BAB 2

TINJAUAN PUSTAKA

2.1 Prediksi Keterlambatan Biaya Kuliah

Prediksi Keterlambatan merupakan suatu proses memperkirakan secara sistematis tentang sesuatu yang akan terjadi diwaktu mendatang berdasarkan informasi yang relevanpadawaktu-waktu sebelumnya melalui suatu metode klasifikasi (Nurmahaludin, 2014). Cara mengetahui pola klasifikasi yang tepat dan terlambat dengan menentukan indikator mana yang paling berpengaruh terhadap keterlambatan dari biaya kuliah tersebut (Apandi dkk., 2019). Berbagai penelitian prediksi keterlambatan pembayaran biaya kuliah yang telah dilakukan, Penelitian relavan ini ditunjukkan pada tabel 2.1 berikut.

Tabel 2.1 Penelitian Relevan

Penulis	Judul	Metode	Hasil
(Apandi dkk., 2019)	Menganalisis Kemungkinan Keterlambatan Pembayaran SPP (Studi Kasus Politeknik TEDC Bandung)	Algoritma C4.5	Memperoleh tingkat hasil akurasi sebesar 75%.
(Prabowo dkk., 2021)	Teknik Klasifikasi Pembayaran SPP Berdasarkan Tingkat Ketepatan Pembayaran	Algoritma *Naïve Bayes*	diperoleh tingkat akurasi sebesar 63,64%.
(Robi Wariyanto Abdullah,	Prediksi Keterlambatan Pembayaran SPP SMK AL-ISLAM	Alogritma *K-Nearest Neighbor*	Perhitungan memproleh hasil akurasi sebesar 86%.

Kusrini, 2019)			

Tabel 2.1 Penelitian Relavan (Lanjutan)

Penulis	Judul	Metode	Hasil
(Rohmayani, 2020)	*Analysis Of Student Tuition Fee Pay Delay Prediction With Particle Swarm Optimation Optimazation (Case Study :* Politeknik Tedc Bandung*)*	Algoritma *Naïve Bayes*	Hasil pengujian model mendapatkan hasil akurasi sebesar 73,94%.
(Muqorobin dkk., 2019)	*Estimation System For Late Payment Of School Tuition Fees*	Komparasi Algoritma *Naïve Bayes* dan KNN	Naïve Bayes mendapatkan nilai akurasi 85% dan KNN memperoleh akurasi 81%.

2.2 *Data Mining*

Data mining adalah pengumpulan informasi penting dari suatu set data berukuran besar dengan menggunakan teknik tertentu. Informasi yang didapat dari *data mining* tersebut dapat dipakai dalam memperbaiki pengambilan keputusan. (B. Santosa & Umam, 2018).

Istilah dari data mining juga dapat diartikan sebagai penguraian penemuan pengetahuan di dalam database. Data mining sendiri ialah proses yang menggunakan teknis statistik, kecerdasan buatan, matematika, dan machine learning untuk mengidentifikasi informasi mana yang bermanfaat dan pengetahuan yang didapat dari berbagai database besar. Dari penjelasan tentang data mining di atas, maka data mining merupakan pengetahuan yang tersimpan

di dalam database yang telah di proses untuk menemukan bentuk dan teknik machine learning untuk mengidentifikasi informasi pengetahuan yang didapat melalui database tersebut (Utomo & Mesran, 2020)

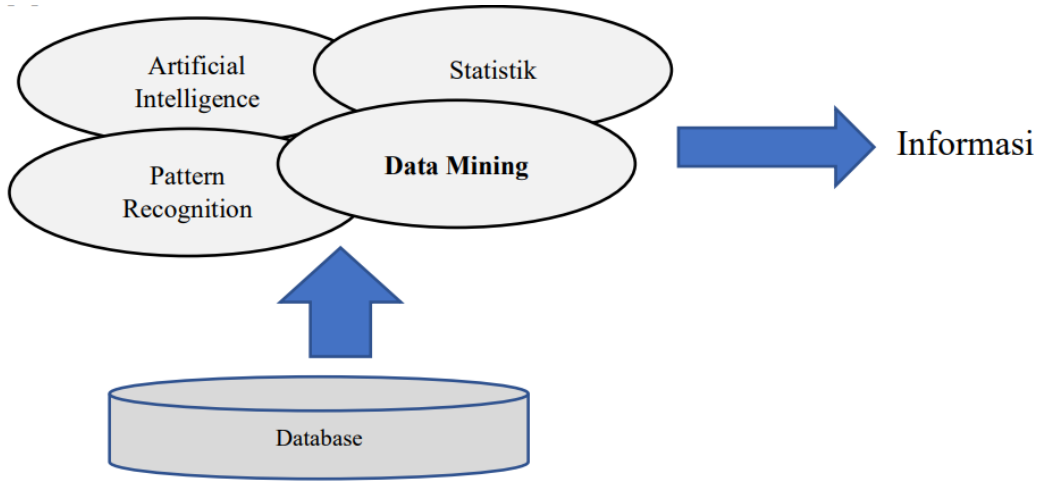


Figure 2.1 Akar Ilmu Data mining

(Sumber: Utomo & Mesran, 2020)

Pada data mining terdapat istilah lain yang mempunyai arti yang sama dengan data mining yaitu *Knowledge Discovery in Database (KDD). Data mining* dan KDD memiliki tujuan yang sama yaitu menggunakan data yang telah ada pada basis data kemudian mengolah data untuk mendapatkan sebuah informasi baru yang bermanfaat. *Knowledge Discovery in Database (KDD)* sendiri merupakan proses ekstraksi pengetahuan, analisis pola / data, arkeologi data, dan pengerukan data.(E. Prasetyo, 2014)

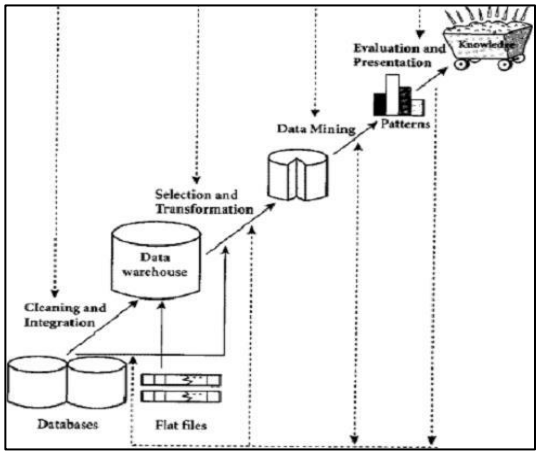


Figure 2.2 Tahapan Proses Knowledge Discovery in Database (KDD)
(Sumber: E. Prasetyo, 2014)

2.3 Teknik Resampling

Resampling ialah teknik yang dilakukan ketika memproses data yang tidak seimbang atau *imbalance*, ketidakseimbangan pada data terjadi ketika pada suatu kelas atau kategori tertentu memiliki data yang lebih banyak dibandingan dengan kelas atau kategori lainnya. Ketidakseimbangan pada data tersebut sangat mempengaruhi akurasi pada proses klasifikasi data (Amelia dkk., 2021)

Dalam melakukan resampling data terdapat 2 teknik yaitu oversampling dan undersampling, cara kerja oversampling yaitu dengan menambahkan sejumlah data pada *minority class,* pada teknik undersampling yaitu dengan cara mengurangi jumlah data pada *majority class* (Indrawati, 2021)

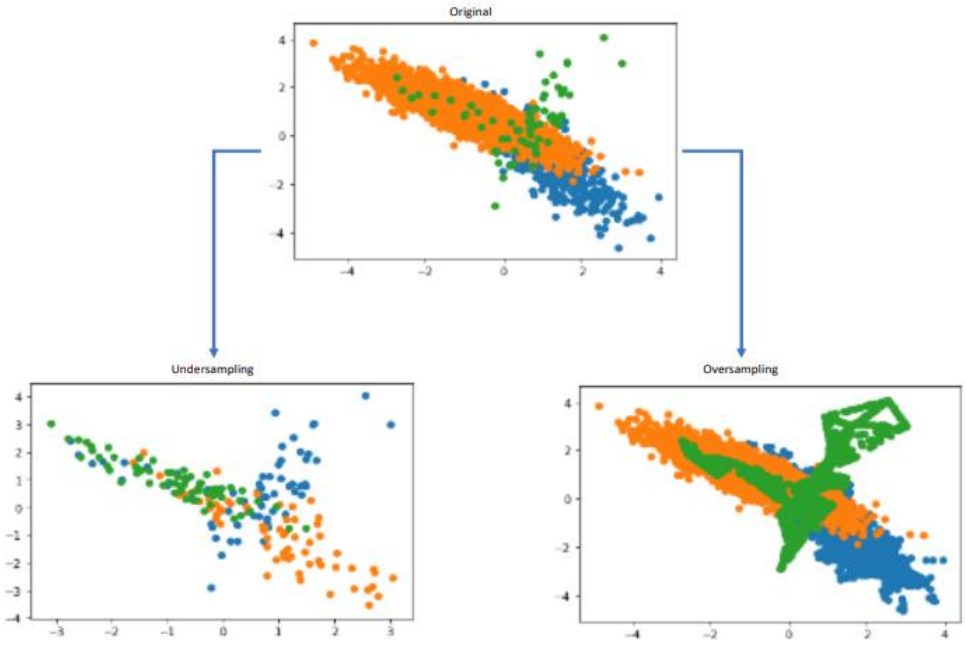
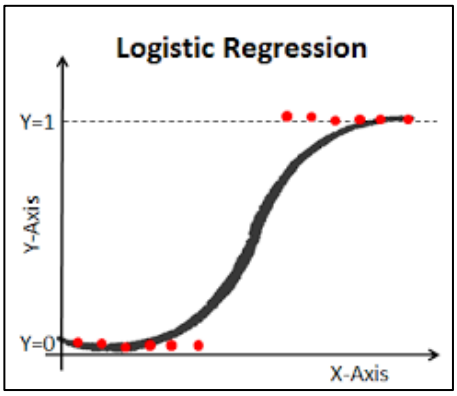


Figure 2.3 Ilustrasi oversampling dan undersampling
(Sumber: Indrawati, 2021)

2.4 *Logistic Regression*

Logistic Regression merupakan hubungan variabel X (Bebas) dan Y (Terikat) tidak mempunyai hubungan yang tidak *linear*. Variabel yang terikat berupa skala dengan dua kategori, misalnya: ya dan tidak, baik dan buruk atau tinggi dan rendah

yang dimaksud dengan *binary classification* menggunakan metode prediksi probalitas (Primartha, 2021).



Gambar 2.1 Pola Kurva
Sumber: (Tyasnurita & Pamungkas, 2020)

Pada gambar 2.1 di atas menunjukkan jika kurva munuju ke positif, nilai y (output) maka akan diprediksi menjadi 1, jika kurva menuju ke negatif, nilai y (output) maka diprediksi menjadi 0. Jika dirumuskan :

$$p \geq 0.5, class = 1 \quad (2.1)$$

$$p < 0.5, class = 0 \quad (2.2)$$

Apabila jumlah variabel tidak dibatasi, Persamaan *Logistic Regression* dinyatakan dengan rumus sebagai berikut :

$$P(Y = 1) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \quad (2.3)$$

Atau

$$logit(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (2.4)$$

Keterangan :

Ln : Logaritma Natural

β_0 : Konstanta

β_1 : Koefisien masing-masing variabel

P : Probabilitas logistik

Mengubah bentuk algoritma (Ln) menjadi exponsial (e) atau probabilitas *Logistic Regression* sebagai berikut :

$$p(Y = 1) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad (2.5)$$

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x \quad (2.6)$$

Rumus didapat melalui buku (Primartha, 2021) yang berjudul Algoritma *Machine Learning.*

Algoritma *Logistic Regression* telah banyak dipakai pada penelitian sebelumnya pada kasus tentang prediksi dengan hasil akurasi yang cukup baik, maka dari itu peneliti mencoba membuat prediksi tentang keterlambatan biaya kuliah. Berikut ini adalah tabel 2.2 penelitian terdahulu tentang prediksi dengan menggunakan metode *Logistic Regression.*

Tabel 2.2 Penelitian Prediksi

Penulis	Judul	Metode	Hasil
(Pambudi dkk., 2020)	prediksi status pengiriman barang menggunakan metode *machine learning*	*Logistic Regression*	Menghasilkan akurasi sebesar 73,81%.
(Maulana dkk., 2018)	*Influence Models For Prediction and Analysis of Diabetes Risk Factors*	*Logistic Regression*	Uji validasi diperoleh akurasi sebesar 94,77%.
(Pakgohar & Kazemi, 2015)	*An Examination of Crash Severity Differences Between Male and Female Drivers*	*Logistic Regression*	Akurasi model yang didapat sebesar 91%.
(Kurniawati, 2018)	Prediksi Kelulusan Mahasiswa Dengan Menggunakan Metode Machine Learning	*Logistic Regression*	Akurasi yang diperoleh adalah 64,73%.
(Nahib, 2016)	Prediksi Spasial Dinamika Areal Terbangun Kota Semarang	*Logistic Regression*	Hasil akurasi model ini adalah 78,21 %.

2.5 *Confusion Matrix*

Confusion Matrix adalah alat pengukuran yang digunakan untuk mengevaluasi kinerja pada model klasifikasi untuk membandingkan nilai sebenarnya dengan nilai hasil prediksi. Ketika sedang melakukan pengklasifikasian dan memiliki data bernilai benar, maka nilai *True-Positive* dan *True-Negative* berfungsi untuk memberikan informasi tersebut. Jika pengklasifikasi memiliki kesalahan saat mengklasifikasi data, maka nilai dari *False-Positive* dan *False-*

Negative akan memberikan informasi tersebut (Primartha, 2021). Tabel *confusion matrix* ditunjukkan dalam tabel 2.3.

Tabel 2.3 Confusion Matrix

Class		Prediction	
		TRUE	*FALSE*
Actual	*TRUE*	TP	FP
	FALSE	FN	TN

Sumber: (Primartha, 2021)

Keterangan :

TP = Jumlah kelas positif yang diklasifikasikan sebagai positif
TN = Jumlah kelas negatif yang diklasifikasikan sebagai negatif
FP = Jumlah kelas negatif yang diklasifikasikan sebagai positif
FN = Jumlah kelas negatif yang diklasifikasikan sebagai negatif

Berikut adalah perhitungan akurasi nya :

1) Nilai akurasi menunjukkan seberapa akurat model dalam melakukan klasifikasi.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.7)$$

2) Nilai presisi menggambarkan akurasi antara jumlah data kategori positif dibagi dengan total data yang diklasifikasi positif

$$Presisi = \frac{TP}{TP+FP} \quad (2.8)$$