BAB 2

TINJAUAN PUSTAKA

2.1 Kinerja

Prawirosentono (1999) dalam buku manajemen kinerja adalah hasil kerja yang dicapai oleh seseorang atau kelompok orang dalam suatu organisasi, sesuai dengan wewenang serta tanggung jawab masing-masing, untuk mencapai tujuan organisasi bersangkutan secara legal, tidak melanggar hukum serta sesuai dengan moral dan etika (Fauzi & Nugroho A, 2020).

2.2 *Data Mining*

Data mining merupakan kegiatan untuk mengekstrak informasi atau pengetahuan yang penting dari suatu set data besar dengan menggunakan teknik tertentu. Informasi yang dihasilkan dari *data mining* bisa digunakan untuk memperbaiki pengambilan keputusan. *Data mining* disebut juga penambangan data karena proses penemuan informasi dalam data set dilakukan dengan cara kegiatan penambangan (Umam, 2018).

Dalam data mining terdapat 3 metodologi yang dapat diterapkan untuk industri maupun penulisan ilmiah. Metode tersebut yaitu *KDD (Knowledge Discovery From Data)*, *SEMMA (Sample, Explore, Modify,Model, And Assess)*, dan *CRISP-DM (Cross Industry Standard Proses For Data Mining)* (Suntoro, 2019)

Terdapat 5 proses pemecahan masalah dalam beberapa pengelompookan data mining yaitu:

a. *Estimatio*n (Estimasi), merupakan teknik melakukan estimasi terhadap data baru yang tidak mempunyai keputusan berdasarkan histori data yang ada.
b. *Forecasting* (Prediksi/Peramalan), merupakan teknik yang digunakan untuk memperkirakan suatu kejadian sebelum peristiwa tertentu terjadi.
c. *Classification* (Klasifikasi), merupakan teknik yang digunakan dengan cara melihat kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini juga dapat memberikan klasifikasi pada data baru dengan cara

memanipulasi data yang sudah diklasifikasi menggunakan hasilnya untuk memberikan sejumlah aturan.

d. *Clustering* (Klastering), merupakan teknik untuk menganalisis pengelompokan yang berbeda terhadap data, sama seperti klasifikasi, akan tetapi pengelompokannya belum di definisikan sebelum dijalankannya *tool data mining*.
e. *Association* (Asosiasi), merupakan teknik yang digunakan untuk mengenali kelakuan dari kejadian khusus atau proses hubungan asosiasi yang muncul pada setiap kejadian (Nofriansyah & Nurcahyo, 2015).

2.3 *Prepocessing Data*

Prepocessing data merupakan teknik untuk mendapatkan hasil analisis yang lebih akurat dalam *machine learning* dan *data mining. Prepocessing data* bermanfaat untuk pengurangan waktu komputasi untuk *large scale problem* dan bisa membuat nilai data menjadi kecil tanpa mengubah informasi yang dikandungnya (Umam, 2018)

Prepocessing Data menurut Nofriansyah & Nurcahyo (2015) merupakan hal terpenting pada data mining, hal ini termasuk antara lain :

a. *Data Selection*

   Merupakan data kasus dalam proses operasional *data mining*. Pada data yang ada, kolom yang diambil adalah hasil yang disebut dengan atribut keputusan, sedangkan kolom yang diambil dalam pembentukan pohon keputusan adalah atribut penentuan.

b. *Data Cleaning*

   Data yang diterapkan untuk menambah isi atribut yang hilang dan merubah data yang tidak konsisten.

c. *Data Transformation*

   Data yang ditransfer ke dalam bentuk yang sesuai pada proses *data mining*.

d. *Data Reduction*

   Data yang dilakukan dengan cara menghilangkan atribut yang tidak di perlukan

hingga ukuran dari database menjadi kecil dan hanya menyertakan atribut yang diperlukan pada proses *data mining*.

2.4 Klasifikasi

Klasifikasi adalah sebuah proses training suatu fungsi tujuan yang digunakan untuk memetakan himpunan atribut suatu objek ke satu dari label kelas tertentu yang telah didefinisikan sebelumnya. Dalam mendeskripsikan data set dengan tipe data suatu himpunan data biner dan nominal, teknik klasifikasi ini sangat cocok untuk di gunakan (Nofriansyah & Nurcahyo, 2015).

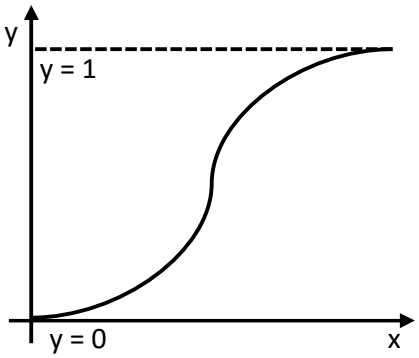
Klasifikasi Muflikhah et al., (2018) merupakan peran pada *data mining* yang menggunakan metode pendekatan prediktif, yang didefinisikan sebagai berikut:

a. Jika terdapat sekumpulan *record (training set)* yang setiap *record* terdiri atas sekumpulan atribut dan satu atribut merupakan kelas.
b. Menentukan suatu model pada atribut kelas sebagai fungsi nilai dari atribut lain.
c. Tujuan *record* yang tidak terlihat sebelumnya di tentukan oleh suatu kelas seakurat mungkin.
d. Kumpulan data uji digunakan untuk menentukan keakuratan model.
e. Data set yang diberikan akan dibagi kedalam sekumpulan data latih dan data uji untuk pengujiannya.

2.5 Algoritma *Logistic Regression*

Logistic regression merupakan algoritma yang dapat memisahkan dataset menjadi dua bagian yang disebut dengan *binary classification* menggunakan metode prediksi probabilitas. *Logistic regression* menghasilkan output yang bersifat kualitatif dan kategori (Primartha, 2021).

Logistic regression masuk dalam kategori *suppervised learning* yang bisa digunakan untuk menyelesaikan berbagai macam permasalahan *binary classification*. Data yang digunakan pada algortima *logistic regression* tunduk pada ketentuan dataset untuk *supervised learning*. Dataset ini berpasangan (*input/output*) yang disebut dengan dataset berlabel (*labeled dataset*).



Gambar 2. 1 *Logistic Regression*

Sumber : Dios Kurniawan (2020)

Grafik ini membagi dataset menjadi dua class = 1 dan class = 0 tepat di tengah, yaitu saat Y = 0.5. Class merupakan prediksi probabilitas (*p* atau *P*) yang di rumuskan:

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

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

Probabilitas regresi logistik sebagai berikut:

$$p = \frac{e^{\beta_0 + \beta_1 + \varepsilon_i}}{1 + e^{\beta_0 + \beta_1 X + \varepsilon_i}} \quad (2.3)$$

Atau

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X + \varepsilon_i)}} \quad (2.4)$$

Di mana:

P : Probabilitas

e : Fungsi exponen

2.6 *Confusion Matrix*

Confusion Matrix merupakan teknik yang digunakan untuk mengetahui seberapa akurat model *classification*. Untuk menganalisis seberapa akurat suatu model *classification* menggunakan tabel *Confusion Matrix* (Primartha, 2021).

Tabel 2. 1 *Confusion Matrix*

		Kenyataan	
		Positif	Negatif
Prediksi	Positif	TP	FP
	Negatif	FN	TN

Sumber: Dios Kurniawan (2020)

Terdapat sejumlah ukuran yang di gunakan untuk menilai atau mengevaluasi model klasifikasi, yaitu:

a. TP (*True Positif)* model sukses memprediksi positif "ya".
b. TN (*True Negative)* model sukses memprediksi negatif "tidak".
c. FP (*False Positif)* model prediksi positif "ya", tetapi salah karena kenyataannya negatif "tidak".
d. FN (*False Negative)* model prediksi negatif "ya", tetapi salah karena kenyataannya positif "ya".

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.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.5)$$

Dimana :

TP = *True Positif*

TN = *True Negative*

FP = *False Positif*

FN = *False Negative*

b. *Precision* (ketepatan)

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

Precision digunakan untuk memberi petunjuk seberapa baik model dapat "menangkap" prediksi yang positif. Semakin banyak FP atau model yang sering salah memprediksi kemunculan data positif, maka angka *precision* semakin rendah.

$$precision = \frac{TP}{Total\ yang\ diprediksi\ Positif} = \frac{TP}{TP+FP} \quad (2.6)$$

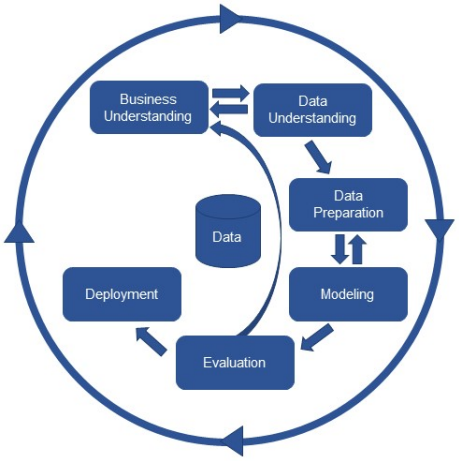
c. *Sensivity/recall*

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

$$sensivity = \frac{TP}{Total\ semua\ Kenyataan\ Positif} = \frac{TP}{TP+FN} \quad (2.7)$$

2.7 *CRISP-DM*

CRISP-DM pertama kali dikenalkan pada tahun 1999 oleh 4 perusahaan besar yaitu, Daimler-Benz (perusahaan pembuat mobil), NRC Crop (produsen perangkat keras dan perangkat lunak), OHRA (penyedia asuransi), dan SPSS (perusahaan pembuat perangkat statistik) (Suntoro, 2019).



Gambar 2. 2 CRISP-DM

Sumber:Taylor (2017)

Terdapat 6 tahap didalam CRISP-DM yaitu:

1. *Business understanding* merupakan tahapan yang berisi tentang cara menentukan tujuan, menilai, dan menetapkan tujuan yang dilakukan pada

data mining.

2. *Data understanding* merupakan kegiatan persiapan, mengevaluasi persyaratan data, dan pengumpulan data.
3. *Data preparation* merupakan data yang perlu diidentifikasi, dibersihkan, dipilih, kemudian dibangun ke dalam format yang diinginkan.
4. *Modeling* merupakan algoritma untuk mencari, mengidentifikasi, dan menampilkan pola.
5. *Evaluation* digunakan untuk membantu pengukuran evaluasi pada model.
6. *Deployment* merupakan tahapan yang digunakan untuk otomatisasi model atau pengembangan aplikasi.

2.8 Peneliti terdahulu

Setiap mahasiswa mempunyai hasil evaluasi yang berbeda ataupun hasil yang sama dengan berbagai faktor penyebab. Seperti penelitian komparasi *support vector machine, logistic regression,* dan *artificial neural network* dalam prediksi penyakit jantung. Dari hasil penelitian didapatkan hasil akurasi tertinggi pada metode algoritma *logistic regression* sebesar 86% menggunakan pembagian data 80:20 (Handayani, 2021).

Prediksi loyalitas pelanggan telekomunikasi menggunakan *logistic regression* dengan seleksi fitur *particle swarm optimization.* Dari hasil penelitian yang di peroleh delapan model regresi logistik, yaitu model A-H. Nilai akurasi tertinggi untuk jalur prestasi ialah model F (dataset 2008-2013) dengan akurasi prediksi 73,73%. Hasil akurasi tertinggi untuk jalur non prestasi ialah dengan model D (dataset 2008- 2011) dan E (dataset 2008-2012) dengan akurasi prediksi 56,76% (Santosa & Artanto, 2015).

Perbandingan Model *Logistic Regression* dan *Artificial Neural Network* pada Prediksi Pembatalan Hotel. Dengan nilai *accuracy* sebesar 79.77%, nilai *precision* 85.86% dan nilai *recall* 55.07% (Putra & Azhar, 2021).

Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode *Grid Seacrh* pada Algoritma *Logistic Regression*. Model algoritma yang di gunakan ialah *Logistic Regression.* Diperoleh model *Logistic*

Regression dengan *Grid Search* pada *Classification Report* memiliki rata-rata akurasi model sekitar 79% dan akurasi data check sebesar 83,33% (Gunawan et al., 2020).

Perbandingan Algoritma *Random Forest Classifier*, *Support Vector Machine* dan *Logistic Regression Clasifier* Pada Masalah High Dimension (Studi Kasus: Klasifikasi Fake News). Nilai performa *Sensitivity* terbaik dari semua skenario percobaan *Logistic Regression Classifier* dengan persentase 99,96 % dibandingkan dengan yang lain (Rini, 2021).

Berdasarkan penelitian terdahulu yang telah di jabarkan di atas terdapat beberapa perbedaan dalam data dan studi kasus yang di lakukan. Dalam beberapa penelitian terdapat hasil akurasi yang tinggi untuk mengetahui prediksi yaitu dalam peneliti Handayani (2021) komparasi *support vector machine, logistic regression,* dan *artificial neural network* dalam prediksi penyakit jantung, dengan nilai *accuracy* 86%. Maka dari itu peneliti ingin mencoba membuktikan algoritma tersebut dalam studi kasus prediksi kinerja mahasiswa dalam perkuliahan daring berbasis *e-learning* apakah akan menghasilkan nilai akurasi yang tinggi dengan menggunakan algoritma *logisitc regression.*