BAB III

LANDASAN TEORI

3.1 Statistika Deskriptif

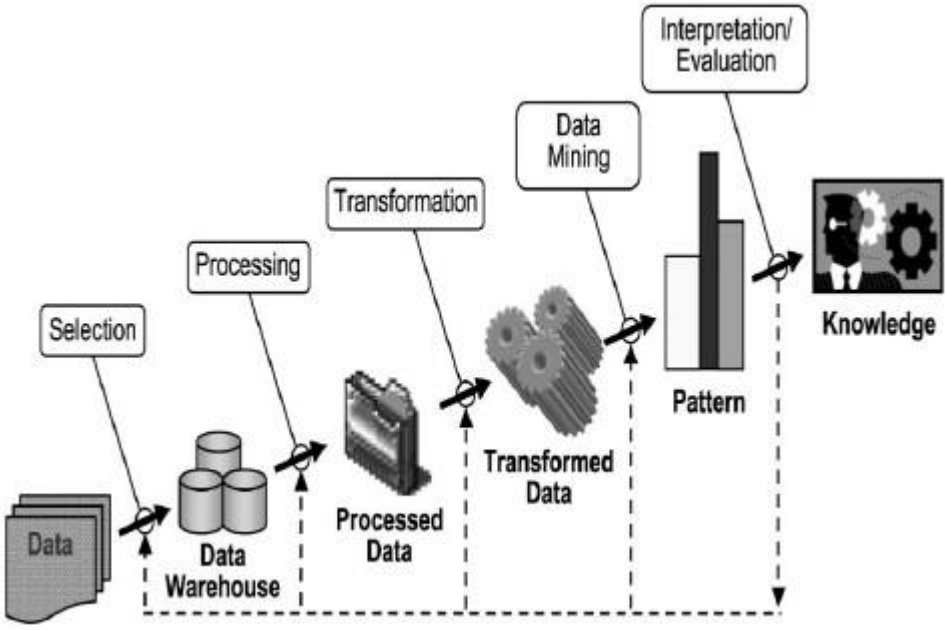
Metode statistik adalah prosedur-prosedur yang digunakan dalam pengumpulan, penyajian, analisis dan penafsiran data. Metode tersebut dibagi menjadi dua, yaitu statistika deskriptif dan statistik inferensial (Walpole dkk, 1995). Statistika deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna (Nugraha, 2013) Pada statistika deskriptif, yang perlu disajikan adalah:

1. Ukuran pemusatan data (*measures of central tendency).* Ukuran pemusatan data yang sering digunakan adalah distribusi frekuensi. Ukuran statistik ini cocok untuk data nominal dan data ordinal (data kategorik). Sementara nilai *mean* adalah ukuran pemusatan data yang cocok untuk data *continuous*. Ukuran deskriptif lain untuk pemusatan data adalah median (nilai tengah) dan modus (nilai yang paling sering muncul).
2. Ukuran penyebaran data (*measures of spread*). Ukuran penyebaran data yang sering digunakan adalah standar deviasi. Ukuran penyebaran data ini cocok digunakan untuk data numerik atau *continuous*. Sementara untuk data kategorik, nilai *range* merupakan ukuran yang cocok.

3.2 Knowledge Discovery in Database (KDD)

Istilah data mining dan *knowledge discovery in database (KDD)* seringkali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Salah satu tahapan

dalam keseluruhan proses KDD adalah data *mining*. Proses KDD secara garis besar dapat dijelaskan sebagai berikut. Fayyad (1996).



Gambar 3.1 Penambangan Data Sebagai Tahapan Dalam Proses KDD

a. *Data Selection*

   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

b. *Pre-Prosescing / Cleaning*

   Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.

c. *Data Transformation*

Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses penambangan data. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

d. Data Mining

Data *mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data *mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

e. Interpretation / Evaluation

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

3.3 Devinisi Data Mining

Data mining adalah proses menemukan korelasi baru yang bermakna, dengan memilih pola dan tren melalui tempat penyimpanan data dalam jumlah besar, menggunakan teknologi pengenalan pola serta statistik dan teknik matematika Larose (2005). Terdapat beberapa definisi lain tentang data *mining*:

a. Data *Mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai data*base* besar. Turban dkk (2005).
b. Menurut Hand, data *mining* adalah analisis pengamatan set data untuk menemukan hubungan yang tidak terduga dan meringkas data dengan cara baru yang sama-sama dimengerti dan berguna untuk pemilik data. Larose (2005).

c. Menurut Evangelos Simoudis, data *mining* adalah bidang *interdisipliner* yang menyatukan teknik dari pembelajaran komputer, pengenalan pola, statistik, data*base*, dan untuk mengatasi masalah ekstraksi informasi dari data*base* yang besar. Larose (2005).

d. Hermawati (2013), data *mining* adalah proses yang mempelerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis.

e. Menurut Syaifullah (2010), data *mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstrasi dan mengedintifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai data*base* besar.

3.4 Teknik-Teknik Data Mining

Adapun teknik teknik data mining sebagai berikut:

a. Association

Association juga disebut sebagai *Market Basket Analysis*. Sebuah problem bisnis yang khas adalah menganalisa tabel transaksi penjualan dan mengindentifikasi produk-produk yang seringkali dibeli bersamaan oleh *customer*, misalnya apabila orang membeli sambal, biasanya juga dia membeli kecap. Kesamaan yang ada dari data pembelian digunakan untuk mengidentifikasi kelompok kesamaan dari produk dan kebiasaan apa yang terjadi guna kepentingan *cross-selling* (Hermawati, 2013).

b. Klasifikasi (*Classification*)

Klasifikasi merupakan proses pembelajaran suatu fungsi tujuan (target) yang memetakan tiap himpunan atribut x sebagai *input* ke satu dari label kelas y yang didefinisikan sebelumnya sebagai *output.* Fungsi target disebut juga model klasifikasi. Beberapa algoritma klasifikasi antara lain pohon keputusan, *nearest neighbor, naïve bayes, neural networks* dan *support vector machines* (Hermawati, 2013).

c. Pengelompokan (*Clustering)*

Analisa *cluster* menemukan kumpulan objek hingga objek-objek dalam satu kelompok sama (atau punya hubungan) dengan yang lain dan berbeda (atau tidak berhubungan) dengan objek-objek dalam kelompok lain. Tujuan dari analisa *cluster* adalah meminimalkan jarak di dalam *cluster* dan memaksimalkan jarak antara *cluster* (Hermawati, 2013).

d. Regresi

Regresi ini biasanya digunakan untuk memprediksi nilai dari suatu variabel kontinyu yang diberikan berdasarkan nilai dari variabel lain, dengan mengasumsikan sebuah model ketergantungan linier atau nonlinier. Teknik ini banyak dipelajari dalam statistika, bidang jaringan syaraf tiruan (*neural network)* (Hermawati, 2013).

3.5 *Frequent Pattern Growth (FP-Growth)*

Frequent Pattern Growth (FP-Growth) adalah salah satu alternatif algoritma yang dapat digunakan untuk menentukan himpunan data yang paling sering muncul (*frequent itemset*) dalam sebuah kumpulan data. Samuel (2008). Penggalian *itemset* yang *frequent* dengan menggunakan algoritma *FP-Growth* akan dilakukan dengan cara membangkitkan struktur data *tree* atau disebut dengan *FP Tree*. Pembuatan *tree* ini dilakukan dengan melakukan *scanning data* dari tabel transaksi seperti pada Gambar 3.2 hanya saja item-item dari tiap transaksi tersebut harus diurutkan kembali berdasarkan jumlah *count*-nya Gambar 3.3

TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Gambar 3.2 Tabel Data Transaksi

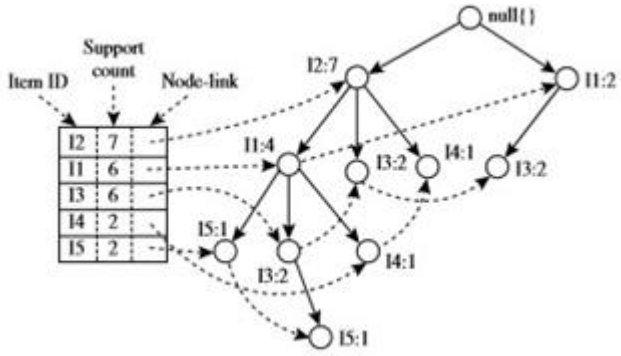
(Sumber : Han et al. 2006)

Itemset	Sup. count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Gambar 3.3 Tabel Daftar Support Count Tiap Item

(Sumber : Han et al. 2006)

Pada TID T100 daftar *item*-nya akan berubah menjadi {12,11,15}, T200 menjadi {12,14}, T300 {12,13}, T400 {12,11,14}, T500 {11,13}, T600 {12,13}, T700 {11,13}, T800 {12,11,13,15}, T900 {12,11,13}, setelah data *list item* tersebut diurutkan, dibuatlah data transaksi tersebut kedalam bentuk *tree* seperti Gambar 3.4.



Gambar 3.4 Pembuatan *FP Tree*

(Sumber : Han et al. 2006)

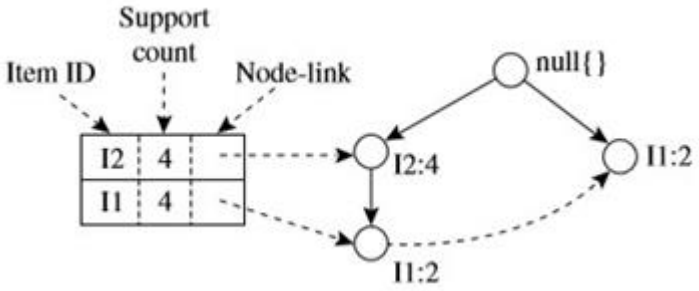
Cara pembuatan *FP Tree* dilakukan dengan cara membaca satu persatu dari transaksi pertama. Misalnya untuk TID T100 daftar *item*-nya adalah {12,11,15}, maka untuk dibuat kedalam *FP Tree* buatlah 3 *node* untuk 12,11 dan 15 beserta *path* sehingga menjadi null ➔ 12 ➔ 11 ➔ 15 dengan *count* untuk 12,11 dan 15 adalah 1. Selanjutnya untuk TID T200 dengan daftar *item* {12,14}, maka dibuat 2 *node* untuk 12 dan 14 beserta *path*-nya *null* ➔ 12 ➔ 14 TID T100 dan T200 memiliki *prefix* yang sama yaitu 12. Maka *count* 12 bertambah 2.

Metode *FP Growth* dapat dibagi menjadi 3 tahapan utama (Han et al. 2006). Ketiga tahapan ini akan dilakukan secara berulang-ulang untuk setiap *item* di *header table* yang diurutkan berdasarkan frekuensinya.

a. Tahapan Pembangkitan *Conditional Pattern Base*

   Conditional pattern base merupakan sub data*base* yang berisi *prefix path* (lintasan *prefix*) dan *suffix pattern* (pola akhiran). Pembangkitan *conditional pattern base* didapatkan melalui *FP Tree* yang telah dibangun sebelumnya. Contoh berikut adalah proses pencarian *conditional pattern base* untuk *item* 13. Keberadaan *node* 13 didalam *tree* dapat dengan mudah ditelusuri dari *link* yang menghubungkan *tree* dan *headertable* sudah dibuat sebelumnya. Setelah menemukan *node* tersebut, maka dapat ditelusuri *node-node* apa saja

yang dilalui dari 13 sampai ke *root. Node-node* yang dilewati tersebut akan menjadi sebuah lintasan. Lintasa-lintasan yang terbentuk untuk *node* 13 adalah {12,11:2}, {1:2} dan {11:2}. Lintasan-lintasan tersebutlah yang akan menjadi *conditional pattern base.*



Gambar 3.5 Sub Database *Node* 13

b. Tahap Pembangkitan *Conditional FP Tree*

Pada tahapan ini *support count* dari setiap item pada setiap *conditional pattern base* dijumlahkan, lalu setiap *item* yang memiliki jumlah *support count* lebih besar sama dengan *minimal support count* yang akan dibangkitkan dengan *conditional FP Tree.* Berdasarkan hasil *conditional pattern base* untu *node* 13 yang telah dijelaskan pada taham sebelumnya, dapat dihitung *support count* dari setiap *item*-nya adalah{12:4, 11:2} dan {11:2}.

c. Tahapan pencarian *frequent itemset*

Pada tahapan ini, apabila *conditional FP Tree* merupakan *single path*, maka akan di dapatkan *frequent itemset* dengan melakukan kombinasi item untuk setiap *conditional FP Tree.* Jika bukan *single path* maka, akan dilakukan pembangkitan FP Growth secara rekursif. Untuk pencarian *frequent itemset* pada *node* 13 akan dilakukan rekursif karena *conditional FP Tree*-nya bukan merupakan *single path* melainkan bercabang. Untuk setiap *single path* akan dikombinasikan dan hasil *frequent pattern*-nya adalah [{12,13:4}, {11,13:4}, {12,11,13:2}].

3.6 Lift Ratio

Salah satu cara yang lebih baik untuk melihat kuat tidaknya aturan asosiasi adalah denan menghitung *lift ratio.* Cara kerja metode ini adalah membagi *confidence* dengan *expected confidence. Confidence* dapat dihitung dngan rumus 3.1. *Anteseden* merupakan sebab yang menjadikan item *konsekuen.* Sedangkan *konsekuen* adalah sebuah akibat atau juga item yang akan dibeli setelah membeli a*nteseden.* Jika didapatkan aturan asosiasi A➔ B maka A sebagai a*nteseden* dan B sebagai *konsekuen.* Nilai dari *expected confidence* dapat dihitung dengan rumus 3.2.

$$Confidence = \frac{\text{jumlah transaksi yang mengandung} anteseden \text{ dan } konsekuen}{\text{jumlah transaksi yang mengandung} anteseden} \ldots . (3.1)$$

$$Expected\ confidence\ = \frac{\text{jumlah transaksi yang mengandung } konsekuen}{\text{jumlah transaksi dalam data base}} \ldots . (3.2)$$

Lift ratio dapat dihitung dengan cara membandingkan antara *confidence* untuk suatu aturan dibagi dengan *expected confidence.* Berikut rumus dari *lift ratio :*

$$Lift\ ratio = \frac{Confidence}{Expected\ confidence} \ldots . (3.3)$$

Nilai *lift ratio* lebih besar dari 1 menunjukan adanya manfaat dari aturan tersebut. Lebih tinggi nilai *lift ratio,* lebih besar kekuatan asosiasinya (Santosa, 2007). Jika nilai *lift ratio* < 1 maka kemunculan A berkorelasi negatif dengan kemunculan B, artinya kemunculan salah satu item mempengaruhi hal yang sebaliknya pada kemunculan item lainnya. Contoh dari korelasi negatif adalah jika penjualan item A naik maka mempengaruhi jumlah penjualan B menjadi menurun. Jika didapatkan *lift ratio* > 1 maka kemunculan A berkorelasi positif dengan kemunculan B, artinya kemunculan A ini berhubungan dengan kemunculan B. Contoh dari korelasi positif adalah jika item A dibeli maka item B juga akan dibeli. Sedangkan jika *lift ratio* = 1 maka kemunculan item A dan B *independent* dan tidak ada korelasi diantara kedua item tersebut (Han et al. 2006).

3.7 Pengertian Kecelakaan Lalu Lintas

Definisi kecelakaan menurut Peraturan Pemerintah Nomor : 43 tahun 1993 pasal 93 tentang Prasarana dan Lalu Lintas Jalan adalah : suatu peristiwa dijalan yang tidak disangka-sangka dan tidak sengaja melibatkan kendaraan dengan atau tanpa pemakai jalan lainnya, mengakibatkan korban manusia atau kerugian harta benda. Korban kecelakaan lalu lintas sebagaimana dimaksud dalam hal ini adalah terbagi menjadi 3 yaitu: korban mati, korban luka berat dan korban luka ringan Pamungkas (2014).

3.8 Jenis dan dampak Kecelakaan Lalu Lintas

Menurut Dirjen Perhubungan Darat, dalam Hakim (2015), Karakteristik kecelakaan menurut jumlah kendaraan yang terlibat digolongkan menjadi:

1. Kecelakaan tungal, yaitu kecelakaan yang hanya melibatkan satu kendaraan bermotor, dan tidak melibatkan pemakai jalan lain, contohnya seperti menabrakan pohon, kendaraan tergelincir, dan terguling akibat ban pecah.
2. Kecelakaan ganda, yaitu kecelakaan yang melibatkan lebih dari satu kendaraan atau kendaraan dengan pejalan kaki yang mengalami kecelakaan di waktu dan tempat yang bersamaan.

Menurut Peraturan Pemerintah No. 43 Tahun 1993, dampak yang ditimbulkan akibat kecelakaan lalu lintas dapat menimpa sakaligus atau hanya beberapa diantaranya. Berikut beberapa kondisi yang digunakan untuk mengklasifikasikan korban kecelakaan lalu lintas, yaitu Pamungkas (2014):

1. Kecelakaan lalu lintas meninggal dunia adalah saat korban kecelakaan dipastikan meninggal dunia sebagai akibat kecelakaan lalu lintas dalam jangka waktu paling lama 30 hari setelah kecelakaan tersebut.
2. Kecelakaan lalu lintas luka berat adalah saat korban kecelakaan yang karena luka-lukanya menderita cacat tetap atau harus dirawat inap di rumah sakit dalam jangka waktu lebih dari 30 hari sejak terjadi kecelakaan.

3. Kecelakaan lalu lintas luka ringan adalah saat korban kecelakaan mengalami luka-luka yang tidak memerlukan rawat inap atau yang harus dirawat di rumah sakit kurang dari 30 hari.