

BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1. Tinjauan Pustaka

Pada tahun 2015 pernah dilakukan penelitian dengan topik pengklasifikasian penerima dan bukan penerima Kartu Identitas Miskin (KIM) menggunakan metode “*Naive Bayes* Berbasis Desktop ” penelitian ini dilakukan oleh Andi Bekto Rahardjo, digunakan untuk melakukan klasifikasi data dalam menentukan penerima Kartu Identitas Miskin (KIM). Dalam melakukan pengklasifikasian data maka menggunakan kriteria status perkawinan, jenis kelamin, alamat, pekerjaan, luas bangunan per m², jenis lantai bangunan, ketersediaan wc, sumber penerangan, sumber air minum, bahan bakar, frekuensi makan daging, rata-rata pakaian baru dalam setahun, frekuensi makan dalam sehari, kesanggupan biaya pengobatan, penghasilan KK, pendidikan terakhir KK, dan jumlah harta yang dimiliki.

Pada tahun 2017 pernah dilakukan penelitian dengan topik pengklasifikasian penerima program bantuan pemerintah menggunakan metode “*Algoritma Naive Bayesian* Berbasis Desktop ” penelitian ini dilakukan oleh Masna Wati dan Abdul Hadi, digunakan untuk melakukan mengklasifikasikan data dalam menentukan penerima program bantuan pemerintah. Dalam melakukan pengklasifikasian data maka menggunakan kondisi hunian, kesehatan dan gizi, tingkat pengetahuan, kondisi ekonomi, penggunaan komunikasi, lingkungan sosial politik, transportasi, kondisi religius, makanan dan pakaian, keamanan.

Pada tahun 2016 pernah dilakukan penelitian dengan topik pengklasifikasian status kesejahteraan rumah tangga menggunakan metode “*Naive Bayes* Berbasis Desktop ” penelitian ini dilakukan oleh *Erfan Karyadiputra*, digunakan untuk melakukan klasifikasi data dalam menentukan status kesejahteraan rumah tangga. Dalam melakukan pengklasifikasian data maka menggunakan kriteria status kesejahteraan, jenis kelamin kepala keluarga, umur kepala keluarga, jumlah keluarga, jumlah anggota keluarga, pendidikan kepala keluarga, lapangan usaha kepala keluarga, status kependudukan dalam pekerjaan kepala keluarga, status pengusaan bangunan tempat tinggal, jenis atap, kualitas atap, jenis dinding, kualitas dinding, jenis lantai, sumber air minum, sumber penerangan utama, dan bahan bakar utama masak.

Pada tahun 2016 pernah dilakukan penelitian dengan topik pengklasifikasian rumah tidak layak huni menggunakan metode “*Learning Vector Quantization* dan *Naive Bayes* Berbasis Desktop ” penelitian ini dilakukan oleh Fitri Juniaty Simatupang, digunakan untuk melakukan klasifikasi data dalam menentukan rumah tidak layak huni. Dalam melakukan pengklasifikasian data maka menggunakan kriteria luas lantai perkapita $< 10 \text{ m}^2$, jenis lantai rumah adalah tanah, jenis dinding rumah terbuat dari bambu/lainnya, jenis atap rumah terbuat dari daun/lainnya, sumber penerangan bukan listrik, sumber air minum tidak layak, tidak mempunyai fasilitas buang air besar, tidak mempunyai tempat pembuangan akhir tinja berupa tangki septik.

Pada tahun 2014 pernah dilakukan penelitian dengan topik perancangan sistem identifikasi dan pemetaan potensi kemiskinan untuk optimalisasi program

kemiskinan menggunakan metode “*AHP (Analytical Hierarchy Process)* Berbasis web ” penelitian ini dilakukan oleh Sri Redjeki, M. Guntara, dan Pius Anggoro, digunakan untuk melakukan pengoptimalan dalam menentukan pemberian bantuan yang tepat bagi warga miskin. Dalam melakukan proses ini maka menggunakan data yang telah dilengkapi dengan indikator-indikator dalam mencari hasil akhir dari proses pengambilan keputusan ini. Setelah semua data telah dikategorikan, data tersebut akan dimasukkan kedalam sistem informasi geografis.

Untuk penilaian yang diusulkan pada tahun 2017 dengan topik “Pengklasifikasian tingkat prioritas bantuan pada data rumah tidak layak huni dengan basis Desktop menggunakan metode *Naive Bayes*”. Pada penelitian ini menggunakan kriteria yaitu jenis atap, lantai, dinding, penghasilan dan pendidikan.

Tabel 2.1 Tabel Perbandingan Tinjauan Pustaka

Peneliti, Tahun	Objek/ Data	Metode	Teknologi	Hasil
Andi Bekto Rahardjo, 2015	Data warga miskin Kota Semarang Kelurahan Sumurrejo Kecamatan Gunungpati tahun 2013	<i>Naive Bayes Classifier</i>	Desktop	Penerima dan bukan penerima kartu KIM
Masna Wati dan Abdul Hadi, 2017	Data warga Desa Selangkau Kecamatan Kaliorang Kutai Timur	<i>Naive Bayes</i>	Desktop	Peluang sejahtera dan peluang tidak sejahtera

Tabel 2.1 Tabel Lanjutan

Peneliti, Tahun	Objek/ Data	Metode	Teknologi	Hasil
<i>Erfan Karyadiputra, 2016</i>	Data PPLS	<i>Naive Bayes</i>	Desktop	Rumah tangga miskin dan rumah tangga sangat miskin
Sri Redjeki, M. Guntara, dan Pius Anggoro, 2014	Data kemiskinan wilayah Kabupaten Bantul	<i>AHP (Analytical Hierarchy Process)</i>	Web	GIS pengkategorian kemiskinan
Fitri Juniaty Simatupang, 2016	Data rumah layak huni dan rumah tidak layak huni	<i>Learning Vector Quantization, Naive Bayes</i>	Desktop	Analisa perbandingan metode LVQ dan <i>Naive Bayes</i>
Usulan	Data rumah tidak layak huni	<i>Naive Bayes</i>	Desktop, API google maps	Prioritas tinggi, prioritas sedang dan prioritas rendah

2.2. Dasar Teori

2.2.1 Kemiskinan

Kata "miskin" yang artinya tidak berharta benda dan serba kekurangan. Departemen Sosial dan Biro Pusat Statistik, mendefinisikan kemiskinan dari perspektif kebutuhan dasar. Kemiskinan sebagai ketidakmampuan individu dalam memenuhi kebutuhan dasar minimal untuk hidup layak. Lebih lanjut Nurhadi

menyebutkan kemiskinan merupakan sebuah kondisi yang berada di bawah garis nilai standar kebutuhan minimum, baik untuk makanan dan non-makanan yang disebut garis kemiskinan (*poverty line*) atau batas kemiskinan (*poverty tresshold*). Garis kemiskinan adalah sejumlah rupiah yang diperlukan oleh setiap individu untuk dapat membayar kebutuhan makanan secara 2.100 kilo kalori per orang per hari dan kebutuhan non-makanan yang terdiri dari perumahan, pakaian, kesehatan, pendidikan, transportasi, serta aneka barang dan jasa lainnya. (Nurhadi, 2007: 13).

2.2.2 Data Mining

Data mining istilah yang digunakan dalam menguraikan penemuan pengetahuan di dalam database. Data mining adalah metode yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning yang bermanfaat dan pengetahuan yang terkait dari berbagai database yang besar. Data mining dibagi menjadi beberapa teknik yaitu(Khalida dan Noor, 2016):

1. Deskripsi

Peneliti dan analis terkadang mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang ada dalam data. Deskripsi dari pola dan kecenderungan memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

2. Estimasi

Estimasi sama dengan klasifikasi, variabel estimasi lebih ke arah numerik dari pada ke kategori. Model yang dibangun menggunakan *record* lengkap yang menyediakan nilai prediksi sebagai nilai dari variabel target.

3. Prediksi

Prediksi mirip dengan klasifikasi dan estimasi, namun dalam prediksi nilai dari hasilnya akan ada di masa datang.

4. Klasifikasi

Dalam klasifikasi terdapat target variabel kategori. Data mining memeriksa kumpulan *record*, setiap record mempunyai informasi pada variabel target dan variabel predictor.

5. Pengklusteran

Pengklusteran adalah pengelompokan record, pengamatan, dan membentuk kelas dari setiap objek yang mempunyai kemiripan. Kluster merupakan kumpulan *record* yang mempunyai kemiripan satu dengan yang lain, namun tidak mirip dengan *record* pada kluster lain.

6. Asosiasi

Asosiasi adalah menemukan atribut yang muncul bersamaan dalam satu waktu.

Tahap-tahap Data Mining adalah sebagai berikut(Alfa, 2015):

1. Pembersihan data (*Data Cleaning*)

Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan.

2. Integrasi data (*Data Integration*)

Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru.

3. Seleksi data (*Data Selection*)

Data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari database.

4. Transformasi data (*Data Transformation*)

Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam Data Mining.

5. Proses Mining

Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data. Beberapa metode yang dapat digunakan berdasarkan pengelompokan Data Mining.

6. Evaluasi pola (*Pattern Evaluation*)

Untuk mengidentifikasi pola-pola menarik ke dalam knowledge based yang ditemukan.

7. Presentasi pengetahuan (*Knowledge Presentation*)

Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

2.2.3 Klasifikasi

Menurut Prasetyo (2012), klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi terdapat dua proses yang dilakukan yaitu dengan membangun model untuk disimpan sebagai memori dan menggunakan model

tersebut untuk melakukan pengenalan atau klasifikasi atau prediksi pada suatu data lain supaya diketahui di kelas mana objek data tersebut dimasukkan berdasarkan model yang telah disimpan dalam memori (Riyan, 2014).

Sistem dalam klasifikasi diharapkan mampu melakukan klasifikasi semua set data dengan benar, namun tidak dapat dipungkiri bahwa kesalahan akan terjadi dalam proses pengklasifikasian tersebut sehingga perlunya dilakukan pengukuran kinerja dari sistem klasifikasi tersebut. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi (confusion matrix). Matriks konfusi merupakan tabel pencatat hasil kerja klasifikasi. Contoh dari matriks konfusi untuk dua kelas (biner) dapat dilihat pada Tabel 1.

Tabel 2.2 Matriks Konfusi untuk Klasifikasi Dua Kelas

f_{ij}		Kelas hasil prediksi (j)	
		Kelas = 1	Kelas = 2
Kelas asli (i)	Kelas = 1	f_{11}	f_{12}
	Kelas = 2	f_{21}	f_{22}

Setiap sel f_{ij} dalam matriks menyatakan jumlah rekod atau data dari kelas i yang hasil prediksinya masuk ke kelas j . Dari matriks konfusi dapat diketahui jumlah data pemetaan yang diprediksi benar dengan cara menjumlahkan nilai f_{11} dan f_{22} ($f_{11} + f_{22}$) dan jumlah data pemetaan yang diprediksi salah dengan menjumlahkan nilai f_{21} dan f_{12} ($f_{21} + f_{12}$). Akurasi hasil prediksi dapat dihitung ketika jumlah data yang diklasifikasi secara benar maupun salah telah diketahui. Untuk menghitung akurasi digunakan formula:

$$\text{Akurasi} = \frac{\text{jumlah data yang diprediksi secara benar}}{\text{jumlah prediksi yang dilakukan}} = \frac{f_{11} + f_{22}}{f_{11} + f_{12} + f_{21} + f_{22}} \quad (\text{persamaan 2.1})$$

2.2.4 Naive Bayes

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma menggunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Definisi lain mengatakan *Naive Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Alfa, 2015).

Model Naïve Bayes

Persamaan dari teorema Bayes adalah (Alfa, 2015) :

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)} \quad (\text{persamaan 2.2})$$

Di mana :

X : Data dengan class yang belum diketahui

H : Hipotesis data merupakan suatu class spesifik

$P(H|X)$: Probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas)

$P(H)$: Probabilitas hipotesis H (prior probabilitas)

$P(X|H)$: Probabilitas X berdasarkan kondisi pada hipotesis H

$P(X)$: Probabilitas X

Untuk menjelaskan metode *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, metode *Naive Bayes* di atas disesuaikan sebagai berikut:

$$P(C|F_1 \dots F_n) = \frac{P(C)P(F_1 \dots F_n|C)}{P(F_1 \dots F_n)} \quad (\text{persamaan 2.3})$$

Di mana Variabel C merepresentasikan kelas, sementara variabel F1 ... Fn merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (Posterior) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut prior), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C (disebut juga likelihood), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global (disebut juga evidence). Karena itu, rumus di atas dapat pula ditulis secara sederhana sebagai berikut:

$$\textit{Posterior} = \frac{\textit{prior x likelihood}}{\textit{evidence}} \quad (\text{persamaan 2.4})$$

2.2.5 MYSQL

MYSQL sering didefinisikan sebagai kumpulan data yang terkait. Secara teknis, yang berada dalam sebuah database adalah sekumpulan tabel atau objek lain (indeks, view, dan lain-lain). Tujuan utama pembuatan database adalah untuk memudahkan dalam mengakses data. (Abdul Kadir, 2009).

2.2.6 Geographic Information System (GIS)

Geographic Information System (GIS) adalah suatu sistem informasi yang dirancang untuk bekerja dengan data yang bereferensi spasial atau berkoordinat geografis atau dengan kata lain suatu GIS adalah suatu sistem basis data dengan kemampuan khusus untuk menangani data yang bereferensi keruangan (spasial) bersamaan dengan seperangkat operasi kerja (Barus dan Wiradisastra, 2000). Sedangkan menurut Prahasta (2002) sistem informasi geografis adalah suatu sistem informasi yang dapat memadukan antara data grafis (spasial) dengan data teks (atribut) objek yang dihubungkan secara geografis di bumi (georeference). Disamping itu, GIS juga dapat menggabungkan data, mengatur data dan melakukan analisis data yang akhirnya akan menghasilkan keluaran yang dapat dijadikan acuan dalam pengambilan keputusan pada masalah yang berhubungan dengan geografis (Antonio, 2013).