

BAB II

LANDASAN TEORI

2.1 Kredit

Kredit adalah pemberian uang atau tagihan berdasarkan suatu kontrak atau perjanjian pinjam meminjam antara bank dengan pihak lain, dimana debitur berkewajiban untuk membayar kembali hutangnya dengan bunga setelah jangka waktu tertentu. Kesempatan ini dapat digunakan oleh debitur untuk memenuhi berbagai kebutuhan debitur (Annur, 2018).

Kurniawan dan Kriestanto (2016) menyebutkan penyaluran kredit merupakan kegiatan utama dari bank, karena hal tersebut adalah sumber pendapatan utama bank. Semakin besar pemberian jumlah kredit, maka semakin besar risiko yang ditanggung oleh bank. Kredit yang diajukan oleh debitur dapat berisiko, karena sejumlah besar debitur mengajukan aplikasi dan beberapa debitur kemungkinan memiliki masalah pembayaran, sehingga menyebabkan munculnya kredit macet.

Kredit macet merupakan debitur yang gagal membayar kembali pinjaman kepada perusahaan kartu kredit, tidak mematuhi kontrak yang telah disepakati dan pada akhirnya menyebabkan kerugian bagi instansi itu sendiri (Mardhiyah, Siregar, dan Palupiningsih, 2020). Hal yang dapat dilakukan untuk mengurangi kredit macet adalah dengan melakukan analisis kredit terhadap debitur sebelum menyetujui pinjaman yang diajukan debitur untuk menentukan apakah permohonan kredit akan disetujui. Analisis kredit adalah studi tentang faktor-faktor yang mempengaruhi lancar atau tidaknya pengembalian kredit (Kurniawan dan Kriestanto, 2016). Faktor permasalahan kredit macet yang sering terjadi terutama berasal dari nasabah akibat kegagalan usaha, ketidaktepatan bank dalam menganalisis data calon nasabah yang berkarakter kurang baik yang muncul saat melakukan analisis dan terbatasnya pengetahuan pihak bank saat melakukan analisis kelayakan usaha debitur. Akibatnya, analisis kredit tidak tepat dan munculnya masalah kredit macet (Nirwana, Siregar, dan Rahmat, 2022).

2.2 Data Mining

Data mining adalah proses ekstraksi informasi atau menarik dari data yang ada dalam database besar. *Data mining* juga dikenal sebagai *Knowledge Discovery in Database (KDD)*. *Data mining* didefinisikan sebagai seperangkat teknik yang digunakan secara otomatis untuk mengeksplorasi secara keseluruhan dan membawa ke permukaan hubungan yang kompleks dalam kumpulan data yang sangat besar (Siregar dan Puspabhuana, 2017).

Data mining umumnya digunakan untuk mengelompokkan data, memprediksi, mengestimasi dan menentukan aturan pemetaan untuk data yang ada. Kebutuhan akan data mining didorong oleh banyaknya data yang dapat digunakan untuk menghasilkan informasi dan pengetahuan yang bermanfaat. Informasi dan pengetahuan yang diperoleh dapat digunakan untuk menemukan pola data dalam jumlah yang besar dan kebutuhan untuk mengubah data menjadi informasi yang bermanfaat. Berbagai jenis metode *data mining* disesuaikan dengan kebutuhan (Kristanto, 2014). *Data mining* memiliki akar yang panjang dalam bidang keilmuan seperti kecerdasan buatan (*artificial intelligence*), pembelajaran mesin (*machine learning*), statistik dan basis data. Teknik yang biasa disebutkan dalam literatur *data mining* termasuk *clustering*, *classification*, *association rule mining*, *neural network* dan *genetic algorithm* (Handoko, 2016).

2.3 Pembelajaran Mesin (*Machine Learning*)

Pembelajaran mesin (*machine learning*) adalah salah satu bidang dengan perkembangan tercepat dalam ilmu komputer dan salah satu teknologi yang saat ini paling berpengaruh di dunia. *Big data* adalah salah satu pendorong atau alasan *machine learning* memiliki dampak yang begitu besar. *Machine learning* dapat dimanfaatkan untuk menganalisis dan menemukan pola tersembunyi di dalam data. Pola dan pengetahuan yang tersembunyi terkait masalah dapat digunakan untuk memprediksi kejadian di masa depan dan membuat segala macam keputusan yang kompleks. *Machine learning*

merupakan salah satu alat yang mengubah informasi menjadi pengetahuan (Ardilla, *et al*, 2021).

Machine learning adalah cabang dari studi kecerdasan buatan yang memiliki fokus pada membangun dan menyelidiki sistem sehingga dapat mengambil keputusan dari data yang telah diperoleh. Data diperlukan jika ingin menerapkan teknik *machine learning*. Karena jika tanpa data, maka algoritma pada *machine learning* tidak dapat bekerja. Secara umum, data dibagi menjadi dua jenis, yaitu data latih dan data uji. Data latih diperlukan untuk melatih algoritma, sedangkan data uji diperlukan untuk menentukan kinerjanya ketika menemukan data baru yang belum pernah dilihat oleh algoritma yang dilatih sebelumnya (Fikriya, Irawan, dan Soetrisno, 2017).

Pembelajaran mesin dapat dibedakan menjadi dua kategori yaitu *supervised machine learning* dan *unsupervised machine learning*. *Supervised machine learning* adalah suatu teknik yang dimanfaatkan untuk menentukan korelasi antara variabel input dan variabel target. Klasifikasi merupakan salah satu kategori dari *supervised machine learning* (Savitri, Rahman, Venyutzky, dan Rakhmawati, 2021). Sedangkan *unsupervised machine learning* adalah teknik pembelajaran mesin yang tidak mengharuskan mesin dikendalikan berdasarkan jawaban yang benar. Sebaliknya, mesin dibiarkan bekerja sendiri untuk mencari informasi. *Clustering, association* dan *dimensionality reduction* merupakan salah satu kategori dari algoritma *unsupervised machine learning* (Kristiawan dan Widjaja, 2021).

2.4 Klasifikasi

Klasifikasi dalam *data mining* adalah suatu proses penentuan kesamaan karakteristik suatu kelompok atau kelas (*class*). Klasifikasi *data mining* merupakan salah satu metode yang paling umum. Tujuan dari metode ini adalah untuk mengestimasi suatu kelas objek yang belum diketahui labelnya (Maqfiroh dan Mujiyono, 2022).

Klasifikasi termasuk kedalam pembelajaran dengan jenis *supervised learning*. Selain *supervised learning*, terdapat jenis lain yaitu *unsupervised*

learning yang disebut dengan *clustering*. Dalam *supervised learning*, data latih berisi pasangan data input dan keluaran yang diharapkan, sedangkan pada *unsupervised learning* tidak perlu mencapai target keluaran yang diinginkan (Darujati dan Gumelar, 2012). Terdapat dua tahapan dalam pengklasifikasian data, yaitu tahap pembelajaran dan tahap pengklasifikasian. Tahap pembelajaran merupakan tahap pembentukan model klasifikasi. Sedangkan tahap pengklasifikasian merupakan tahap dimana model klasifikasi digunakan untuk memprediksi label kelas dari suatu data (Sartika dan Sensuse, 2017).

2.5 Preprocessing Data

Preprocessing data merupakan teknik yang digunakan untuk mengubah data mentah menjadi sesuai format yang berguna dan efisien. *Preprocessing data* digunakan karena data mentah yang tidak lengkap dan memiliki format yang tidak konsisten. *Preprocessing data* bertujuan untuk mengubah data input mentah menjadi format yang sesuai untuk dianalisis lebih lanjut (Amalia, 2020). Salah satu masalah paling umum yang dihadapi selama langkah *processing data* adalah menangani adanya nilai yang hilang (*missing value*).

Missing value dapat terjadi ketika bagian dari informasi tidak tersedia pada suatu subjek (kasus), yang dapat disebabkan oleh kesalahan entri data, informasi tentang subjek yang tidak diberikan atau tidak tersedia. Data boleh dihapus apabila persentase data *missing value* melebihi 30%, sedangkan apabila persentase data *missing value* kurang dari 30%, maka data yang hilang diperhitungkan nilai rata-ratanya (*mean*) apabila menggunakan data kuantitatif dan modus apabila menggunakan data kualitatif (Sugesti, Mukid, dan Tarno, 2019).

Preprocessing data yang meliputi pembersihan data (*data cleaning*), integrasi data (*data integration*), seleksi data (*data selection*) dan transformasi data (*data transformation*) dilakukan untuk meningkatkan kualitas dari model klasifikasi. Pembersihan data adalah proses menghilangkan *noise* dan data yang inkonsisten, integrasi data adalah proses menggabungkan beberapa sumber data menjadi basis data, seleksi data adalah pemilihan data yang digunakan untuk

proses analisis data dan transformasi data adalah proses mengubah dan penggabungan data yang digunakan untuk proses analisis. Pada tahap ini, format data diubah menjadi format yang sesuai dengan analisis yang digunakan (Kholifah dan Insani, 2016).

2.6 Data Tidak Seimbang (*Imbalanced Data*)

Data tidak seimbang adalah suatu kondisi dimana sebaran kelas data tidak seimbang, jumlah kelas suatu data lebih sedikit atau lebih banyak jika dibandingkan jumlah kelas data lainnya (Qadrini, Hikmah dan Megasari, 2022). Kelompok kelas data yang lebih sedikit disebut kelompok minoritas (*minority*), sedangkan kelompok kelas data lainnya disebut kelompok mayoritas (*majority*) (Qadrini, Hikmah, dan Megasari, 2022). Tabel derajat ketidakseimbangan pada proporsi kelompok minoritas (Depari, Widiastiwi, dan Santoni, 2022) ditunjukkan pada Tabel 2.1.

Tabel 2.1 Tabel Derajat Ketidakseimbangan Proporsi Kelompok Minoritas

Derajat Ketidakseimbangan	Proporsi Kelompok Minoritas
Ringan	20%-40% dari kumpulan data
Sedang	1%-20% dari kumpulan data
Ekstrim	<1% dari kumpulan data

Prasetya (2022) mengatakan, apabila terdapat data tidak seimbang, ada tiga cara untuk mengatasi ketidakseimbangan tersebut, yaitu: i) Pendekatan tingkat algoritma, yaitu menangani masalah distribusi kelas dengan memodifikasi langkah-langkah proses *learning*; ii) Pendekatan *ensemble*, dimana kelas mayoritas dibagi menjadi beberapa kelompok bagian ukuran yang sama besarnya dengan kelas minoritas, kemudian sub kelompok tersebut dilatih secara terpisah, sehingga masing-masing menghasilkan model *learning* dan model *learning* terbaik ditentukan oleh keputusan mayoritas (*majority voting*); iii) Pendekatan level data, dimana rasio kelas dalam kumpulan data disesuaikan untuk mencapai distribusi kelas yang seimbang. Pendekatan level data terdiri

dari dua teknik, yaitu teknik *Random Oversampling* dan teknik *Random Undersampling*. *Random Oversampling* adalah teknik untuk menggandakan atau menghasilkan sampel di kelas minoritas, sedangkan *Random Undersampling* adalah teknik untuk menghilangkan sampel di kelas mayoritas. Upaya menghilangkan sampel dari kelas mayoritas baik untuk menghindari kemungkinan adanya *noise* yang dapat memperburuk hasil klasifikasi.

2.7 Confusion Matrix

Confusion matrix merupakan metode yang biasa digunakan untuk menghitung tingkat akurasi dalam *data mining*. *Confusion matrix* berisikan informasi terkait klasifikasi yang diprediksi dengan benar oleh suatu sistem klasifikasi (Gunawan, Pratiwi dan Pratama, 2018). *Confusion matrix* dilakukan untuk memprediksikan objek yang benar dan salah (Dinata, Safwandi, Hasdyna dan Azizah, 2020). Contoh tabel *confusion matrix* untuk klasifikasi biner ditunjukkan pada Tabel 2.2 (Normawati dan Prayogi, 2021).

Tabel 2.2 *Confusion Matrix*

		Nilai Prediksi	
		<i>Negative (0)</i>	<i>Positive (1)</i>
Nilai Aktual	<i>Negative (0)</i>	TN	FP
	<i>Positive (1)</i>	FN	TP

dengan:

- TP (*True Positive*) = data *positive* yang terdeteksi dengan benar
- TN (*True Negative*) = data *negative* yang terdeteksi dengan benar
- FP (*False Positive*) = data *negative* yang terdeteksi sebagai data *positive*
- FN (*False Negative*) = data *positive* yang terdeteksi sebagai data *negative*

Muafa dan Iswari (2022) menyebutkan terdapat lima parameter pada *confusion matrix* yang perlu dihitung, yaitu akurasi, presisi, *recall/sensitivity*, *specificity* dan *F1-Score*. Akurasi digunakan untuk mengukur total keseluruhan

dari model klasifikasi yang bernilai *positive*, presisi digunakan untuk mengukur prediksi yang bernilai benar jika model memprediksi *positive*, *recall/sensitivity* digunakan untuk menghitung kebenaran prediksi dari data seluruh data *positive*, *specificity* digunakan untuk mengukur kebenaran prediksi dari data seluruh data *negative* dan *F1-Score* digunakan untuk menghitung rata-rata dari presisi dan *recall/sensitivity*. Mahmud, *et al.* (2021) menyebutkan model dikatakan dapat dengan baik mengklasifikasi status data dengan tingkat akurasi, presisi, *sensitivity/recall specificity* dan *F1-Score* yang keseluruhannya memiliki nilai diatas 50%. Akurasi, presisi, *recall/sensitivity*, *specificity* dan *F1-Score* dapat dihitung menggunakan rumus *confusion matrix* melalui persamaan berikut (Muafa dan Iswari, 2022):

$$Akurasi = \frac{TP+TN}{Total} \times 100\% \quad (2.1)$$

$$Presisi = \frac{TP}{TP+FP} \times 100\% \quad (2.2)$$

$$Recall/sensitivity = \frac{TP}{TP+FN} \times 100\% \quad (2.3)$$

$$Specificity = \frac{TN}{TN+FP} \times 100\% \quad (2.4)$$

$$F1 - Score = 2 \times \frac{Recall \times Presisi}{Recall + Presisi} \times 100\% \quad (2.5)$$

dengan:

TP (<i>True Positive</i>)	= data <i>positive</i> yang terdeteksi dengan benar
TN (<i>True Negative</i>)	= data <i>negative</i> yang terdeteksi dengan benar
FP (<i>False Positive</i>)	= data <i>negative</i> yang terdeteksi sebagai data <i>positive</i>
FN (<i>False Negative</i>)	= data <i>positive</i> yang terdeteksi sebagai data <i>negative</i>