastikan protein spesifik diinternalisasi. Dan protein ini juga telah terbukti berpartisipasi dalam daur ulang vesikel sinaptik melalui interaksi dengan synaptotagmin 1 yang diperlukan untuk transmisi neurot.
9.	X218951_s_at	PLCXD1	Gen ini adalah gen protein-coding dalam pseudoautosomal (PAR) pada kromosom X dan Y.
10.	X220559_at	EN1	Gen yang mengandung homeobox diduga memiliki peran dalam mengendalikan perkembangan. Dalam Drosophila, gen 'terukir' (en) memainkan peran penting selama segmentasi, dimana diperlukan untuk pembentukan kompartemen posterior. Mutasi yang berbeda pada homolog tikus, En1 dan En2, menghasilkan cacat perkembangan yang berbeda yang seringkali mematikan. Homolog yang berukir manusia 1 dan 2 mengkodekan protein yang mengandung homeodomain yang telah terlibat dalam kontrol pembentukan pola selama pengembangan sistem saraf pusat.

5.5 Analisis *Random Forest*

Tahapan pertama Analisis *Random Forest* adalah menentukan berapa banyak pohon yang akan terbentuk (*ntree*) dan menentukan seberapa banyak *random sample* yang akan di ambil untuk setiap percobaan (*mtry*).

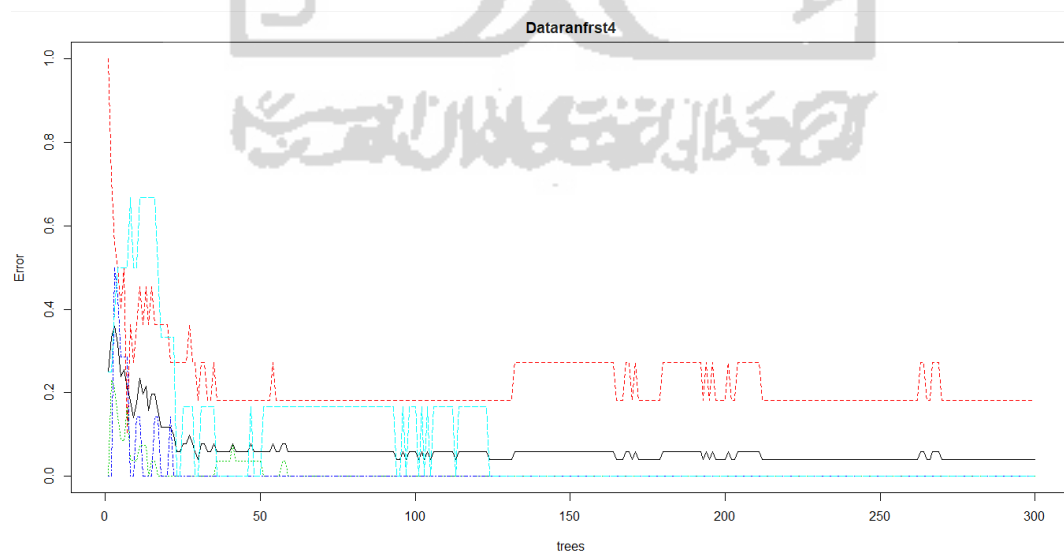
5.5.1 Hasil *Mtry* dan *Ntree*

Berikut adalah nilai *error* yang dihasilkan dari masing-masing nilai *Mtry*.

Tabel 5.10 Nilai *Error* dari Masing-Masing *Mtry*

<i>Mtry</i>	<i>Error</i>
34	0.058
68	0.039
136	0.098

Dari **Tabel 5.10** dapat diketahui bahwa nilai *mtry* yang optimum adalah sebesar 68, hal ini dikarenakan nilai *error* yang dihasilkan paling kecil dibandingkan nilai *mtry* yang lainnya. Setelah mendapatkan nilai *mtry* maka lanjut ketahap selanjutnya yaitu mencari nilai *ntree* yang optimum. Dalam penelitian ini *ntree* optimum adalah 300 dengan nilai *error* terkecil sebesar 3.92% untuk lebih lanjut berikut grafiknya.



Gambar 5. 7 *Ntree* Optimum

Dapat dilihat pada **Gambar 5.7** bahwa dari *ntree* 270 an keatas grafik yang dihasilkan dapat dikatakan stabil maka dari itu peneliti dalam penelitian kali ini memutuskan untuk menggunakan *ntree* sebesar 300.

5.5.2 Confussion Matrix Random Forest

Proses klasifikasi pada analisis *Random Forest* dilakukan dengan melakukan penggabungan pohon dengan menggunakan data train pada sampel yang digunakan. Penggunaan pohon menjadi pengaruh dalam klasifikasi ini karena semakin banyak pohon yang digunakan maka akan mempengaruhi hasil akurasi yang didapat menjadi lebih Pemenang dari *tree* yang terbentuk ditentukan dengan *vote* terbanyak yang didalamnya terdapat empat subgrup yaitu SHH,WNT,Group 4 dan Group 3. Hasil yang dapat dilihat pada **Tabel 5.11**

Tabel 5.11 Confussion Matrix Random Forest Data Uji (*Test*)

Prediksi	Data Uji (<i>Test</i>)			
	Group 3	Group 4	WNT	SHH
Group 3	4	1	0	0
Group 4	1	11	0	0
WNT	0	0	3	0
SHH	0	0	0	2
Akurasi : 90.92%				

Berdasarkan **Tabel 5.11** hasil akurasi yang didiperoleh yaitu 0.9092 atau 90.92%. Hasil ini menunjukkan bahwa metode random forest dalam melakukan prediksi untuk data baru dengan hasil prediksi yang tepat adalah sebesar 90.92% atau bisa dikatakan dapat memprediksi dengan bagus.

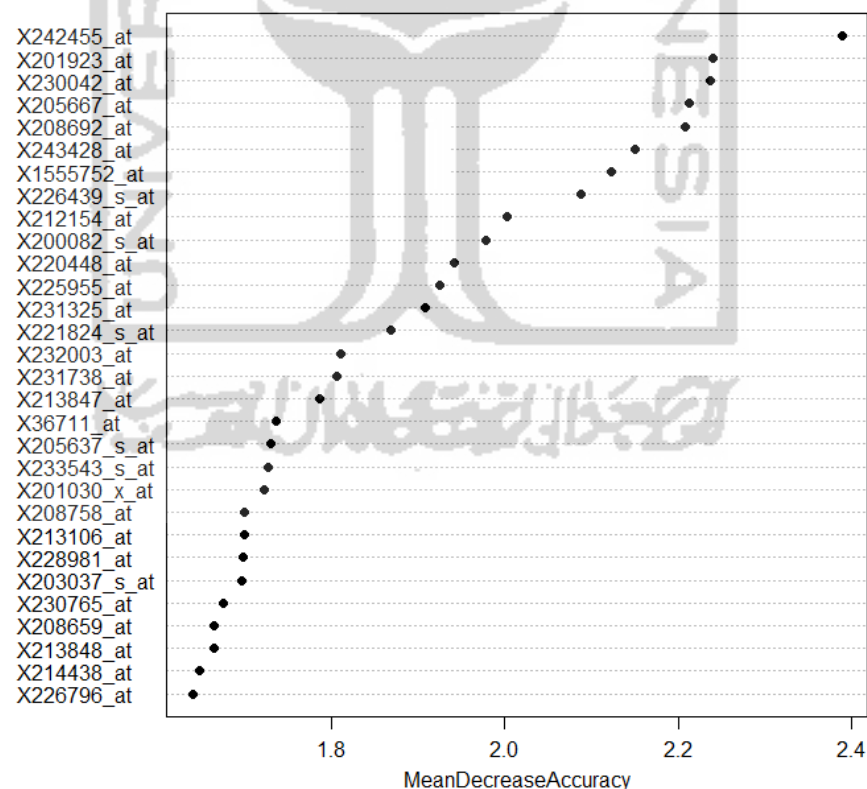
5.5.3 ROC Curve Random Forest

Nilai akurasi yang didapatkan dari model sebelumnya menggunakan data uji yaitu sebesar 90.91% dan hasil ini akan ditinjau kembali dengan melihat *area under ROC curve* (AUC).

Nilai *multi-class* AUC yang di dapatkan dalam penelitian ini yaitu sebesar 0.9764, nilai ini dapat dikatakan tinggi karena pada dasarnya nilai maksimal AUC sendiri yaitu 1. Jadi dalam penelitian ini didapatkan nilai akurasi yaitu 90.92% dan nilai *multi-class* AUC sebesar 0.9764 atau 97.64% maka dapat dikatakan jika model yang diperoleh baik karena nilai AUC menunjukkan akurasi prediksi dari model.

5.5.4 Importance Variable

Pada analisis *random forest* untuk menentukan *importance variable* dilakukan menggunakan data asli.



Gambar 5. 8 Importance Variable

Gambar 5.8 menjelaskan tentang *importance variable* dari klasifikasi yang telah dilakukan menggunakan metode *random forest*. *Importance variable* ini menunjukkan hubungan variabel itu sendiri dalam memengaruhi hasil analisis prediksi. Semakin besar hasil yang diperoleh maka semakin besar peran variabel tersebut dalam hasil analisis. Adapun yang paling besar peranan variabel nya yaitu X242455_at, X201923_at, X230042_at, X205667_at, X208692_at, X243428_at, X226439_s_at, X212154_at, X200082_s_at, X220448_at.

Dari *Probe ID* ini dapat diketahui seperti tabel **Gambar 5.8** menunjukkan bahwa probe berasal dari gen protein yang mengandung *domain zinc finger*. Merujuk pada situs NCBI bahwa setiap *probe* memiliki pengaruh yang berbeda-beda. *Probe* X242455_at merupakan anggota gen POU kelas 3 dalam faktor transkripsi saraf. Protein yang terdapat didalamnya terlibat dalam diferensiasi neuron dan meningkatkan aktivitas gen yang diatur oleh hormon pelepas kortikotropin. Ekspresi yang berlebih dari protein ini dikaitkan dengan peningkatan proliferasi sel melanoma.

Tabel 5.12 Keterangan *Probe ID* dari Hasil Analisis *Random Forest*

No.	<i>Probe ID</i>	Simbol	Keterangan
1.	X242455_at	POU3F2	merupakan anggota gen POU kelas 3 dalam faktor transkripsi saraf. Protein yang terdapat didalamnya terlibat dalam diferensiasi neuron dan meningkatkan aktivitas gen yang diatur oleh hormon pelepas kortikotropin. Ekspresi yang berlebih dari protein ini dikaitkan dengan peningkatan proliferasi sel melanoma
2.	X201923_at	PRDX4	Protein yang dikodekan dalam gen ini adalah enzim antioksidan dan merupakan milik keluarga peroxiredoxin. Protein dalam gen ini terlokalisasi ke sitoplasma. Peroksidase yang merupakan keluarga dari peroxiredoxin mengurangi hidrogen peroksida dan alkil hidroperoksida menjadi air dan alkohol dengan menggunakan pengurangan setara yang berasal dari molekul donor yang mengandung tiol. Protein ini telah ditemukan memainkan peran pengaturan dalam aktivasi faktor transkripsi NF-kappaB.
3.	X230042_at	ITSN1	Protein yang terdapat dalam gen ini adalah protein terkait membran sitoplasma yang secara tidak langsung berfungsi mengkoordinasikan lalu lintas

No.	Probe ID	Simbol	Keterangan
			membran endositik dengan mesin rakitan aktin. Selain itu, juga dapat mengatur pembentukan vesikel berlapis clathrin dan dapat terlibat dalam faur ulang vesikel sinaptik. Protein ini telah terbukti berinteraksi dengan synamin, CDC42, SNAP23, SNAP25, SPIN90, EPS15, EPN1, EPN2, DAN STN2. Beberapa varian transkrip ang mengkode isoform berbeda telah ditemukan untuk gen ini, tetapi sifat penuh dari hanya dua dari mereka telah dikarakteristik sejauh ini.
4.	X205667_at	WRN	Gen ini merupakan anggota sub-famili RecQ dari protein helicase DNA. Protein nuklir yang dikodekan penting dalam pemeliharaan stabilitas genom dan berperan dalam perbaikan, replikasi, transkripsi, dan pemeliharaan telomer DNA. Protein ini mengandung domain eksonuklease Nterminal 3 hingga 5, domain helicase bergantung ATP, dan domain RQC(RecQ helicase conserved region) di wilayah pusatnya, dan domain C-terminal HDRC(helicase Rnase D C-terminal) dan sinyal lokalisasi nuklir. Cacat pada gen ini adalah penyebab sindrom werner, kelainan resesif autosom yang ditandai dengan penuaan yang cepat dan peningkatan kanker tertentu.
5.	X208692_at	RPS3	Ribosom, organel yang mengkatalisasi sintesis protein, terdiri dari subunit 40S kecil dan subunit 60S besar. Bersama-sama subunit ini terdiri dari 4 spesies RNA dan sekitar 80 protein yang berbeda secara struktural. Protein ribosom yang merupakan komponen subunit 40S berperan membentuk bagian domain tempat enzim dimulai. Protein milik keluarga S3P dari protein ribosom. Studi tentang tikus dan telah menunjukkan bahwa protein menunjukkan bahwa protein memiliki peran ekstraribosomal sebagai endonuklease yang terlibat dalam perbaikan kerusakan DNA akibat imbas UV. Protein tampaknya terletak di sitoplasma dan nukleus tetapi tidak di nukleolus. Tingkat ekspresi gen ini yang lebih tinggi pada adenokarsinoma usus besar dan polip adenomatosa dibandingkan dengan mukosa kolon normal yang berdekatan telah diamati. Gen ini ditranskripsi bersama dengan gen RNA nukleolar kecil U15A dan U15B, yang masing-masing terletak pada intron pertama dan kelima. Sebagaimana tipikal untuk gen yang mengkode protein ribosom, ada beberapa pseudogen yang diproses dari gen ini yang tersebar melalui genom. Beberapa varian transkrip disambung alternatif pengkodean isoform yang berbeda telah ditemukan untuk gen ini.

No.	Probe ID	Simbol	Keterangan
6.	X243428_at	KCNQ1OT1	Wilayah kromosom manusia 11p15.5 berisi dua kelompok gen yang diatur secara epigenetik yang diekspresikan hanya dari satu kromosom dengan cara parent-of-origin. Setiap cluster atau domain yang dicetak, diatur oleh imprinting control region independen (ICR) yang fungsional. Domain CDKN1C/KCNW1OT1 manusia diatur oleh ICR yang terletak diintron KCNQ1, dan mengandung setidaknya delapan gen yang diekspresikan secara eksklusif atau istimewa dari alel yang diturunkan secara maternal. DNA ICR secara khusus dimetilasi pada kromosom yang diwariskan secara maternal, dan tidak temetilasi pada kromosom yang diwariskan paternal. ICR berisi promotor gen KCNQ1OT1 yang di ekspresikan secara eksklusif dari alel paternal. Transkrip KCNQ1OT1 adalah antisense untuk gen KCNQ1 dan merupakan RNA non-coding yang panjang dan tanpa pewangi. Ini berinteraksi dari kromatin dan mengatur transkripsi beberapa gen target melalui modifikasi epigenetik. Transkrip diekspresikan secara abnormal dari kedua kromosom pada kebanyakan pasien dengan sindrom Beckwith-Wiedemann, dan transkrip juga memainkan peran penting dalam karsinogenesis kolorektal.
7.	X226439_s_at	NBEA	Gen ini mengkodekan anggota kelompok besar, beragam protein jangkar A-kinase yang menargetkan aktivitas protein kinase A kesitus subseluler tertentu dengan mengikat subunit pengatur tipe II. Ekspresi spesifik otak dan perekrutan selaput protein seperti protein yang sangat mirip pada tikus menunjukkan keterlibatan dalam lalu lintas membran post-golgi neuronal. Mutasi pada gen ini dapat dikaitkan dengan bentuk autisme. Gen ini dengan ekspresinya sering terganggu pada pasien dengan multiple myeloma. Penyambung alternatif menghasilkan beberapa varian transkrip yang mengkode isoform yang berbeda. Varian transkrip tambahan mungkin ada, tetapi sifat lengkapnya belum ditentukan.
8.	X212154_at	SDC2	Protein yang terdapat dalam gen ini adalah transmembran (tipe I) heparan sulfat proteoglikan dan merupakan anggota keluarga proteoglikan sindecan. Syndecans memediasi pengikatan sel, pensinyalan sel, dan organisasi sitoskeletal dan reseptor sindrom diperlukan untuk internalisasi protein HIV-1 tat. Protein sindrom-2 berfungsi sebagai protein membran integral dan berpartisipasi dalam proliferasi sel, migrasi sel, dan interaksi matriks-sel melalui reseptornya untuk protein matriks ekstraseluler. ekspresi sindrom-2

No.	Probe ID	Simbol	Keterangan
			yang diubah telah terdeteksi pada beberapa jenis tumor berbeda.
9.	X200082_s_at	RPS7	Ribosom, organel yang mengkatalisasi sintesis protein, terdiri dari subunit 40S kecil dan subunit 60S besar. Bersama-sama subunit ini terdiri dari 4 spesies RNA dan sekitaar 80 protein yang berbeda secara struktural. Gen ini mengkodekan protein ribosom yang merupakan komponen dari subunit 40S. Protein ini memiliki keluarga S7E dari protein ribosom. Terletak di sitoplasma. Seperti tipikal untuk gen yang mengkode protein ribosom, ada beberapa pseudogen yang diproses dari gen ini yang disebrkan melalui genom.
10.	X220448_at	KCNK12	Gen ini mengkodekan salah satu anggota superfamili protein saluran kalium yang mengandung dua domain P pembentuk pori. Produk gen ini belum terbukti sebagai saluran fungsional. Namun, mungkin memerlukan protein pembentuk non-pori lainnya untuk aktivitas.

5.6 Perbandingan Metode *Support Vector Machine* (SVM) Dan *Random Forest*

Setelah penulis melakukan analisis menggunakan metode *support vector machine* (SVM) dengan kernel linear, polynomial, dan sigmoid serta menggunakan metode random forest, didapatkan nilai akurasi yang berbeda-beda dari masing-masing metode. Berikut adalah perbandingan tingkat akurasi dari kedua metode seperti berikut :

Tabel 5.13 Perbandingan Akurasi SVM dan *Random Forest*

<i>Support Vector Machine</i> (SVM)			<i>Random Forest</i>
Kernel Linear	Kernel Polynomial	Kernel Sigmoid	
90.92 %	77.27%	86.36%	90.92%

Dari **Tabel 5.13** dapat dilihat perbandingan nilai akurasi yang dihasilkan antara kalsifikasi dengan metode SVM dan *Random Forest*. Dalam studi kasus *novel mutation target distinct subgroups of medulloblastoma in America* adalah

klasifikasi menggunakan metode *support vector machine* (SVM) dengan kernel linear dengan akurasi sebesar 90.92%, dimana metode SVM dengan kernel linear menggunakan nilai cost sebesar 0.1 dan dapat menggunakan metode *random forest* dimana nilai akurasi yang didapatkan sebesar 90.92% dengan *mtry* dan *ntree* yang digunakan sebesar 68 dan 300.



BAB VI

PENUTUP

6.1 Kesimpulan

Berdasarkan hasil analisis yang telah dilakukan oleh penulis, diperoleh beberapa kesimpulan yaitu sebagai berikut:

1. Nilai optimal atau nilai akurasi yang didapatkan ketika menggunakan metode *support vector machine* (SVM) menggunakan *kernel linear* menghasilkan nilai akurasi sebesar 90.92%, untuk *kernel polynomial* menghasilkan nilai akurasi sebesar 77.27%, dan untuk *kernel sigmoid* menghasilkan nilai akurasi sebesar 86.36%, serta menggunakan metode *random forest* menghasilkan nilai akurasi sebesar 90.92%.
2. Hasil dari perbandingan klasifikasi antara metode *Support Vector Machine* (SVM) dan *Random Forest* yang terbaik adalah metode *Support Vector Machine* (SVM) dengan *kernel linear* dan *Random Forest*, dimana masing-masing metode ini menghasilkan tingkat akurasi yaitu sebesar 90.92%.

6.2 Saran

Saran yang dapat diberikan dari hasil penelitian yang telah dilakukan penulis adalah :

1. Bobot yang terbentuk pada setiap variabel dalam penelitian ini dapat menjadi salah satu target dalam melakukan *treatmen* pada pasien.
2. Data yang digunakan lebih diperhatikan lagi guna untuk membagi data *training* dan data *latih* yang akan digunakan dalam prediksi atau analisis.
3. Karena dalam penelitian ini hasil yang didapatkan sama. Maka, peneliti menyarankan untuk mencoba kedua metode karena dalam kasus *medulloblastoma* ini penyebab pasti terjadinya belum diketahui.

DAFTAR PUSTAKA

- Ahmad, A. (2017). Mengenal Artificial Inteligence, Machine Learning,, Neural Network, dan Deep Learning. *Yayasan Cahaya Islman Jurnal Teknologi Indonesia*.
- Aliady, H., Tuasikal, N. J., & Widodo, E. (2018). Implementasi Support Vector Machine (SVM) dan Random Forest pada Diagnosis Kanker Payudara. *Seminar Nasional Teknologi Informasi dan Komunikasi 2018(SENTIKA 2018)*, 278-285.
- Alpaydin, E. (2010). *Introduction to Machine Learning, Secon Edition*. London: The MIT Press, cambridge, Massachusetts.
- American Society of Clinical Oncology ® Cancer.Net, Doctor-Approved Patient Information. (2019, Januari 02). *www.cancer.net*. Diambil kembali dari *ancer-types/medulloblastoma-childhood/statistics*:
<https://www.cancer.net/cancer-types/medulloblastoma-childhood/statistics>
- Bedy Purnama, S. (2019). *Pengantar Mechine Learning*. Bandung: Informatika Bandung.
- Bolstad, B. (2004). *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. University of California, Berkeley: Ph.D. Dissertation.
- Breiman, L. (2001). Random Forest. *Machine Laerning*, 45 pp.5-32.
- Breiman, L., & Cutler, A. (2003, Mei 2). Manual--Setting Up, Using, and Understanding Random Forest V4.0. Diambil kembali dari https://www.stat.berkeley.edu/~breiman/Using_random_forest_v4.0.pdf
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. New York: Chapman & Hall.
- Cahyo, L. B. (2018, Januari 03). Implementasi Metode Support Vector Machine untuk Melakukan Klasifikasi Pada Data Bioinformatika. 27-39. Diambil kembali dari GABUNGAN.pdf?sequence:

<https://dspace.uui.ac.id/bitstream/handle/123456789/5558/GABUNGAN.pdf?sequence=1&isAllowed=y>

Cancer Council. (2019, September 6). *www.cancer.org.au*. Diambil kembali dari types-of-cancer/childhood-cancers: <https://www.cancer.org.au/about-cancer/types-of-cancer/childhood-cancers/types.html>

David Murphey, P., Lina Guzman, P., & Alicia Torres, P. (2014). *America's Hispanic Children: Gaining Ground, Looking Forward*. Amerika Serikat: Child Trends.

Diani, R., Wisesty, U. N., & Annisa Aditsania. (2017). Analisis Pengaruh Kernel Support Vector Machine (SVM) pada Klasifikasi Data. *Ind. Journal on Computing*, 109-118.

Gautier, L., Cope, L., Bolstad, B. M., and, & Irizarry, R. A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 307-315.

Gentleman, R., Carey, V., Huber, W., & Hahne, d. F. (2019). Package 'genefilter'. 19.

Gentlemen, R., Carey, V., Huber, W., & Hahne, F. (2020). genefilter: methods for filtering genes from high-throughput experiments. 1-35.

Ghahramani, Z. (2004). Unsupervised Learning. *in advanced lectures on machine learning*, 72-112.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning : Data Mining , Inference, and Prediction*. California: Springer.

Honakan, A. (2018). Analisis dan Implementasi Support Vector Machine dengan String Kernel dalam Melakukan Klasifikasi Berita Berbahasa Indonesia. *Eproach*.

Huljanah, M., Rustam, Z., Utama, S., & Siswantining, a. T. (2019). Feature Selection Using Random Forest Classifier for Predicting Prostate Cancer. *9th Annual Basic Science International Conference 2019 (BaSIC 2019)*, 1-8.

Institute, N. C. (2019, Mei 22). *www.cancer.gov*. Diambil kembali dari rare-brain-spine-tumor/tumors/medulloblastoma: <https://www.cancer.gov/rare-brain-spine-tumor/tumors/medulloblastoma>

- K, A. A., Aljahdali, S., & Hussain, S. N. (2013). Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques. *International Journal of Computer Applications*, 12-16.
- K, S., S, K., & K, N. (2016). Data Mining : Analysis of student database using calasification techniques. *International Juornal of Computer Applications*, 22-27.
- Karabulut, E. M., Ozel, S. A., & Ibrikci, T. (2011). A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology*, 323-327.
- Klassen, M. (2010). Learning microarray cancer dataset by random forest and support vector mechines. *IEEE*.
- Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*. Tylor & Francis Group, LLC .
- Kuncoro. (2009). *Metode Riset Untuk Bisnis & Ekonomi*. Jakarta: Erlangga.
- Mardi, Y. (2014). Data Mining : Klasifikasi Menggunakan Algoritm c4.5. *Jurnal Edik Informatika*, 213-219.
- Mase, J., Furqon, M. T., & Rahayudi, d. B. (2018). Klasifikasi Penyakit Gigi dan Mulut Menggunakan Metode Support Vector Mechine. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 3648-3654.
- Modern Cancer Hospital. (2020, Januari 12). *Modern Cancer Hospital*. Diambil kembali dari Kanker Otak: https://www.moderncancerhospital.co.id/kanker_otak/
- Morfouace, M., Cheepala, S., Jackson, S., Fukuda, Y., Patel, Y. T., Fatima, S., . . . Roussel, a. M. (2015). ABCG2 Transporter Expression Impacts Group 3 Medulloblastoma Response to Chemotherapy. *American Association for Cancer Research Jurnals*, 3879-3889.
- NCBI. (2019, Maret 25). *NCBI*. Diambil kembali dari www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37418: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37418>

- NCBI. (2020, Juli 05). *NCBI*. Diambil kembali dari www.ncbi.nlm.gov/gene:https://www.ncbi.nlm.gov/gene?LinkName=geoprofiles_gene&from_uid=58606042
- Northcott, A. P., Korshunov, A., Witt, H., Hielscher, T., Eberhart, G. C., Mack, S., . . . D.Taylor, & M. (2011). Medulloblastoma Comprises Four Distinct Molecular Variants. *Journal of Clinical Oncology*, 1408-1414.
- Octaviani, T. L., & Rustam, d. Z. (2018). Random Forest for Breast Cancer Prediction. *Proceedings of the 4th International Symposium on Current Progress in Mathematics and Sciences (ISCPMS2018)*, 1-6.
- Park, H. S., Goo, M. J., & Jo, & H. (2004). Receiver Operating Characteristic(ROC) Curve :Partial Review for Radiologist. *Korean J Radiol* 5(1), 11-18.
- Robinson, G., Parker, M., Kranenburg, T. A., Lu, C., & Xiang. (2015). Novel mutations target distinct subgroups of medulloblastoma. *HHS Public Access*, 1-16.
- S.K.Agarwal. (2008). *BIOINFORMATICS*. Ansari Road, Darya Ganj New Delhi-110002: A P H Publishing Corporation.
- Serin, A. (2011). *Biclustering Analysis for Large Scale Data*. Jerman: Universitas Berlin.
- Siang, T. C., Soon, T. W., Kasim, S., Mohamad, M. S., Howe, C. W., Deris, S., . . . Ibrahim, a. Z. (2015). A review of cancer classification software for gene expression data. *International Journal of Bio-Science and Bio- Technology*, 89-108.
- Sumathi, K., Kannan, S., & Nagarajan, d. K. (2016). Data Mining Analysis of Student Database Using Clasifcation Techniques. *International Journal of Computer Applications*, 22-27.
- Sutton, C. (2005). Classification and Ragression Tree, Bagging, and Boosting. *Handbook of Statistics*, 303-329.
- Suyanto. (2018). *Machine Learning Tingkat Dasar dan Lanjut* . Bandung: Informatika Bandung.

Tsuda, K. (2000). Overview of Support Vector Machine. *Journal of IEICE*, pp.460-466.



LAMPIRAN

Lampiran 1 Membaca Data *Microarray*

```
inputdata<- list.celfiles("E:/TUGAS AKHIR/databaru/GSE37418_RAW", full.names=T)
memory.limit()
memory.limit(150000)
affy.data = ReadAffy(filename=inputdata)
affy.data
class(affy.data)
```

Lampiran 2 Melihat dan Menyimpan *Pheno* Data

```
geoset.data<-getGEO(GEO="GSE37418",GSEMatrix =TRUE)
str(geoset.data)
data.geosetdata<- exprs(geoset.data[[1]])
class(data.geosetdata)
dim(data.geosetdata)
data.geosetdata[1:5,1:4]
view(data.geosetdata)
phenodata<- pData(phenodata(geoset.data[[1]]))
str(phenodata)
varLabels(phenodata(geoset.data[[1]]))
dim(phenodata)
view(phenodata)
setwd("E:/TUGAS AKHIR/databaru")
#save pheno in txt file
writeMatrixToFile <- function(tmpMatrix, tmpFileName, blnRowNames, blnColNames)
{ output <- file(tmpFileName, "at")
  utils::write.table(tmpMatrix, output, sep = "\t", quote = FALSE,
    row.names = blnRowNames, col.names = blnColNames)
  close(output)}
writeMatrixToFile(tmpMatrix=phenodata, tmpFileName="phenodata.txt",
  blnRowNames=TRUE, blnColNames=TRUE)

getwd()
## PHENO DATA
phenodata1 <- read.AnnotatedDataFrame(file.path("phenodata.txt"), sep = "\t", header = TRUE)
class(phenodata1)
head(phenodata1)
rownames(phenodata1)
phenodata(affy.data) <- phenodata1
#get pheno data dari affy data
phenodata = pData(phenodata(affy.data))
varMetadata(phenodata1)

newgeosetdata<-get.array.subset.affybatch(affy.data,
  "characteristics_ch1", c("subgroup: SHH", "subgroup: G4",
    "subgroup: G3", "subgroup: WNT"))

#selected pheno sample to use\
```

```
#selected pheno sample to use\\
selected<- c('GSM918578','GSM918579','GSM918580',
'GSM918581','GSM918582','GSM918583','GSM918584',
'GSM918585','GSM918586','GSM918587','GSM918588',
'GSM918589','GSM918590','GSM918591','GSM918592',
'GSM918593','GSM918594','GSM918595','GSM918596',
'GSM918597','GSM918598','GSM918599','GSM918600',
'GSM918601','GSM918602','GSM918603','GSM918604',
'GSM918605','GSM918606','GSM918607','GSM918608',
'GSM918609','GSM918610','GSM918611','GSM918612',
'GSM918613','GSM918614','GSM918615','GSM918616',
'GSM918617','GSM918619','GSM918620','GSM918621',
'GSM918622','GSM918623','GSM918624','GSM918625',
'GSM918626','GSM918627','GSM918629','GSM918630',
'GSM918631','GSM918632','GSM918633','GSM918634',
'GSM918635','GSM918636','GSM918637','GSM918638',
'GSM918639','GSM918640','GSM918641','GSM918642',
'GSM918643','GSM918645','GSM918646','GSM918647',
'GSM918648','GSM918649','GSM918650','GSM918651',
'GSM918652','GSM918653')
phenodata1<-phenodata1[selected,]
phenoData(newgeosetdata)<-phenodata1
phenodata<-pdata(phenodata(newgeosetdata))
str(phenodata)
dim(phenodata)
view(phenodata)
```

Lampiran 3 Visualisasi data

```
##piechart subtype
desc1<-table(phenodata$ subgroup.ch1)
percent1<-round(desc1/sum(desc1)*100)
desc1<-as.data.frame(desc1)
lbls1<-paste(desc1$var1,'-',percent1,'% ',sep='')
pie(desc1$Freq,label=lbls1,col=c('green','blue','red','orange'))

####barplot gender
desc2<-table(phenodata$sex.ch1)
desc2<-as.data.frame(desc2)
barplot(desc2$Freq,xlab='Gender',ylab='Pasien',names.arg=c('Male','Female'))

##piechart ethnic
desc3<-table(phenodata$ethnic.ch1)
percent3<-round(desc3/sum(desc3)*100)
desc3<-as.data.frame(desc3)
lbls3<-paste(desc3$var1,'-',percent3,'% ',sep='')
pie(desc3$Freq,label=lbls3,col=c('green','blue','red','orange','yellow','grey',4,2))

####pie chart age
library(plyr)
desc4<-count(phenodata,"age.ch1")
desc4$age.ch1_2<-as.character(gsub("age:", "", desc4$age.ch1))
percent4<-round(desc4$freq/sum(desc4$freq)*100)
desc4<-as.data.frame(desc4)
lbls4<-paste(desc4$age.ch1_2,'-',percent4,'% ',sep='')
pie(desc4$freq,main="Age",label=lbls4,col=c("red","orange","grey","green","blue"),
border="white",cex=0.5)
```

Lampiran 4 Tahap *Preprocessing*

```
###preprocessing data
library(affyPLM)
eset.dchipdata<-threestep(newgeosetdata, background.method = "RMA.2",
normalize.method = "quantile",summary.method = "median.polish")
ekspressdata<-exprs(eset.dchipdata)
dim(eset.dchipdata)

#boxplot after preprocessing
###ini juga di ss plotnya
dev.new(wight=4+dim(geoset.data)[[2]]/5, height=6)
par(mar=c(2+round(max(nchar(sampleNames(geoset.data)))/2),4,2,1))
after.PreProcessing<-paste("GSE37418","/",'GPL570',"selected samples",sep='')
boxplot(ekspressdata, boxwex=0.7, notch=T, main=after.PreProcessing, outline=FALSE, las=2)
```

Lampiran 5 Tahap *Filtering*

```
####filtering data
filtdata<-nsFilter(eset.dchipdata, require.entrez=T,var.func=IQR,
                  remove.dupEntrez=T,var.cutoff=0.6,feature.exclude="^AFFX")
logdatavika<-filtdata$filter.log
esetdatavika<-filtdata$eset
featureNames(esetdatavika)<-make.names(featureNames(esetdatavika))
####melihat hasil esetdata
dim(esetdatavika)
head(esetdatavika)
class(esetdatavika)
esetdatavika
####membuat data microarray menjadi matrix/data frame
databaruvika<-exprs(esetdatavika)
dim(databaruvika)
head(databaruvika)
class(databaruvika)
view(databaruvika)
```

Lampiran 6 Membuat Label Klasifikasi

```
####kelas vektor ada 4 (SHH,WNT,G3,G4)
datacl<-c(1,1,3,1,2,1,1,1,0,0,0,0,1,1,1,3,
          0,1,1,1,1,0,0,1,3,1,1,2,2,2,2,
          1,0,0,1,0,1,1,1,2,2,2,1,1,0,3,1,
          1,0,1,1,1,1,1,1,0,0,3,0,0,3,3,3,
          1,1,1,1,2,1,2,1,1)
length(datacl)
class(datacl)
datafac<-factor(datacl,levels=0:3,labels = c("G3","G4","SHH","WNT"))
datafac
```

Lampiran 7 *Feature selection*

```
##filtering dengan menggunakan multtest
library(multtest)
datatestvika<-mt.teststat(databaruvika,datacl,test = "f")
class(datatestvika)
length(datatestvika)
qqnorm(datatestvika)
qqline(datatestvika)

###adjusting p-value(untuk melihat p-value yang sesuai dan tidak sesuai)
rawp=2*(1-pnorm(abs(datatestvika)))
prosedur=c("Benforroni","Holm","Hochberg","BH","BY")
adjusted=mt.rawp2adjp(rawp,prosedur)
data0<-adjusted$adjp[,]
data01<-data0[order(adjusted$index),]
head(data01)

####mengambil kolom rawp
ffs<-data01[,1]
class(ffs)
length(ffs)
ffs[1:10]
view(ffs)

#adjusting rawp
datarawp<-data.frame(databaruvika,ffs)
row.names(datarawp)<-row.names(databaruvika)
class(datarawp)
head(datarawp)
dim(datarawp)
library(dplyr)
datadatarawpfilt<-filter(datarawp,ffs<0.00000005)
datarawpfiltfinal<-subset(datarawp,ffs<0.00000005)
rownames(datarawpfiltfinal)
class(datarawpfiltfinal)
dim(datarawpfiltfinal)
head(datarawpfiltfinal)

datadef<-datarawpfiltfinal[,1:73]
head(datadef)
dim(datadef)
summary(datadef)
colnames(datadef)
```

Lampiran 8 Membuat data untuk di analisis

```
#usepackages(analysis)
library(e1071)
library(pROC)
dataaa=as.data.frame(t((datadef)))
head(dataaa)
dataY=as.factor(dataa1)
dataY
datause=as.data.frame(cbind(dataaa,dataY))
datause
dim(datause)
head(datause)

#### dengan memperhatikan unbalance data ####
G3<-dplyr::filter(datause, datause$dataY==0)
dim(G3)
G4<-dplyr::filter(datause, datause$dataY==1)
dim(G4)
WNT<-dplyr::filter(datause, datause$dataY==2)
dim(WNT)
SHH<-dplyr::filter(datause, datause$dataY==3)
dim(SHH)

#penentuan ukuran data tranning dan testing
set.seed(986)
rasioo=7/10
n<-round(nrow(G3)*0.7)
sampe1_G3<-sample(1:nrow(G3),n)
m<-round(nrow(G4)*0.7)
sampe1_G4<-sample(1:nrow(G4),m)
x<-round(nrow(WNT)*0.7)
sampe1_WNT<-sample(1:nrow(WNT),x)
z<-round(nrow(SHH)*0.7)
sampe1_SHH<-sample(1:nrow(SHH),z)

#data training dan testing
#training
training_G3=G3[sampe1_G3,]
dim(training_G3)
training_G4=G4[sampe1_G4,]
dim(training_G4)
training_WNT=WNT[sampe1_WNT,]
dim(training_WNT)
training_SHH=SHH[sampe1_SHH,]
dim(training_SHH)
#testing
testing_G3=G3[-sampe1_G3,]
dim(testing_G3)
testing_G4=G4[-sampe1_G4,]
dim(testing_G4)
testing_WNT=WNT[-sampe1_WNT,]
dim(testing_WNT)
testing_SHH=SHH[-sampe1_SHH,]
dim(testing_SHH)
data_train=rbind(training_G3,training_G4,training_WNT,training_SHH)
data_test=rbind(testing_G3,testing_G4,testing_WNT,testing_SHH)
dim(data_train)
view(data_training)
```

Lampiran 9 Analisis SVM

```
tuninglinear<-tune(svm,dataY~.,data = data_train,kernel="linear",types="c-clasifikasi",
  ranges = list(cost=c(0.1,0.01,0.001,1,10,100)))
summary(tuninglinear)
modellinear <- svm(dataY~., data_train, kernel = "linear", cost=0.1 ,
  scale=T, types = "c-clasifikasi",decision.value=T)
summary(modellinear)
predtrainlinear <- predict(modellinear, data_train)
table(predtrainlinear, data_train$dataY)
predtestlinear<-predict(modellinear, data_test)
table(predtestlinear, data_test$dataY)
mean(predtestlinear==data_test$dataY)
```

```

tuning_polynomial<-tune(svm,dataY~.,data = data_training,kernel="polynomial",types="c-clasification",
  ranges = list(cost=c(10,100,200,300,400,500)))
summary(tuning_polynomial)
model_polinom <- svm(dataY~., data_train, kernel = "polynomial", cost=10,
  scale=T, types = "c-clasification",decision.value=T)
pred_train_lpolinom <- predict(model_polinom, data_train)
table(pred_train_lpolinom, data_train$datay)
pred_test_polinom<-predict(model_polinom, data_test)
table(pred_test_polinom, data_test$datay)
mean(pred_test_polinom==data_test$datay)

tuning_sigmoid<-tune(svm,dataY~.,data = data_training,kernel="sigmoid",types="c-clasification",
  ranges = list(cost=c(0.1,1,10,100),gamma=c(0.1,1,2,3)),tunecontrol = tune.control(cross = 10))
summary(tuning_sigmoid)
model_sigmoid <- svm(dataY~., data_train, kernel="sigmoid", cost=0.1,
  gamma= 0.1, types="c-clasification", decision.value=T)
pred_train_sigmoid<- predict(model_sigmoid, data_train)
table(pred_train_sigmoid, data_train$datay)
pred_test_sigmoid<-predict(model_sigmoid, data_test)
table(pred_test_sigmoid, data_test$datay)
mean(pred_test_sigmoid==data_test$datay)

###ROC###
library(pROC)
forauc<-as.numeric(predict(modellinear,data_test))
auc<-multiclass.roc(data_test$datay,forauc)
auc$auc

###BOBOT###
weight <- t (modellinear$coefs) %%% modellinear$SV
weight <- apply(weight,2,function(v){sqrt(sum(v^2))})
weight<- sort (weight, decreasing = T)
X <- as.data.frame(weight)
view(X)
all(colnames(weight)==colnames(data_train))

```

Lampiran 10 Analisis Random Forest

```

library(ROSE)
library(randomForest)
library(caret)
randomfrst<-tunerf(data_train[,-4696],data_train[,4696],stepFactor = 2,
  fImprove = 0.05,trace = TRUE,plot = TRUE,doBest = FALSE)

randomfrst
Dataranfrst <- randomForest(dataY~.,data=data_train,importance=T, ntree=50,mtry=68)
Dataranfrst2 <- randomForest(dataY~.,data=data_train,importance=T, ntree=100,mtry=68)
Dataranfrst3 <- randomForest(dataY~.,data=data_train,importance=T, ntree=200,mtry=68)
Dataranfrst4 <- randomForest(dataY~.,data=data_train,importance=T, ntree=300,mtry=68)
Dataranfrst5 <- randomForest(dataY~.,data=data_train,importance=T, ntree=400,mtry=68)
Dataranfrst6 <- randomForest(dataY~.,data=data_train,importance=T, ntree=500,mtry=68)
Dataranfrst
Dataranfrst2
Dataranfrst3
Dataranfrst4
Dataranfrst5
Dataranfrst6
plot(Dataranfrst4)

plot(Dataranfrst)
plot(Dataranfrst2)
plot(Dataranfrst3)
plot(Dataranfrst5)
plot(Dataranfrst6)

Predtrainrf <-predict(Dataranfrst4,data_train)
confusionMatrix(Predtrainrf,data_train$datay)
Predtestrandomfrst4<-predict(Dataranfrst4,data_test)
confusionMatrix(Predtestrandomfrst4,data_test$datay)

## ROC ##
library(pROC)
foraucRF<-as.numeric(predict(Dataranfrst4,data_test))
aucRF<-multiclass.roc(data_test$datay,foraucRF)
aucRF$auc
###importance variable random forest
view(importance(Dataranfrst5),"GENENAME")
varImpPlot(Dataranfrst4,pch=19, main = "importance variable of moptimum")

```