BAB 2

TINJAUAN PUSTAKA

2.1. Prediksi Keterlambatan Mahasiswa

Prediksi keterlambatan mahasiswa untuk mengetahui pola klasikasi yang tepat atau terlambat dengan mengetahui indikator mana yang paling berpengaruh (Apandi dkk., 2019). Prediksi kemungkinan mahasiswa yang terlambat dalam melakukan pembayaran SPP dapat dijadikan sebagai rekomendasi dengan menggunakan teknik data mining salah satunya klasifikasi, kemudian dari klasifikasi tersebut akan digunakan sebagai dasar untuk prediksi pembayaran SPP di semester berikutnya (Rohmayani, 2020). Berikut ini beberapa penelitian terdahulu yang berkaitan tentang prediksi keterlambatan pembayaran SPP yang dapat dilihat pada Tabel 2.1.

Tabel 2.1 Penelitian Prediksi Keterlambatan Pembayaran SPP

No.	Penulis	Judul	Metode	Hasil
1	(Muqorobin dkk., 2020)	*Estimation System For Late Payment Of School Tuition Fees*	Algoritma *Naïve Bayes* dan *K-Nearest Neighbors*	Hasil akurasi untuki algoritma *Naïve Bayes* memperoleh tingkat akurasi 64% dan algoritma *K-Nearest Neighbors* memperoleh tingkat akurasi 63%.
2	(Rohmayani, 2020)	*Analysis of Student Tuition Fee Pay Delay Prediction Using Naive Bayes Algorithm With Particle Swarm Optimation Optimazation* (*Case Study* : Politeknik Tedc Bandung)	Algoritma *Naïve Bayes*	Hasil dari penelitian ini memperoleh nilai akurasi 73,94%, presisi 78,50%, recall 69%, dan AUC 0.771
	(Abdullah, 2019)	Sistem Prediksi Keterlambatan Pembayaran SPP Sekolah	Algoritma *K-Nearest Neighbor*	Hasil penelitian ini memperoleh nilai akurasi 65% menggunakan 5 fold

Tabel 2.1 Penelitian Prediksi Keterlambatan Pembayaran SPP (Lanjutan)

No.	Penulis	Judul	Metode	Hasil
		dengan Metode *K-Nearest Neighbor* (Studi Kasus: SMK Al-Islam Surakarta)		dengan nilai presisi 63%, recall 59%, dan F-measure 62%. Pada pengujian k=3 akurasi paling tinggi 68%, presisi 65%, dan recall 73%.
4	(Istiana, 2018)	Aplikasi Prediksi Pembayaran Bulanan Santri Dengan Menggunakan Algoritma C4.5 (Studi Kasus Pondok Pesantren Assalafi Al Fithrah Meteseh Semarang)	Algoritma C4.5	Hasil dari algoritma C4.5 memperoleh akurasi 81,15%, precision 77,62% dan recall 91,90%.
5	(Muqorobin dkk., 2019)	Optimasi Metode *Naive Bayes* dengan *Feature Selection Information Gain* Untuk Prediksi Keterlambatan Pembayaran Spp Sekolah	Algoritma *Naive Bayes* dan *Information Gain*	Hasil dari algoritma *Naive Bayes* memperoleh nilai akurasi 80%, presisi 75%, dan recall 75%. Sedangkan untuk penerapan algoritma *Naive Bayes* dengn *information gain* memperoleh nilai akurasi 90%, presisi 75%, dan recall 100%.

2.2. *Data Mining*

Data *mining* merupakan metode yang digunakan untuk penggalian sebagai nilai tambah suatu informasi yang tidak diketahui secara manual dari *database*. Sehingga dapat menghasilkan sebuah Informasi yang berasal dari ektraksi dan

identifikasi pola penting terhadap data yang terkandung dalam *database* (Vulandari, 2017).

Algoritma dan metode data mining adalah sebagai berikut (Nofriansyah & Nurcahyo, 2015).

1. Estimasi dapat dilakukan pada data baru yang belum ditentukan keputusannya berdasarkan data masa lalu.
2. Asosiasi untuk mendeteksi suatu peristiwa tertentu atau sering terdapat di setiap peristiwa. Metode yang dapat digunakan diantaranya algoritma apriori.
3. Klasifikasi merupakan teknik yang mempertimbangkan perilaku dan variabel yang telah ditentukan. Teknik ini memungkinkan anda saat melakukan klasifikasi data baru melalui manipulasi data dengan menunjukkan bahwa hasilnya akan berupa aturan-aturan. Metode klasifikas sering digunakan diantaranya *decision tree* karena mudah di interprestasikan yaitu algoritma C4.5, ID3, dll.
4. Klastering dapat dilakukan untuk analisis kelompok data yang berbeda. Klastering mirip seperti klasifikasi, tetapi untuk pengelompokkannya tidak terbentuk sebelum menggunakan *tools* data *mining.* Metode klastering diantaranya *neural network* maupun statistik.
5. Prediksi merupakan suatu proses dalam melakukan perkiraan atau memprediksi peristiwa sebelum peristiwa itu terjadi. Metode umum yang digunakan diantaranya *rough set*

2.3. *Decision Tree*

Decision Tree merupakan salah satu teknik klasifikasi untuk membuat sebuah pola pohon keputusan, sehingga memperoleh suatu jawaban atas suatu masalah yang telah dmasukan. Dari Pohon keputusan tersebut digunakan untuk memudahkan dalam mengidentifikasi hubungan melalui indikator yang berpengaruh terhadap masalah untuk menemukan jalan keluar dengan memperhitungkan indikator tersebut (Rufiyanto dkk., 2021). Pohon keputusan telah berkembang jauh dan untuk algoritma yang umum digunakan diantaranya ID3, C4.5, dan C5.0 (Jollyta dkk., 2020)

Algoritma C4.5 merupakan salah satu algoritma untuk membangun pohon keputusan melalui kriteria pembangunan keputusan (Nofriansyah, 2017). Algoritma C4.5 mirip seperti pohon yang terdapat node internal (bukan daun) dengan memiliki atribut di masing-masing cabang untuk atribut yang diuji dan tiap daun untuk kelas (Rohman, 2021). Penggunaan algoritma C4.5 pada penelitian sebelumnya telah banyak digunakan dalam berbagai studi kasus terdapat pada Tabel 2.2.

Tabel 2.2 Penelitian Terdahulu Menggunakan Algoritma C4.5

No	Penulis	Judul	Metode	Hasil
1	(Irmayansyah & Lastrini, 2021)	Penerapan Metode Algoritma C4.5 Untuk Prediksi Mahaiswa Non Aktif	Algoritma C4.5	Hasil dari uji kelayakan memperoleh nilai sebesar 87,50%, sedangkan rumus *confusion matrix* memperoleh tingkat akurasi sebesar 81%.
2	(Etriyanti dkk., 2020)	Implementasi Data Mining Menggunakan Algoritme *Naive Bayes Classifier* dan C4.5 untuk Memprediksi Kelulusan Mahasiswa	Algoritma *Naïve Bayes* dan C4.5	Hasil akurasi dari algoritma *Naïve Bayes* memperoleh akurasi 78,46%, sedangkan algorima C4.5 memperoleh tingkat akurasi 79,08%. Sehingga algoritma C4.5 lebih efektif untuk prediksi kelulusan mahasiswa.
3	(Haryoto dkk., 2021)	Algoritma C4.5 Dalam Data Mining Untuk Menentukan Klasifikasi Penerimaan Calon Mahasiswa Baru	Algoritma C4.5	Hasil penelitian ini mendapatkan tingkat akurasi 81,32%.
4	(Astuti, 2017)	Prediksi Ketepatan Waktu Kelulusan Dengan Algoritma Data Mining C4.5	Algoritma C4.5	Hasil penilitian untuk prediksi waktu kelulusan mahasiswa memperoleh tingkat akurasi 82%.

Tabel 2.2 Penelitian Terdahulu Menggunakan Algoritma C4.5 (Lanjutan)

No.	Penulis	Judul	Metode	Hasil
5	(Desiani dkk., 2020)	Prediksi Tingkat Indeks Prestasi Kumulatif Akademik Mahasiswa dengan Menggunakan Teknik Data Mining	Algoritma *Naïve Bayes* dan C4.5	Hasil penilitian algoritma *Naïve Bayes* mendapatkan tingkat akurasi 74,47% dan algoritma C4.5 mendapatkan tingkat akurasi 75,18%. Sehingga algoritma C4.5 lebih efektif dalam prediksi tingkat kumulatif akademik mahasiswa.
6	(Siallagan & Fitriyani, 2021)	Prediksi Penyakit Diabetes Mellitus Menggunakan Algoritma C4.5	Algoritma C4.5	Hasil dari penelitian ini mendapatkan tingkat akurasi 91,82%.
7	(Karlia, 2021)	Penerapan Algoritma C4. 5 untuk Klasifikasi Keterlambatan Pembayaran Premi Asuransi	Algoritma C4.5	Hasil penelitian ini mampu memperoleh tingkat akurasi 88%.
8	(Wahono & Riana, 2020)	Prediksi Calon Pendonor Darah Potensial Dengan Algoritma *Naïve Bayes*, *K-Nearest Neighbors* dan *Decision Tree* C4.5	Algoritma *Naïve Bayes*, *K-Nearest Neighbors* dan C4.5	Algoritma *Naïve Bayes* memperoleh nilai akurasi 85,15% dan nilai AUC 0.927, Untuk algoritma *K-Nearest Neighbors* mendapatkan tingkat akurasi 84,10% dan nilai AUC 0.816, kemudian algoritma C4.5 mendapatkan tingkat akurasi 93,83% dan nilai AUC 0.978. maka disimpulkan algoritma C4.5 lebih efektif dalam prediksi calon pendonor darah.

Tabel 2.2 Penelitian Terdahulu Menggunakan Algoritma C4.5 (Lanjutan)

No.	Penulis	Judul	Metode	Hasil
9	(Wahyono & Nugroho, 2018)	Penerapan Algoritma C4.5 Untuk Prediksi Tingkat Kompetensi Karyawan Pt Multistrada Arah Sarana	Algoritma C4.5	Hasil dari penelitian mendapatkan tingkat akurasi 80,39%, precision 85%, dan recall 70,83%, sedangkan nilai AUC Optimistic sebesar 0,907.
	(Khasanah, 2017)	Penerapan Algortima C4.5 Untuk Penentuan Kelayakan Kredit	Algoritma C4.5	Hasil penelitian penentuan kelayakan kredit calon nasabah melakukan pembayaran memperoleh tingkat akurasi 88,52%.

Algoritma C4.5 untuk membentuk pohon keputusan dilakukan sebagai berikut (Nofriansyah, 2017).

1. Pilih atribut sebagai akar

   Pemilihanan atribut berdasarkan nilai gain tertinggi, untuk perhitungannya menggunakan persaman 2.1.

$$Gaint\ (S,A) = Entropy(S) \sum_{i=1}^{n} \frac{|\ S_i\ |}{|\ S\ |} * entropy(Si) \qquad (2.1)$$

   Keterangan :

   S : Himpunan kasus

   A : Atributs

   n : Jumlah partisi atribut A

   |Si| : Jumlah kasus pada partisi ke-i

   |S| : Jumlah kasus dalam S

   Sebelumnya menghitung nilai gain dilakukan perhitungan nilai entropi terlebih dahulu menggunakan persamaan 2.3.

$$Entropy(S) = \sum_{i=1}^{n} -pi^{*} log_2 pi \qquad (2.2)$$

Keterangan :

S : Himpunan kasus

A : Fitur

n : Jumlah partisi S

pi : Proporsi dari Si terhadap S

2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang
4. Ulangi proses setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

2.4. Prapemrosesan Data

Prapemrosesan data dilakukan untuk pembersihan data, integrasi data, reduksi data, Penambahan data, normalisasi data, dan diskrikitisasi data dengan penjelasan sebagai berikut (Suyanto, 2017).

1. Pembersihan Data

   Sebuah data dikatakan tidak bersih jika mengandung kotoran seperti adanya nilai kosong, derau, pencilan, dan inkonsistensi. Pada data *mining* jika ada data yang kotor (tidak dibersihkan lebih dulu) umumnya memberikan hasil yang kurang baik.

2. Integrasi Data

   Pada data *mining*, secara praktis integrasi atau penggabungan sejumlah basisdata berbeda seringkali dilakukan. Integrasi data yang baik akan menghasilkan data gabungan dengan sedikit redundansi dan inkonsistensi sehingga meningkatkan akurasi dan kecepatan proses data *mining.*

3. Reduksi Data

   Data dapat direduksi menjadi jauh lebih kecil dengan tetap menjaga integritas yang terdapat pada data asli. Analisis dan penambangan pada data yang sudah direduksi akan lebih efisien dan hasilnya akan sama (hampir sama) dengan hasil analisis yang dilakukan menggunakan data asli.

4. Normalisasi Data

   Nilai dari variabel data dengan rentang yang berbeda kadang-kadang

perlu dinormalisasi atau distandarisasi untuk menghindari terjadinya bias. Normalisasi data biasanya dilakukan menggunakan rentang yang sempit diantaranya [0,1] dan [-1,1], maka dari masing-masing variabel akan mempunyai nilai yang sama.

5. Diskritisasi Data

   Pada beberapa kasus mungkin perlu dilakukannnya transformasi data sehingga sesuai dengan proses data mining seperti yang bertipe numerik ditranformasi menjadi data yang bertipe kategorik dan juga perlu ditranformasi dari nilai-nilai kontinu menjadi diskrit.

2.5. *Confusion matrix*

Confusion matrix merupakan salah satu teknik untuk mengetahui seberapa akurat model *classification* (Primartha, 2021). *Confusion matrix* menunjukan hasil prediksi yang dihasilkan menggunakan tabel *confusion matrix* terdapat di Tabel 2.3 (Daqiqil Id, 2021).

Tabel 2.3 Confusion Matrix

Class	*Actual = Yes*	*Actual =No*
Predicted = Yes	TP	FP
Predicted = No	FN	TN

Sumber: (Daqiqil Id, 2021)

Pada Tabel 2.3 merupakan tabel *confusion matrix* dengan penjelasan sebagai berikut.

1. TP (*True Positive)* dimana data mendapatkan label *Yes* yang memang sebenarnya labelnya *Yes*.
2. TN (*True Negative)* dimana data mendapatkan label *No* yang memang sebenarnya labelnya *No*.
3. FP (*False Positive)* dimana data mendapatkan label *Yes* yang memang sebenarnya labelnya *No*.
4. FN (*False Negative)* dimana data mendapatkan label *No* yang memang sebenarnya labelnya *Yes*.

Confusion matrix dapat digunakan untuk mengukur kinerja metode klasifikasi dengan rumus sebagai berikut (Kurniawan, 2020):

1. *Accuracy* (Akurasi)

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

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.3)$$

2. *Precision* (Ketepatan)

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

$$precision = \frac{TP}{TP + FP} \quad (2.4)$$

3. *Sensivity/recall*

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

$$sensivity = \frac{TP}{TP + FN} \quad (2.5)$$