BAB 2

KAJIAN LITERATUR

2.1 TINJAUAN PUSTAKA

Bagian ini berisi landasan teori terkait teori – teori yang digunakan dan pekerjaan yang sudah dilakukan oleh penelitian sebelumnya untuk mendukung penyelesaian penelitian yang akan dilakukan.

2.2 Data Mining

Dengan bertambah banyaknya jumlah data yang ada dalam basis data maka peran analis untuk menganalisa data secara manual perlu digantikan dengan aplikasi yang berbasis komputer yang dapat menganalisa data secara otomatis menggunakan alat yang lebih kompleks dan canggih. Pada dasarnya data mining berhubungan erat dengan analisa data dan pengunaan perangkat lunak untuk mencari pola dan kesamaan dalam sekumpulan data. *Data mining* merupakan prinsip dasar dalam menyusun atau mengurutkan data dalam jumlah yang sangat banyak dan mengambil informasi–informasi yang berkaitan dengan apa yang diperlukan seperti apa yang biasa dilakukan oleh seorang analis (Rully, 2009).

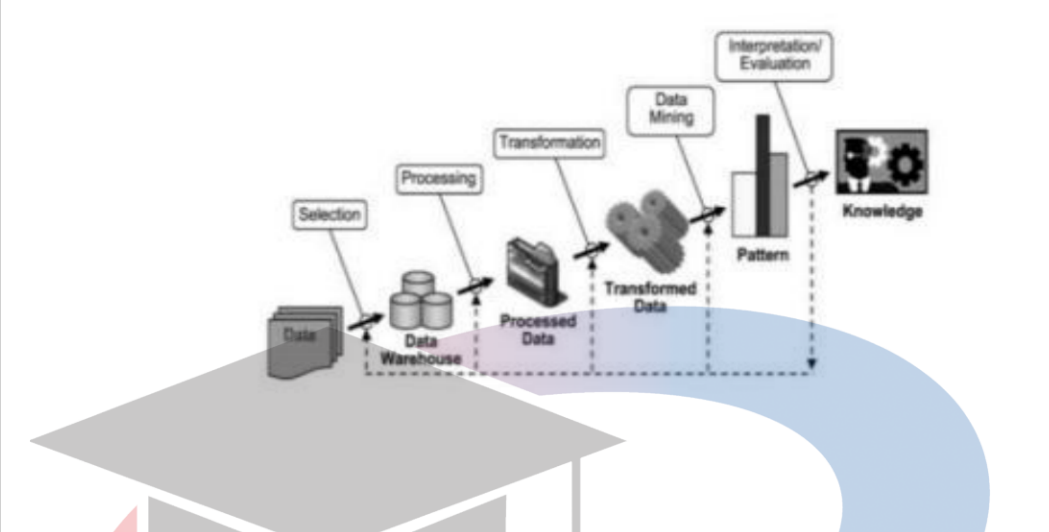
Data Mining adalah ekstraksi informasi yang implisit, sebelumnya tidak diketahui, dan berpotensi berguna dari data. Idenya adalah untuk membangun program komputer yang menyaring melalui database secara otomatis, mencari keteraturan atau pola. Pola yang kuat, jika ditemukan, kemungkinan akan menggeneralisasi untuk membuat prediksi akurat pada data masa depan. Apa pun yang ditemukan akan tidak tepat: Akan ada pengecualian untuk setiap aturan dan kasus yang tidak dicakup oleh aturan apa pun. Algoritma harus cukup kuat untuk mengatasi data yang tidak sempurna dan mengekstrak keteraturan yang tidak tepat tetapi bermanfaat (Ian & Eibe 2005).

Ada beberapa pengertian data mining menurut para ahli Didalam bukunya Kusrini dan Luthfi, Algoritma Data mining, 2009, yaitu :

- Menurut Larose data mining adalah suatu istilah yang digunakan untuk menemukan pengetahuan yan tersembunyi didalam database. Data mining merupakan proses semi otomatik yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi pengetahuan yang terkait dari berbagai database besar.
- Menurut Gartner Group data mining adalah suatu proses menemukan hubungan yang berarti, pola, dan kecendrungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika.

Sebagian ahli berpendapat bahwa data mining adalah langkah analisis terhadap proses penemuan pengetahuan didalam basis data atau *Knowledge discovery in database* yang disingkat KDD (Fayyed et al. 1996). Data mining merupakan inti dari proses Knowledge Discovery Databases (KDD), meliputi dugaan algoritma yang mengeksplor data, membangun model dan menemukan pola yang belum diketahui. KDD merupakan penyelesaian masalah dengan menganalisa data yang ada pada database dengan data tersimpan secara elektronik dan pencariannya dilakukan otomatis seperti pada computer (Vulandari 2016).

Tahapan pada proses KDD pada database menurut Vulandari (2016) seperti Gambar pada 2.1 :



Gambar 2.1 Tahapan Proses KDD

Sumber : Vulandari, 2016

- Data Selection Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan suatu berkas, terpisah dari basis data operasional.
- Pre-processing/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, memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakuakan proses enrichment, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.
- Transformation Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

- 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.
- 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 hipotesa yang ada sebelumnya. Langkah terakhir KDD adalah mempresentasikan pengetahuan dalam bentuk yang mudah dipahami pengguna.

Menurut Turban dalam bukunya yang berjudul ”Decision Support Systems and Intelligent Systems”, data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam basis data. Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar.

Gambar 2.2 Data mining merupakan irisan dari berbagai disiplin

Sumber : Turban et al 2005

Menurut CRISP-DM (proses standar industri lintas untuk Data Mining) bahwa data mining adalah metodologi dan model proses penambangan data yang komprehensif yang memberikan kesempatan bagi siapapun baik pemula atau para ahli untuk melakukan proyek penambangan data. CRISP-DM memecah siklus hidup proyek data mining menjadi enam fase (Larose, 2006). Enam fase CRISP-DM (Cross Industry Standard Process for Data Mining) (Larose, 2006).

1. Fase Pemahaman Bisnis *(Business Understanding Phase)*

- Penentuan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan.
- Menerjemahkan tujuan dan batasan menjadi formula dari permasalahan data mining.
- Menyiapkan strategi awal untuk mencapai tujuan.

2. Fase Pemahaman Data *(Data Understanding Phase)*

- Mengumpulkan data.
- Menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal.
- Mengevaluasi kualitas data.
- Jika diinginkan, pilih sebagian kecil kelompok data yang mungkin mengandung pola dari permasalahan

3. Fase Pengolahan Data *(Data Preparation Phase)*

- Siapkan dari data awal, kumpulan data yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan berat yang perlu dilaksanakan secara intensif.
- Pilih kasus dan variabel yang ingin dianalisis dan yang sesuai analisis yang akan dilakukan.
- Lakukan perubahan pada beberapa variabel jika dibutuhkan.
- Siapkan data awal sehingga siap untuk perangkat pemodelan.

4. Fase Pemodelan *(Modeling Phase)*

- Pilih dan aplikasikan teknik pemodelan yang sesuai.
- Kalibrasi aturan model untuk mengoptimalkan hasil.
- Perlu diperhatikan bahwa beberapa teknik mungkin untuk digunakan pada permasalahan data mining yang sama.
- Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk menjadikan data ke dalam bentuk yang sesuai dengan spesifikasi kebutuhan teknik data mining tertentu.

5. Fase Evaluasi *(Evaluation Phase)*

- Mengevaluasi satu atau lebih model yang digunakan pada fasepemodelan atau Evaluation Pattern
- Menetapkan apakah model tadi sudah sesuai dengan tujuan pada fase awal
- Perlu diperhatikan bahwa beberapa teknik mungkin untuk digunakan pada permasalahan data mining yang sama. Menentukan apakah terdapat permasalahan penting .
- Mengambil keputusan berkaitan dengan penggunaan data mining

6. Fase Penyebaran *(Devolopment Phase)*

- Menggunakan model yang dihasilkan dan yang dipresentasika

2.2.1 Normalisasi

Variabel atribut dengan nilai yang besar memiliki pengaruh yang lebih besar dalam melakukan prediksi klasifikasi daripada variabel dengan nilai yang kecil. Untuk mengatasi masalah tersebut, digunakan teknik normalisasi sehingga semua variabel berada pada jangkauan yang sama dan tidak ada variabel yang memiliki pengaruh dominan terhadap variabel lainnya. Untuk menghitung normalisasi data digunakan rumus:

$$Data_{normalisasi} = \frac{Data\ ke\ i - \min Data}{Maks\ Data - Min\ Data}$$

Dimana Nilai min data adalah nilai minimal dari dataset pada atribut data ke-i, max data adalah nilai maksimum dari dataset pada atribut data .

2.3 Teknik – Teknik Data Mining

Sebagian kalangan kesulitan membedakan data mining dengan machine learning. Hal ini dapat dimaklumi mengingat kedua istilah tersebut memang memiliki tumpang tindih uyang tinggi. Dalam banyak literatur, sebagian ahli mendefinisikan data mining sebagai sains yang menggunakan beberapa teknik. Yang sebagian di pelajari dengan Machine Learning, untuk mengekstrak pola – pola penting dan berguna dari kumpulan-kumpulan data berukuran besar (*big data*). Dengan kata lain pembelajaran mesin (machine learning) merupakan teknik – teknik yang mendukung dan paling banyak digunakan dalam penggalian data (data mining). Sementara itu, machine learning sering kali disebut sebut secara berdampingan dengan istilah deep learning dan artificial Intelligence. Pada awalnya munculnya machine learning ditahun 1980-an para ahli banyak mengembangkan teknik – teknik pembelajaran yang dapat dikelompokan kedalam 3 kategori besar yaitu Supervised Learning dan Unsupervised Learning (Dr.Suyanto 2019).

2.3.1 Supervised Learning

Supervised Learning merupakan teknik pembelajaran mesin yang membuat suatu fungsi berdasarkan data latihan yang sudah ada, dalam hal ini dapat dikatakan untuk teknik ini sudah tersedia data latihan secara detil dan terklasifikasi dengan baik yang akan dijadikan sebuah model data saat dilakukan proses ujicoba dengan data tes yang baru dan menghasilkan hasil keluaran yang sesuai diharapkan sebelumnya berdasarkan data latihan yang ada (Shwartz dan David, 2014).

Supervised Learning adalah tipe *learning* di mana kita mempunyai variable input dan variable output, dan menggunakan satu algoritma atau lebih untuk mempelajari fungsi pemetaan dari input ke output. *Goal*-nya adalah untuk memperkirakan fungsi pemetaannya, sehingga ketika kita mempunya input baru, kita

dapat memprediksi output untuk input tersebut (Pang-Ning Tan, et all 2006). Proses dari sebuah algoritma belajar dari *training dataset* dapat diumpamakan sebagai seorang guru yang mengawasi (*supervising*) proses belajar. Kita tahu jawaban yang benar, dan algoritma secara iteratif membuat prediksi pada data latih (*training data*) dan dikoreksi oleh 'guru'nya. *Learning* berhenti ketika algoritma mencapai level performansi yang diterima. Permasalahan *Supervised Learning* dapat dikelompokkan menjadi masalah regresi (*regression problem*) dan masalah klasifikasi (*classification problems*) (N. J. Nilson, 1996). Model regresi memetakan ruang input menjadi domain nilai riil. Misalnya, seorang regressor dapat memprediksi permintaan untuk suatu hal tertentuproduk yang diberikan karakteristiknya. Di sisi lain, *classifier* memetakan input ruang ke dalam kelas yang telah ditentukan(Minsky &Papert 1961). Misalnya, pengklasifikasi dapat digunakan untuk mengklasifikasikan hipotekkonsumen baik (pengembalian penuh hipotek tepat waktu) dan buruk (pengembalian tertunda). Ada banyak alternatif untuk mewakili pengklasifikasi, misalnya, dukunganmesin vektor, pohon keputusan, ringkasan probabilistik, fungsi aljabar, dll. Seiring dengan regresi dan estimasi probabilitas, klasifikasi adalah salah satu yang paling banyakmempelajari model, mungkin satu dengan relevansi praktis terbesar. Manfaat potensial dari kemajuan dalam klasifikasi sangat besar karena teknik ini memiliki dampak yang besar di area lain, baik di dalam Penambangan Data dan dalam penerapannya. Berikut beberapa proses dari konsep pembelajaran *Supervised Learning* menurut Larose(2005) :

a. Estimasi (*Estimation*)

   Estimasi memiliki makna yang hampir samadengan prediksi, yang menjadi perbedaannya adalah estimasi lebih berorientasi kepada angka atau bilangan daripada mengarah kepada kategori, kelas atau kelompok. Menggunakan record lengkap model dibangun yang menyediakan nilai dari variabel target sebagai nilai dari prediksi tersebut. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.

b. Prediksi(*Prediction)*

Prediksi memiliki makna yang menyerupai klasifikasi, perkiraan apa yang akan terjadi dikemudian hari adalah hasil dari pengklasifikasian data berdasarkan perilaku dan nilai. Contoh dari tugas prediksi misalnya untuk prediksi harga saham yang akan terjadi diminggu atau masa yang akan dating. Teknik yang sering digunakan dalam proses prediksi dapat dikatakan sama dengan teknik yang digunakan dalam proses estimasi.

c. Klasifikasi (*Classification)*

Dalam menemukan suatu model atau fungsi untuk menjelaskan atau mendiskripsikan dan membedakan data kedalam kelas adalah fungsi utama dari klasifikasi. Didalam kelas yang sudah defenisikan sebelumnya pemeriksaan dari objek dan memasukkan objek kedalam kelas tersebut akan melibatkan proses klasifikasi. Teknik yang digunakan dalam klasifikasi yaitu:

1. Decision tree
2. ANN (Artificial Neural, Network)
3. SPV (Support Vector Machine)
4. K-NN (K- Nearest Neighbour)
5. NWK-NN (Neighbour weighted K- Nearest Neighbour)

d. Asosiasi (*Assosiation*)

Asosiasi adalah proses dalam data mining yang terdapat dalam kedua metode baik *Supervised Learning* maupun *Unsupervised Learning*. Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam suatu waktu.Dalam dunia bisnis lebih umum disebut analisis keranjang belanja (market basket analisis).Tugas asosiasi berusaha untuk mengungkap aturan untuk mengukur hubungan antara dua atau lebih atribut.

2.3.1.1 Decision Tree

Decision Tree adalah pohon yang digunakan sebagai prosedur penalaran untuk mendapatkan jawaban dari masalah yang dimasukkan. Pohon yang dibentuk tidak selalu berupa pohon biner. Jika semua fitur dalam data set menggunakan 2 macam

nilai kategorikal maka bentuk pohon yang didapatkan berupa pohon biner. Jika dalam fitur berisi lebih dari 2 macam nilai kategorikal atau menggunakan tipe numeric maka bentuk pohon yang didapatkan biasanya tidak berupa pohon biner. Kefleksibelan membuat metode ini atraktif, khususnya Karen memberikan keuntungan berupa visualisasi sasaran (dalam bentuk decision tree) yang membuat prosedur prediksinya dapat diamati (Gorunescu, 2011). Karakteristik dari decision tree dibentuk sejumlah elemen sebagai berikut (Tan, 2006):

a. Node Akar, tidak mempunyai lengan masukan dan mempunyai nol atau lebih lengan keluaran.
b. Node internal, setiap node yang bukan daun (*nonterminal*) yang mempunyai tepat satu lengan masukan dan dua atau lebih lengan keluaran. Node ini menyatakan pengujian yang didasarkan pada nilai fitur.
c. Lengan, setiap cabang menyatakan nilai hasil pengujian di node bukan daun.
d. Node daun (*terminal*), node yang mempunyai tepat satu lengan masukan dan tidak mempunyai lengan keluaran. Node ini menyatakan label kelas (keputusan)

Ada banyak pilihan algoritma untuk menginduksi decision tree, seperti: Hunt, CART (C&RT), ID3, C4.5, SLIQ, SPRINT, QUEST, DTREG, THAID, CHAID, dan sebagainya.

2.3.1.2 ANN (Artificial Neural Network)

Artificial Neural Network / Jaringan Saraf Tiruan (JST) adalah paradigm pengolahan informasi yang terinspirasi oleh sistem saraf secara biologis, seperti proses informasi pada otak manusia. Elemen kunci dari paradigma ini adalah struktur dari sistem pengolahan informasi yang terdiri dari sejumlah besar elemen pemrosesan yang saling berhubungan (neuron), bekerja serentak untuk menyelesaikan masalah tertentu. Cara kerja JST seperti cara kerja manusia, yaitu belajar melalui contoh. Lapisan-lapisan penyusun JST dibagi menjadi 3, yaitu lapisan input (input layer), lapisan tersembunyi (hidden layer), dan lapisan output (ouput layer) (Sutojo, 2010).

Jaringan neural network digunakan untuk membangun sistem informasi yang digunakan untuk analisis dan asesment risiko banjir yang disebabkan oleh hujan. Sistem informasi ini digunakan sebagian tambahan bagi implementasi teknologi mobile untuk menampilkan peringatan. Teknologi ini digunakan untuk pencegahan dampak kerugian bagi manusia yang disebabkan fenomena alam yang tidak terduga. Level peringatan yang digunakan adalah Normal (Menggunakan warna hijau), Warning (menggunakan warna oranye) dan Alert (menggunakan warna merah atau coklat) (Vivian dkk., 2012).

2.3.1.3 SVM (Support Vector Machine)

Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, dan Vapnik, pertama kali diperkenalkan pada tahun 1992 di Annual Workshop on Computational Learning Theory. Konsep dasar metode SVM sebenarnya merupakan gabungan atau kombinasi dari teori-teori komputasi yang telah ada pada tahun sebelumnya, seperti marginhyperplane (Dyda dan Hart, 1973; Cover, 1965; Vapnik, 1964), kernel diperkenalkan oleh Aronszajn tahun 1950, Lagrange Multiplier yang ditemukan oleh Joseph Louis Lagrange pada tahun 1766, dan demikian juga dengan konsep-konsep pendukung lain.

Menurut Fachrurrazi (2011) SVM merupakan suatu teknik untuk melakukan prediksi, baik prediksi dalam kasus regresi maupun klasifikasi. Teknik SVM digunakan untuk mendapatkan fungsi pemisah (hyperplane) yang optimal untuk memisahkan observasi yang memiliki nilai variabel target yang berbeda (William, 2011). Hyperplane ini dapat berupa line pada two dimension dan dapat berupa flat plane pada multiple dimension. Karakteristik SVM secara umum dirangkum sebagai berikut (Nugroho, 2003):

1. Secara prinsip SVM adalah linear classifier.
2. Pattern recognition dilakukan dengan mentransformasikan data pada ruang input (input space) ke ruang yang berdimensi lebih tinggi (feature space), dan optimisasi dilakukan pada ruang vector yang baru tersebut. Hal ini

membedakan SVM dari solusi pattern recognition pada umumnya, yang melakukan optimisasi parameter pada hasil transformasi yang berdimensi lebih rendah daripada dimensi input space.

3. Menerapkan strategi Structural Risk Minimization (SRM).
4. Prinsip kerja SVM pada dasarnya hanya mampu menangani klasifikasi dua kelas, namun telah dikembangkan untuk klasifikasi lebih dari dua kelas dengan adanya pattern recognition.

2.3.1.4 KNN (K-Nearest Neighbor)

Klasifikasi berbasis Nearest Neighbor (NN) menjadi salah satu metode dalam top sepuluh metode data mining yang paling popular digunakan (Wu dan Kumar, 2009). Metode NN yang murni termasuk dalam klasifikasi yang *lazy learner* karena menunda proses pelatihan (atau bahkan tidak melakukan sama sekali) sampai ada data uji yang ingin diketahui label kelasnya, maka metode baru akan menjalankan algoritmanya. Algoritma Nearest Neighbor melakukan klasifikasi berdasarkan kemiripan suatu data dengan data yang lain (Tan *et al.*, 2005).

Metode K-Nearest Neighbor (K-NN) menjadi salah satu metode berbasis NN yang paling tua dan popular. Nilai K yang digunakan di sini menyatakan jumlah tetangga terdekt yang dilibatkan dalam penentuan prediksi label kelas pada data uji. Dari K tetangga terdekat yang terpilih kemudian dilakukan voting kelas dari K tetangga terdekat tersebut. Kelas dengan jumlah suara tetangga terbanyaklah yang diberikan sebagai label kelas hasil prediksi pada data uji tersebut (Tan *et al*, 2005). Berikut algoritma prediksi dengan K-NN (Tan et al, 2005):

1. Z = (x',y') adalah data uji dengan data x' dan label kelas y' yang belum diketahui
2. C adalah himpunan label kelas data.
3. Hitung jarak d(x',x) jarak di antara data uji z ke setiap vector data latih, simpan dalam D.

4. Pilih Dz ∈ D, yaitu K tetangga terdekat dari z

2.3.2 Unsupervised Learning

Unsupervised Learning merupakan teknik pembelajaran mesin yang berusaha untuk melakukan representasi pola sebuah input yang berasal dari data latihan dan salah satu yang menjadi perbedaan dengan Supervised Learning adalah tidak adanya pengklasifikasian dari input data. Dalam Machine Learning teknik Unsupervised Learning menjadi esensial karena sistem kerja yang diberikan sama dengan cara kerja otak manusia dimana dalam proses pembelajaran tidak ada role model atau informasi dan contoh yang tersedia untuk dijadikan sebagai model dalam melakukan proses ujicoba untuk penyelesaian sebuah masalah dengan data yang baru (Shwartz dan David, 2014).

Berikut beberapa proses dari konsep pembelajaran *Supervised Learning* menurut Larose(2005) :

a. Deskripsi (*Description*)

Deskripsi bertujuan untuk mengidentifikasi pola yang muncul secara berulang pada suatu data dan mengubah pola tersebut menjadi aturan dan kriteria yang dapat mudah dimengerti oleh para ahli pada domain aplikasinya. Mengidentifikasi pola yang terlihat secara berulang pada suatu data dan pola tersebut diubah menjadi ketentuan tertentu dan parameter yang dapat dengan mudah dipahami oleh para ahli sesuai dengan aplikasi yang digunakannya merupakan tujuan dari Deskripsi. Ketentuan yang dihasilkan harus mudah dipahami agar dapat dengan efisien menaikkan nilai tambah tingkat pengetahuan (*knowledge*) pada sistem.Tugas deskriptif pada *data mining* sekali waktu dibutuhkan pada *teknik postprocessing* untuk melakukan verifikasi dan menjelaskan hasil pengolahan data menggunakan *data mining.*

Postprocessing merupakan proses yang digunakan untuk memastikan hanya hasil yang terverifikasi dan berguna yang dapat digunakan oleh pihak yang bersangkutan atau yang diijinkan.

b. Pengelompokan (*Clustering*)

Pengelompokandata tanpa berlandaskan kelompok atau kategori data tertentu kedalam kelas objek yang cocok merupakan tugas dari *clustering.*

Sebuah kluster atau kelompok adalah kumpulan record yang memiliki kesaman atau kesesuaian suatu data dengan yang lainnya dan memiliki ketidaksamaan dengan record dalam kluster lain. Targetnya adalah untuk menciptakan pengelompokan objek yang sesuai atau memiliki kesaman satu sama lain dalam kelompok-kelompok. Semakin besar kesamaan objek dalam suatu cluster dan semakin besar perbedaan tiap cluster maka kualitas analisis cluster semakin baik. Teknik yang digunakan dalam *clustering* seperti :

1. *K-Means*
2. Hierarchical Clustering
3. *Fuzzy C-Means*

2.3.2.1 Clustering *K-Means*

Ada banyak pekerjaan yang dilakukan di bidang segmentasi gambar dengan menggunakan metode yang berbeda. Dan banyak dilakukan berdasarkan berbagai aplikasi segmentasi gambar. Algoritma K-means adalah salah satu dari pengelompokan yang paling sederhana algoritma dan ada banyak metode yang diimplementasikan sejauh ini dengan metode yang berbeda untuk menginisialisasi pusat. Dan banyak peneliti juga berusaha menghasilkan metode baru yang lebih efisien daripada metode yang ada, dan menunjukkan hasil tersegmentasi lebih baik (Nameirakpam Dhanachandra et *al*, 2015).

Pallavi Purohit dan Ritesh Joshi memperkenalkan pendekatan efisien baru terhadap algoritma pengelompokan K-means. Mereka mengusulkan metode baru

untuk menghasilkan pusat cluster dengan mengurangi rata-rata kesalahan

Kuadrat dari cluster akhir tanpa kenaikan besar dalam waktu eksekusi. Ini mengurangi kesalahan kuadrat berarti tanpa mengorbankan waktu eksekusi. Banyak perbandingan telah dilakukan dan dapat disimpulkan bahwa keakuratan lebih untuk dataset padat daripada dataset jarang (Pallavi Purohit dan Ritesh Joshi, 2013)

Dalam penyelesaiannya, algoritma *K-Means*akan menghasilkan titik*centroid* yang dijadikan tujuan dari algoritma *K-Means*. Setelah iterasi *K-Means*berhenti , setiap objek dalam dataset menjadi anggota dari suatu*cluster*. Nilai *cluster* ditentukan dengan mencari seluruh objek untukmenemukan *cluster* dengan jarak terdekat ke objek. Algoritma *K–means*akan mengelompokan item data dalam suatu dataset ke suatu *cluster*berdasarkan jarak terdekat (Bangoria et al., 2013). Nilai *centroid* awal yangdipilih secara acak yang menjadi titik pusat awal, akan dihitung jarakdengan semua data menggunakan rumus Euclidean Distance. Data yangmemiliki jarak pendek terhadap *centroid* akan membuat sebuah *cluster*.Proses ini berkelanjutan sampai tidak terjadi perubahan pada setiap kelompok (Agrawal & Gupta, 2013, Chaturved & Rajavat, 2013, Bhatia &Khurana, 2013). kelompok (Agrawal & Gupta, 2013, Chaturved & Rajavat, 2013, Bhatia & Khurana, 2013).

Berikut ini merupakan alur dari algoritma *K-Means* (Poteras, et al. 2014):

1. Tentukan banyaknya k, dimana k adalah jumlah cluster yang akan dibentuk, Untuk menentukan banyaknya *cluster* k dilakukan dengan beberapa pertimbangan seperti pertimbangan teoritis dan konseptual yang mungkin diusulkan untuk menentukan berapa banyak *cluster*. Pada penelitian ini untuk menentukan k akan menggunakan algoritma *kd tree*
2. Tetapkan titik pusat cluster secara acak atau random, titik pusat cluster sering disebut dengan nama *centroid*.
3. Setelah menentukan *centroid* awal, maka setiap data akan menemukan *centroid* terdekatnya yaitu dengan menghitung jarak setiap data ke masing-

masing *centroid* menggunakan rumus korelasi antar dua objek yaitu *Euclidean Distance*.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (2.1)$$

Keterangan:

D_e= Euclidean Distance

i = banyaknya objek,

(x,y)= merupakan koordinat objek, dan

(s,t)= merupakan koordinat centroid (titik pusat cluster)

4. Kemudian alokasikan masing-masing objek ke dalam cluster berdasarkan jarak minimum (Goyal & Kumar 2014). Tentukan centroid baru dengan persamaan berikut

$$\bar{v}_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj} \quad (2.2)$$

Keterangan

$\bar{v}$ = centroid/rata-rata *cluster* ke-i untuk variabel ke-j

Ni = jumlah data yang menjadi anggota cluster ke-i i,

k= indeks dari *cluster*

j= indeks dari variabel

x_{kj}= nilai data ke-k yang ada di dalam *cluster* tersebut untuk variabel ke-j

5. Kembali ke langkah no 3, 4 dan 5. Jika pada iterasi kedua, anggota cluster tidak ada yang berpindah ke cluster lain maka iterasi berhenti tetapi jika ada anggota cluster yang berpindah ke *cluster*lain maka kembali ke langkah nomor 3, 4 dan 5. Lakukan iterasi berikutnya sampai tidak ada anggota *cluster* yang berpindah ke *cluster* lain (Poteras, et al. 2014).

Algoritma *K-Means* juga memiliki keuntungan yaitu :

1. Dalam implementasi menyelesaikan masalah, algoritma *K-Means* sangat *simple* serta *fleksibel*. Artinya perhitungan komputasinya tidak terlalu rumit dan algoritma ini dapat diimplementasikan pada segala bidang.
2. Algoritma *K-Means* sangat mudah untuk dipahami, terutama dalam implementasi data yang sangat besar serta dapat mengurangi kompleksitas data yang dimiliki (Bangoria et al., 2013)

Kelemahan yang dimiliki oleh algoritma *K-Means* yaitu :

1. Di Algoritma *K-Meansuser* memerlukan angka yang tepat dalam menentukan jumlah *cluster* sebanyak k karena terkadang pusat *cluster*awal dapat berubah sehingga kejadian ini bisa mengakibatkan pengelompokan data menjadi tidak stabil (Joshi & Nalwade, 2013).
2. Algoritma K-Means tidak bisa maksimal dalam menentukan atau menginisialkan nilai *centroid* awalnya, karena pada pengelompokan data dengan algoritma K-Means sangat bergantung pada nilai *centroid*nya (Ahmed & Ashour, 2011)
3. Output dari K-Means tergantung pada nilai – nilai pusat yang dipilihpada *clustering*. Sehingga pada algoritma ini nilai awal titik pusat *cluster* menjadi dasar dalam penentuan *cluster*. Pemilihan *centroidcluster* awal secara acak akan memberikan pengaruh terhadap kinerja *cluster* tersebut (Singh & Kaur, 2013; Sujatha & Sona, 2013)

Beberapa penelitian dilakukan untuk mengatasi kelemahan yang ada padaAlgoritma *K-Means* yaitu:

1. Kaur et al., (2013) mengusulkan perbaikan pada algoritma *K-Means*klasik untuk menghasilkan *cluster* yang lebih akurat. Algoritma yang diusulkan terdiri dari metode berdasarkan pemisahan data, untuk menemukan *centroid* awal sesuai dengan distribusi data. Hasilpenelitian ini menunjukkan bahwa algoritma yang diusulkan menghasilkan *cluster* yang lebih baik dalam waktu perhitungan yang singkat.

2. Kodinariya & Makwana, (2013) mengusulkan beberapa cara untuk menentukan nilai k sebagai jumlah *cluster* yang dibentuk secaradinamis, salah satunya adalah dengan cara metode *elbow*. Penelitian ini menyatakan bahwa metode *elbow* akan menentukan jumlah *cluster*yang sebenarnya pada satu data set. Nilai k akan terus meningkat pada setiap langkahnya dan suatu saat nilai k akan mengalami penurunan dengan nilai yang besar, saat seperti itulah akan terbentuk siku dar semua nilai k yang didapat dan siku tersebut menjadi nilai k yang diinginkan.

2.3.2.2 Hierarchical Clustering

Dalam statistik, pengelompokan berbasis hirarki adalah metode analisis cluster yang berusaha untuk membangun sebuah hiearki cluster. Strategi untuk pengelompokan berbasis hierarki umumnya jatuh ke dalam dua jenis, yaitu aglomeratif dan divisive. Pembahasan di bab ini dibatasi hanya pada agglomerative Hierarchical Clustering (AHC). Aglomeratif merupakan metode pengelompokan berbasis hierarki dengan pendekatan buttom up, yaitu proses pengelompokan dimulai dari masing-masing data sebagai satu buah cluster, kemudian secara rekursif mencari cluster terdekat sebagai pasangan untuk bergabung sebagai satu cluster yang lebih besar. Proses tersebut diulang terus sehingga tampak bergerak ke atas membentuk hierarki. Cara ini membutuhkan suatu parameter kedekatan cluster (*cluster proximity*) (Presetyo, 2014).

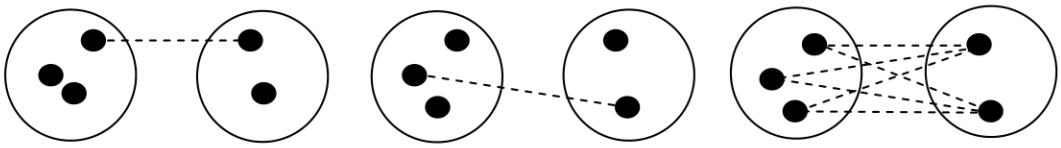
Divisif merupakan metode pengelompokan berbasis hierarki dengan pendekatan top down, yaitu proses pengelompokan dimulai dari satu cluster yang berisi semua data, kemudian secara rekursif memecah cluster menjadi dua cluster sampai setiap cluster hanya berisi satu data tunggal (data itu sendiri). Untuk cara ini, yang dibutuhkan adalah keputusan cluster yang manakah yang akan dipecah pada setiap langkah dan bagaimana cara memecahkannya. Pengelompokan berbasis hierarki sering ditampilkan dalam bentuk grafis menggunakan diagram yang mirip pohon (*tree*) yang disebut dengan *dendrogram*. Dendrogram merupakan diagram yang menampilkan hubungan cluster dan subcluster-nya dalam urutan yang mana

cluster yang digabung (*agglomerative view*) atau dipecah (*divisive view*). Algoritma AHC dijabarkan dalam Algoritma berikut (Prasetyo, 2014):

Algoritma: Agglomerative Hierarchical Clustering

1. Hitung jarak dari semua objek. Nyatakan hasil perhitungan jarak ke dalam matriks jarak.
2. Lakukan pencarian disemua sel matriks jarak untuk menemukan dua cluster/objek yang paling mirip/serupa.
3. Gabungkan dua cluster/objek terdekat berdasarkan parameter kedekatan yang ditentukan untuk menghasilkan sebuah cluster yang memiliki minimal 2 objek.
4. Perbarui matriks jarak dengan menghitung jarak antara cluster baru dan semua cluster yang lain.
5. Ulangi langkah 2 sampai semua objek masuk ke dalam satu cluster.

Pada metode single linkage (MIN), kedekatan di antara dua cluster ditentukan dari jarak terdekat (terkecil) di antara pasangan di antara dua data dari dua cluster berbeda (satu dari cluster pertama satu dari cluster yang lain) atau disebut juga nilai kemiripan yang paling maksimal. Maka, dengan cara ini kita memulainya dari masing-masing data sebagai cluster, kemudian mencari tetangga terdekat dan menggunakan single linkage untuk menggabungkan dua cluster berikutnya sampai semuanya bergabung menjadi satu cluster. Metode ini bagus untuk menangani set data yang bentuk distribusi datanya non-elips (*non-elliptical shapes*), tapi sangat sensitive terhadap noise dan outlier (Prasetyo, 2014).



a. Single linkage b. complete linkage c. average linkage

Gambar 2.2. Kedekatan Agglomerative Hierarchical Clustering (Prasetyo, 2014)

Pada metode complete linkage (MAX), kedekatan di antara dua cluster diterntukan dari jarak terjauh (terbesar) di antara pasangan di antara dua data dari dua cluster berbeda (satu dari cluster pertama satu dari cluster yang lain) atau disebut juga nilai kemiripan yang paling minimal. Maka dengan cara ini kita memulainya dari masing-masing data sebagai cluster, kemudian mencari tetangga terdekat dan menggunakan complete linkage untuk menggabungkan dua cluster berikutnya sampai semuanya bergabung menjadi satu cluster. Metode ini kurang peka terhadap noise dan outlier, tetapi bagus untuk data yang mempunyai distribusi bentuk bulat (Prasetyo, 2014)

Pada metode average linkage (AVERAGE), kedekatan di antara dua cluster ditentukan dari jarak rata-rata di antara pasangan di antara dua data dari dua cluster berbeda (satu dari cluster pertama satu dari cluster yang lain) atau disebut juga nilai rata-rata di antara single linkage dan complete linkage. Maka, dengan cara ini kita memulainya dari masing-masing data sebagai cluster, kemudian mencari tetangga terdekat dan menggunakan *average linkage* untuk menggabungkan dua cluster berikutnya sampai semuanya bergabung menjadi satu cluster. Metode ini merupakan pendekatan yang mengambil pertengahan di antara *single linkage* dan *complete linkage* (Prasetyo, 2014).

2.3.2.3 Fuzzy C-Means

Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981 (Kusumadewi, 2006). *Fuzzy Cluster Means* (FCM) merupakan algoritma yang digunakan untuk melakukan clustering data sesuai berdasarkan keberadaan tiap-tiap titik data sesuai dengan derajat keanggotaannya (Ahmadi dan Hartati, 2013). Algoritma ini merupakan salah satu teknik *soft clustering* yang paling popular dengan menggunakan pendekatan data *point* dimana titik pusat*cluster* akan selalu diperbaharui sesuai dengan nilai keanggotaan dari datayang ada dan selain itu algoritma *Fuzzy C-Means* juga merupakan algoritmayang bekerja dengan

menggunakan model *fuzzy* sehingga memungkinkansemua data dari semua anggota kelompok terbentuk dengan derajat keanggotaan yang berbeda antara 0 dan 1 (Bora dan Gupta, 2014; Sanmorino, 2012). Metode *Fuzzy C-Means* pada dasarnya memiliki tujuan meminimalisasikan fungsi serta mendapatkan pusat *cluster* yang nantinyaakan digunakan untuk mengetahui data yang masuk ke dalam sebuah *cluster* (Simbolon et al., 2013).

Fuzzy C-Means berhubungan dengan konsep kesamaan fungsi objek yang berdekatan dan menemukan titik pusat *cluster* sebagai *prototype*. Untuk beberapa objek data tidak memiliki batasan pada salah satu kelas saja tetapi data tersebut dapat dikelompokan berdasarkan derajat keanggotaan yaitu antara 0 dan 1 yang menunjukkan keanggotaan parsial dari data tersebut (Phukon dan Baruah, 2013). Beberapa contoh dalam penerapan *Fuzzy C-Means* adalah masalah pengelompokan data nyata yang telah dibuktikan dengan menghasilkan karateristik data yang baik (Phukon dan Baruah, 2013). Algoritma ini dimulai dengan menentukan jumlah *cluster* yang diinginkan serta menginisialisasikan nilai keanggotaan yang berisikan semua data kemudian akan dikelompokan berdasarkan *cluster*nya. Pusat pusat *cluster* dihitung dari jarak terdekat ke titik-titik yang memiliki nilai keanggotaan lebih besar. Dengan kata lain, nilai-nilai keanggotaan tersebut akan bertindak sebagai nilai bobot sementara pada suatu *cluster* (K.G dan Patnaik, 2006).

Algoritma *Fuzzy C-Means* memiliki keuntungan yaitu:

1. Dalam implementasi menyelesaikan masalah algoritma *Fuzzy C-Means* dapat memahami karakteristik data yang kabur atau data yang tidak terdefinisikan.
2. Memiliki kemampuan dalam mengelompokan data yang besar
3. Lebih kokoh terhadap data *outlier*/ data dengan karakter yang berbedaatau *value* yang berbeda dalam satu atau beberapa variabel
4. Penentuan titik *cluster* yang optimal (Ali et al., 2008; Suganya dan Shanthi, 2012; Martino dan Sessa, 2009)
5. Dapat melakukan *clustering* lebih dari satu variabel secara sekaligus(Simbolon et al., 2013).

Beberapa kelemahan yang dimiliki oleh algoritma *Fuzzy C-Means* yaitu :

1. Pada algoritma *Fuzzy C-Means* user memerlukan lebih banyak waktuuntuk proses perhitungan komputasinya dalam menentukan *cluster*pada setiap anggota di suatu dataset (Bora dan Gupta, 2014)
2. Masih terpengaruh terhadap cara pembagian data yang sering dipergunakan pada data yang sama dan sangat sensitif terhadap kondisi awal seperti jumlah *cluster* dan titik pusat *cluster* pada pengelompokan data (Lu et al., 2013).

2.4 KLASIFIKASI

Bagian yang sangat penting dalam data mining adalah teknik klasifikasi, yaitu bagaimana mempelajari sekumpulan data sehingga dihasilkan aturan yang bisa mengklasifikasi atau mengenali data – data baru yang belum pernah dipelajari. Klasifikasi dapat didefenisikan sebagai proses untuk menyatakan suatu objek data sebagai salah satu kategori atau kelas yang telah didefenisikan sebelumnya (Zaki *et al* 2013). Klasifikasi adalah salah satu metode yang ada di dalam data mining. Di dalam klasifikasi, label dari setiap kelas sudah ditentukan terlebih dahulu. Proses klasifikasi sendiri merupakan proses untuk menemukan model atau membedakan kelas atau data yang bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. Klasifikasi merupakan bentuk analisis data yang dapat menggambarkan ekstrak model dari suatu data yang penting (Jiawei, et al., 2012).

Klasifikasi adalah proses penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. (Leidiyana Henny, 2013) Algoritma klasifikasi yang banyak digunakan secara luas, yaitu *Decision/ classification* trees, *Bayesian classifiers/ Naïve Bayes classifiers*, *Neural networks*, Analisa Statistik, Algoritma Genetika, *Rough sets*, *k-nearest neighbor*, Metode *Rule Based*, *Memory based reasoning*, dan *Support vector machines* (SVM).

Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah *learning* (fase

training), dimana algoritma klasifikasi dibuat untuk menganalisa data *training* lalu direpresentasikan dalam bentuk *rule* klasifikasi. Proses kedua adalah klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari *rule* klasifikasi. (Leidiyana Henny, 2013)

Proses klasifikasi didasarkan pada empat komponen :

1. Kelas

   Variabel dependen yang berupa kategorikal yang merepresentasikan 'label' yang terdapat pada objek. Contohnya: risiko penyakit jantung, risiko kartu kredit, *customer loyalty*, jenis gempa.

2. *Predictor*

   Variabel independen yang direpresentasikan oleh karakteristik (atribut) data. Contohnya: merokok, minum alkohol, tekanan darah, tabungan, aset, gaji.

3. *Training dataset*

   Satu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk menentukan kelas yang cocok berdasarkan *predictor.*

4. *Testing dataset*

   Berisi data baru yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi dievaluasi

2.5 FEATURE SELECTION

Feature Selection atau seleksi fitur adalah sebuah proses yang biasa digunakan pada Machine Learning dimana sekumpulan dari fitur yang dimiliki oleh data digunakan untuk pembelajaran algoritma. Feature selection telah menjadi bidang penelitian aktif dalam pengenalan pola, statistik, dan Data Mining (Oded Maimon, 2010). Seleksi fitur adalah salah satu faktor yang paling penting yang dapat mempengaruhi tingkat akurasi klasifikasi karena jika dataset berisi sejumlah fitur, dimensi dataset akan menjadi besar hal ini membuat rendahnya nilai akurasi klasifikasi. Masalah dalam

seleksi fitur adalah pengurangan dimensi, dimana awalanya semua atribut diperlukan untuk memperoleh akurasi yang maksimal. Empat alasan utama untuk melakukan pengurangan dimensi (Oded Maimon, 2010):

1. Decreasing the learning cost atau penurunan biaya pembelajaran.
2. Increasing the learning performance atau meningkatkan kinerja pembelajaran.
3. Reducing irrelevant dimensions atau mengurangi dimensi yang tidak relevan.
4. Reducing redundant dimensions atau mengurangi dimensi yang berlebihan.

Ide utama dari Feature Selection adalah memilih subset dari fitur yang ada tanpa transformasi karena tidak semua fitur/atribut relevan dengan masalah. Bahkan beberapa dari fitur atau atribut tersebut mengganggu dan mengurangi akurasi. Noisy Features atau fitur yang tidak terpakai tersebut harus dihapus untuk meningkatkan akurasi. Selain itu dengan fitur atau atribut yang sangat banyak akan memperlambat proses komputasi.

2.5.1 FORWARD SELECTION

Forward Selection merupakan salah satu metode pemodelan (pembangunan model linier) untuk menemukan kombinasi peubah yang "terbaik" dari suatu gugus peubah. Dalam Prosedur *Forward selection*, sekalinya variable masuk kedalam persamaan maka tidak bisa dihilangkan. Selain itu, *forward selection* dapat berarti memasukkan variabel bebas yang memiliki korelasi yang paling erat dengan variabel tak bebasnya (variabel yang paling potensial untuk memiliki hubungan linier dengan Y). kemudian secara bertahap memasukkan variabel bebas yang potensial berikutnya dan nanti akan terhenti sampai tidak ada lagi variabel bebas yang potensial. (Kusuma Widya Intan, 2011)

Forward Selection merupakan pendekatan *wrapper* yang sering digunakan, mulai dengan fitur himpunan kosong dan *storage* dan meningkatkan kecepatan algoritma, menghapus fitur yang tidak relevan, mengembangkan dan menambah

kualitas data, mempercepat waktu *running algoritma learning,* mengembangkan dan menambah kualitas data, serta meningkatkan performa dan akurasi model. (L. Ladha, 2011).

Metode *forward selection* dilakukan dengan cara memasukkan prediktor secara bertahap, prediktor ini berdasarkan korelasi parsial terbesar. Dalam metode *forward selection*, variabel prediktor yang dimasukkan dalam model tidak akan dapat dikeluarkan lagi. Proses tersebut dihentikan ketika prediktor-prediktor baru tidak bisa meningkatkan berpengaruh secara signifikan (sig di bawah 0.05) terhadap variable respon. Karena itulah prosedur *forward selection* menjadi salah satu prosedur pemilihan model terbaik dalam regresi dengan eliminasi variabel bebas yang membangun model secara bertahap. Ada beberapa cara yang dapat digunakan dalam pengujian dengan metode *forward selection* ini (Khrisna 2017).

Forward Selection merupakan menyeleksi variabel berdasrkan koefisien korelasi dan meregresikan variabel-variabel bebas X satu demi satu sampai diperoleh persamaan yang sempurna (khairul Saleh 2010). Forward Selection dimulai dengan pemilihan atribut yang kosong serta dalam setiap putaran ia menambahkan setiap atribut yang tidak terpakai sebagai contoh set. Untuk setiap atribut yang ditambahkan diperkirakan mengugunakan kinerja operator batin, misalnya cross validasi. Yang dimana hanya memberikan atribut yang kinerjanya tinggi untuk ditambahkan ke seleksi. Kemudian memulai babak baru untuk pemilihan dengan memodifikasi (Rian, 2012).

Untuk prosedur Forward Selection dapat di rumuskan sebagai berikut (Fared, 2012) :

- Menentukan model awal $\hat{y} = b0$
  - Memasukan variabel respon dengan setiap variabel berprediktor, misalnya X1, X2, …. Xn yang terkait dengan $\hat{y}$. Misalkan X1 sehingga membentuk model $\hat{y} = b0 + b1X1$.
  - Uji F terhadap peubah pertama yang terpilih. Jika Fhitung< Ftabel maka

peubah terpilih dibuang dan proses dihentikan. Apa bila Fhitung> Ftabelmaka peubah terpilih memiliki pengaruh nyata terhadap peubah terkait y, sehingga layak untuk di perhitungkan di dalam model.

- Masukan peubah bebas terpilih (yang paling signitifikan) ke dalam model. Misalkan x2, sehingga membentuk suatu model $\hat{y}$ = b0+ b1X1+b2X2 e. Uji F, jika Fhitung< Ftabel maka proses dihentikan dan model terbaik adalah model sebelumnya. Namun jika Fhitung ≥ Ftabel, variabel peubah bebas layak untuk dimasukan ke dalam model dan kembali ke langkah c. Proses akan berakhir jika tidak ada lagi peubah yang tersisa yang bisa dimasukan ke dalam model.

Pada penelitian sebelumnya Forward selection ini digunakan untuk menyeleksi setiap fitur yang tidak terpakai saat memulai iterasi fitur (zainuddin Sidik, 2019). Dan dalam penelitian ini juga Berdasarkan hasil pelanggan kartu kredit yang diklasifikasikan akan menggunakan seleksi fitur forward selection untuk menyeleksi satiap atribut yang tidak terpakai dan dapat diharapkan menghasilkan tingkat akurasi yang lebih baik dengan menggunakan algoritma Naighbor Weighted K-Nearest Neighbor.

2.6 Metode Evaluasi Klasifikator

Matrik kinerja klasifikator akan memberikan nilai dari suatu klasifikator. Nilai – nilai tersebut didapatkan dengan menguji set data yang ada. Untuk itu label kelas selama fase pengujian sistem harus diketahuai karena label kelas yang didapatkan dari prediksi harus dibandingkan dengan label kelas yang sebenarnya dimiliki oleh data yang diuijikan tersebut. Pengukuran kinerja model pada set data uji sangat penting untuk dilakukan karena ukuran yang didapatkan memberikan perkiraan tidak bias dari generalisasi eror. Akurasi atau eror yang dihitung dari set data uji juga bisa digunakan untuk membandingkan kinerja relatif dari klasifikator berbeda pada domain yang sama (Eko Prasetyo, 2014). Berikut ini sejumlah metode pengujian yang dapat digunakan untuk mengukur kinerja klasifikator.

2.6.1 Hold – Out

Dalam metode Hold Out, set data (yang sudah diketahui label aslinya) dipecahkan menjadi dbm bagian terpisah, yaitu set data latih dan set data uji. Model kalsifikasi kemudian dibangun berdasarkan set data latih dan kemudian kinerja diukur berdasarkan set data uji. Proses pembagian dari set data latih dan set data uji biasanya diskret misalnya 50/50 (artinya 50% sebagai set data latih dan 50% sebagai set data uji) atau 60/40 (artinya 60% sebagai set data latih dan 40% sebagai set data uji). Akurasi atau metrik yang lain dihitung berdasarkan hasil pengjuian prediksi menggunakan set data uji pada model (seperti Decision tree atau bobot pada ANN) yang dibangun berdasarkan set data latih (Tan *et al* ,2006).

Metode Hold-out merupakan metode yang paling sederhana tetapi mempunyai beberapa keterbatasan. Pertama jumlah data latih lebih sedikit yang tersedia untuk pelatihan karena sebagian harus digunakan untuk pengujian. Akibatnya model dibangun mungkin tidak sebagus ketika semua data digunakan sebagai data latih. Kedua model yang dibangun sangat tergantung pada komposisi pemecahan set data latih dan set data uji. Semakin dikit data latih maka varian dari model semakin besar. Di sisi lain, Jika set data latih terlalu besar maka akurasi yang didapatkan berasal dari set data uji yang lebih sedikit dan itu kuran dipercaya (*reliable*). Yang ketiga set data latih dan set data uji tidak lagi bebas satu sama lain karena keduanya adalah bagian dari set data asli, semnetara kelas yang kelebihan perwakilan di set data latih akan kukurangan perwakilan diset data uji, begitu pula sebaliknya (Eko Prasetyo, 2014).

2.6.2 Random Subsampling

Metode Random subsampling melakukan metode hold-out bebearapa kali (misalnya *k* kali) untuk meningkatkan perkiraan kinerja klasifikator. Metode Random

subsampling masi mengalami masalah seperti pada hold-out kaena tidak menggunakan banyak data untuk pelatihan. Metode ini juga tidak mengontrol berlebihnya jumlah berapa kalinya setiap data untuk menjadi bagian dar pelatiahn dan pengujian. Akibatnya beberapa baris data bisa saja digunakan untuk pelatihan lebih banyak dari pada yang lain (Tan *et al* ,2006).

2.6.3 *Confusion matrix*

Metode ini menggunakan tabel matriks seperti pada Tabel 1 jika data set hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif. (Leidiyana Henny, 2013).

Tabel 1 Confusion Matrix

Classificat Ion	Predicted Class	
	Class=Yes	Class=No
Class=Yes	A(True Positive-TP)	B (False Negative-FN)
Class=No	C (False Positive-FP)	D (True Negative-TN)

True positif adalah jumlah *record positif* yang diklasifikasikan sebagai positif dan *false positives* adalah jumlah record negatif yang diklasifikasikan sebagai positif, *false negatives* adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true negatives* adalah jumlah *record* negatif yang diklasifikasikan sebagai negative, kemudian masukkan data uji. Setelah data uji dimasukkan ke dalam *confusion matrix*, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *sensitivity* (*recall*), *specificity*, *precision* dan *accuracy*. *Sensitivity* digunakan untuk membandingkan jumlah TP terhadap jumlah *record* yang positif sedangkan *specificity* adalah perbandingan jumlah TN terhadap jumlah *record* yang negative (Zainuddin sidik, 2019).

2.6.4 *K-Fold Cross Validation*

K-fold cross-validation secara luas diadopsi sebagai kriteria pemilihan model. Dalam *K-fold cross validation*, lipatan digunakan untuk konstruksi model dan lipatan *hold-out* dialokasikan untuk validasi model. Ini menyiratkan konstruksi model lebih ditekankan daripada prosedur validasi model. Namun, beberapa penelitian telah mengungkapkan bahwa lebih banyak penekanan pada prosedur validasi dapat menghasilkan pemilihan model yang lebih baik. Secara khusus, *leave-m-outcross validation* dengan n sampel dapat mencapai konsistensi pemilihan variabel ketika m / n mendekati ke 1 (Jung, 2017).

K-Fold Cross Validation merupakan teknik validasi yang membagi data ke dalam *k* bagian dan kemudian masing- masing bagian akan dilakukan proses klasifikasi. Dengan menggunakan *K- Fold Cross Validation* akan dilakukan percobaan sebanyak *k*. Tiap percobaan akan menggunakan satu data testing dan *k-1* bagian akan menjadi data training, kemudian data testing itu akan ditukar dengan satu buah data training sehingga untuk tiap percobaan akan didapatkan data testing yang berbeda-beda. (Anto, 2015).

K-Fold Cross Validation merupakan alternatif baik dari pada Random Subsampling. Pada pendekatan ini setiap data digunakan dalam jumlah yang sama untuk pelatihan dan tepat 1 kali untuk pengujian. Ilustrasinya sebagai berikut. Andaikan set data dipecah menjadi dbm bagian dengan ukuran yang sama. Pertama dipilih 1 bagian pelatihan sedangkan lainnya untuk pengujian. Selanjutnya adalah menukar peran, bagian yang tadinya menjadi set data latihsekarang diktukar menjadi set data uji, begitu pula sebaliknya. Pendekatan seperti ini disebut dengan *two-fold cros-validation.* Total eror didapatkan dengan menjumlahkan eror yang didapat dari dua kali proses tersebut. Setiap data berkesempatan satu kali menjadi data ujidan satu kali menjadi data latih (Tan *etal*,2006).

2.7 WEIGHTED K NEAREST NEIGHBOR (WKNN)

Metode WKNN merupakan pengembangan dari metode KNN. Metode ini menggunakan prinsip pembobotan. Bobot akan diberi lebih sedikit ke jumlah k tetangga yang berasal dari kelas mayoritas, dan sebaliknya untuk kelas minoritas. Metode *Weighted* k-*nearest neighbor* (WKNN) mampu melakukan klasifikasi dengan baik, karena metode ini cocok untuk diimplementasikan ke data yang tidak terdistribusi secara rata (Indriati dan Ridok, 2016). Langkah algoritma pada metode *Weighted* k-*nearest neighbor* (WKNN) tidak jauh berbeda dengan langkah algoritma KNN, yang membedakan adalah adanya pembobotan untuk setiap jenis/kelas dan proses perhitungan skor untuk menentukan klasifikasi terhadap data uji (Faldy, 2014). Langkah-langkah dalam algoritma *Weighted* k-*nearest neighbor* (WKNN) menurut D.A Adeniyi et al (2016) adalah sebagai berikut :

1. Menentukan nilai variabel K 2. Menghitung nilai kedekatan ketetanggan antara data uji terhadap data latih menggunakan Persamaan Euclidean Distance atau Cosine Similiraty (CosSim)

2. Menghitung ketetanggaan terdekat yaitu dengan menghitung jarak antara data latih terhadap data uji menggunakan rumus Euclidean Distance dapat digunakan Persamaan 1 :

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^{n} (x_2 - x_1) 2} \quad (1)$$

Keterangan :

X1 = nilai data latih

X2 = nilai data uji

n = jumlah data

i = data ke-i

3. Menghitung ketetanggaan terdekat yaitu dengan menghitung kedekatan menggunakan rumus Cosine Simimilary dapat digunakan Persamaan 2 :

$$(q, d_j) = \frac{\bar{d}_j \cdot \bar{q}}{|\bar{d}_j| \cdot |\bar{q}|} = \frac{\sum_{i=1}^{m}(w_{ij} \cdot w_{ig})}{\sqrt{\sum_{i=1}^{m} w_{ij}{}^2} \cdot \sum_{i=1}^{m} w_{ig}{}^2}$$

Keterangan :

Cosine Similarity (q,di) = nilai similaritas data uji q dengan data latih dj

q = data uji

dj = data latih

$\overrightarrow{d_j} \blacksquare \vec{q}$ = hasil total perkalian vektor antara data latih dengan data uji

$|\overrightarrow{d_j}| \blacksquare |\vec{q}|$ = hasil total perkalian vektor antara normalisasi data latih dengan data uji

wij = bobot nilai i pada data latih j

wiq = bobot nilai i pada data uji q

m = banyaknya jumlah nilai

3. Mengurutkan hasil perhitungan jarak atau kedekatan kedalam kelompok yang mempunyai kedekatan jarak atau similarity.

4. Mengumpulkan kategori klasifikasi nearest neighbor

5. Perhitungan bobot digunakan Persamaan 2.3

$$weight_i = \frac{1}{\left(\frac{Num\left(c_i^d\right)}{Min\left\{Num\left(c_j^d\right)\, j=1,..k*\right\}}\right)^{1/\exp}} \quad (3)$$

Keterangan :

$Num\ (c_i^d)$ = Banyaknya data latih d pada kelas i

$Num\ (c_j^d)$ = Banyaknya data latih d pada kelas j, dimana j terdapat dalam himpunan k tetangga terdekat

Exp = Eksponen (nilai exp lebih dari 1) Setiap data yang telah dihitung nilai bobotnya akan digunakan untuk menghitung nilai skor. Dimana hasil nilai bobot akan dikalikan dengan Persamaan hasil skor.

Rumus hasil skor dihitung dengan Persamaan 2.4 dan 2.5.

$$Skor(X, C_i) = Weight_i \cdot \left(\sum_{djKNN}((\sqrt{\sum_{i=1}^{n}(x_{zi} - x_{ki})^2}\right) \cdot \delta(d_j, c_i)$$

atau

$$Skor(X, C_i) = Weight_i \cdot \left(\sum_{djKNN}((\sqrt{\sum_{i=1}^{n}(x_{zi} - x_{ki})^2}\right) \cdot \delta(d_j, c_i)$$

Keterangan :

Weighti = bobot jenis/kelas i

djNKWNN(x) = data latih dj pada kumpulan tetangga terdekat dari data uji x

$\sqrt{\sum_{i=1}^{n}(x_{2i} - x_{1i})^2}$ = jarak antara data uji dan data latih

$\delta(dj,Ci)$ = akan bernilai 1 jika nilai jarak ∈ Ci dan bernilai 0 jika nilai jarak ∉Ci.

Sim (q,dj) = nilai Cosine Similiraty antara data uji dan data latih

Ci = jenis atau kelas i

Pada penelitian sebelumnya bahwa penilaian kartu kredit adalah metode kuantitatif untuk mengevaluasi risiko kartu kredit dari aplikasi kartu kartu kredit. Metode yang digunakan dalam menganalisis risiko kartu kredit untuk membantu mereka memutuskan apakah pengajuan kartu kredit disetujui atau tidak. Metode ini bertujuan untuk mengklasifikasi prilaku masa depan dalam hal risiko kartu kredit berdasarkan pengalaman masa lalu pengajuan. Dan penelitian ini menggunakan metode *Weighted* k-*nearest neighbor* (WKNN) untuk penilaian kartu kredit dengan mempertimbangkan beberapa atribut dan data set yang tersedia.

2.8 Kartu Kartu kredit

Menurut Suryohadibroto dan Prakoso, kartukartu kredit adalah alat pembayaran sebagaipengganti uang tunai yang sewaktu-waktudapat digunakan konsumen untuk ditukarkandengan produk barang dan jasa yangdiinginkannya pada tempat-tempat yangmenerima kartu kartu kredit (merchant) atau bisadigunakan konsumen untuk menguangkankepada bank penerbit atau jaringannya (cashadvance)(Hermansyah, 2011). Dasar Hukum Kartu Kartu kredit Dasar hukum atas legalisasi pelaksanaan kegiatan kartu kartu kredit di Indonesia adalahsebagai berikut (Fuady, 1999):

Dasar hukum atas legalisasi pelaksanaan kegiatan kartu kartu kredit di Indonesia adalah sebagai berikut (Fuady, 1999):

- Perjanjian antara para pihak disesuaikan dengan sistem hukum di Indonesia yang menganut asas kebebasan berkontrak yang terdapat dalam Pasal 1338 Ayat (1) KUH Perdata.
- Perundang-Undangan yang dengan tegas menyebut dan memberi landasan hukum terhadap penerbitan dan pengoperasian kartu kartu kredit ini, yaitu sebagai berikut : (a) Peraturan Presiden Republik Indonesia Nomor 9 Tahun 2009 tentang Lembaga Pembiayaan; (b) Keputusan Menteri Keuangan Republik Indonesia Nomor 1251/KMK.013/1988 Tentang Ketentuan dan Tata Cara Pelaksanaan Lembaga Pembiayaan, yang telah diubah dengan Keputusan Menteri Keuangan Republik Indonesia Nomor 448/KMK.017/2000.
- Undang-Undang Nomor 7 Tahun 1992 yang telah diubah dengan Undang- Undang Nomor 10 Tahun 1998 tentang Perbankan.
- Peraturan Bank Indonesia Nomor 7/52/PBI/2005 tanggal 28 Desember 2005 yang diperbaharui dengan Peraturan Bank Indonesia Nomor 10/8/PBI/2008 tentang Penyelenggaraan Kegiatan Alat Pembayaran dengan Menggunakan Kartu.

Perjanjian penggunaan kartu kartu kredit adalah perjanjian tiga pihak antara pemegang kartu kartu kredit sebagai pembeli, perusahaan dagang sebagai penjual dan penerbit sebagai pembayar. Perjanjian ini merupakan perjanjian *accesoir* dari perjanjian

penerbitan kartu kartu kredit sebagai perjanjian pokok. Perjanjian ini digolongkan ke dalam perjanjian jual-beli yang diatur dalam Pasal 1457-1518 KUH Perdata tetapi pelaksanaan pembayaran digantungkan pada syarat yang disepakati dalam perjanjian pokok, yaitu perjanjian penerbitan kartu kartu kredit (Sunaryo, 2008).