BAB 2

TINJAUAN PUSTAKA

2.1. *Data Mining*

Data mining merupakan kegiatan mengestrak informasi penting dari suatu set data yang berukuran besar menggunakan metode tertentu (Santoso & Umam, 2018). Sedangkan menurut Nofiansyah & Gunadi Widi Nurcahyo (2015) *data mining* merupakan analisis otomatis dari data yang berjumlah besar. Adapun tugas – tugas yang biasa dilakukan *data mining* adalah sebagai berikut (Santoso & Umam 2018):

a. Klastering adalah suatu pengelompokkan objek ke dalam beberapa kelompok berdasarkan kesamaan antar objek, dimana dalam satu klaster harus berisi objek yang saling menyerupai dan klaster antar objek saling tidak mirip.
b. Klasifikasi adalah suatu pengelompokan objek berdasarkan kelompok yang telah ada. Namun berbeda dengan klastering, klasifikasi ini memerlukan *data training* yang telah diberi label kelompok atau kelas.
c. Regresi atau Estimasi adalah pada dasarnya menyerupai klasifikasi, namun membutuhkan *data training* yang telah diberi label target. Perbedaanya terletak pada output dari klasifikasi yaitu nilai diskrit, sedangkan output dari regresi yaitu nilai kontinyu.
d. Asosiasi adalah melakukan asosiasi antar objek dalam suatu set data, biasanya data transaksional.

2.2. Prediksi Keterlambatan Biaya Kuliah

Prediksi keterlambatan biaya kuliah adalah mengetahui pola klasikasi yang tepat atau terlambat dengan mengetahui indikator mana yang paling berpengaruh (Apandi dkk, 2019). Dimana mahasiswa yang terlambat dalam melakukan pembayaran Sumbangan Pembinaan Pendidikan dapat diminimalisir dengan menggunakan teknik data mining yaitu klasifikasi, kemudian dari klasifikasi tersebut akan dijadikan sebagai dasar untuk prediksi pembayaran SPP di semester

berikutnya. Berikut adalah tabel penelitian terkait yang membahas tentang prediksi keterlambatan biaya yang tertera pada tabel 2.1.

Tabel 2. 1 Penelitian Terkait Keterlambatan Pembayaran

No	Nama Peneliti	Tema Penelitian	Metode	Hasil
1.	(Abdullah dkk 2019)	Prediksi keterlambatan pembayaran Sumbangan Pembinaan Pendidikan di SMK Al-Islam Surakarta	*K-Nearest Neighbor*	Hasil akurasi yang didapatkan k=3 sebesar 86%
2.	(Apandi dkk 2019)	Menganalisis kemungkinan keterlambatan pembayaran sumbangan pembinaan pendidikan di Politeknik Tedc Bandung	C4.5	Hasil akurasi terbaik didapatkan sebesar 75%
3.	(Rosiana, 2020)	Menganalisis kemungkinan keterlambatan pembayaran Sumbangan pembinaan pendidikan di SMP Perintis 2 Bandar	SVM	Hasil akurasi terbaik yang didapatkan sebesar 97.8469%, *error rate* 0.021531, *false positive rate* 0.005208, *recall* 0.964602, *specificity* 0.994792, dan *precision* 0.99543
4.	(Muqorobin dkk 2019)	Prediksi keterlambatan pembayaran biaya pendidikan sekolah	Naïve Bayes	Hasil akurasi terbaik didapatkan sebesar 90%

2.3. Algoritma *K-Nearest Neighbor*

Pertama kali secara formal menggunakan metode klasifikasi yang menjadi cikal bakal *K-Nearest Neighbor* adalah Alhazen (Ibn al-Haytham), seorang ilmuan hidup diantara tahun 956 hingga 1040 (Primartha, 2021). Algoritma *K- Nearest Neighbor* bekerja dengan cara mencari nilai *k* objek atau pola data (dari semua

pola training yang tersedia) yang paling terdekat dengan pola masukan dan memilih kelas dengan jumlah pola terbanyak di antara nilai *k* pola tersebut (Suyanto, 2017).

Rumus *Euclidien Distance* :

$$Euclidien\ distance = \sqrt{\sum_{i=1}^{P}(a_{k-}\ b_k)^2} \quad (2.1)$$

Sumber : (Jatmiko Indriyanto, 2021)

Keterangan :

a_k = Sampel data

b_k = Data uji atau *testing*

p = Dimensi data

i = Variabel data

Algoritma *K- Nearest Neighbor* adalah salah satu teknik klasifikasi yang digunakan untuk penyelesaian masalah pada bidang data mining dengan pendekatan untuk memilih kasus baru kemudian menghitung kedekatan dengan kasus lama (Nofiansyah & Gunadi Widi Nurcahyo, 2015). Berikut adalah tabel penelitian terdahulu yang menggunakan algoritma *K-Nearest Neighbor* seperti ditunjukkan pada tabel 2.2.

Tabel 2. 2 Penelitian Tentang *K-Nearest Neighbor*

No	Nama Peneliti	Tema Penelitian	Metode	Hasil
1.	(Farkhina Dwi Utari, Amril Mutoi Siregar, 2020)	Prediksi hasil produksi mesin di PT. Showa Indonesia	*K-Nearest Neighboar*	Hasil akurasi yang didapatkan k=5 sebesar 100%
2.	(Mutiara Ayu Banjarsari, H. Irwan Budiman, 2015)	Prediksi kelulusan tepat waktu mahasiswa di Program Studi Ilmu Komputer Fmipa Universitas Lambung Mangkurat	*K-Nearest Neighboar*	Hasil akurasi yang didapatkan k=5 sebesar 80,00%.
3.	(Sukamto dkk, 2020)	Prediksi Kelompok UKT Mahasiswa Univeristas Riau	*K-Nearest Neighboar*	Hasil akurasi yang didapatkan k=13 sebesar 84,21%.

Tabel 2. 2 Penelitian Tentang *K-Nearest Neighbor* (Lanjutan)

No	Nama Peneliti	Tema Penelitian	Metode	Hasil
4.	(Rani dkk, 2019)	Prediksi kelulusan siswa SMK Anak Bangsa	*K-Nearest Neighboar*	Hasil akurasi yang didapatkan k=5 sebesar 93,55%
5.	(Prasetyawan & Gatra, 2022)	Prediksi Prestasi Mahasiswa Berdasarkan Latar Belakang Pendidikan dan Ekonomi Studi kasus UIN Sunan Kalijaga	*K-Nearest Neighboar*	Hasil akurasi yang didapatkan sebesar 95,85%
6.	(Dinata dkk, 2020)	Prediksi Predikat Prestasi Mahasiswa studi kasus UIN Sultan Syarif Kasim Riau	*K-Nearest Neighboar*	Hasil akurasi yang didapatkan sebesar 82%.
7.	(Salam et dkk, 2020)	Prediksi Mahasiswa non Aktif studi kasus Universitas Dian Nuswantoro Semarang	*K-Nearest Neighboar*	Hasil akurasi yang didapatkan sebesar 97.27%.
8.	(Jasmir, Dodo Zaenal Abidin, Siti Nurmaini, 2017)	Prediksi Masa Studi Mahasiswa di STIKOM Dinamika Bangsa	*K-Nearest Neighboar*	Hasil akurasi yang didapatkan k=10 sebesar 83%
9.	(Irma Darmayanti, Pungkas Subarkah, Luky Rafi Anunggilarso, 2021)	Prediksi Potensi Siswa Putus Sekolah di SMP Banyumas Akibat Pandemi Covid-19	*K-Nearest Neighboar*	Hasil akurasi yang didapatkan sebesar 87.4214%
10.	(Şengür & Turhan, 2018)	Prediksi Tingkat Identifikasi Tindakan Berbasis Guru Tentang Komitmen Organisasi Dan Kepuasan Kerja	*K-Nearest Neighboar*	Hasil akurasi yang didapatkan sebesar 93,6%

Dari uraian penelitian yang telah dilakukan sebelumnya, algoritma *K-Nearest Neighbor* telah digunakan untuk melakukan prediksi dengan berbagai studi kasus. Terdapat sembilan penelitian yang dilakukan di bidang pendidikan dan satu penelitian dibidang industri. Penulis pada penelitian ini membahas tentang prediksi keterlambatan biaya kuliah di Universitas Muhammadiyah Kalimantan Timur dengan menggunakan metode yang sama dari lima penelitian tersebut. Namun, yang menjadi pembeda pada penelitian ini dan penelitian sebelumnya yaitu studi kasus, parameter atribut dan jumlah atribut yang digunakan. Sehingga penelitian ini memunculkan pembahasan penelitain baru yang belum pernah diteliti sebelumya.

2.4. Persiapan Data

Persiapan data bisa dilakukan pembersihan, integrasi, reduksi, penambahan, dan transformasi data, adapun penjelasan tahapannya sebagai berikut (Suyanto, 2017):

1) Pembersihan data

   Pada tahapan pembersihan data ini dilakukan pada entri data yang hilang, data yang salah, dan data yang tidak konsisten dihapus dari data.

2) Integrasi data

   Pada tahapan integrasi data yaitu proses menggabungkan dua atau lebih data dari berbagai sumber agar meningkat akurasi proses kecepatan pada data mining.

3) 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).

4) Penambahan data

Strategi penambahan data adalah meramu algoritma atau eksponensial secara tepat terhadap dua bilangan kecil akan memperbesar jarak keduanya sedangkan eksponensial akan memperkecil.

2.5. *Confusion Matrix*

Confusion matrix memberikan keputusan yang diperoleh selama pelatihan dan pengujian, dan *confusion matrix* juga memberikan penilaian kinerja klasifikasi berdasarkan apakah objek itu benar atau salah (Jatmiko Indriyanto, 2021). Terdapat empat kolom pada *confusion matrix* yaitu :

Tabel 2. 3 Confusion Matrix

		Predicted	
		Negatif	Positif
Aktual	Negatif	***TP***	***FP***
	Positif	***FN***	***TN***

Sumber: (Daqiqil, 2021)

a. TP (True Positive) adalah total data point berlabel yes dan nilainya diidentifikasi benar
b. TN (True Negative) adalah total data point berlabel no dan nilainya diidentifikasi salah.
c. FP (False Positive) adalah total data point berlabel yes dan nilai sebenarnya diidentifikasi salah.
d. FN (False Negative) adalah total data point berlabel no dan nilai sebenarnya teridentifikasi benar.

Confusion Matrix dapat dihitung dengan menggunakan perhitungan sebagi berikut (Daqiqil Id, 2021):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.2)$$

Dimana :

TP = *True Positive*

TN = *True Negative*

FP = *False Positive*

FN = *False Negative*