BAB II
TINJAUAN PUSTAKA

2.1 Kanker

Kanker adalah suatu penyakit yang disebabkan oleh pertumbuhan sel-sel jaringan tubuh yang tidak normal. Sel-sel kanker akan berkembang dengan cepat, tidak terkendali, dan terus membelah diri, selanjutnya menyusup ke jaringan disekitarnya (*invasive*) dan terus menyebar melalui jaringan ikat, darah, dan menyerang organ-organ penting serta saraf tulang belakang. Dalam keadaan normal, sel hanya akan membelah diri jika ada perggantian sel-sel yang telah mati dan rusak. Sebaliknya, sel kanker akan membelah terus meskipun tubuh tidak memerlukannya, sehingga akan terjadi penumpukan sel-sel baru. Penumpukan sel tersebut mendesak dan merusak jaringan normal, sehingga mengganggu organ yang ditempatinya [MAN-09].

Umumnya sebelum kanker meluas atau merusak jaringan di sekitarnya, penderita tidak merasakan adanya keluhan atau pun gejala, bila sudah ada keluhan atau gejala biasanya penyakit berada pada taraf stadium lanjut. Awalanya kanker tidak menimbulkan keluhan karena hanya melibatkan beberapa sel. Bila sel kanker bertambah, maka keadaan bergantung pada orang yang terkena. Misalnya pada usus rongga besar, tumor harus mencapai ukuran besar sebelum memicu keluhan. Pada taraf stadium lanjut sel kanker menyebar sampai ke organ vital seperti otak atau paru lalu mengambil nutrisi yang dibutuhkan oleh organ tersebut, akibatnya organ itu rusak dan mati. Penyakit kanker sendiri dapat melemahkan penderitanya, penyakit tersebut serta pengobatannya dapat menurunkan gairah hidup dan kemampuan tubuh untuk melawan penyakit [LUM-09].

Kanker adalah penyakit genetik. Bukan berarti bahwa kanker adalah selalu penyakit keturunan, tetapi karena penyebab transformasi suatu sel menjadi ganas terletak pada materi genetikanya (DNA). Penyebab kanker bervariasi dan tidak dapat diketahui dengan pasti. Pola insiden kanker bervariasi sesuai jenis kelamin, ras, dan letak geografik. Beberapa kanker dapat dipengaruhi faktor genetik keluarga, namun yang sering terjadi karena faktor lingkungan dan gaya hidup.

Promotor kanker, yang disebut karsinogen seperti bahan kimia, virus serta faktor lingkungan dan gaya hidup [STU-02].

Kondisi dan penanganan kanker dapat menimbulkan stres, sehingga tidak saja mempengaruhi kondisi fisik tetapi juga kondisi psikologis pasien. Meskipun reaksi psikologis terhadap diagnosis penyakit dan penanganan kanker sangat beragam dan keadaan serta kemampuan masing-masing penderita tergantung pada banyak faktor, namun ada enam reaksi psikologis yang utama yaitu kecemasan, depresi, perasaaan kehilangan kontrol, gangguan kognitif, atau status mental (*impairment*), gangguan seksual serta penolakan terhadap kenyataan (*denial*). Jay, Elliot & Varni (1986) menyatakan bahwa profil psikologis pasien yang datang pada pemeriksaan medis menunjukan tingginya kecemasan, rasa marah dan keterasingan.

Kanker dapat menyebabkan banyak gejala yang berbeda, bergantung pada lokasinya dan karakter dari keganasan dan apakah ada metastatis (penyebaran). Sebuah diagnosis biasanya membutuhkan pemeriksaan mikroskopik jaringan yang diperoleh dengan *biopsy*. Setelah didiagnosis, pasien kanker biasanya dirawat dengan operasi, kemoterapi dan/atau radiasi. Kebanyakan pasien kanker dapat dirawat dan banyak disembuhkan, terutama bila perawatan dimulai sejak awal. Bila tidak terawat, kebanyakan kanker menyebabkan kematian pada pasien [HAS-09].

2.1.1 Kanker payudara

Ada berbagai jenis kanker yang telah teridentifikasi, salah satunya adalah kanker payudara. Kanker payudara adalah momok menakutkan yang mengintai para wanita. Payudara merupakan salah satu organ yang menjadi indentitas kesempurnaan seorang wanita. Jika organ tersebut terserang kanker maka kesempurnaan seorang wanita menjadi berkurang. Sehingga, seseorang yang terserang kanker payudara akan berusaha mencari pengobatan yang bisa menyembuhkan penyakitnya [MAH-12].

Kanker payudara adalah tumor ganas yang tumbuh di dalam jaringan payudara. Kanker bisa mulai tumbuh di dalam kelenjar susu, saluran susu, jaringan lemak maupun jaringan ikat pada payudara. Kanker payudara tidak

menyerang kulit payudara yang berfungsi sebagai pembungkus [MAR-04]. Peningkatan jumlah sel yang tumbuh secara tidak normal dan tidak terkontrol pada jaringan organ manusia akan menyebabkan benjolan. Benjolan ini seringkali mengindikasikan adanya kanker. Berdasarkan kategorinya kanker dibedakan menjadi dua jenis yakni kanker jinak(*benign*) dan kanker ganas (*malignant*). Kanker jinak hanya tumbuh dan membesar secara lokal dan tidak menyebar ke jaringan organ yang lain. Berbeda dengan kanker ganas yang dapat menyebar ke jaringan organ lain melalui sistem peredaran darah dan limfa.

Kanker payudara adalah salah satu jenis kanker yang terdapat pada wanita dan masih merupakan masalah kesehatan pada wanita, karena selain salah satu penyakit keganasan kedua terbanyak juga sering menyebabkan kematian. Kanker payudara berasal dari *parenchyma* atau dari *stroma mamma.* Penyakit ini oleh *World Health Organization (WHO)* dimasukkan dalam *International of Diseases (ICD).* Di Indonesia, kanker payudara merupakan kanker terbanyak kedua pada wanita setelah kanker leher rahim (*serviks*). Kanker payudara umumnya menyerang wanita yang telah berumur lebih dari 40 tahun. Namun demikian, wanita muda pun bisa terserang kanker ini [WAH-06].

Beberapa faktor resiko pada kanker payudara yang sudah diterima secara luas oleh kalangan "*oncologist*" di dunia adalah sebagai berikut [RAM-02]:

1. Umur lebih dari 30 tahun mempunyai kemungkinan yang lebih besar untuk mendapat kanker payudara dan resiko ini akan bertambah sampai umur 50 tahun dan setelah menopause.
2. Tidak kawin/Nulippara resikonya 2-4 kali lebih tinggi daripada wanita yang kawin dan punya anak.
3. Anak pertama lahir setelah 35 tahun resikonya 2 kali lebih besar.
4. "*Menarche*" kurang dari 12 tahun resikonya 1,7-3,4 kali lebih tinggi daripada wanita dengan "*Menarche*" yang datang pada usia normal atau lebih dari 12 tahun.
5. Menopause datang terlambat lebih dari 55 tahun, resikonya 2,5-5 kali lebih tinggi.
6. Pernah mengalami infeksi, trauma atau operasi tumor jinak payudara, resikonya 3-9 kali lebih besar.

7. Adanya kanker pada payudara kontralateral, resikonya 3-9 kali lebih besar.
8. Pernah mengalami operasi ginekologis-tumor ovarium, resikonya 3-4 kali lebih tinggi.
9. Yang mengalami radiasi di dinding dada resikonya 2-3 kali lebih tinggi.
10. Riwayat keluarga ada yang menderita kanker payudara pada ibu, saudara perempuan ibu, saudara perempuan, adik/kakak, resikonya 2-3 kali lebih tinggi.
11. Kontrasepsi oral pada penderita tumor payudara jinak seperti kelainan fibrokistik yang ganas akan meningkatkan resiko untuk mendapatkan kanker payudara 11 kali lebih tinggi.

Seiring dengan berkembangnya teknologi di dunia medis, maka ditemukan beberapa cara pengobatan kanker payudara. Setiap jenis pengobatan terhadap penyakit ini dapat menimbulkan masalah fisiologis, psikologis dan sosial bagi pasien. Salah satu jenis pengobatan tersebut adalah dengan cara mastektomi. Mastektomi adalah pengobatan kanker payudara dengan cara mengangkat seluruh jaringan payudara. Efek jangka panjang dari mastektomi berpengaruh sangat besar terhadap kualitas hidup karena rasa sakit dan ketidak nyamanan berikutnya.

Perempuan yang telah berjuang melawan kanker payudara dan selamat melalui mastektomi memiliki kekuatan dan semangat untuk bertahan. Kekuatan baru mereka yang mereka dapatkan kemudian diterapkan ke area lain dari kehidupan mereka dan mengakibatkan pengambilan resiko serta kepercayaan diri ketika berhadapan dengan tantangan dan kesulitan meningkat.

Hawari (2004) menyatakan bahwa wanita yang menjalani operasi mastektomi menunjukan ekspresi yang mencerminkan kecemasan dan depresi serta sikap penolakan. Arroyo dan Lopez (2011) yang menemukan bahwa wanita pasca mastektomi akan merasa dirinya tidak menarik, takut akan ditinggalkan dan juga khawatir dengan kesehatannya selanjutnya [MAH-12].

2.2 Data Mining

2.2.1 Pengertian Data Mining

Beberapa pengertian *data mining* dari beberapa pendapat adalah sebagai berikut [HAI-04] :

1. *Data mining* secara sederhana didefinisikan sebagai ekstraksi informasi atau pola yang penting atau menarik dari data yang ada dalam database yang besar sehingga menjadi informasi yang sangat berharga[HAI-04].
2. *Data mining* merupakan proses penemuan yang efisien sebuah pola terbaik yang dapat menghasilkan sesuatu yang bernilai dai suatu koleksi data yang sangat besar[HAI-04].
3. *Data mining* adalah suatu pola yang menguntungkan dalam melakukan *search* pada sebuah database yang terdapat pada sebuah model. Proses ini dilakukan berulang-ulang (iterasi) sehingga didapat satu set pola yang memuaskan yang dapat berfungsi sesuai yang diharapkan[HAI-04].
4. *Data mining* adalah sebuah *class* dari suatu aplikasi *database* yang mencari pola-pola yang tersembunyi di dalam sebuah *group* data yang digunakan untuk memprediksi prilaku yang akan datang[HAI-04].

Berdasarkan beberapa definisi tersebut, maka definisi *data mining* adalah sebuah proses pengumpulan suatu informasi atau pola dari suatu dataset yang berukuran besar, yang kemudian informasi atau pola tersebut dapat digunakan untuk memperbaiki sistem pengambilan keputusan[HAI-04].

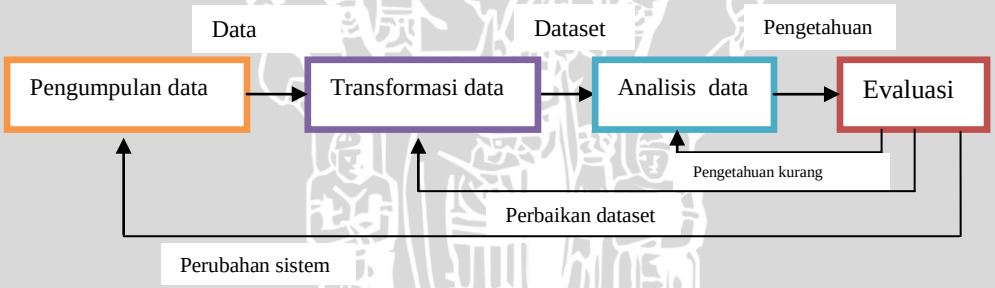
Data mining dapat digunakan untuk menunjang berbagai aplikasi bisnis diantaranya adalah menentukan target pemasaran, *work flow management*, mengatur tata letak toko, dan lainnya. Teknik *data mining* dikembangkan untuk menjelajah database yang besar untuk menemukan pola yang mungkin tidak diketahui. *Data mining* juga menawarkan kemampuan dalam memprediksi *item* yang akan dibeli konsumen ketika berbelanja[WIT-11].

Pada *data mining*, data disimpan secara elektronik dan pencarian dilakukan secara otomatis atau setidaknya ditambah dengan komputer, ekonomi, statistik, peramal, dan insinyur komunikasi telah lama bekerja dengan gagasan bahwa pola dalam data dapat dicari secara otomatis, diidentifikasi, divalidasi, dan

digunakan untuk prediksi. *Data mining* adalah tentang memecahkan masalah dengan menganalisis data yang sudah ada dalam database. *Data mining* didefinisikan sebagai proses menemukan pola dalam data. Proses harus otomatis atau (biasanya) semi-otomatis [WIT-11].

2.2.2 Proses Data Mining

Kerangka proses *data mining* yang akan dibahas tersusun atas tiga tahapan, yaitu pengumpulan data (*data collection*), transformasi data (*data transformation*), dan analisis data (*data analysis*). Proses tersebut diawali dengan *preprocessing* yang terdiri atas pengumpulan data untuk menghasilkan data mentah (*raw data*) yang dibutuhkan oleh *data mining*, yang kemudian dilanjutkan dengan transformasi data untuk mengubah data mentah menjadi format yang dapat diproses oleh *data mining*. Hasil transformasi data akan digunakan oleh anlisis data untuk membangkitkan pengetahuan dengan menggunakan teknik seperti analisis statistik, *machine learning*, dan visualisasi informasi[AYU-07].



Gambar 2.1 Aliran informasi dalam *data mining*

Gambar 2.1 menggambarkan aliran informasi dalam proses *data mining*, yang ditunjukkan sebagai proses *iterative*. Hasil evaluasi pengetahuan yang dihasilkan *data mining* dapat menimbulkan kebutuhan pengetahuan yang lebih lengkap, perbaikan kumpulan data (*dataset*) atau perubahan pada system [AYU-07].

2.2.3 Fungsionalitas Data Mining

Data mining dapat diklasifikasikan berdasarkan fungsi yang dilakukan atau berdasarkan kelas aplikasi yang digunakan [HAR-05].

1. *Association*

   Tipe pola yang penting yang dapat ditemukan dari basis data adalah sebuah aturan. *Association rule* mempunyai bentuk LHS => RHS dengan interpretasi jika setiap item dalam LHS (*Left Hand Side*) diperlukan maka item dalam RHS (*Right Hand Side*) juga diperlukan.

2. Pola *Sequential*

   Sequential Pattern adalah *sequence* dari *itemset* yang dibeli oleh pembeli yang sama.

3. *Bayesian Network*

   Tipe dari *rule* yang dibahas menggambarkan asisiasi dalam basis data dan tidak implisit relasi kausal. *Bayesian network* adalah model grafis yang dapat merepresentasikan relasi kausal.

4. *Classification* dan *Regression*

   Classification dan *Regression rule* adalah *rule* umum yang melibatkan atribut numerik dan kategori.

5. *Decision tree*

   Classification data *regression rule* sering direpresentasikan dalam bentuk *tree.* Sebuah *tree* yang merepresentasikan koleksi *classification rule* disebut *decision tree. Decision tree* dibangun secara *greedy top-down.*

6. *Clustering*

   Clustering bertujuan untuk melakukan partisi sebuah koleksi *record* dalam kelompok yang disebut klaster yang sama dan *record* yang tidak mirip terdapat dalam klaster yang beda. Kemiripan biasanya berdasa pada fungsi jarak.

7. Pencarian *Query* dalam *Sequence*

   Kemiripan *query* dalam *Sequence* artinya adalah dapat melakukan *query* secara tepat yang dan juga mendapatkan hasil yang agak berbeda dari hasil *query* yang tepat. Sebuah *sequence* adalah rangkaian terurut dari angka. Pengukuran perbedaan diantara dua *sequence* dapat dilakukan dengan menghitung jarak *eucledian* diantara *sequence.*

2.3 Klasifikasi

Kusnawi (2007) menjelaskan bahwa *classification* adalah suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah diklasifikasi. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Teknik ini menggunakan *supervised induction,* yang memanfaatkan kumpulan pengujian dari *record* yang terklasifikasi untuk menemukan kelas-kelas tambahan [KUS-07]. Menurut pramudiono (2003), proses klasifikasi biasanya dibagi menjadi dua faset, yaitu [PRA-03]:

1. *Learning*

   Pada fase ini sebagian data yang telah diketahui kelas datanya diumpamakan untuk membentuk model perkiraan.

2. *Test*

   Pada fase ini model yang sudah terbentuk diuji dengan sebagian data lainnya untuk mengetahui akurasi dari model tersebut. Bila akurasinya mencukupi, model ini dapat dipakai untuk prediksi kelas data yang belum diketahui.

Terdapat beberapa metode klasifikasi, antara lain *decision tree, Bayesian, fuzzy, neural network, support vector machine* (SVM) dan *k-nearest neighbor*.

2.4 *Clustering*

Analisa klaster atau disebut juga *clustering* merupakan pengelompokan sekumpulan objek sehingga bisa berada dalam satu kelompok yang sama yang disebut klaster. Obyek-obyek dalam sebuah klaster memiliki tingkat kemiripan yang tinggi. Dan antar klaster memiliki tingkat kemiripan yang rendah. *Clustering* merupakan teknik yang umum digunakan dalam menganalisa data statistik untuk berbagai bidang, misalnya *machine learning*, *pattern analysis* , *image analysis*, information *retrieval* dan bio informatika. Analisa klaster bisa dilakukan dengan beberapa algoritma dengan kelebihan dan kekurangan masing-masing. Sebuah algoritma bisa membentuk klaster-klaster yang detail dan akurat, namun memiliki kekurangan karena memerlukan *resource* komputer yang sangat tinggi dan hal ini berlaku kebalikan untuk jenis algoritma yang lain. Tipe dokumen yang diklaster

juga merupakan salah satu faktor untuk menentukan algoritma *clustering* yang digunakan [BUD-13].

2.5 K-Means

2.5.1 Pengertian K-Means

Data *clustering* merupakan salah satu metode *Data Mining* yang bersifat arahan *(unsupervised).* Ada dua jenis data clustering yang sering digunakan dalam proses pengelompokan data yaitu hirarki data dan non hirarki data *clustering.* K-Means merupakan salah satu metode data *clustering* non hirarki yang mempartisi data yang ada ke dalam satu klaster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam klaster yang lain [AGU-07].

2.5.2 Algoritma K-Means

Algoritma K-Means merupakan sebuah algoritma *clustering* dimana membagi data berdasarkan jarak antara data ke jumlah kelompok yang telah ditetapkan (asalkan ada cukup banyak kasus yang berbeda). Algoritma berbasis jarak ini bergantung pada jarak metrik (fungsi) untuk mengukur kesamaan antara titik data. Untuk menghitung jarak metrik biasa digunakan jarak *eucledian, consine,* atau jarak *fast consine*. Data dimasukkan ke kelompok terdekat sesuai dengan hasil jarak metrik yang digunakan [SAN-07].

Pengelompokan menggunakan K-Means bermaksud untuk mempartisi n obyek ke dalam kelompok k didasari pada jarak yang disebut dengan apriori dan harus dihitung dari data yang ada. Tujuan dari metode ini adalah meminimalkan jumlah varian antar klaster. Dengan fungsi kesalahan kuadrat sebagai berikut :

$$J= \sum_{i=1}^{K} \sum_{n \in Sj} (x_n-\mu_j)^2 \quad \text{.....(2-1)}$$

Dimana k adalah jumlah kelompok S_i, (i = 1,2,….k), μ_j adalah titik *centroid* atau rata-rata semua x_n poin dalam S_i.

Untuk menghitung *centroid* digunakan perhitungan dengan mencari nilai tengah dari kumpulan data dalam sebuah kelompok. Perhitungan ini didasarkan pada rumus *average* :

$$\text{average} = \sum_{i=1}^{n} \frac{x_i}{n} \quad \text{.....(2-2)}$$

Langkah-langkah untuk melakukan pengelompokan dengan tujuan menghasilkan suatu data yang terkelompok adalah sebagai berikut [SAN-07] :

1. Standarisasi data yang akan dikelompokkan (menentukan bobot dari data mentah yang telah didapatkan). Hal ini dilakukan agar data mempunyai skala yang sama, sehingga pengelompokkan akan stabil.

   $$X_{y=} \frac{X_y - XMin_j}{XMax_j - XMin_j}, i=1,2\ldots\ldots,30;\ j=1,2,\ldots\ldots,4 \quad \text{.....(2-3)}$$

   Diamana :

   X_y = Data yang distandarisasi

   $XMin_j$ = Nilai min pada table ke-j

   $XMax_j$ = Nilai max pada table ke-j

2. Melakukan pengelompokan dengan metode K-Means *clustering*.
   1. Pilih jumlah klaster k.
   2. Inisialisasi k pusat klaster ini bisa dilakukan dengan berbagai cara. Yang paling sering dilakukan adalah dengan cara *random*. Pusat-pusat klaster diberi nilai dengan angka-angka *random*.
   3. Tempatkan setiap data/obyek ke klaster terdekat. Kedekatan dua obyek ditentukan berdasarkan jarak kedua obyek tersebut. Demikian juga kedekatan suatu data ke klaster tertentu ditentukan jarak antara data dengan pusat klaster. Dalam tahap ini perlu dihitung jarak tiap data dengan data ke tiap pusat klaster. Jarak paling dekat antara satu data dengan data satu klaster teretentu akan menetukan suatu data masuk dalam klaster mana. Menentukan ukuran kemiripan atau ketidakmiripan antar data dengan metode jarak *Euclidean*. Rumusnya adalah sebagai berikut:

      $$d(x,y) = |x-y|^{2=} \sqrt{\sum_{i=1}^{n} (x_1 - y_1)^2} \quad \text{.....(2-4)}$$

Dimana :

d((x,y) = Ukuran ketidakmiripan.

x = (x1,x2,……xj) adalah variabel data.

y = (y1,y2,……yj) adalah variabel pada titik pusat.

4. Hitung kembali pusat klaster dengan keanggotaan klaster yang sekarang. Pusat klaster adalah rata-rata dari semua data/obyek dalam klaster tertentu. Jika dikehendaki bisa juga memakai median dari klaster tersebut. Jadi rata-rata (*mean*) bukan satu-satunya ukuran yang bisa dipakai.
5. Tugaskan lagi setiap obyek dengan memakai pusat klaster yang baru. Jika pusat klaster sudah tidak berubah lagi, maka proses pengklasteran selesai. Atau, kembali lagi ke langkah nomor 3 sampai pusat klaster tidak berubah lagi.

2.6 Naïve Bayes

2.6.1 Algoritma Naïve Bayes

Naïve Bayes merupakan metode klasifikasi yang dapat memprediksi probabilitas sebuah kelas, sehingga dapat menghasilkan keputusan berdasarkan data pembelajaran. Naïve Bayes adalah algoritma pembelajaran sederhana yang menggunakan aturan bayes dengan asumsi yang kuat bahwa atribut adalah independen secara bersyarat yang diberikan dengan label kelas y. Meskipun asumsi independen ini sering diabaikan dalam praktek, Naïve Bayes tetap memberikan akurasi klasifikasi yang kompetitif. Ditambah dengan efisiensi komputasi dan banyak fitur yang diinginkan lainnya, ini menyebabkan Naïve Bayes banyak diterapkan dalam praktek [SAM-11].

Naïve Bayes menyediakan mekanisme untuk menggunakan informasi dari sata sampel untuk memperkirakan probabilitas posterios P(y|x) dari setiap kelas y terhadap obyek x yang diberikan. Setelah memiliki perkiraan tersebut, dapat digunakan untuk klasifikasi. Aturan bayes dinyatakan pada persamaan (2-5).

$$P(y|x)= \frac{P(y)P(x|y)}{P(x)} \quad \text{……..(2-5)}$$

Dimana :

P(y|x) = Probabilitas hipotesis y jika diberi *evidence* x.

P(x|y) = Probabilitas munculnya *evidence* x jika diketahui hipotesis y.

P(y) = Probabilitas y tanpa mengandung *evidence* apapun

P(x) = Probabilitas *evidence* x.

Asumsi independen bersyarat dapat dinyatakan dalam persamaan (2-6).

$$P(x|y)= \prod_{i=1}^{n} P(x_i|y) \quad \text{.....(2-6)}$$

Dengan x_i adalah nilai ke i dari atribut x, dan n adalah jumlah atribut.

$$P(x)= \prod_{i=1}^{k} P(c_i)P(x|c_i) \quad \text{.....(2-7)}$$

Dengan k adalah jumlah *class* dan c_i adalah *class* ke i. Dengan demikian, persamaan (1) dapat dihitung dengan *numericators* dari sisi kanan persamaan.

Jika atribut kategorikal maka $P(x_i|C_i)$ adalah jumlah tupel kelas C_i di data *training* yang memiliki nikai x_k untuk atribut ke k, dibagi dengan jumlah tupel kelas C_i dalam data *training*[25].

Jika atribut bernilai kontinu, maka perlu melakukan pekerjaan sedikit lebih, tetapi perhitungan ini cukup sederhana. Sebuah atribut bernilai kontinu biasanya diasumsikan memiliki distribusi Gaussian dengan mean dan standar deviasi, ditetapkan pada persamaan (2-8).

$$P(x_i|C_i)= \frac{1}{\sqrt{2\mu\sigma_{ij}}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{.....(2-8)}$$

2.6.2 Metode Klasifikasi Naïve Bayes

Metode klasifikasi Naïve Bayes merupakan salah satu *machine learning* yang dapat digunakan untuk mengklasifikasikan kategori suatu dokumen, teorema bayes berawal dari persamaan (2-9) berikut :

$$P(A|B)=\frac{P(B\cap A)}{B(B)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2-9)}$$

Dimana P(A|B) artinya peluang A jika diketahui B. kemudian dari persaman 2-9 didapatkan persamaan 2-10.

$$P(B\cap A)=P(B|A)P(A) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(2-10)}$$

Sehingga didapatkan teorema bayes seperti persamaan yang ditunjukan pada persamaan 2-11.

$$P(A|B) \ \frac{P(B|A)P(A)}{P(B)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(2-11)}$$

Metode klasifikasi Naïve Bayes atau Naïve Bayes *classifier* termasuk dalam algoritma pembelajaran bayes. Algoritma pembelajaran bayes menghitung probabilitas eksplisit untuk menggambarkan hipotesa yang dicari. Suatu data pada Naïve Bayes *classifier* direpresentasikan dengan konjungsi dari nilai-nilai atribut dari sebuah fungsi target *f(x)* yang dapat memiliki nilai apapun dari himpunan set domain V [DUM-98]. Sistem dilatih menggunakan data latih lengkap berupa pasangan nilai-nilai atribut dari nilai target kemudian sistem akan diberikan sebuah data baru dalam bentuk $<a_1,a_2,a_3,\ldots,a_n>$ dan sistem diberi tugas untuk menebak nilai fungsi target dari data tersebut [MIT-97].

Naïve Bayes *classifier* memberikan nilai target kepada data baru dengan menggunakan nilai V_{MAX}, yaitu nilai kemungkinan yang tertinggi dari keseluruhan anggota himpunan set domain V_{MAX} Hal ini dirumuskan pada persamaan 2-12.

$$V_{map} = \text{argmax}_{v_j \in V} P(v_j|a_1,a_2,a_3,\ldots,a_n) \ \ldots\ldots\ldots\ldots\text{(2-12)}$$

Kemudian teorema bayes digunakan untuk menulis ulang persamaan 2-12 menjadi persamaan baru yaitu persamaan 2-13.

$$V_{map}=\text{argmax}_{v\,j\,\square\,v}\frac{P(a_1,a_2,a_3,\ldots,a_n|v_j)\,P(a_i|v_j)}{P(a_1,a_2,a_3,\ldots,a_n)} \quad \ldots\ldots\ldots\ldots\ldots\ldots(2\text{-}13)$$

Karena $P(a_1,a_2,a_3,\ldots,a_n)$ nilainya konstan untuk semua v_j, sehingga persamaan 2-13 dapat ditulis menjadi persamaan 2-14.

$$V_{map} = \text{argmax}_{v\,j\,\epsilon\,v}\,P(a_1,a_2,a_3,\ldots,a_n|v_j)\,P(a_i|v_j) \quad \ldots\ldots\ldots(2\text{-}14)$$

Tingkat kesulitan menghitung $P(a_1,a_2,a_3,\ldots,a_n|v_j)$ menjadi tinggi karena jumlah term $P(a_1,a_2,a_3,\ldots,a_n|v_j)$ bisa jadi akan sangat besar, hal ini disebabkan jumlah term (kata) tersebut sama dengan jumlah kombinasi posisi kata dikalikan kategori. Naïve Bayes *classifier* menyederhanakan hal ini dengan bekerja dengan dasar asumsi bahwa atirbut-atribut yang digunakan bersifat *conditionally independent* antara satu dan lainnya, dengan kata lain dalam setiap kategori, serta setiap kata *independent* satu sama lainnya. Sehingga pada persamaan (2-15).

$$P(a_1,a_2,a_3,\ldots,a_n|v_j) = \prod_i P(a_i|v_j) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(2\text{-}15)$$

Substitusi persamaan 2-14 dengan 2-15 menjadi persamaan 2-16

$$V_{NB} = \text{argmax}_{v\,j\,\epsilon\,v}\,P(v_j)\prod_i P(a_i|v_j) \quad \ldots\ldots\ldots\ldots\ldots\ldots(2\text{-}16)$$

V_{NB} adalah nilai probabilitas hasil perhitungan Naïve Bayes *classifier* untuk nilai fungsi target yang bersangkutan. Frekuensi kemunculan kata menjadi dasar perhitungan nilai dari $P(v_j)$ dan $P(a_i|v_j)$. Himpunan set dari nilai-nilai probabilitas ini berkorespondensi dengan hipotesa yang ingin dipelajari. Hipotesa kemudian digunakan untuk mengklasifikasikan data-data baru.

$$P(V_j)=\frac{D_j}{D} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2\text{-}17)$$

$$P(a_k|v_j)=\frac{n_f+1}{n+n_j} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2\text{-}18)$$

Dimana :

D_j = Jumlah dokumen yang memiliki kategori j.

D = Jumlah keseluruhan dokumen.

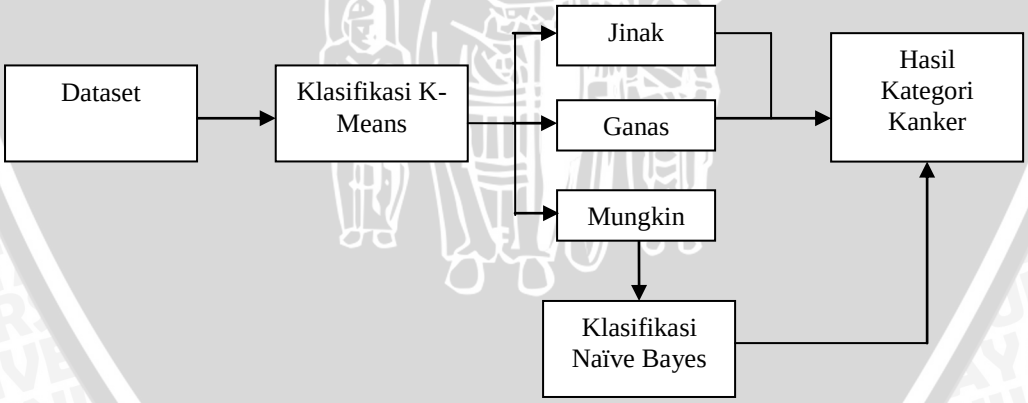
n_f = Nilai frekuensi term/kata.

n = Jumlah keseluruhan term/kata.

n_j = Jumlah keseluruhan term/kata dalam kategori j.

2.7 Gabungan Naïve Bayes dengan K-Means

K-Means Naïve Bayes (KMNB) dibentuk oleh kombinasi teknik pengklasteran dan klasifikasi. Teknik pengklasteran K-Means digunakan untuk mengelompokkan data berdasarkan tiga kategori yaitu ganas, jinak, dan mungkin(belum masuk klaster jinak/ganas). Berikutnya untuk data 'mungkin' yang belum masuk dalam pengelompokan akan diklasifikasikan ke dalam kategori ganas atau jinak menggunakan pengklasifikasian Naïve Bayes. Dalam hal ini, data yang belum masuk klaster jinak/ganas dengan algoritma K-Means akan diklasifikasikan menggunakan algoritma Naïve Bayes.



Gambar 2.2 : Arsitektur sistem

Prosedur yang digunakan dalam pengklasifikasian data dapat dilihat pada langkah-langkah berikut :

1. Masukkan dataset.

2. Kelompokkan data dengan K-Means dan kembali dengan hasil jinak, ganas, dan mungkin (belum masuk klaster jinak/ganas).
3. Jika hasil jinak maka keluaran adalah jinak.
   Jika hasil ganas maka keluaran adalah ganas.
   Jika hasil mungkin maka dilanjutkan ke langkah selanjutnya.
4. Data yang belum masuk klaster jinak/ganas (mungkin) akan diklasifikasikan dengan Naïve Bayes.

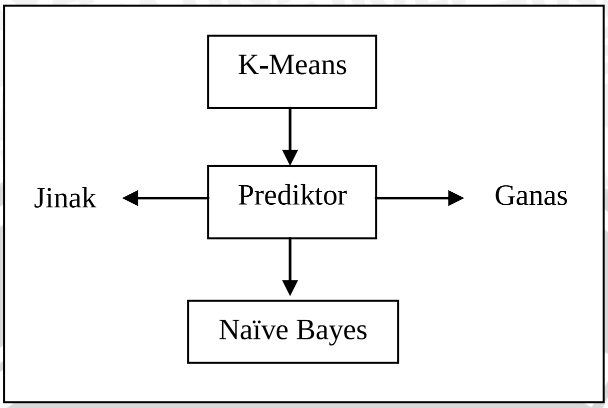
K-Means adalah salah satu metode yang paling sederhana dan populer dari algoritma pembelajaran yang bersifat arahan (*unsupervised*) untuk memecahkan masalah pengelompokan. Dalam statistik dan data mining, K-Means adalah sebuah metode analisis klaster yang bertujuan untuk mempartisi n obyek ke dalam kelompok k didasari pada jarak yang disebut dengan apriori dan harus dihitung dari data yang ada. Deskripsi : Diberikan satu set obyek (x1, x2, ..., xn), di mana setiap obyek adalah vektor sebenarnya d-dimensi. K-Means bertujuan untuk partisi obyek n ke k set ($k \leq n$) S = {S 1, S 2, ..., S k} sehingga dapat meminimalkan kuadara jarak di dalam klaster (*within-cluster sum of squares* (WCSS)):

$$J= \sum_{j-1}^{k} \sum_{i-1}^{x} ||x_i^{(f)} - c_j||^2 \quad \text{.....(2-19)}$$

Algoritma K-Means direpresentasikan dengan langkah-langkah berikut:

1. Tempatkan titik sejumlah k ke ruang yang dinyatakan oleh obyek yang akan diklaster. Titik ini menyatakan inisial *centroid.*
2. Tandai setiap obyek ke grup di *centroid* terdekat.
3. Saat semua obyek telah ditandai, hitung lagi posisi k *centroid.*
4. Ulangi langkah 2 dan 3 sampai dengan *centroid* tidak mengalami perubahan.

Untuk memperjelas algoritma K-Means dan Naïve Bayes dapat dilihat pada gambar 2.3.



Gambar 2.3 : Proses K-Means

Pada gambar 2.3 tersebut, algoritma Naïve Bayes akan diterapkan pada klaster mungkin untuk mendeteksi kanker jinak atau ganas.

Algoritma Naïve Bayes berdasarkan pada sistem penyimpanan data dan mesin pembelajaran. Algoritma ini berguna untuk membuat model yang lebih prediktif dan menjadikan cara baru untuk mengeksplorasi dan mengerti akan data. Teorema bayes, misalkan: A adalah kesimpulan sementara, F adalah bukti/penelitian, E adalah hubungan dari F dan E. Probabilitas ini ditunjukkan dengan P(F|E) :

$$P(E/F)=[P(E/F)P(F)]/[P(E)] \quad \text{.............................(2-20)}$$

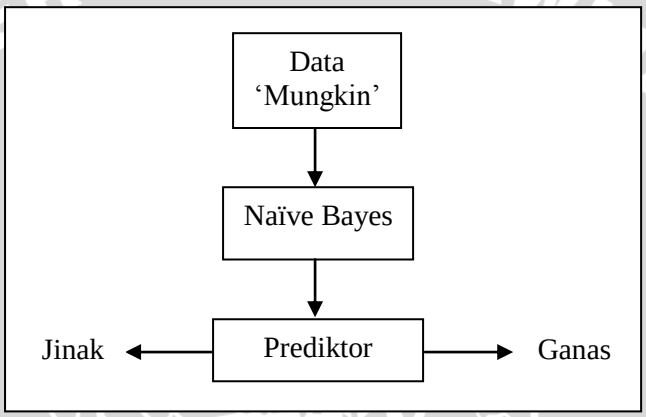
Naïve bayes adalah salah satu dari deteksi tingkat tinggi pada pemahaman teks dokumen. Fitur (a1, a2, a3, a4) yang ada di dataset adalah independen dari yang lainnya. Setiap fitur ($1 \le i \le 4$) *binary* teks akan menunjukan properti dari dataset tersebut. Probabilitas yang ditentukan terdiri dari dua kelas m (m_1 : Jinak dan m_2 : Ganas) seperti dibawah ini :

$$P(m_1/A) = (P(m_1)*P(A/m_i)/P(A)) \quad \text{........................(2-21)}$$

Dalam menyelesaikan prediksi kategori dari sampel A yang belum diketahui, dapat dikalkulasi secara bersamaan dengan nilai dari $P(a_i|m_1)$ dan $P(m_i)$. Sampel A akan masuk ke dalam kategori m_i hanya jinak [MEI-09]:

$$P(m_i|A) > P(m_j|A) \qquad 1 \leq j \leq m, j \neq i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2\text{-}22)$$

Sehingga didapatkan hasil berupa kategori jinak maupun ganas. Untuk memperjelas alur dari penggunaan algoritma Naïve Bayes dapat dilihat pada gambar 2.4.



Gambar 2.4: Proses Naïve Bayes

Kombinasi dari pengklasteran K-Means dan pengklasifikasian Naïve Bayes menunjukkan peningkatan dibandingkan dengan pengklasifikasian tunggal dengan Naïve Bayes, karena ini dapat meningkatkan akurasi, nilai deteksi, dan mengurangi peringatan salah [MAN-14].