BAB III
LANDASAN TEORI

3.1. Etanol

3.1.1 Definisi

Etanol merupakan zat kimia yang tidak sulit ditemui dalam kehidupan sehari hari. Disekitar kita umumnya dapat ditemui wujud etanol yang berupa cairan jernih (mirip seperti air mineral) tidak memiliki warna, etanol juga memiliki sifat yang mudah untuk menguap dan sangat sensitif sehinnga mudah terbakar. Wujud etanol mirip seperti yang tidak memiliki warna dan jernih (air mineral) sehingga terkadang agak sulit membedakannya dengan zat kimia lain yang memiliki wujud serupa dengan etanol seperti air, methanol, eter, kloroform dan aseton. Etanol juga sering dijuga dikenal dengan nama etil alkohol yang mana memiliki rumus kimia C_2H_5OH atau CH_3CH_2OH dimana memiliki titik didih 78,4° C. Menurut Rama (2008) etanol dapat dikelompokan menjadi 2 salah satunya adalah etanol sintetik seperti *methanol*. *Methanol* terbuat dari etilen yang merupakan salah satu dervat minyak bumi atau batubara yang dapat hasilkan dari proses sintesis zat kimia dengan nama hidrasi. Kemudian selanjutnya Bioetanol, bioethanol dapat dibuat dari bahan berupa tanaman. Sesuai dengan namanya etanol jenis ini dihasilkan melalui proses biologi yaitu peragian karbohidrat yang terdapat pada malt dan beberapa buah-buahan seperti hop, anggur dan sebagainya oleh mikroba atau melalui sintesis dari etilen dan alkohol melalui proses biologi seperti enzimatik dan fermentasi (Dewi, 2009).

3.1.2 Kegunaan

Dalam bidang kesehatan seperti yang berhubungan dengan farmasi atau kedokteran dan kimia seperti pabrik kimia misalnya, etanol umumnya digunakan sebagai bahan pelarut pada zat kimia lain seperti obat obatan dan senyawa kimia lainnya. Alkohol merupakan pelarut yang paling bermanfaat dalam bidang farmasi, digunakan sebagai pelarut utama untuk banyak senyawa organik, serta sebagai

bakterisida (pembasmi bakteri) terutama sebagai pembersih kulit sebelum injeksi. Etanol 60-80 % berkhasiat sebagai bakterisida yang kuat dan cepat terhadap bakteri bakteri, sebagai germisida alat-alat, sebagai obat sedatif dan depresan sistem saraf pusat yang memberikan efek tenang dan euforia (Makiyah, et al., 2005)

3.1.3 Efek Samping Pada Tubuh

Menurut BPOM (2016) etanol dapat mengiritasi mata. Terhirupnya uap etanol dalam konsentrasi tinggi dapat menyebabkan iritasi pada mata dan saluran pernapasan. Paparan etanol dalam jangka pendek dapat menyebabkan korban mengalami gangguan emosional, gangguan koordinasi motorik (gangguan keseimbangan, bicara kurang jelas), gangguan sensorik (vertigo, pandangan ganda), wajah kemerahan, detak jantung cepat, berkeringat, mual, muntah, mengantuk, pingsan, hingga koma. Korban juga dapat mengalami kejang yang disebabkan oleh kondisi hipoglikemia. Pada keracunan etanol ringan hingga sedang, korban/pasien dapat mengalami gejala-gejala seperti rasa gembira yang berlebihan, gangguan keseimbangan, nystagmus (bola mata bergerak tidak beraturan), berkurangnya ketajaman penglihatan, hilangnya rasa malu/batasan moral, perilaku agresif, mual, muntah, kulit kemerahan, dan dapat terjadi takiaritmia supraventrikular. Sementara pada keracunan yang berat, korban/pasien dapat mengalami koma, depresi sistem pernapasan, aspirasi paru, hipoglikemia, dan hipotermia.

3.1.4 Golongan Minuman Etanol

Menurut Keputusan Presiden Republik Indonesia No. 3/1997, minuman beralkohol dibedakan menjadi 3 (tiga) golongan. Minuman beralkohol golongan A adalah minuman beralkohol dengan kadar etanol 1% sampai 5%, misalnya bir. Minuman beralkohol golongan B adalah minuman beralkohol dengan kadar etanol 5% sampai 20%, misalnya anggur. Minuman beralkohol golongan C adalah minuman beralkohol dengan kadar etanol 20% sampai 55%, misalnya wiski dan brendi (Keppres RI, 1997)

3.1.5 Kosentrasi BAC

BAC (*Blood Acurate Consentartion*) adalah Konsentrasi Alkohol Darah (BAC) mengacu pada persen alkohol (etil alkohol atau etanol) dalam aliran darah seseorang. BAC sebesar 0,10% berarti bahwa suplai darah seseorang mengandung satu bagian alkohol untuk setiap 1000 bagian darah, Seseorang dikatakan mabuk ketika memiliki nilai 0,08% atau lebih tinggi. Beberapa factor yang mempengaruhi BAC adalah jumlah minuman standar (dalam satuan tertentu), jumlah waktu minuman yang dikonsumsi, berat badan, seks biologis, obat obatan, makanan (sedikit banyak). Menurut *Stanford Unversity* (2017) beberapa efek BAC dalam tubuh manusia pada konsentrassi tertentu seperti tabel 3.1

Tabel 3.1 Efek BAC Pada Tubuh Manusia

BAC (dalam %)	Efek Fisik dan Mental
0,01 - 0,03	Tidak ada efek yang terlihat. Peningkatan mood yang sedikit.
.04 - .06	Perasaan santai. Sensasi kehangatan. Gangguan penalaran dan memori minor.
.07 - .09	Gangguan keseimbangan, bicara, penglihatan dan kontrol ringan dilarang untuk mengemudi atau bersepeda pada level ini.
.10 - .12	Kerusakan yang signifikan dari koordinasi motorik dan kehilangan penilaian. Bicara mungkin tidak jelas.
.13 - .15	Penurunan kontrol motorik kotor. Visi kabur dan kehilangan keseimbangan. *Onset disforia* (kecemasan, gelisah).
.16 - .20	*Dysphoria* mendominasi. Mual mungkin muncul. Peminum memiliki penampilan "mabuk ceroboh."
0,25 - .30	Keracunan parah. Butuh bantuan berjalan. Kebingungan mental. Disforia dengan mual dan muntah.
.35 - .40	Hilang kesadaran. Jurang koma.
0,40 dan lebih tinggi	Timbulnya koma. Kemungkinan kematian karena gagal napas.

Adapun untuk menghitung konsentrasi *alcohol* dalam darah manusia maka dapat menggunakan perhitungan berikut :

$$BAC = \left(\frac{0.806\ X\ SD\ X\ 1.2}{BW\ X\ w_t} - MR - DP \right) X\ 10 \qquad (3.1)$$

dimana:

- 0,806 adalah konstanta untuk air tubuh dalam darah (rata-rata 80,6%),
- SD adalah jumlah minuman standar , yang menjadi 10 gram etanol masing-masing
- 1.2 adalah faktor untuk mengonversi jumlah dalam gram ke standar Swedia yang ditetapkan oleh Institut Kesehatan Publik Nasional Swedia,
- BW adalah konstanta air tubuh (0,58 untuk pria dan 0,49 untuk wanita),
- Berat badan (kilogram),
- MR adalah konstanta metabolisme (0,015 untuk pria dan 0,017 untuk wanita) dan
- DP adalah periode minum dalam hitungan jam.
- 10 mengubah hasilnya menjadi permillage alkohol

Dalam penelitian Kupfer (2013) dilakukan penelitian terhadap pasien yang mengkonsumsi minuman beralkohol. Hasilnya didapat nilai ekspresi gen yang berbeda beda pada masing masing pasien dengan cara perlakuan pasien yang berbeda beda. Pada kondisi *rising %* dimana nilai ekspresi gen pada kadar *alcohol* dalam darah pasien diperoleh saat BAC sedang meningkat kesatuan persen (%), pada kondisi *declining %* dimana nilai ekspresi gen pada kadar *alcohol* dalam darah pasien diperoleh saat BAC sedang menurun kesatuan %, sedangkan pada kondisi OJ time diperoleh nilai ekspresi gen saat waktu tertentu sehingga tidak berdasarkan konsentrasi BAC

3.2. Ekspresi Gen

Sebuah rangkaian *system* yang bertujuan untuk menterjemahkan informasi *genetic* dalam bentuk urutan basa DNA menjadi protein merupakan pengertian dari *gene epression* atau ekspresi gen. Dikatakan juga bahwa ekspresi gen adalah proses penentuan sifat dari suatu organisme oleh gen. Dengan kata lain ekspresi gen

merupakan wujud implementasi terjemahan genetic yang dilakukan dari fenotipik gen atau bisa siebut juga dengan gen melalui serangkaian proses transkripsi dan translasi *genetic*. (Anonim, 2017).

Pada proses ekrpresi gen di mana instruksi dalam DNA akan diubah menjadi produk fungsional, seperti protein. Dalam ekspresi gen proses mengubah DNA diatur secara ketat yang memungkinkan sel merespons lingkungannya yang berubah. Karena berfungsi sebagai saklar on / off untuk mengontrol ketika protein dibuat dan juga kontrol volume yang menambah atau mengurangi jumlah protein yang dibuat. Ada dua langkah kunci yang terlibat dalam membuat protein, transkripsi dan terjemahan (Silva dan Perera 2017).

Transkripsi adalah proses ketika DNA dalam suatu gen di *copy* atau disalin untuk memproduksi RNA atau biasa disebut *messenger* RNA (mRNA). mRNA dibawa oleh enzim yang sering disebut RNA polimerase yang terdapat pada nukleus dari satu sel untuk membentuk RNA. RNA secara kimia serupa dalam struktur dan sifat DNA.

Translasi adalah proses sintesis protein, terjadi setelah mRNA membawa informasi yang ditranskripsikan dari DNA ke ribosom (tempat pembuatan protein di dalam sel). Informasi yang dibawa oleh mRNA dibaca oleh molekul pembawa yang disebut *transfer* RNA (tRNA). mRNA membaca tiga huruf (kodon) pada suatu waktu. Setiap kodon menentukan asam amino tertentu. Misal tiga basis 'GGU' kode untuk asam amino yang disebut glisin.

Karena hanya ada 20 asam amino tetapi 64 kombinasi potensial kodon, lebih dari satu kodon dapat mengkode asam amino yang sama. Misalkan kodon ‘GGU’ dan ‘GGC’ merupakan kode untuk glisin. Setiap asam amino melekat khusus pada molekul tRNAnya sendiri. Ketika urutan mRNA dibaca, setiap molekul tRNA mengirimkan asam amino ke ribosom dan mengikat sementara ke kodon yang sesuai pada molekul mRNA. Setelah tRNA terikat, ia melepaskan asam amino dan asam amino yang berdekatan, semua bergabung bersama menjadi rantai panjang yang disebut polipeptida. Proses ini berlanjut sampai protein terbentuk. Data yang terbentuk sebagai hasil dari proses ekspresi gen sangat penting dalam mengidentifikasi mutasi dan perubahan dalam gen.

3.3. Bioinformatika

Bioinformatika terdiri dari dua akar kata yakni “bio” merujuk pada istilah biologi yang merupakan disiplin ilmu yang mengkaji mengenai makhluk hidup dan “informatika” merujuk pada istilah yang digunakan dalam dunia teknologi untuk mengintrepetasikan informasi yang terdapat dari/dalam suatu data yang diperoleh. Bioinformatika seringkali dianggap sebagai bentuk implementasi metode komputasi untuk menganalisa fenomena fenomena data yang didpat dalam penelitian biologi. Awal penggunaaan dari istilah bioinformatik sendiri sudah ada sejak tahun 80-an ketika pertama kali diperkenalkan istilah bioinformatik dikemukakan sebagai disiplin ilmu yang mengacu pada penerapan komputer dalam biologi (Fatchiyah, 2009).

Dalam penelitian Luscombe, Greenbaum dan Gerstein (2000) mengatakan sebagai hasil dari kajian informatika umumnya data yang ditangani umumnya dimensi data dalam jumlah besar hal tersebut menyimpan informasi mengenai dinamika kompleks yang diamati di alam. Menurut Kerlavage (199) dalam penelitian Luscombe, Greenbaum dan Gerstein (2000) sebuah laboratorium eksperimental dapat dengan mudah menghasilkan lebih dari 100 gigabytes data dalam sehari. Sehingga analisis yang dilakukan bergantung pada kemampuan dalam pemrosesan data yang memungkinkan penghitungan yang lebih cepat, penyimpanan data yang lebih baik, dan merevolusi metode untuk mengakses dan bertukar data.

Adapun tujuan bioinformatika adalah : Pertama, bioinformatika yang paling sederhana mengatur data dengan cara yang memungkinkan para peneliti untuk mengakses informasi yang ada dan untuk mengirimkan entri baru ketika mereka diproduksi, misalnya Bank Data Protein untuk struktur makromolekul 3D Berstein et al (1997). sehingga tujuan bioinformatika jauh melampaui kontrol volume belaka. Tujuan kedua adalah mengembangkan alat dan sumber daya yang membantu dalam analisis data. Sebagai contoh, setelah mengurutkan protein tertentu, langkah selanjutnya adalah untuk membandingkan data dengan urutan yang ditandai sebelumnya. Hal ini membutuhkan lebih dari sekadar pencarian basis

data langsung. Sehingga harus mempertimbangkan apa yang menjadi fakor kemiripan signifikan secara biologis. Tujuan ketiga adalah menggunakannya untuk menganalisis data dan menafsirkan hasilnya dengan cara yang bermakna secara biologis. Secara tradisional, studi biologi memeriksa sistem individu secara rinci, dan sering membandingkannya dengan beberapa yang terkait. Dalam bioinformatika, dapat melakukan analisis global terhadap semua data yang tersedia dengan tujuan mengungkap prinsip-prinsip umum yang berlaku di banyak sistem dan menyoroti fitur-fitur yang unik bagi sebagian orang. (Molekul) bio-informatika: bioinformatika adalah konsep biologi dalam hal molekul (dalam arti kimia fisik) dan menerapkan "teknik informatika" (berasal dari disiplin ilmu seperti matematika terapan, ilmu komputer dan statistik) untuk memahami dan mengatur informasi yang terkait dengan molekul-molekul dalam skala besar (Luscombe, Greenbaum dan Gerstein, 2000).

.

3.4. *Microarray*

Microarray merupakan suatu *chip* mikroskop yang berisikan serangkaian sample berupa DNA, RNA, protein dan jaringan (Pasanen, 2014). *Chip* gen dari *microarray* terbuat dari silikon atau kaca dimana bahan genetik akan ditempatkan dan memiliki struktur seperti grid, setiap spot yang mengandung rangkaian *nukleotide* tunggal yang berbeda disebut sebagai *probe* dan setiap *spot* akan mempunyai jutaan salinan *probe* (Santamaria, 2009).

Dalam topik kajian ilmu bioinformatika data dengan bentuk *microarray* merupakan data yang sangat lazim ditemukan. Pada data *microarray* umumnya memiliki ukuran dimensi data yang sangat besar beberapa diantara bahkan diatas puluhan ribu jumlah data. Hal ini disebabkan data yang dimuat merupakan data gen pada manusia yang jumlahnya sangat banyak sehingga memiliki informasi terkait gen dan memiliki *feature* yang sangat banyak. Data *microarray* sangat menarik untuk dianlisis menggunakan pendekatan data mining Selain itu *microarray* juga terbagi atas samplenya: protein *miroarray* dan DNA/RNA *miroarray* (Kokoh 2006).

3.5. *Preproccesing*

Preproccesing adalah langkah awal yang dilakukan bertujuan melakukan penyesuaian dari data yang dimiliki serta melakukan konversi data *affybatch* ke dalam bentuk *expression set.* Pada langkah yang dilakukan ketika *preprocessing* juga dilakukan beberapa proses seperti *background correction*, *normalization* dan *summarization*.

Tujuan dari langkah *background correction* adalah untuk menghilangkan *noise* pada latar belakang secara keseluruhan. Setiap array dibagi menjadi satu set daerah, maka nilai latar belakang untuk itu adalah perkiraan *grid*. Kemudian setiap intensitas probe disesuaikan berdasarkan rata-rata tertimbang dari masing-masing nilai latar belakang. Bobot tergantung pada jarak dari pusat massa masing-masing grid. Secara khusus, bobotnya adalah

$$\text{wk (x, y)} = \frac{1}{d_k^2\,(x,y)+smooth} \qquad (3.2)$$

Dimana dk (x, y) adalah jarak euclidean dari lokasi x, y ke pusat massa wilayah k. Default nilai smooth adalah 100. Perhatian khusus diberikan untuk menghindari nilai negatif atau proble numerik lainnya (Bolstad, 2004)

Langkah selanjutna adalah normalisasi adalah proses menghilangkan variasi non-biologis yang tidak diinginkan yang mungkin ada di antara chip dalam percobaan *microarray*. Diketahui bahwa variabilitas bisa ada antara *array*. Beberapa kepentingan biologis dan kepentingan non-biologis, Dua jenis variasi ini dikategorikan menarik atau tidak jelas oleh Hartemink, et al. (2001). Variasi yang tersembunyi ini yang akan dihapus saat menormalkan *array*. Sumber untuk mengaburkan variasi dapat mencakup perbedaan pengaturan pemindai, jumlah mRNA yang digabungkan dan banyak lainnya faktor-faktor. Hartemink, et al. (2001). Dimana langkah langkah sesuai dengan fungsi sebagai berikut :

1. Terdapat n *array* dengan panjang p, berbentuk X dari dimensi p × n di mana setiap *array* adalah kolom.

2. Urutkan setiap kolom X untuk memberikan Xsort.
3. Ambil rata-rata melintasi baris-baris Xsort dan tetapkan nilai ini untuk setiap elemen di baris untuk mendapatkan quantile equalized X 'menyortir.
4. Dapatkan Xnormalisasi dengan mengatur ulang setiap kolom X ' urutkan agar memiliki urutan yang sama dengan X asli.

Terdapat dua pendekatan yang digunakan pada *summarization* yatitu model *single chip* dan model *multi chips*. Model *Single chip* adalah Metode ini hanya menggunakan informasi *probe* pada *array* individual untuk menghitung ringkasan ekspresi untuk *array*. Nilai ekspresi untuk setiap larik dihitung terpisah dari informasi dalam *array* lainnya. Sedangkan model multi chip adalah metode yang bekerja dengan memeriksa pola respons *probe* di seluruh *array*. Hal itu dapat terjadi dengan mengamati bahwa variabilitas antara *probe* yang berbeda lebih besar daripada variabilitas *probe* tunggal melintasi *array mulitple*. Karena tidak sedikit juga ditemukan pada *array* individu yang berperilaku berbeda kemungkinan disebabkan oleh faktor non-biologis. Adapun parameter yang digunakan pada *summarization* adalah median polish dengan model seperti berikut :

$$log_2\left(y_{ij}^{(n)}\right) = \mu^n + \theta_j^n + \alpha_j^n + \varepsilon_{ij}^n \tag{3.3}$$

dengan median (θ_j) = median (α_j) = 0 dan median i (ε_{ij}) = median j (ε_{ij}) = 0. Log2 nilai ekspresi diberikan oleh $\hat{\beta}_j^n = \hat{\mu}_j^n + \hat{\theta}_j^n$. algoritma ini menghasilkan langkah langkah sebagai berikut: pertama-tama sebuah matriks dibentuk untuk setiap *probe* n sehingga *probe* berada dalam baris dan *array* berada di kolom. Matriks ini ditambah dengan efek baris dan kolom yang memberikan bentuk matriks.

$$\begin{matrix} e_{11} & \dots & e_{1NA} & a_1 \\ \vdots & & \vdots & \vdots \\ e_{In1} & \dots & e_{InNA} & a_{In} \\ b_1 & \dots & b_{NA} & m \end{matrix} \tag{3.4}$$

dimana awalnya $e_{ij} = y_{ij}^{n}$, dan $a_i = b_j = m = 0$. Selanjutnya, setiap baris diseleksi dengan mengambil nilai median melalui kolom (mengabaikan kolom terakhir dari efek baris) mengurangkannya dari setiap elemen di baris itu dan menambahkannya ke kolom terakhir (a_1, ..., a_{In}, m). Kemudian kolom dilakukan dengan cara yang sama dengan mengambil nilai median melalui baris, kemudian mengurangi dari setiap elemen dalam baris itu dan setelahnya menambahkan ke baris bawah (b_1, ..., b_{NA}, m). Prosedur berlanjut, iterasi seleksi baris diikuti oleh kolom, sehingga perubahannya menjadi kecil atau nol. Pada akhir prosedur ini $\hat{\mu}$ = m, $\hat{\theta}_j$ = b_j dan $\alpha_i = a_i$. Elemen e_{ij} akan menjadi nilai residu. Prosedur ini mungkin menyatu estimasi parameter yang berbeda tergantung pada apakah baris atau kolom disapu terlebih dahulu. Dalam analisis dalam disertasi ini, baris selalu disapu terlebih dahulu. Satu kelemahan dari prosedur poles median adalah bahwa hal itu tidak secara alami memberikan kesalahan standar perkiraan. Lain adalah bahwa kita dibatasi untuk model efek kolom-baris seimbang.

3.6. *Filtering*

Filtering merupakan suatu kegiatan atau tahap yang digunkan untuk mengurangi jumlah gen dan meningkatkan kekuatan dalam suatu analisis dan juga untuk memilih data yang berpengaruh dan mengesampingkan data yang tidak berpengaruh dalam analisis data. Semakin banyak jumlah gen dan proporsi dari tingkat *gene expression* sangat rendah maka dapat mengakibatkan rendahnya nilai akurasi dalam suatu analisis. Dimana dapat dinyatakan dalam persamaan sebagai untuk gen *g* dalam sampel *k* di blok *j* dan kelompok *i* , dihasilkan sesuai dengan model

$$Y_{ijkg} = F_{ig} \times I_g + B_{jk} + Z_{ijkg} . \qquad (3.5)$$

Proporsi gen dipilih secara acak untuk memiliki variabel indikator nilai Ig dan sisanya gen memiliki nilai Ig. Dimana $F_{ig} \sim N(\mu, \sigma^2)$ gen dikelompokkan secara acak menjadi n blok dari N gen, ditunjukkan oleh subskrip j dan dengan $B_{jk} \sim N(0, \sigma_b^2)$. Variabel $Z_{ijkg} \sim N(0, \sigma_g^2)$ dimana $\sigma_g^2 \sim$ Uniform (u_{min} , u_{max}) digunakan

untuk memungkinkan varians berbeda di antara gen. Menurut Hackstadt (2009) *filtering* bekerja dengan cara mendeteksi variansi,medeteksi rata-rata sinyal gen.

3.7. Analisis Deskriptif

Statistika deskriptif adalah bagian dari ilmu statistik yang meringkas, menyajikan dan mendeskripsikan data dalam bentuk yang mudah dibaca sehingga memberikan informasi yang jelas dan lengkap. Statistik deskriptif hanya berhubungan dengan hal menguraikan atau memberikan keterangan-keterangan mengenai suatu data atau keadaan atau fenomena. Dengan kata lain hanya melihat gambaran secara umum dari data yang didapatkan (Hasan, 2005). Pada umumnya metode yang biasa digunakan untuk menjelaskan karakteristik suatu data yaitu dalam bentuk tabel, grafik, diagram, atau statistik sampel.

Menurut Sugiyono (2009) analisis deskriptif adalah metode yang berfungsi untuk mendeskripsikan atau memberi gambaran terhadap objek yang diteliti melalui data atau sampel yang telah terkumpul sebagaimana adanya tanpa melakukan analisis dan membuat kesimpulan yang berlaku untuk umum. Contoh dari penyajian data dalam statistika deskriptif adalah tabel, diagram dan grafik (Walpole, 1995).

3.8. *Machine Learning*

Machine Learning merupakan suatu cara atau metode ilmu yang mengkaji terkait solusi permasalahan analisis mengenai bagaimana membuat kecerdasan buatan atau *artificial intelligence*. Adapun pendekatan yang dilakukan adalah dengan menggunakan metode pembelajaran mesin dimana sebuah mesin diberikan algorita khusus untuk mempelajari pola data yang diberikan. Sehingga machine learning merupakan suatu bagian metode ilmu dari *artificial intelegence* atau kecerdasan buatan dimana mesin diberikan algoritma tertentu untuk mempelajari pola data sehingga dapat digunakan menjadi solusi permasalahan. Alpaydin (2010)

Cara kerja algoritma pada *Machine learning* adalah dengan mengoptimal kan data set yang ada untuk meningkatkan kinerja pembelajaran dari *machine learning*. Dimana data yang dimiliki dipartisi menjadi beberapa bagian sebagai

data training dan data *testing*. Pada data training digunakan untuk membuat atau memprediksi parameter model berdasarkan data yang digunakan setelah model didapatkan kemudian menggunakan model yang telah didapat dalam data training untuk memprediksi data testing. Sehingga model yang didapat akan memberikan informasi berdasarkan hasil analisis yang dilakukan. Dalam menentukan model Machine learning menggunakan pendekatan algoritma yang berhubungan dengan terori perhitungan statistic (Alpaydin, Intoduction to Machine Learning 2010). *Machine learning* dibagi menjadi tiga bagian, yaitu sebagai berikut :

1. *Supervised Learning*

   Pada data dengan metode Supervised learning umumnya telah terdapat label data pada dataset yang digunakan label yang sudah dimiliki dari dataset kemudian digunakan untuk mendapatkan model sehingga dapat membuat prediksi dari model yang dimiliki.

2. *Unsupervised Learning*

   Pada data dengan metode Supervised learning umumnya tidak terdapat label data pada dataset yang digunakan sehingga dataset yang sudah dimiliki dilakukan pengelompokan pada data yang tidak memiliki label dengan acuan dari karakterisik data yang dimiliki.

3. *Reinforcement Learning*

   Pada data dengan metode *Reinforcement Learning* umunya yang akan dilakukan oleh mesin adalah melatih dirinya berulang ulang dengan acuan lingkungan yang dipengaruhinya, kemudian ketika model susdah didapat mesin akan menerapkan pengetahuan dari model yang didapat sebagai solusi pada suatu kasus masalah.

Metode analisis *biclutering* bekerja dengan melakukan pengelompokan data pada bagian baris dan kolom secara bersamaan berdasarkan tingkat kemiripan yang paling dekat (homogen) sehingga membedakan antar kelompok *biclustering*. Hal termasuk kedalam pendekatan *machine learning* dengan metode *unsupervised learning* dimana pendekatan pada metode tersebut bekerja dengan cara mengelompokan data yang tidak memiliki label berdasarkan karakterisik-karakteristik yang ditemui (Alpaydin, Intoduction to Machine Learning 2010)

3.9. *Biclustering*

Analisis *cluster* merupakan salah satu metode *machine learning* yang termasuk kedalam algoritma *unsupervised learning*. Sharma (1996), dalam Rahmawati et al. (2010) menyatakan analisis *cluster* atau yang biasa dikenal sebagai *clustering* adalah salah satu metode statistik yang bertujuan untuk mengelompokkan objek ke dalam suatu kelompok sedemikian sehingga objek yang berada dalam satu kelompok akan memiliki kesamaan yang tinggi dibandingkan dengan objek yang berada di kelompok lain.

Struktur data dasar biasanya adalah matriks, dengan variabel M (atau fitur, baris matriks) dan pengamatan N (atau kondisi, sampel, kolom dari matriks) berikut ini.

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix} \tag{3.6}$$

dimana:

X = Nama matriks

n = banyak baris

m = banyak kolom

Clustering menggunakan baris dan kolom dalam matriks data secara terpisah. Misal dalam analisis data ekspresi gen, ketika menggunakan analisis *clustering* pengelompokan setiap gen didefinisikan menggunakan semua kondisi, dan pengelompokan kondisi dicirikan oleh aktivitas semua gen. Sehingga yang didapat adalah *cluster* global, hal ini menyebabkan banyak pola aktivasi gen hanya dibawah kondisi eksperimental tertentu dan banyak informasi yang dikesampingkan. Mengatasi keterbatasan metode *clustering*, dalam analisis data ekspresi gen, dengan pengelompokan gen dan sampel secara bersamaan, ada konsep yang disebut *biclustering*.

3.9.1 Pengertian *Biclustering*

Biclustering pertama kali di kembangkan pada tahun 1972 oleh J.A. Hartigan, dimana Hartigan mengusulkan suatu model dan teknik simultan untuk mengelompokkan cases dan variabel pada dataset voting dan tekniknya disebut

sebagai teknik *Direct Clustering.* Cheng & Church (2000) mengaplikasikan *biclustering* pada bidang bioinformatika dimana *biclustering* digunakan pada data *microarray gene expression*.

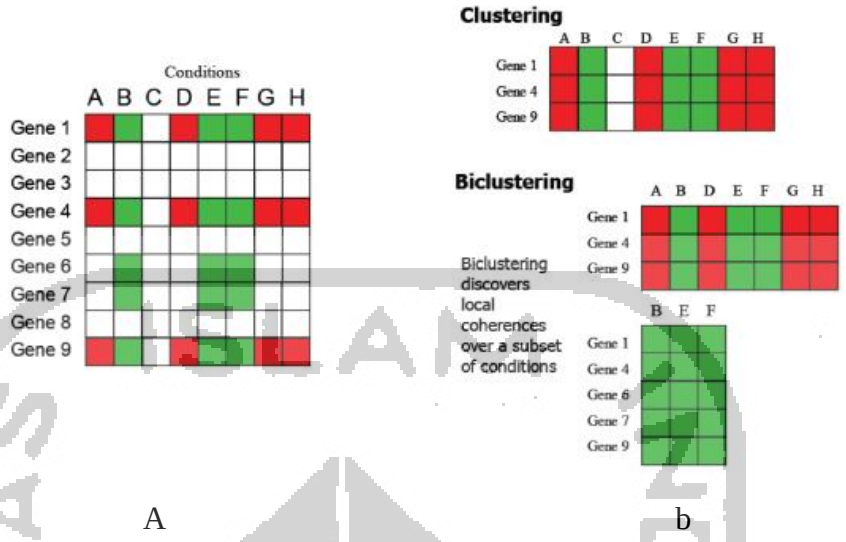
Zhao *et al.,* (2012) mengatakan analisis *biclustering* adalah metodologi yang berfungsi untuk menemukan pola yang berhubungan pada submatriks yang tersembunyi dalam matriks data. Berbeda dengan *clustering*, dimana pada *biclustering* dilakukan pengelompokan secara simultan pada dua dimensi matriks data, yaitu pada baris dan kolom. Dengan menggunakan metode ini, akan terungkap submatriks dengan elemen-elemen yang memanifestasikan perilaku yang sama dari anggota kelompok. Submatriks ini terdiri dari subset kolom yang bisa mencirikan suatu kelompok baris.

Madeira dan Oliveira (2004) dalam papernya mengatakan, *biclustering* dibutuhkan untuk menemukan pola lokal yang dapat membantu mengungkapkan jalur genetik yang tersembunyi. Tujuan metode *biclustering* dalam data ekspresi gen adalah:

1. Mengelompokkan gen sesuai dengan nilai ekspresinya dengan berbagai kondisi.
2. Anotasi pada gen, memberikan klasifikasi yang diketahui pada ekspresi gen.
3. Kondisi kelompok sesuai dengan data ekspresi dari sekelompok gen yang diatur bersama.
4. Klasifikasikan sampel baru.

Cheng & Chruch (2000) mengatakan salah satu tujuan *biclustering* untuk mengungkapkan keterlibatan gen atau kondisi. Hal menarik dalam analisis data ekspresi gen selain ditemukan *bicluster* maksimum, adalah temuan sekelompok gen yang menunjukkan regulasi yang bermiripan dalam beberapa kondisi.

Biclustering pada matriks data ekspresi gen bisa divisualisasikan seperti pada Gambar 3.3. Gambar 3.3a adalah gambaran dari matriks data dan Gambar 3.3b atas mengilustrasikan hasil dari *clustering* biasa, dan Gambar 3.3b bawah mengilustrasikan hasil dari *biclustering*.



A b

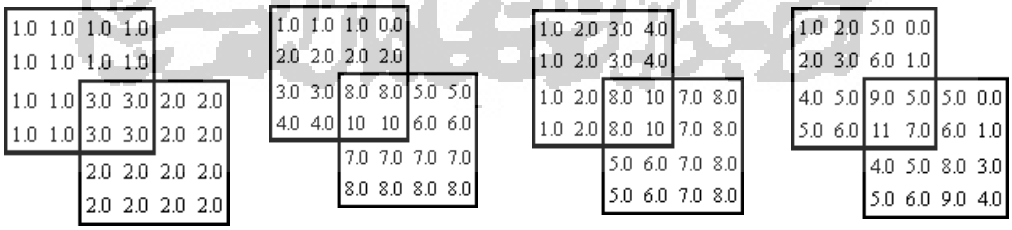
Gambar 3.1 (a) matriks data dan (b) *clustering* dan *biclustering*, sumber: (Mina, 2010)

3.9.2 Tipe-Tipe dan Struktur pada *Bicluster*

Beberapa tipe *biclustering* menurut (Madeira and Oliveira 2004) adalah:

a. *Bicluster* dengan nilai konstan
b. *Bicluster* dengan nilai konstan pada baris
c. *Bicluster* dengan nilai konstan pada kolom
d. *Bicluster* dengan nilai koheren

Algoritma *bicluster* yang paling sederhana mengidentifikasi himpunan bagian dari baris dan himpunan bagian kolom dengan nilai konstan. Berikut adalah tipe-tipe dari *bicluster* yang terbentuk dari suatu matriks.



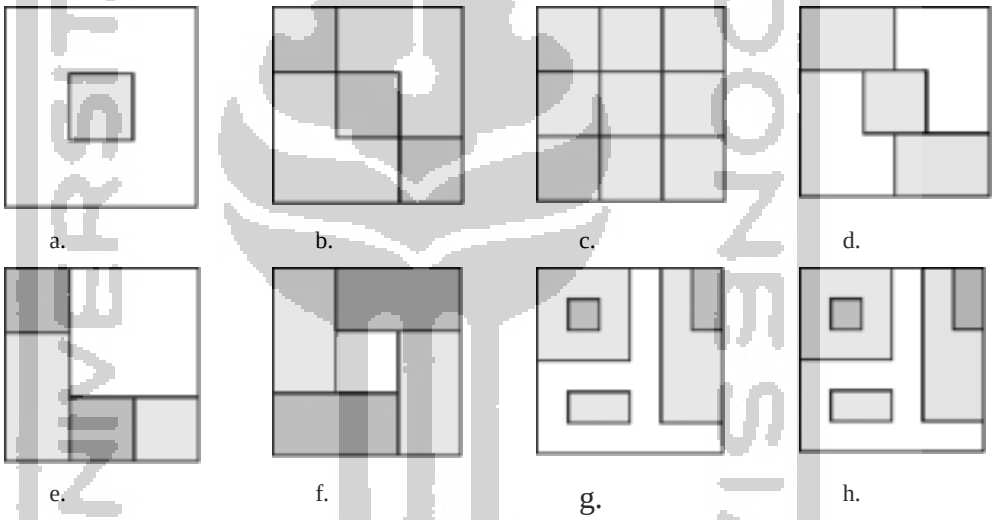
A B c d

Gambar 3.1 (Tipe-tipe *Bicluster*, sumber: (Madeira and Oliveira 2004)

Pada algoritma *Biclustering* terdapat beberapa struktur *bicluster*, hal ini terjadi karena saat algoritma *bicluster* berasumsi adanya beberapa *bicluster* dalam

satu matriks data. Struktur-struktur *bicluster* di visualisasikan pada Gambar 3.3 sebagai berikut:

a) *Bicluster* tunggal
b) *Bicluster* baris dan kolom eksklusif
c) *Biclustering Non-Overlapping* dengan struktur kotak-kotak
d) *Bicluster* eksklusif pada baris
e) *Bicluster* eksklusif pada kolom
f) *Bicluster Non-Overlapping* dengan struktur pohon
g) *Bicluster Overlapping* Hirarki
h) *Biclustering* tumpang tindih



Gambar 3.3 Struktur-struktur *bicluster*, sumber : (Madeira & Oliveira, 2004)

3.9.3 Kelas *Biclustering* dan Algoritma *Biclustering*

Menurut penelitian Pontes, et al. (2015) *biclustering* dapat diklasifikasikan menjadi 9 kategori, seperti *Iterative greedy search, Stochastic iterative greedy search, Nature-inspired meta-heuristics,Clustering-based approaches, Graph-based approaches, One-way clustering-based approaches, Probabilistic models, Linear algebra, Optimal reordering of rows and columns.* Setiap 9 kategori metode tersebut memiliki pendekatan *biclustering* yang berbeda beda, sehingga hasil nya juga tentu akan berbeda beda tergantung pendekatan metode yang digunakan. Pada metode yang digunakan penulis yaitu *Spectral Biclustering*, menggunakan

pendekatan aljabar linier dimana matriks disusun sedimikan rupa kemudian dipartisi baris dan kolom secara bersamaan menggunakan pendekatan metode *Singular Value Decomposition* dan *K-means.* Adapaun algoritma *biclustering* yang juga menggunaan pendekatan aljabar linier yaitu:

Tabel 3.1 Pembagian Algoritma *Biclustering Linear Algebra*

Kelas *Biclustering*	**Algoritma *biclustering***
Linear algebra	*Spectral Biclustering (SB), Iterative Signature Algorithm (ISA), Non-smooth Non-negative Matrix Factorization (nsNMF), Pattern-based Biclustering (BicPAM).*

3.10. Algoritma *Spectral Biclustering*

Menurut (Piscopo and Marina 2017) dasar penggunaan metode *biclustering* spektral dalam penelitian (Kluger 2003) adalah dalam matriks data ekspresi setelah nilai tersebut diatur dengan benar menggunakan pendekatan aljabar linier, melalui Dekomposisi Nilai Singular (SVD). Menggunakan SVD, data matriks D dari dimensi N × M dapat didekomposisi sebagai $D = U\Lambda V^T$. di mana Λ adalah matriks diagonal dengan penurunan entri non-negatif, dan U dan V adalah matriks kolom ortonormal dengan dimensi N × min (N, M) dan M × min (N, M). Masing-masing. Jika matriks data memiliki struktur diagonal blok (dengan semua elemendi luar blok sama dengan nol), maka setiap blok dapat dikaitkan dengan *bicluster*. Khususnya, jika matriks data berbentuk:

$$D = \begin{bmatrix} D1 & 0 & \dots & 0 \\ 0 & D2 & \dots & 0 \\ . & . & & . \\ . & . & & . \\ 0 & 0 & \dots & Dk \end{bmatrix} \quad (3.7)$$

Dimana Di (i = 1, ..., k) adalah matriks arbitrer, kemudian, untuk masing-masing Di, akan ada vektor tunggal berpasangan (ui, vi) sehingga komponen bukan-nol dari ui sesuai dengan baris-baris yang ditempati oleh Di, dan bukan nol komponen vi sesuai dengan kolom yang ditempati oleh Di.

Singular Value Decomposition (SVD) adalah suatu pemfaktoran matrik dengan mengurai suatu matrik ke dalam dua matrik P dan Q. Jika diketahui suatu matrik adalah matrik A berukuran m×n dengan rank r > 0 , maka dekomposisi dari matrik A dinyatakan sebagai

$$A = P \Delta Q^T \tag{3.8}$$

Rank (r) menyatakan banyaknya jumlah baris atau kolom yang saling independen antara baris atau kolom lainnya dalam suatu matrik. P merupakan matrik orthogonal berukuran m×r sedangkan Q merupakan matrik orthogonal berukuran n×r. Δ adalah matrik diagonal berukuran r×r yang elemen diagonalnya merupakan akar positif dari eigenvalue matrik A.

Adapun langkah perhitungan Misal diketahui matrik B berukuran mxm *non singular* (matrik *fullrank* / matrik yang determinannya tidak sama dengan nol). Menghitung matrik B^TB dan BB^T. Misalkan matrik B^TB = matrik Y dan BB^T = matrik Z. Kemudian mencari eigenvalue (λ) dari matrik Y dan Z . Dimana determinan dari matrik Y dan Z dikurangi λ dikalikan dengan matrik identitas (I) sama dengan 0.

$$|Y-\lambda I| = 0 \text{ dan } |Z-\lambda I| = 0 \tag{3.9}$$

Banyaknya eigenvalue (λ) yang akan diperoleh sama dengan ukuran matrik Y dan Z yaitu sebanyak m. Setelah diketahui nilai-nilai λ nya, langkah selanjutnya adalah mencari eigenvektor untuk masing-masing λ. Eigenvektor diperoleh melalui rumus $(Y - \lambda I)\underline{x} = \underline{0}$ dan $(Z - \lambda I)\underline{x} = \underline{0}$. Sehingga akan diperoleh persamaan x dalam bentuk x_1, x_2 hingga x_m ($a_1x_1+a_2x_2+..+a_mx_m=0$). Setelah didapatkan persamaan kemudian dilakukan normalisasi (penormalan) dari tiap-tiap λ dengan mensubsitusikan tiap elemen $\underline{x_1}$. Proses penormalan adalah sebagai berikut:

$$\underline{x_1^*} = \frac{\underline{x_1}}{\left(\underline{x_1}^T \underline{x_1}\right)^{1/2}} = \frac{\begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}}{\left((x_{11}\ x_{12}) \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix} \right)^{1/2}} \tag{3.10}$$

Selajutnya juga dilakukan penormalan seperti contoh di atas untuk eigenvalue (λ) yang lain. Setelah $\underline{x_1^*}$ dan $\underline{x_2^*}$ telah diperoleh elemen-elemennya, selanjutnya adalah menggabungkan ketiga hasil penormalan tersebut ke dalam satu matrik dimana kolom pertama adalah $\underline{x_1^*}$, kolom kedua adalah $\underline{x_2^*}$. Sehingga diperoleh matrik

$$X = \left[x_1^*\ x_2^*\right] = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \tag{3.11}$$

Menentukan D yang merupakan matrik diagonal dengan elemen diagonalnya adalah akar dari eigenvalue matrik Y atau Z.

$$D = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \tag{3.12}$$

Diperoleh SVD dengan mengoperasikan $\underline{P}D\underline{Q}$ dimana hasilnya akan sama dengan matrik B.

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{3.13}$$

Jika $\underline{P}$ adalah eigenvektor dari matrik Z dan $\underline{Q}$ adalah eigenvektor dari matrik Y. Dan ketika dioperasikan kedalam $\underline{P}D\underline{Q}$ maka akan menghasilkan matrik yang sama dengan B.

Pada proses algoritma *spectral* setelah mencari nilai SVD maka langkah selanjutnya adalah memilih beberapa pendekatan. terdapat beberapa pendekatan algoritma *spectral*, beberapa diantara dengan *log normal*, *irrc* (*independent rescaling gene and condition*), dan *bi-stochastization*. Masing-masing pendekatan memiliki karateristik yang berbeda. pendekatan *biclustering* spektral meliputi

normalisasi baris dan kolom sebagai bagian integral dari algoritma. Dimana bekerja secara bersamaan menormalkan kedua gen dan kondisi. Hal ini dapat dilakukan dengan mengulangi prosedur secara iteratif pada skala baris dan kolom sampai didpat nilai konvergensi. Pendekatan tersebut dikenal dengan *bi-stochastization*, dimana menghasilkan matriks *B* persegi panjang yang memiliki struktur ganda, semua baris dijumlahkan ke konstanta dan semua kolom dijumlahkan ke konstanta berbeda. Menurut untuk teorema *Sinkhorn*, matriks *B* kemudian dapat ditulis sebagai persamaaan :

$$B = D_1 A D_2 \tag{3.14}$$

Dimana D_1 dan D_2 adalah matriks diagonal (Bapat dan Raghavan 1997). Secara umum *B* dapat dihitung dengan normalisasi berulang baris dan kolom (dengan matriks normalisasi sebagai R^{-1} dan C^{-1} atau $R^{-½}$ dan $C^{-½}$). D_1 dan D_2 kemudian akan mewakili nilai dari semua normalisasi. Setelah nilai D_1 dan D_2 ditemukan, selanjutnya menerapkan SVD ke matriks *B* tanpa normalisasi lebih lanjut untuk mengungkapkan struktur blok pada matriks. Penyesuaian pada baris dan kolom dari pada setiap baris dan kolom dapat dilakukan secara bersamaan dengan formula. Mendefinisikan $\bar{L}_{i.} = \frac{1}{m} \sum_{j=1}^{m} L_{ij}$ menjadi rata-rata dari baris ke- *i* , $\bar{L}_{.j} = \frac{1}{n} \sum_{j=1}^{n} L_{ij}$ menjadi rata-rata kolom ke- *j* , dan $\bar{L}_{..} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} L_{ij}$ menjadi rata-rata seluruh matriks, hasil penyesuaian ini adalah matriks *interaksi K = (Kij)*, dihitung dengan rumus :

$$Kij = L_{ij} - \bar{L}_{i.} - \bar{L}_{.j} + \bar{L}_{..} \tag{3.15}$$

Jika dataset pada matriks memiliki struktur "kotak-kotak", setidaknya ada sepasang pasangan yang konstan vektor eigen *u* dan *v* yang sesuai dengan nilai eigen yang sama. Vektor eigen yang sesuai dengan nilai eigen nontrivial terbesar akan memberikan partisi optimal dalam algoritma pendekatan spektral untuk pengelompokan (Shi & Malik 1997).

Pada klasifikasi vektor eigen terdapat kemungkinan bukan milik nilai eigen nontrivial terbesar, dan beberapa pemeriksaan vektor eigen yang sesuai dengan nilai eigen terbesar pertama. Partisi vektor eigen biasanya dihubungkan dengan salah satu dari nilai eigen yang terbesar, Untuk mengekstraksi informasi partisi dari sistem eigen, hal yang dilakukan adalah dengan memeriksa semua vektor eigen dengan memasangkannya ke vektor konstan piecewise. Hal ini dilakukan dengan menyortir entri masing-masing vektor eigen, pengujian semua ambang batas yang mungkin, dan memilih vektor eigen dengan partisi yang diperkirakan oleh piecewise vektor konstan (Memilih satu ambang batas mem-partisi entri dalam vektor eigen yang diurutkan menjadi dua himpunan bagian, dua ambang batas menjadi tiga himpunan bagian, dan sebagainya.).

Mempartisi vektor eigen menjadi dua, diperlukan pertimbangan n-1 ambang yang berbeda, untuk mempartisi menjadi tiga, memerlukan pemeriksaan (n-1) (n-2)/2, berbeda ambang batas dan sebagainya. Prosedur tersebut menggunakan penerapan algoritma k-means untuk satu dimensi vektor eigen. Dimana persamaan *k-means* sebagai berikut :

$$d\left(X_i, X_j\right) = \left(\left|x_{i1} - x_{j1}\right|^g + \left|x_{i2} - x_{j2}\right|^g + \ \ldots \ + \left|x_{ip} - x_{jp}\right|^g\right)^{1/g} \quad (3.16)$$

Dimana untuk menentukan titik centroid digunakan fungsi sebagai berikut :

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_p \quad (3.17)$$

Dalam pengelompokan *spectral* hal yang dilakukan adalah melakukan langkah pengelompokan akhir data diproyeksikan ke sejumlah kecil vektor eigen, bukan sekedar mengelompokkan masing-masing vektor eigen secara individual (Shi dan Malik 1997). Langkah pengelompokan akhir dengan menerapkan k-means dan algoritma pemotongan yang dinormalisasi untuk data yang diproyeksikan ke dua atau tiga vektor eigen terbaik. Metode pengelompokan yang dilakukan tidak

hanya menyediakan pembagian ke dalam kelompok, tetapi juga peringkat tingkat keanggotaan gen (dan kondisi) ke m-masing klaster sesuai dengan nilai aktual di mempartisi vektor eigen yang diurutkan. Setiap vektor diurutkan eigen partisi dapat diperkirakan dengan langkah-seperti (piecewise konstan) struktur, tetapi nilai vektor eigen yang diurutkan dalam setiap langkah adalah monoton menurun. Nilai-nilai ini dapat digunakan untuk memberi peringkat atau mewakili transisi bertahap dalam kelompok.