BAB II
LANDASAN TEORI

2.1. Tinjauan Pustaka

Dalam penelitian ini, penulis mengacu kepada penelitian lain sebagai referensi, salah satu penelitian yang sejenis yang dilakukan oleh Resti Hutami dkk (2016) dengan judul "Implementasi Metode *K-Nearest Neighbor* Untuk Prediksi Penjualan Furniture Pada CV. Octo Agung Jepara". Penelitian ini membahas tentang metode *K-Nearest Neighbor* yang diimplementasikan untuk prediksi penjualan furniture. Penelitian tersebut melakukan prediksi penjualan furniture dengan teknologi data mining untuk menganalisis volume data penjualan. Metode *K-Nearest Neighbor* digunakan karena memiliki akurasi yang tinggi dengan rasio kesalahan yang minim. Hasil dari prediksi tersebut menunjukan bahwa metode *K-Nearest Neighbor* berhasil diimplementasikan untuk menyelesaikan kasus prediksi penjualan dengan tingkat eror sebesar 6% dan akurasi sebesar 94%. (Resti Hutami dkk, 2016).

Penelitian selanjutnya yang dilakukan oleh Agus Panoto dkk (2017) dengan judul "Penerapan *K-Nearest Neighbor* Untuk Prediksi Kelulusan Mahasiswa pada STMIK Sinar Nusantara Surakarta". Penelitian ini membahas tentang penerapan *K-Nearest Neighbor* untuk mengetahui tingkat kelulusan mahasiswa dan memprediksi mahasiswa yang lulus tepat waktu. Hasil dari penelitian menunjukan bahwa prediksi menggunakan metode *K-Nearest Neighbor* (KNN) berhasil diterapkan, didapatkan dari 20 data testing hasilnya 18 benar dan hanya 2 yang salah. Pada penelitian ini prediksi menggunakan metode KNN mempunyai tingkat akurasi sebesar 90%. (Agus Panoto dkk, 2017).

Penelitian selanjutnya adalah yang dilakukan oleh Christian Yonathan Sillueta (2016) dengan judul "Implementasi *Data Mining* Untuk Memprediksi Kelulusan Mahasiswa Dengan Metode Klasifikasi Dan *Algoritma K-Nearest Neighbor* Berbasis *Desktop*". Pada penelitian ini, analisis yang digunakan merupakan analisis prediksi dengan metode Klasifikasi dan Algoritma *K-Nearest Neighbor*. Atribut yang paling penting pada penelitian ini meliputi nilai IPS

mahasiswa 6 (enam) semester pertama. Setelah melakukan pengujian terhadap aplikasi yang sudah dibuat, didapatkan hasil yaitu pada pengujian tunggal memiliki akurasi tertinggi sebesar 70% dengan memiliki rata-rata 61.11%, lalu pada pengujian jamak memiliki akurasi tertinggi sebesar 75.36% dengan memiliki rata-rata 61.88%. (Christian Yonathan Sillueta dkk, 2016)

Penelitian selanjutnya yang dilakukan oleh Mustakim dkk (2016), dengan judul "Algoritma *K-Nearest Neighbor Classification* Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa". Pada penelitian ini, penerapan algoritma KNN dapat dilakukan sebuah prediksi berdasarkan kedekatan dari histori data lama (*training*) dengan data baru (*testing*). Penentuan atribut ini berdasarkan hasil penelitian terdahulu yang memiliki kesamaan dalam kasus prediksi mahasiswa yang selanjutnya divalidasi oleh bagian Akademik Fakultas Sains dan Teknologi. Proses prediksi dilakukan terhadap mahasiswa Program Studi Sistem Informasi angkatan 2014/2015 sebagai data *testing* dengan jumlah 50 data, serta berdasarkan dari data angkatan 2012/2013 sebagai data *training* dengan jumlah 165 data yang menghasilkan pengujian akurasi sebesar 82%. Hasil dari perhitungan algoritma KNN diimplememetasikan terhadap sebuah *Early Warning System* (EWS). *Output* dari sistem yang dibangun dapat dijadikan sebagai acuan bagi Mahasiswa untuk meningkatkan prestasi dan predikat perkuliahan dimasa yang akan datang. (Mustakim dkk, 2016)

Penelitian selanjutnya yang dilakukan oleh Andi Gita Novianti dkk (2017), dengan judul "Penerapan Algoritma *K-Nearest Neighbor* (KNN) Untuk Prediksi Waktu Kelulusan Mahasiswa". Pada penelitian ini menerapkan Algoritma *K-Nearest Neighbor* (KNN) dan Fungsi *Similarity* untuk menghitung kemiripan data dalam sebuah perangkat lunak yang dapat memberikan prediksi waktu kelulusan mahasiswa. Hasil pengujian menggunakan aplikasi prediksi waktu kelulusan dengan 7 (tujuh) kriteria yaitu IPS1-IPS4, jumlah SKS lulus sampai semester 4, jurusan SLTA, program studi, asal suku, penghasilan orang tua dan jenis kelamin di dapat akurasi untuk Program Studi Teknik Informatika S1 sebesar 84% sedangkan Program Studi Sistem Informasi S1 sebesar 87% (Andi Gita Novianti dkk, 2017) .

Penelitian selanjutnya yang dilakukan oleh Hendri Risman dkk (2015) dengan judul "Penerapan Metode *K-Nearest Neighbor* Pada Aplikasi Penentu Penerima Beasiswa Mahasiswa di STMIK Sinar Nusantara Surakarta". Pada penelitian ini membahas tentang penentuan calon penerima beasiswa dengan menggunakan metode *K-Nearest Neighbor*. Dari pengujian yang dilakukan terhadap 22 data sampel yang dijadikan acuan dalam perhitungan *K-Nearest Neighbor* dalam menghasilkan keputusan diperoleh nilai keakuratan sebesar 90,90%. (Hendri Risman dkk, 2015)

Penelitian selanjutnya yang dilakukan oleh Ferry Hermawan dkk (2017), dengan judul "Implementasi Metode *K-Nearest Neighbor* Pada Aplikasi Data Penjualan PT. Multitek Mitra Sejati". Pada penelitian ini membahas tentang memprediksi penjualan berdasarkan ketegori barang. Hasil dari penelitian ini adalah *K-Nearest Neighbor* dapat memprediksi penjualan di tahun 2015 berdasarkan data penjualan barang dari tahun 2012-2014 dengan menggunakan *Euclidean Distance*, dengan tingkat keberhasilan metode 58,33% pada nilai toleransi *error* 10% dan rata-rata keakuratan prediksi 88,54% yang tergolong memiliki kinerja bagus dan memprediksi penjualan berdasarkan kategori barang dengan tingkat keberhasilan algoritma 70% pada nilai toleransi *error* 10% dan rata-rata keakuratan prediksi 85,91% yang tergolong memiliki kinerja bagus. (Ferry Hermawan dkk, 2017)

Berdasarkan penjelasan sebelumnya tentang perbedaan dari beberapa penelitian yang telah dilaksanakan sebalumnya, maka perbedaan yang dimiliki dari penelitian ini adalah Penerapan Data Mining Untuk Prediksi Pendapatan Parkir di Kabupaten Karanganyar Menggunakan Metode *K-Nearest Neighbor*. Aplikasi *data mining* yang dipakai adalah *Rapidminer* dan hasil yang diharapkan dari penelitian ini yaitu untuk memprediksi pendapatan retribusi parkir.

2.2. Data

Data merupakan komponen dasar dari informasi yang akan diproses lebih lanjut untuk menghasilkan informasi. Sedangkan, menurut Longkutoy dalam bukunya "Pengenalan komputer", Data adalah suatu istilah majemuk yang berarti

fakta atau bagian dari fakta yang mengandung arti yang digubungkan dengan kenyataan, simbol-simbol, gambar- gambar, angka-angka, huruf-huruf, atau simbol-simbol yang menunjukkan suatu ide, objek, kondisi, atau situasi dan lain-lain. (Al-bahra bin Ladjamudin, 2005)

Data adalah dapat berupa angka-angka, huruf-huruf, gambar-gambar atau simbol-simbol apapun yang dapat dimasukan (*input*) ke komputer dan dikeluarkan (*output*) dari komputer, karena komputer itu benda mati yang tidak memiliki kemampuan apapun termasuk kemampuan untuk mengenali mana huruf, angka, data dan informasi. (Wahyudi, 2008).

Dari beberapa defenisi Data dari para ahli dapat disimpulkan bahwa Data adalah suatu fakta yang bisa berupa simbol, gambar, angka, huruf dan lain-lain yang dapat diproses lebih lanjut guna menghasilkan informasi.

2.3. Basis Data

Basis data adalah kumpulan data yang terorganisir, relasi antar data, dan objektifnya. (Fathansyah, 2004)

Basis data adalah kumpulan data yang saling berhubungan secara logis dan di desain untuk mendapatkan data yang dibutuhkan oleh suatu organisasi. (Indrajani, 2015)

Basis data (*database*) adalah sistem terkomputerisasi yang tujuan utamanya adalah memelihara data yang sudah diolah atau informasi dan membuat informasi tersedia saat dibutuhkan. Pada intinya basis data adalah media untuk menyimpan data agar dapat diakses dengan mudah dapn cepat. (Sukamto & Shalahuddin, 2013)

Berdasarkan definisi basis data menurut para ahli maka dapat dirangkum definisi basis data adalah kumpulan data-data yang berada pada sebuah media penyimpanan data yang saling terhubung dan berguna bagi pemakai ataupun organisasi.

2.4. Data Mining

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

Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. *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 database besar. (Turban dkk, 2005)

Berdasarkan definisi-definisi di atas tentang *Data mining* dapat disimpulkan bahwa *data mining* adalah sebuah proses pencarian secara otomatis untuk menemukan pola atau model dari suatu database yang besar.

2.4.1. Operasi *Data Mining*

Operasi *data mining* menurut sifatnya dibedakan menjadi 2, yaitu bersifat (1) prediksi (*prediction driven*) untuk menjawab pertanyaan apa dan sesuatu yang bersifat abstrak atau transparan. Operasi prediksi digunakan untuk validasi *hipotesis*, *querying* dan pelaporan. (2) penemuan (*discovery driven*) bersifat transparan dan untuk menjawab pertanyaan ”mengapa?”. Operasi penemuan digunakan untuk analisis data eksplorasi, pemodelan prediktif, segmentasi *database*, analisis keterkaitan (*link analysis*) dan deteksi deviasi. (Hermawati, 2013)

2.4.2. Teknik *Data Mining*

Beberapa teknik dan sifat *data mining* adalah sebagai berikut :

1. Klasterisasi. Adalah mempartisi *data-set* menjadi beberapa *sub-net* atau kelompok sedemikian rupa sehingga elemen-elemen dari suatu kelompok tertentu memiliki *set property* yang di *share* bersama, dengan tingkat similaritas yang tinggi dalam suatu kelompok yang rendah. Disebut juga dengan “*unsupervised learning*”.
2. Regresi. Adalah memprediksi nilai dari suatu variabel kontinyu yang diberikan berdasarkan nilai adari variabel yang lain, dengan mengasumsikan sebuah model ketergantungan linier atau nonlinier.
3. Klasifikasi. Adalah menentukan sebuah *record* data baru ke salah satu dari beberapa kategori (kelas) yang telah didefinisikan sebelumnya dan disebut juga

dengan "*supervised learning*".

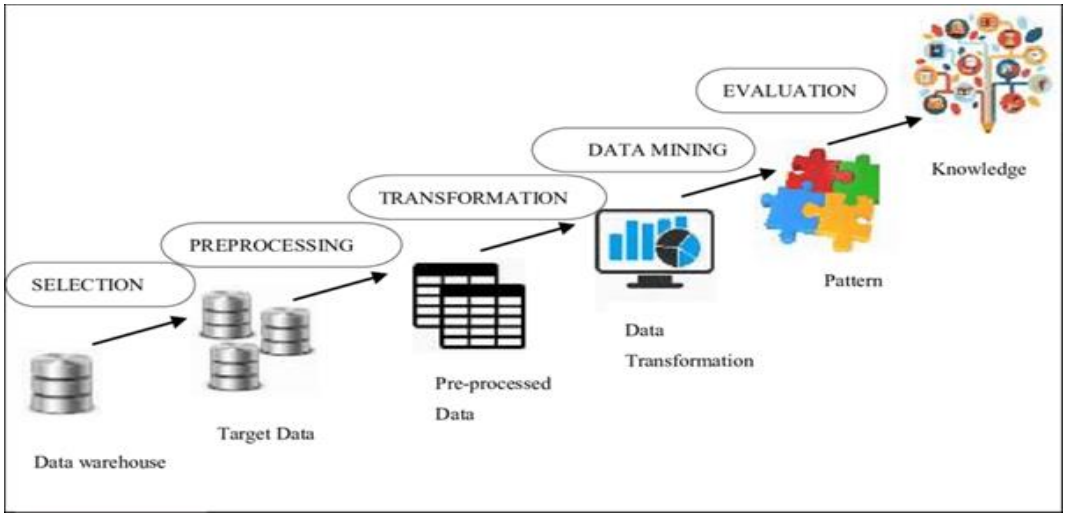
4. Kaidah Asosiasi (*association rule*). Adalah mendeteksi kumpulan atribut- atribut yang muncul bersamaan (*co-occur*) dalam frekuensi yang sering dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut (Hermawati, 2013).

2.5. Prediksi/forcasting

Prediksi/*forcasting* adalah menetukan jumlah kebutuhan bulan mendatang terkait dengan dukungan data historis (*historical data*) atau serangkaian waktu/periode yang dianalisis sehingga dapat diperhitungkan untuk memprediksi jumlah kebutuhan pada bulan mendatang. Prediksi juga dapat digunakan dalam pengklasifikasian, tidak hanya untuk memprediksi *time series*, karena sifatnya yang bisa menghasilkan *class* berdasarkan atribut yang ada.

2.6. Knowledge Discovery in Database

Knowledge Discovery in Database (KDD) adalah proses menentukan informasi yang berguna serta pola-pola yang ada dalam data. Informasi ini terkandung dalam basis data yang berukuran besar yang sebelumnya tidak diketahui dan potensial bermanfaat. *Data Mining* merupakan salah satu langkah dari serangkaian proses *iterative* KDD (Kusrini dkk, 2009). Berikut tahapan proses KDD dapat dilihat pada gambar 2.1.



Gambar 2.1. Tahapan dalam KDD

Tahapan proses KDD terdiri dari:

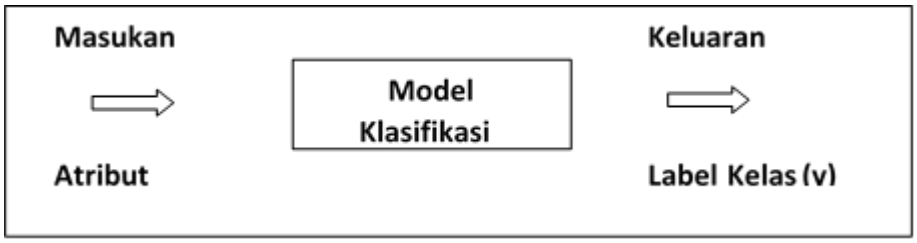
1. Data *Selection*. Pada proses ini dilakukan pemilihan himpunan data, menciptakan himpunan data target, atau memfokuskan pada subset variable (sampel data) dimana penemuan (*discovery*) akan dilakukan. Hasil seleksi disimpan dalam suatu berkas yang terpisah dari basis data operasional.
2. Pre-*Processing* dan *Cleaning* Data. Pre-*Processing* dan *Cleaning* Data dilakukan membuang data yang tidak konsisten dan *noise*, duplikasi data, memperbaiki kesalahan data, dan bisa diperkaya dengan data *eksternal* yang relevan.
3. *Transformation*. Proses ini mentransformasikan atau menggabungkan data ke dalam yang lebih tepat untuk melakukan proses *mining* dengan cara melakukan peringkasan (agregasi).
4. *Data Mining*. Proses *Data Mining* yaitu proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik, metode atau algoritma tertentu sesuai dengan tujuan dari proses KDD secara keseluruhan.
5. *Interpretation*/Evaluasi. Proses untuk menerjemahkan pola-pola yang dihasilkan dari *Data Mining*. Mengevaluasi (menguji) apakah pola atau informasi yang ditemukan bersesuaian atau bertentangan dengan fakta atau hipotesa sebelumnya. Pengetahuan yang diperoleh dari pola-pola yang terbentuk dipresentasikan dalam bentuk visualisasi.

2.7. Klasifikasi

Klasifikasi merupakan proses pemberlajaran suatu fungsi tujuan (*target*) f yang memetakan tiap himpunan atribut x ke satu dari label kelas yang didefinisikan sebelumnya. Fungsi target disebut juga model klasifikasi. (Hermawati, 2013)

Klasifikasi adalah sebuah proses untuk menemukan model yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui. Di dalam klasifikasi diberikan sejumlah *record* yang dinamakan *training set*, yang terdiri dari beberapa atribut, atribut dapat berupa kontinyu ataupun kategoris, salah satu atribut menunjukkan kelas untuk *record*. (Tan, et all., 2004)

Berikut adalah konsep klasifikasi seperti yang ditunjukkan pada Gambar 2.2



(Sumber: Tan *et all*, 2004)

Gambar 2.2. Konsep Klasifikasi

Ada dua jenis model klasifikasi, yaitu:

1. Pemodelan deskriptif (*descriptive modelling*), yaitu model klasifikasi yang dapat berfungsi sebagai suatu alat penjelasan untuk membedakan objek-objek dalam kelas-kelas yang berbeda.
2. Pemodelan prediktif (*predictive modelling*), yaitu klasifikasi yang dapat digunakan untuk memprediksi label kelas *record* yang tidak ketahui.

2.8. Konsep K-Nearest Neighbor

K-Nearest Neighbor (KNN) menjadi salah satu metode dalam top 10 metode data mining yang paling popular . Metode KNN murni termasuk dalam klasifikasi yang *lazy learner* karena menunda proses pelatihan (atau bahkan tidak melakukan pelatihan sama sekali) sampai ada data uji yang ingin diketahui label kelasnya, maka metode baru akan menjalankan algoritmanya. Algoritma KNN melakukan klasifikasi berdasarkan kemiripan suatu data dengan data yang lain (Tan, et all., 2004).

2.9. K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* (KNN) merupakan sebuah metode untuk melakukan klasifikasi terhadap obyek baru berdasarkan (K) tetangga terdekatnya (Gorunescu, 2011). KNN termasuk algoritma *supervised learning*, yang mana hasi dari *query instance* baru, diklasifikasikan berdasarkan mayoritas dari kategori pada

KNN. Kelas yang paling banyak muncul, yang akan menjadi kelas hasil klasifikasi (Gorunescu, 2011).

Pada algoritma KNN terdapat 5 (lima) cara, untuk mencari tetangga terdekat yaitu:

1. Jarak *Euclidean*
2. Jarak *Manhattan*
3. Jarak *Cosine*
4. Jarak *Correlation*
5. Jarak *Hamming*

Pada penelitian ini penulis menggunakan jarak *Euclidean*, maka rumus perhitungan jarak dengan *Euclidean* seperti di bawah ini :

$$\sqrt{\sum_{i=1}^{K}(Xi - Yi)^2}$$

Gambar 2.3. Rumus Perhitungan Jarak Euclidean

Nilai Xi merupakan nilai yang ada pada data *training*, sedangkan nilai Yi merupakan nilai yang ada pada data *testing* . Nilai K merupakan dimensi atribut.

Langkah-langkah untuk menghitung algoritma K-NN:

1. Menentukan nilai k.
2. Menghitung kuadrat jarak *euclid* (*query instance*) masing-masing objek terhadap *training data* yang diberikan.
3. Kemudian mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak *euclid* terkecil.
4. Mengumpulkan label *class* Y (klasifikasi *Nearest Neighbor*).
5. Dengan menggunakan kategori *Nearest Neighbor* yang paling mayoritas maka dapat dipredeksikan nilai *query instance* yang telah dihitung.

2.10. Rapidminer

Rapidminer (YALE) adalah perangkat lunak open source untuk *knowledge discovery* dan *data mining*. *Rapidminer* memiliki kurang lebih 400 prosedur

(operator) data mining termasuk operator untuk masukan, *output*, data *preprocessing* dan visualisasi (Retno Tri Wulandari, 2017).

Beberapa fitur dari rapidminer, antara lain :

1. Berlisensi gratis (*open source*).
2. Multiplatform karena diprogram dalam bahasa Java.
3. Internal data berbasis XML sehingga memudahkan pertukaran data eksperimen.
4. Dilengkapi dengan *scripting language* untuk otomatisasi eksperimen.
5. Memiliki GUI (*Graphical User Interface*), *command line mode (batch mode)*, dan Java API yang dapat dipanggil dari program lain.
6. Dapat dikembangkan dengan menambahkan *plugin* dan *ekstension*.
7. Fasilitas *plotting* untuk visualisasi data multidimensi dan model.