BAB 2

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Amalia, N, dkk. (2017) melakukan prediksi ketepatan waktu lulus mahasiswa menggunakan algoritma *Naive Bayes Classifier* dan atribut yang digunakan adalah kota asal, sks, IPK, Kelompok Keahlian, TA1, dan TA2. Kelas targetnya yaitu tepat waktu dengan 2 klasifikasi ya dan tidak dan diperoleh tingat akurasi sebesar 86%.

Nugroho, M.F. dan Setyoningsih, W. (2017) melakukan seleksi *forward selection* untuk menentukan atribut yang berpengaruh pada klasifikasi kelulusan mahasiswa menggunakan algoritma *Naive Bayes* dan diperoleh kesimpulan bahwa atribut yang berpengaruh dalam penentuan kelulusan mahasiswa tepat waktu adalah status pekerjaan dan IPK semester 4 dengan akurasi 97,14% .

Mauriza, A. F. (2014) melakukan prediksi kelulusan mahasiswa menggunakan metode *Naive Bayes* dan atribut yang digunakan yaitu jurusan asal sekolah, *gender*, daerah asal sekolah, rata SKS, rata MK, asisten Lab dan untuk klasifikasinya adalah tepat apabila memiliki nilai lulus ≤ 4 tahun dan terlambat apabila memiliki nilai lulus > 4 tahun.

Hananto, V. R. (2017) melakukan perbandingan akurasi antar metode *data mining* untuk memprediksi kelulusan mahasiswa sebagai penunjang angka efisiensi edukasi. Metode yang diguunakan yaitu *Naïve Bayes, Multilayer Perceptron, SMO, J48,* dan *REPTree*. Hasil yang diperoleh dari pengujian bahwa metode Naïve Bayes mempunyai tingkat akurasi tertinggi dan *error rate* terkecil dengan tingkat akurasi

57,3% menggunakan tahapan CRISP-DM. Atribut yang digunakan adalah Dosen Wali, IPK, SKS Kumulatif, Status Mahasiswa(Tugas Akhir), Status menepuh Kerja Praktik, Dosen Pembimbing1,dan Dosen Pembimbing2. Kelas target yang digunakan adalah Lulus Tepat Waktu yang terdiri dari 2 klasifikasi yaitu Ya dan Tidak.

Yulianti, S. (2018) melakukan implementasi *data mining* untuk memperkirakan masa studi mahasiswa dengan metode *K- Nearest Neighbor (K-NN)* dan atribut yang digunakan yaitu indeks prestasi semester 1 sampai 4 dan jumlah sks dari semester 1 sampai semester 4 dengan tingkat akurasi sebesar 70 %.

Selanjutnya skripsi yang dibuat adlah implementasi metode klasifikasi *Naïve Bayes* dalam memprediksi penentuan kelulusan mahasiswa tepat waktu dengan kriteria yang digunakan adalah indeks prestasi semester 1, indeks prestasi kumulatif sampai semester4, total sks yang ditempuh sampai semester 4, jurusan asal sekolah, jumlah nilai D dan E dalam sks. Perbedaan dengan penelitian sebelumnya adalah atribut dan metode yang digunakan yaitu menggunakan *Naïve Bayes Classier*. Alasan menggunakan metode ini karena data yang akan dilakukan pengujian adalah data kategorikal, data yang bermula dari data numerik diubah kedalam bentuk data kategorikal. Data ini diubah kedalam bentuk kategorikal karena dibutuhkan untuk megetahui penyebab kenapa mahasiswa tidak bisa lulus tepat waktu, misalkan apakah karena nilai ipknya rendah atau total sksnya kurang atau jumlah nilai D-nya banyak atau mungkin jumlah nilai E-nya banyak. Alasan pemilihan atribut dengan memprediksi mahasiswa pada semester 4 supaya mahasiswa dapat melakukan perubahan prestasi akademik untuk mencapai target kelulusan yang diinginkan. Tabel perbandingan dari penelitian sebelumnya dapat dilihat pada Tabel 2.1.

Tabel 2.1 Perbandingan Tinjauan Pustaka

No	Penulis	Objek Penelitian	Metode	Atribut
1.	Amalia, N, dkk(2017)	Universitas Telkom	*Naive Bayes Classifier*	kota asal, sks, IPK, KK(kelompok keahlian), TA1, dan TA2.
2.	Nugroho, M.F. dan Setyoningsih, W. (2017)	Fakultas Ilmu Komputer UNAKI Semarang	*Naive Bayes Classifier*	status pekerjaan dan IPK semester 4
3.	Mauriza, A. F. (2014)	Fakultas Komunikasi Dan Informatika UMS	*Naïve Bayes*	jurusan asal sekolah, *gender*, asal sekolah, rata SKS, rata MK, asisten Lab
4.	Hananto, V. R. (2017)	Prodi Sistem Informasi-S1 Stikom Surabaya	*Naïve Bayes, Multilayer Perceptron, SMO, J48,* dan *REPTree*	Dosen Wali, IPK, SKSK(SKS Kumulatif), STS(Status Mahasiswa(Tugas Akhir)), STS_Tempuh_KP(Status menepuh Kerja Praktik), Dosen1(Dosen Pembimbing1,) dan Dosen2(Dosen Pembimbing2)
5.	Yulianti, S. (2018)	Program Studi Sarjana STMIK Akakom Yogyakarta	*K-NN*	indeks prestasi semester 1 sampai 4 dan jumlah sks dari semester 1 sampai semester 4
6.	Rochmana (2019)	Program Studi Sistem Informasi STMIK Akakom Yogyakarta	*Naïve Bayes Classifier*	indeks prestasi semester 1, indeks prestasi kumulatif sampai semester4, total sks sampai semester 4, dan jurusan asal sekolah, jumlah nilai D salam sks, jumlah nilai E dalam sks.

2.2 Dasar Teori

2.2.1 *Data mining*

Data mining adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan. (Santosa, 2007).

Data mining adalah proses menemukan pengetahuan yang menarik dari jumlah data besar yang disimpan dalam database, data warehouse (gudang data), atau repositori informasi lainnya. (Han & Kamber, 2006).

Data mining adalah teknik untuk menemukan dan mendeskripsikan pola-pola yang ada dalam data sebagai sebuah alat untuk membantu menjelaskan data tersebut dan membuat prakiraan dari data itu (Witten & Frank, 2005)

Proses dan teknik penyaringan data menentukan mutu pengetahuan dan informasi yang akan diperoleh. Istilah lain untuk data mining adalah *Knowledge Discovery in Databases* (KDD). KDD merupakan sebuah proses yang terdiri dari serangkaian proses interasi yang terurut, dan *data mining* merupakan salah satu langkah dalam proses KDD. Urutan langkah dalam KDD adalah sebagai berikut:

a. Data Selection

Data Selection menciptakan himpunan data target atau pemilihan himpunan data, dimana penemuan *(discovery)* akan dilakukan. Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai.

b. Pre-processing/ Cleaning

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

c. Transformation

Proses transformasi adalah proses pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada *goal* yang ingin dicapai. Proses transformasi dilakukan pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

d. Data Mining

Data mining dalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode atau algoritma dalam data mining sangat bervariasi.

e. Interpretation/Evaluation

Tahap *interpretation/evaluation* merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

2.2.2 *Naïve bayes*

Naïve bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan *teorema Bayes* (aturan Bayes) dengan asumsi *independensi* (ketidaktergantungan) yang kuat *(naif).* Dengan kata lain, dalam *Naive Bayes* model yang digunakan adalah “model fitur independen” . Fitur independensi yang kuat adalah sebuah fitur pada sebuah data tidak ada kaitannya dengan adanya atau tidak adanya fitur yang lain dalam data yang sama atau dengan kata lain *Na¨ıve Bayesian Classifiers* menganggap bahwa efek dari nilai atribut

pada kelas tertentu tidak bergantung pada nilai atribut lainnya. (Han & Kamber, 2012).

Teori keputusan *Bayes* adalah pendekatan statistik yang fundamental dalam pengenalan pola *(pattern recognition)*. Pendekatan ini didasarkan pada kuantifikasi *trade-off* antara berbagai keputusan klasifikasi dengan menggunakan probabilitas dan ongkos yang ditimbulkan dalam keputusan-keputusan tersebut. Ide dasar dari bayes adalah menangani masalah yang bersifat hipotesis yakni mendesain suatu klasifikasi untuk memisahkan objek (Amalia, N, dkk. 2017)

Bayesian classification adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class*. *Bayesian classification* terbukti memiliki akurasai dan kecepatan yang tinggi saat diaplikasikan ke dalam data base dengan data yang besar (Kusrini dan Luthfi ,2009).

Prediksi *bayes* didasarkan pada *Naïve Bayes Classifier* dengan bentuk umum seperti yang ada pada persamaan (1):

$$P(C_i|\,X) = \frac{p(X\,|C_i)P(C_i)}{p(X)} \qquad \text{.....Persamaan (1)}$$

Keterangan :

- X : Data dengan Class yang belum diketahui
- C_i : Suatu variabel yang harus dideskripsikan secara probabilistik
- $P(C_i|X)$: Probabilitas hipotesis C_i berdasarkan kondisi X (posteriori probability)
- $P(C_i)$: Probabilitas hipotesis C_i (prior probability)
- $P(X|C_i)$: Probabilitas X berdasar kondisi pada hipotesis C_i
- $P(X)$: Probabilitas dari X

Berikut ini merupakan contoh *data training* sebanyak 14 data dengan *output* main sepak bola atau tidak. Setiap data ditandai dengan atribut cuaca, temperatur, kelembaban, dan angin yang dapat dilihat pada Tabel 2.2.

Tabel 2.2 *Data Training* Cuaca dan Keputusan Main atau Tidak

Cuaca x1	**Temperatur x2**	**Kelembaban x3**	**Angin x4**	**Main atau Tidak y**
Cerah	Panas	Tinggi	Kecil	Tidak
Cerah	Panas	Tinggi	Besar	Tidak
Mendung	Panas	Tinggi	Kecil	Ya
Hujan	Sedang	Tinggi	Kecil	Ya
Hujan	Dingin	Normal	Kecil	Ya
Hujan	Dingin	normal	Besar	Tidak
Mendung	Dingin	normal	Besar	Ya
Cerah	Sedang	tinggi	Kecil	Tidak
Cerah	Dingin	normal	Kecil	Ya
Hujan	Sedang	normal	Kecil	Ya
Cerah	Sedang	normal	Besar	Ya
Mendung	Sedang	tinggi	Besar	Ya
Mendung	Panas	normal	Kecil	Ya
Hujan	Sedang	tinggi	Besar	Tidak

Berikut ini merupakan contoh perhitungan metode *Naïve Bayes* yang menggunakan rumus pada persamaan (1) untuk menentukan kelas dari data baru berikut:

(*Cuaca= cerah, Temperatur= dingin, Kelembaban= tinggi, Angin= besar)*

Langkah – langkah perhitungan yang dikerjakan sebagai berikut:

a. Mencari *probabilitas prior*

Probabilitas prior ini menyatakan berapa peluang munculnya keputusan main sepak bola $P(C_1)$ dan peluang keptusan tidak main sepak bola $P(C_2)$. Misalkan N adalah jumlah total keputusan main sepak bola dan tidak, N_1 menyatakan jumlah keputusan main sepak bola dan N_2 menyatakan jumlah keputusan tidak main sepak bola. Berdasarkan pada Tabel 2.2 jumlah total keputusan main sepak bola dan tidak

sebanyak 14, jumlah keputusan main sepak bola sebanyak 9, dan jumlah keputusan tidak main sepak bola sebanyak 5. Didapatkan rumus sebagai berikut :

$$P(C_1) = \frac{N_1}{N} \qquad P(C_2) = \frac{N_2}{N}$$
$$= \frac{9}{14} \qquad = \frac{5}{14}$$
$$= 0{,}64 \qquad = 0{,}36$$

b. Mencari *Probabilitas bersyarat (likelihoad)*

Probabilitas bersyarat (likelihoad), P(X|Cᵢ) ini menyatakan peluang munculnya *X* jika diketahui C_i . Misalkan :

X : (*Cuaca=cerah,Temperatur=dingin,Kelembaban=tinggi,Angin=besar)*

C_i: C_1 adalah main sepak bola dan C_2 adalah tidak main sepak bola.

Sehingga *likelihoad* dapat dihitung dengan mengalikan hasil dari masing – masing nilai probabilitas per-atribut yang berdasarkan pada C_i.

Langkah – langkah menghitung *likelihood* :

- Menghitung probabilitas per-atribut

Misalkan menghitung probabilitas Angin = besar dengan C_i adalah C_1 yaitu main sepak bola. Berdasarkan Table 2.2 banyaknya Angin = besar yang keputusannya adalah main sepak bola ada 3 dan banyaknya kelas dengan keputusan main sepak bola ada 9. Maka probabilitasnya adalah

$$P(Angin = Besar \mid main) = \frac{jumlah\ Angin\ Besar\ dengan\ keputusan\ main}{jumlah\ seluruh\ keputusan\ main}$$

$$= \frac{3}{9}$$

$$= 0{,}33$$

Melakukan perhitungan dengan cara yang sama pada probabilitas per-atribut yang lain, sehingga diperoleh hasil :

P(Cuaca=Cerah|Main)	=	2/9	=	0,22
P(Cuaca=Cerah|Tidak)	=	3/5	=	0,60
P(Temperatur=dingin | Main)	=	3/9	=	0,33
P(Temperatur=dingin | Tidak)	=	1/5	=	0,20
P(kelembaban=tinggi |Main)	=	3/9	=	0,33
P(kelembaban=tinggi |tidak)	=	4/5	=	0,80
P(Angin=Besar |Tidak)	=	3/5	=	0,60

- Mengalikan semua probabilitas per-atribut yang berdasarkan pada C_i yaitu C_1 adalah main yang disebut *Likelihoad Ya* atau C_2 adalah tidak main yang disebut *Likelihoad Tidak.*

- *Likelihoad Ya*

P(X|Main)

*= P(Cuaca=Cerah|Main) * P(Temperatur=dingin | Main) * P(kelembaban=tinggi |Main) * P(Angin=Besar |Main)*
=0,22 * 0,33* 0,33 * 0,33

= 0.0080

- *Likelohoad Tidak*

P(X|Tidak)

*= P(Cuaca=Cerah|Tidak) * P(Temperatur=dingin | Tidak) * P(kelembaban=tinggi |Tidak) * P(Angin=Besar |Tidak)*
=0.60 * 0,20 * 0,80 * 0,60

= 0.0576

c. Mengalikan *Likelihoad* dengan *Prior*

Perkalian *Likelihoad* dengan *Prior* merupakan hal yang penting untuk menemukan *posterior*.

- Perkalian *Likelioad* dengan *Prior* pada keputusan Main

  = $P(X|C_i) * P(C_i)$

  = *P(X|Main) * P(Main)*

  =0.0080 * 0,64

  =0.00512

- Perkalian *Likelioad* dengan *Prior* Tidak Main

  = $P(X|C_i) * P(C_i)$

  = *P(X|Tidak) * P(Tidak)*

  =0.0576 * 0,36

  =0.020736

d. Menghitung *probabilitas posterior*

Probabilitas posterior, P(C_i|X) ini menyatakan probabilitas keluarnya hasil C_i jika diketahui nilai *X* tertentu. *Probabilitas posterior, P(C_i|X)* dicari dengan menggunakan rumus pada persamaan(1) :

- Probabilitas *Posterior* Main

$$P(C_i|\,X) \quad = \frac{p(X\,|C_i)P(C_i)}{p(X)}$$

$$posterior = \frac{likelihoad * \ prior}{evidence}$$

$$= \frac{p(X\ |Main) * P(Main)}{P(X\ |Main) * P(Main) + P(X\ |Tidak) * P(Tidak)}$$

$$= \frac{0.00512}{0.00512+0.020736}$$

= 0,198019802

- Probabilitas *Posterior* Tidak Main

$$P(C_i|\ X) \quad = \frac{p(X\ |C_i)P(C_i)}{p(X)}$$

$$posterior = \frac{likelihoad * prior}{evidence}$$

$$= \frac{P(X\ |Tidak) * P(Tidak)}{P(X\ |Tidak) * P(Tidak) + p(X\ |Main) * P(Main)}$$

$$= \frac{0.020736}{0.020736+0.00512}$$

=0,801980198

Nilai *probabilitas posterior* Tidak Main sebesar 0,801980198 lebih besar dari nilai *probabilitas posterior* Main yang bernilai 0,198019802 dan nilai *probabilitas posterior* Tidak Main mendekati nilai 1, sehingga dengan *Naïve Bayes* prediksi dapat disimpulkan Tidak Main untuk data *input* ini yang berdasarkan pada estimasi probabilitas yang dipelajari dari *data training*.