BAB 2

TINJAUAN PUSTAKA

2.1. Prediksi Keterlambatan Biaya Kuliah

Prediksi keterlambatan biaya kuliah merupakan cara untuk mengetahui pola klasifikasi yang tepat atau terlambat dengan melihat indikator mana yang paling berpengaruh (Apandi *dkk.*, 2019). Dimana mahasiswa yang terlambat dalam melakukan pembayaran SPP dapat diminimalisir dengan menggunakan teknik data mining salah satunya adalah klasifikasi, kemudian dari klasifikasi tersebut akan dijadikan sebagai dasar untuk prediksi pembayaran SPP di semester berikutnya (Rohmayani, 2020). Berbagai penelitian telah dilakukan tentang prediksi keterlambatan pembayaran SPP adapun beberapa penelitian seperti pada tabel 2.1.

Tabel 2.1 Penelitian Terkait Keterlambatan SPP

No	penulis	Judul	Metode	Hasil
1	Muqorobin *dkk* (2020)	*Estimation System For Late Payment Of School Tuition Fees*	Komparasi *Naïve Bayes* dan KNN	Metode *Naïve Bayes* mendapatkan nilai akurasi 85% dan metode K-NN 81%
2	Rohmayani (2020)	*Analysis Of Student Tuition Fee Pay Delay Prediction (Case Study : Politeknik Tedc Bandung)*	*naive bayes with particle swarm optimazation*	Hasil akurasi yang didapatkan sebesar 73.94%

Tabel 2.1 Penelitian Terkait Keterlambatan SPP (Lanjutan)

No	penulis	Judul	Metode	Hasil
3	Ginting *dkk* (2020)	prediksi keterlambatan pembayaran sumbangan pembangunan pendidikan sekolah menggunakan *python*	C4.5	Hasil akurasi yang didapatkan sebesar 73%.
4	Istiana (2018)	Aplikasi Prediksi Pembayaran Bulanan Santri (Studi Kasus Pondok Pesantren Assalafi Al Fithrah Meteseh Semarang)	C4.5	Hasil akurasi diperoleh sebesar 81.15%
5	Abdullah *dkk* (2019)	Sistem Prediksi Keterlambatan Pembayaran SPP Sekolah (Studi Kasus: SMK Al-Islam Surakarta)	*K-Nearest Neighbor*	menghasilkan nilai akurasi sebesar 86%

2.2. *Data Mining*

Data mining merupakan proses untuk menggali nilai tambah berupa informasi yang belum diketahui secara manual dari suatu basis data, informasi yang didapat berasal dari hasil mengekstraksi dan mengenali pola yang penting dari data yang terdapat pada basis data, *data mining* biasanya dipergunakan untuk mencari pengetahuan dari suatu basis data yang besar (Vulandari, 2017).

Menurut Nofriansyah & Gunadi Widi Nurcahyo (2015) pada proses pemecahan masalah dan pencarian pengetahuan baru terdapat beberapa klasifikasi secara umum yaitu:

a) Estimasi

   Estimasi biasanya dipakai untuk memperkirakan sebuah data baru yang tidak memiliki keputusan berdasarkan histori data yang telah ada.

b) Asosiasi

   Asosiasi digunakan untuk mengetahui kelakuan dari suatu kejadian khusus

atau proses dimana hubungan asosiasi muncul pada setiap kejadian.

c) Klasifikasi

Klasifikasi merupakan Suatu teknik dengan melihat pada atribut dari kelompok yang telah didefinisikan sebelumnya. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi objek data yang telah diklasifikasikan dan dapat menggunakan hasilnya untuk memberikan sejumlah aturan.

d) Klasterisasi

Klasterisasi dapat dipergunakan untuk menganalisis pengelompokan berbeda terhadap data, mirip dengan klasifikasi, hanya saja pengelompokan belum belum terdefinisi sebelum diproses melalu *tool data mining*

e) Prediksi

Algoritma prediksi biasanya dipergunakan untuk memperkirakan suatu kejadian sebelum kejadian tersebut terjadi.

2.3. Persiapan Data

Persiapan data meliputi beberapa proses seperti pembersihan, integrasi, reduksi, penambahan, dan transformasi data, adapun penjelasan tahapannya sebagai berikut (Suyanto, 2017):

a) Pembersihan data

Sebuah data bisa dikatakan kotor apabila mengandung data yang tidak memiliki atribut lengkap. Semakin banyak kotoran pada suatu data maka semakin buruk pula dataset tersebut.

b) Integrasi data

Integrasi data yang baik akan menghasilkan data gabungan yang lebih informatif sehingga dapat meningkatkan akurasi dari proses kecepatan pada saat data mining.

c) Reduksi data

Reduksi data dibagi menjadi kedalam tiga kelompok yaitu reduksi dimensi adalah mereduksi dimensi (jumlah atribut) data. Selanjutnya reduksi keterbilangan dengan mengganti data asli dengan representasi baru yang

lebih sederhana. Kemudian yang terakhir kompresi data merupakan metode - metode transformasi data yang bisa berupa lossless (data asli dapat di rekonstruksi dari data terkompress tanpa kehilangan informasi) atau lossy (data asli hanya dapat diaproksimasi, dengan kehilangan sebagian informasi).

2.4. Transformasi

Menurut Kurniawan (2020) algoritma machine learning membutuhkan data numerik untuk training dataset, dalam melakukan transformasi metode yang paling umum adalah *One-Hot Encoding, One-Hot Encoding* merupakan proses untuk membuat kolom baru dari variabel kategorikal dimana setiap kategori menjadi kolom baru dengan nilai 0 atau 1 (0 mewakili tidak ada dan 1 mewakili ada). Transformasi Data adalah upaya yang dilakukan dengan tujuan utama untuk mengubah skala pengukuran data asli menjadi bentuk lain sehingga data dapat memenuhi asumsi-asumsi yang mendasari analisis ragam. Transformasi data ada beberapa jenis, antara lain (Daqiqil Id, 2021):

1) Transformasi Logaritma.
2) Transformasi Arcsin.
3) Transformasi Square (Kuadrat).
4) Transformasi Cubic (Pangkat Tiga).
5) Transformasi Inverse (Kebalikan).
6) Transformasi Inverse Square Root (Kebalikan Akar).

2.5. Klasifikasi

Menurut Zaki et al (2013), bagian yang sangat penting dalam *data mining* adalah teknik klasifikasi, yaitu bagaimana mempelajari sekumpulan data agar dapat menghasilkan aturan yang bisa mengenal pola data baru yang belum pernah dipelajari sebelumnya. Klasifikasi bisa diartikan sebagai proses untuk menyatakan suatu objek data menjadi salah satu kategori yang telah didefinisikan sebelumnya (Suyanto, 2017). Sedangkan menurut Nofiansyah & Gunadi Widi Nurcahyo (2015) klasifikasi dapat didefinisikan sebagai berikut:

a) Jika memiliki sekumpulan record dimana setiap record terdiri dari sekumpulan atribut dan satu atribut merupakan kelas.

b) Menentukan suatu model untuk parameter atribut kelas sebagai suatu fungsi nilai dari atribut lain.
c) Adapun tujuannya adalah rekord-rekord yang tidak terlihat sebelumnya ditentukan suatu kelas seakurat mungkin.
d) Suatu kumpulan data uji yang dapat menentukan keakuratan suatu model. Umumnya, data set dibagi menjadi data training dan data testing, dimana data training dapat dipergunakan untuk membentuk model dan data testing digunakan untuk mengujinya.

Menurut Nofiansyah & Gunadi Widi Nurcahyo (2015) terdapat beberapa teknik klasifikasi data yang bisa digunakan sebagai solusi pemecahan kasus sebagai berikut:

a) Algoritma C4.5
b) Algoritma *K-Nearest Neighbor*
c) Algoritma *Naïve Bayes*
d) ID3
e) CART (*Clasification And Regression Tree*)

2.6. *Algoritma Naive Bayes*

Algoritma *Naive Bayes* adalah salah satu algoritma klasifikasi berdasarkan teorema *Bayesian* pada statistika (Suntoro, 2019). Algoritma *Naive Bayes* dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas (Han & Kamber, 2012).

Teorema *Bayesian* menghitung nilai *probability* P(H|X) menggunakan probabilitas P(H), P(X), dan P(X|H) sebagai berikut (Suntoro, 2019):

$$\mathrm{P(H|X)} = \frac{P(X|H).\mathrm{P(H)}}{P(X)} \qquad (2.3)$$

Keterangan:

X : Merupakan *data testing* yang kelasnya belum diketahui.
H : Merupakan hipotesis data X yang kelasnya lebih spesifik.
P(H) : Merupakan peluang dari hipotesa H.
P(X) : Merupakan *predictor prior* yang merupakan probabilitas X.

P(X|H) : Disebut juga dengan *likelihood* yang merupakan probabilitas hipotesis X berdasarkan kondisi H.

Perhitungan Naïve Bayes bertipe numerik maka menggunakan perhitungan distribusi Gaussian, adapun rumusnya sebagai berikut:

a. Menghitung nilai *mean* yaitu:

$$\mu = x = \frac{\sum_i^n x_i}{\mathrm{n}} \quad atau\ \mu = \frac{x_1+x_2+x_3+\cdots+x_n}{n} \tag{2.4}$$

Keterangan:

μ : rata-rata hitung (*mean*)

Xi : nilai sampel ke -i

n : jumlah sampel

b. Menghitung nilai standar deviasi sebagai berikut:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i-\mu)^2}{n-1}} \tag{2.5}$$

Keterangan:

σ : Standar devisi

Xi : nilai x ke -i

μ : rata-rata hitung

n : jumlah sampel

c. Menghitung nilai distribusi *Gaussian*

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi.\sigma}} exp^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{2.6}$$

Keterangan:

X : nilai data

σ : simpangan baku (Standar devisi)

μ : rata-rata hitung (*mean*)

π : bilangan konstan (3,14 atau $\frac{22}{7}$)

Exp : Konstanta bilangan euler (2,178)

Algoritma *Naive Bayes* telah banyak digunakan dalam penelitian sebelumnya seperti pada tabel 2.2.

Tabel 2.2 Penelitian Terkait Algoritma Naïve Bayes

No	Penulis	Judul	Metode	Hasil
1	Putri (2019)	Klasifikasi Data Nasabah Berpotensi Terkena Kredit Macet	*Naïve Bayes*	akurasi yang didapatkan sebesar 77.2819%.
2	Amelia dkk (2017)	Prediksi Masa Studi Mahasiswa	*Naïve Bayes*	Hasil akurasi yang didapatkan sebesar 85.17%.
3	Hasan (2017)	Prediksi Tingkat Kelancaran Pembayaran Kredit Bank	*Naive Bayes* berbasis *Forward Selection*	Akurasi yang dihasilkan mencapai 71,97%.
4	Mustafa & Simpen (2018)	Evaluasi Kinerja Akademik Mahasiswa	*Naïve Bayes*	hasil akurasi yang didapat sebesar 92,3%.
5	Murtopo (2016)	Prediksi Kelulusan Tepat Waktu Mahasiswa STMIK YMI Tegal	*Naïve Bayes*	hasil dari penelitian ini mendapatkan akurasi sebesar 94,34%.
6	Rahmatullah & Utami (2019)	Prediksi Tingkat Kelulusan Tepat Waktu	*Naive Bayes* dan *K-Nearest Neighbor*	Algoritma *Naive Bayes* menghasilkan akurasi 85%, sedangkan algoritma *K-Nearest Neighbor* menghasilkan 68.89%.
7	Prasetyo (2018)	Prediksi Kelulusan Mahasiswa Tepat Waktu	*Naive Bayes Forward Selection*	hasil akurasi sebesar 97,92%.
8	Widaningsih (2019)	Perbandingan Metode Data Mining Untuk Prediksi Nilai dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika	C4.5, *Naïve Bayes*, KNN, dan SVM	Diperoleh Algoritma *Naïve Bayes* merupakan algoritma terbaik dengan hasil akurasi sebesar 76,79%.

Tabel 2.2 Penelitian Terkait Algoritma Naïve Bayes (Lanjutan)

No	Penulis	Judul	Metode	Hasil
9	Koeswara *dkk* (2020)	Penerapan Particle Swarm Optimization (Pso) Dalampemilihan Atribut Untuk Meningkatkan Akurasi Prediksi Diagnosis Penyakit Hepatitis	*Naive Bayes*	Algoritma *Naïve Bayes* mendapatkan akurasi 84,85% sedangkan Optimasi Naive Bayes menggunakan PSO sebesar 92,50%
10	Sulaehani (2016)	Prediksi Keputusan Klien Telemarketing Untuk Deposito Pada Bank Menggunakan Algoritma Naive Bayes Berbasis Backward Elimination	Algoritma Naive Bayes Berbasis Backward Elimination	algoritma Naive Bayes berbasis Backward Elimination mendapatkan akurasi 90,69%

2.7. *Confusion Matrix*

Confusion matrix merupakan suatu teknik yang dapat digunakan untuk mengetahui seberapa akurat model klasifikasi menggunakan tabel *confusion matrix* (Primartha, 2021).

Tabel 2. 3 Confusion Matrix

Class		*Actual*	
		TRUE	*FALSE*
Prediction	*TRUE*	*True Positive (TP)*	*False Positive (FP)*
	FALSE	*False Negative (FN)*	*True Negative (TN)*

Sumber : (Kurniawan, 2020)

Terdapat sejumlah ukuran yang digunakan untuk menilai atau mengevaluasi model klasifikasi (Daqiqil Id, 2021), yaitu:

a. TP (*True Positive*) merupakan jumlah data berlabel *yes* yang nilainya diidentifikasi benar.
b. TP (*True Negative*) merupakan jumlah data berlabel *no* yang nilainya diidentifikasi salah.
c. FP (*False Positive*) merupakan jumlah data berlabel *yes* yang nilai sebenarnya diidentifikasi salah.
d. FN (*False Negative*) merupakan jumlah data berlabel *no* yang nilai sebenarnya teridentifikasi benar.

Mengukur *confusion matrix* dapat menggunakan rumus sebagai berikut (Kurniawan, 2020):

a. *Accuracy* (Akurasi)

Mengukur akurasi model. Rumusnya Jumlah prediksi benar dibagi dengan total seluruh populasi.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.7}$$

b. *Precision* (Ketepatan)

Mengukur jumlah data yang sukses diprediksi positif, dibandingkan dengan seluruh data yang diprediksi positif, yang kenyataannya benar dan tidak benar.

$$precision = \frac{TP}{TP + FP} \tag{2.8}$$

c. *Sensivity (Recall)*

Mengukur banyaknya data yang sukses saat diprediksi sebagai positif dibandingkan dengan seluruh data yang pada kenyataannya positif.

$$sensivity = \frac{TP}{TP + FN} \tag{2.9}$$