BAB 2
TINJAUAN PUSTAKA

2.1. PENERIMAAN MAHASISWA BARU

Penerimaan mahasiswa baru (PMB) adalah salah satu agenda rutin yang sangat penting di perguruan tinggi. PMB merupakan gerbang awal proses bisnis perguruan tinggi, mencari dan menyeleksi calon mahasiswa yang selanjutnya akan dididik untuk menghasilkan sumber daya manusia yang berkualitas sebagai alumni. Dalam penerimaan mahasiswa baru setiap tahunya pasti ada peningkatan dan penurunan untuk mahasiswa yang masuk di suatu perguruan tinggi, hal itu juga bisa di karena kan akreditasi yang kurang, sehingga masyarakat sekitar kadang kurang percaya terhadap perguruan tinggi tersebut dan memilih perguruan tinggi yang lebih baik. Salah satu cara untuk menunjang akreditasi baik dari perguruan tinggi maupun program studi yaitu dengan menyeleksi mahasiswa yang berkompeten di program studinya, sehingga mahasiswa tersebut dapat memberikan inovasi atau gagasan untuk program studi dan perguruan tinggi tersebut. Maka dari itu perlu adanya seleksi untuk calon mahasiswa baru, yang di harapkan di suatu program studi dapat meningkatkan dari segi sumber daya manusia. Masih banyak hal yang belum diketahui terkait dengan proses pemilihan program studi oleh calon mahasiswa baru, seperti yang di tunjukan sebuah penelitian longitudinal, di sisi mahasiswa pun tak jarang terjadi pergeseran keyakinan akan program studi yang telah dipilih sebelumnya, terutama Ketika mahasiswa yang bersangkutan menyadari bahwa kemampuannya untuk mengikuti pembelajaran di bidang tertentu (Pratama et al. 2022).

2.2. DATA MINING

Data mining adalah suatu proses pengumpulan informasi dan data yang penting dalam jumlah yang besar atau *big data* (Sri Widaningsih 2019) . Dalam proses ini sering kali memanfaatkan beberapa metode, seperti matematika, statistika dan pemanfaatan teknologi *artificial intelligence* (AI). Proses penambangan data terdiri dari beberapa tahapan dan teknik, dari adanya *cleansing*

(pembersihan data), integrasi data, seleksi data dan data *transformation* hingga evaluasi pola dalam mendapatkan informasi dari data itu (Lestari, Sunardi & Fadlil 2022). Selain itu fungsi *data mining* terbagi menjadi dua bagian, yakni deskriptif dan prediktif. Deskriptif sendiri berfungsi untuk memahami lebih jauh tentang data yang diamati, dengan melakukan sebuah proses diharap bisa mengetahui perilaku dai sebuah data tersebut (Amelia Lizsara, Oyama & Wardani 2020). Data tersebut itulah yang nantinya dapat digunakan untuk mengetahui karakteristik dari data yang dimaksud. Sedangkan fungsi *predictive* ialah sebuah bagaimana sebuah proses nantinya akan menemukan pola tertentu dari suatu data, pola – pola tersebut dapat diketahui dari berbagai variabel yang ada pada data (Utomo & Mesran 2020). Ketika sudah menemukan pola maka pola yang didapat tersebut bisa digunakan untuk memprediksi variabel lain yang belum diketahui nilai ataupun jenisnya. Selain dua fungsi tersebut *data mining* juga mempunyai fungsi seperti *characterization, discrimination, association, classification, clustering, outlier and trend analysis,* dan lain lain (Aji Prasetya Wibawa 2018)

2.3. SISTEM PERBANDINGAN METODE

Sistem perbandingan metode adalah sistem yang dirancang untuk mengetahui hasil atau akurasi dari model suatu metode (Alim 2021). Dimana hasil dari metode terbaik akan diambil untuk membuat keputusan suatu masalah. Metode yang di gunakan pada sistem ini adalah dari metode *data mining*, yang di mana itu meliputi *Naïve Bayes,* dan *k-Nearest Neighbor*. Metode di atas di pilih karena dirasa cocok dengan data yang ada pada penelitian kali ini, di karena kan data yang di perolehi nantinya akan di klasifikasikan menjadi 2 yaitu TEPAT JURUSAN dan TIDAK TEPAT JURUSAN.

2.4. METODE NAIVE BAYES

Model yang di gunakan untuk sistem perbandingan model yang pertama adalah model *naïve bayes*, model ini adalah algoritma *machine learning* yang di gunakan dalam berbagai klasifikasi. Untuk bisa memahami algoritma ini bisa di

pahami rumus umum *teorema bayes* yang menjadi dasar *naïve bayes* sendiri (Fatmawati 2016)

Dengan menambahkan bobot pada setiap atribut dalam data set, maka yang menjadi dasar ketepatan klasifikasi bukan hanya probabilitas tiap atribut melainkan juga dari bobot setiap atributnya.

Berikut ini merupakan langkah-langkah dalam menentukan klasifikasi menggunakan metode *Weighted Naïve Bayes*:

Menghitung nilai probabilitas tiap kelas.

Rumus:

$$P(C_i) = \frac{\sum C_i}{n} \quad (2.1)$$

Keterangan:

$P(C_i)$: Probabilitas label kelas C_i

$\sum C_i$: Jumlah data dengan label kelas C_i

n : Jumlah total data latih

Menghitung nilai probabilitas tiap fitur.

Rumus:

$$P(x_k|C_i) = \frac{\sum x_k|C_i}{\sum C_i} \quad (2.2)$$

Keterangan:

$P(x_k|C_i)$: Probabilitas fitur x_k dengan label kelas C_i

$\sum x_k|C_i$: Jumlah data fitur x_k dengan label kelas C_i

$\sum C_i$: Jumlah data dengan label kelas C_i

Menghitung nilai probabilitas tiap kelas pada tiap data.

Rumus:

$$P(C_i|X) = P(C_i) \prod_{k=1}^{n} P(x_k|C_i)^{w_k} \quad (2.3)$$

Keterangan:

$P(C_i|X)$: Probabilitas kelas C_i pada data X

$P(C_i)$: Probabilitas label kelas C_i

$P(x_k|C_i)$: Probabilitas fitur x_k dengan label kelas C_i

w_k : Bobot atribut

Menghitung mean data numerik

(2.4) $$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

standar devisiasi

(2.5) $$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

Normal distribution

(2.6) $$f(x) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

2.5. METODE K-NEARST NEIGHBOR

Menurut Eko Prasetyo (2019), Algoritma *K-Nearest Neighbor* merupakan metode klasifikasi berdasarkan tetangga terdekat dengan konsep sederhana, kuat pada data non-linier, dan dapat digunakan dalam kasus Multi-kelas. Algoritma *K-Nearest Neighbor* merupakan metode klasifikasi terhadap sekumpulan data berdasarkan mayoritas, yang bertujuan untuk mengklasifikasikan objek baru berdasarkan atribut dan kategori yang sama dari sampel *data training* (Permana P, dkk. 2022). *K-Nearest Neighbor* termasuk dalam *supervised learning*, yang mana hasil dari *query instance* baru, diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Hasil dari klasifikasi diambil dari kelas yang paling banyak muncul, yang menjadi kelas hasil klasifikasi (Gorunescu, 2011).

Metode pengukuran jarak *Euclidean* juga paling sering digunakan untuk menghitung kesamaan dari dua vektor. Kelebihan dari metode jarak *Euclidean* ini adalah tingkat kemiripan (*similarity*) lebih tinggi dibanding metode yang lain. Maka rumus perhitungan jarak dengan *Euclidean* seperti di bawah ini :

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (X_{training} - Y_{testing})^2} \quad (2.7)$$

Keterangan :

d (x,y) : jarak *Euclidean*

$X_{training}$: *data training* ke-*i*

$Y_{testing}$: data testing

i : record (baris) ke-*i* dari tabel

n : jumlah *data training*

Langkah – langkah dalam menghitung algoritma KNN:

1. Menentukan nilai K.
2. Menghitung kuadrat jarak *Euclidean* masing-masing label dari data testing terhadap *data training* yang diberikan.
3. Mengurutkan nilai dari hasil perhitungan jarak *Euclidean* data testing terhadap data training mulai dari nilai yang terkecil.
4. Mengumpulkan label *class* Y (klasifikasi *nearest neighbor*).
5. Melihat hasil kategori *nearest neighbor* dengan label mayoritas terbanyak untuk dijadikan label *class* hasil klasifikasi.

2.6. CONFUSSION MATRIX

Confusion Matrix dijelaskan sebagai alat evaluasi yang digunakan untuk mengukur kinerja sistem atau model klasifikasi. Matriks ini terdiri dari empat komponen utama: *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN), yang menggambarkan jumlah prediksi yang benar atau salah dalam mengklasifikasikan data. Dengan menggunakan *Confusion Matrix*, metrik evaluasi seperti akurasi, presisi, *recall*, dan F1-score dapat dihitung untuk

memberikan pemahaman yang lebih komprehensif tentang performa sistem atau model. Matriks kebingungan membantu dalam mengidentifikasi kesalahan klasifikasi dan penting dalam situasi di mana kelas data tidak seimbang atau ketika penekanan khusus diberikan pada jenis kesalahan tertentu.

2.6.1. Akurasi

Akurasi merupakan nilai atau ukuran dari suatu objek yang menentukan tingkat kemiripan dari objek tersebut kepada nilai objek aslinya. Nilai dari sebuah akurasi dalam penelitian dirasa penting karena menjadi ukuran seberapa kuat metode tersebut digunakan dalam penelitian. Sebuah penelitian dapat dikatakan baik apabila memiliki nilai akurasi yang tinggi, jika nilai akurasi yang didapat dirasa kurang penelitian tersebut masih dapat dilanjutkan dengan cara mengubah atau menambahkan metode yang digunakan dengan harapan mendapat nilai akurasi yang lebih baik, di mana nilai tersebut dapat menjadi acuan dalam melakukan penelitian selanjutnya.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \qquad (2.8)$$

Keterangan :

TP : Hasil positif yang diklasifikasikan dengan benar

TN : Hasil negatif yang diklasifikasikan dengan benar

FP : Hasil positif yang diklasifikasikan dengan salah

FN : Hasil negatif yang diklasifikasikan dengan salah

2.6.2. Presisi

Presisi adalah salah satu metrik evaluasi yang digunakan untuk mengukur sejauh mana sistem atau model klasifikasi memberikan prediksi yang tepat dan akurat untuk kelas yang positif. Presisi dihitung dengan membandingkan jumlah prediksi positif yang benar (*true positive*) dengan jumlah total prediksi positif (*true positive* + *false positive*). Presisi memberikan informasi tentang proporsi dari prediksi positif yang sebenarnya benar dari keseluruhan prediksi positif yang dilakukan oleh

sistem atau model. Semakin tinggi nilai presisi, semakin sedikit kesalahan dalam mengklasifikasikan data sebagai positif yang sebenarnya negatif. Presisi adalah metrik yang berguna dalam situasi di mana kesalahan *false positive* perlu dihindari, seperti dalam deteksi penyakit atau spam. Namun, presisi harus dievaluasi bersama dengan metrik evaluasi lainnya, seperti *recall*, untuk mendapatkan gambaran yang lebih lengkap tentang kinerja sistem atau model klasifikasi.

2.6.3. *Recall*

Recall, juga dikenal sebagai sensitivitas atau *true positive rate*, adalah metrik evaluasi yang digunakan untuk mengukur sejauh mana sistem atau model klasifikasi dapat mendeteksi dan mengklasifikasikan dengan benar data yang sebenarnya positif. *Recall* dihitung dengan membandingkan jumlah prediksi positif yang benar (*true positive*) dengan jumlah total data yang sebenarnya positif (*true positive* + *false negative*). Recall memberikan informasi tentang proporsi data positif yang berhasil diidentifikasi oleh sistem atau model. Semakin tinggi nilai recall, semakin sedikit kesalahan dalam mengklasifikasikan data positif sebagai negatif (*false negative*). *Recall* sangat penting dalam kasus-kasus di mana menghindari kesalahan *false negative* sangat kritis, seperti dalam diagnosis medis atau deteksi ancaman keamanan. Namun, *recall* perlu dinilai bersama dengan metrik evaluasi lainnya, seperti presisi, untuk memperoleh gambaran yang lebih komprehensif tentang kinerja sistem atau model klasifikasi.

2.6.4. F1 Score

F1-Score adalah metrik evaluasi yang menyatukan presisi (*precision*) dan *recall* dalam satu angka untuk memberikan gambaran komprehensif tentang kinerja sistem atau model klasifikasi. F1-Score dihitung dengan menggabungkan kedua metrik tersebut menggunakan rata-rata harmonis, yang memberikan bobot yang seimbang antara presisi dan *recall*. F1-Score memberikan informasi tentang sejauh mana sistem atau model dapat mencapai keseimbangan antara akurasi dalam mengklasifikasikan data

positif dan kemampuan untuk mendeteksi dengan baik data yang sebenarnya positif. Nilai F1-Score yang tinggi menunjukkan kinerja yang baik dalam menjaga presisi dan *recall* pada tingkat yang seimbang. F1-Score sangat berguna ketika ada *trade-off* antara presisi dan *recall* yang perlu diperhatikan, seperti dalam kasus deteksi penyakit atau evaluasi sistem keamanan. Namun, F1-Score juga harus dinilai bersama dengan metrik evaluasi lainnya dan disesuaikan dengan konteks penggunaan sistem atau model klasifikasi.

2.7. PENELITIAN SEBELUMNYA

Penelitian sebelumnya yang di lakukan (Saifudin 2018) yang telah melakukan penelitian berjudul "Metode *data mining* untuk seleksi calon mahasiswa pada penerimaan mahasiswa baru di Universitas Pamulang". Pada penelitian ini disimpulkan hasil implementasi dan pengukuran algoritma atau model yang di usulkan diperoleh algoritma atau model terbaik, yaitu *Support Vector Machine* (SVM) dengan akurasi 65% berdasarkan *dataset* yang digunakan.

Penelitian kedua di lakukan oleh (Sri Widaningsih 2019) yang berjudul "Perbandingan metode *data mining* untuk prediksi nilai dan waktu kelulusan mahasiswa program studi Teknik informatika dengan algoritma, *naïve bayes, knn,* dan *svm".* Pada penelitian ini disimpulkan dari hasil perbandingan terlihat bahwa algoritma *Naïve bayes* memiliki nilai yang paling baik untuk semua kategori performansi dibandingkan dengan algoritma lainya. Untuk nilai *accuracy* dan AUC nilai terbesar adalah yang terbaik, sedangkan untuk *error* adalah nilai yang terkecil. Nilai AUC untuk *Naïve bayes* dan C4.5 termasuk ke dalam kategori "baik", sedangkan untuk algoritma *SVM* dan *kNN* termasuk ke dalam kategori "cukup".

Penelitian ketiga di lakukan oleh (Alim 2021) yang berjudul "Implementasi *orange data mining* untuk klasifikasi kelulusan mahasiswa dengan model *k-nearest neighbor, decision tree* serta *naïve bayes".* Pada penelitian ini menunjukan bahwa setelah menggunakan model *K-nearest neighbor, Decision tree* serta *naïve bayes* untuk mengklasifikasi status kelulusan mahasiswa Teknik informatika universitas Islam Madura diperoleh hasil bahwa kinerja *naïve bayes* lebih unggul dari *k-nearst neighbor* serta *decision tree*. Terbukti bahwa dari 35 data uji yang digunakan *naïve bayes* memiliki nilai akurasi 89%, presisi 88%, sedangkan *k-nearst neighbor* memiliki nilai akurasi 77%, presisi 76% dan *decision tree* memiliki nilai akurasi 74% dan presisi 84%. Kontribusi riset ini bisa digunakan oleh manajemen program studi Teknik informatika Universitas Islam Madura untuk mendeteksi sejak awal kondisi mahasiswa supaya kelulusannya tidak terlambat dan mempengaruhi nilai akreditasi program studi Teknik informatika Universitas Islam Madura.

Penelitian keempat di lakukan oleh (Maskuri et al. 2021) yang berjudul "Penerapan Algoritma *k-nearst neighbor* untuk Memprediksi Penyakit Stroke" pada penelitian ini, penulis menggunakan 100 data, dengan pembagian data latih dan data uji menggunakan metode *split validation* dengan perbandingan 80% : 20%. Dengan nilai K yang di gunakan untuk memprediksi stroke yaitu k-9, di mana di dapatkan nilai akurasi sebesar 95%. Sehingga dapat disimpulkan bahwa metode *k-nearst neighbor* memiliki tingkat akurasi yang baik dalam memprediksi penyakit stroke.

Penelitian kelima di lakukan oleh (Putro et al. 2020) yang berjudul "Penerapan Metode *Naive Bayes* Untuk Klasifikasi Pelanggan". Pada penelitian ini di dapatkan bahwa metode naive bayes ini di gunakan untuk mengklasifikasi pelanggan dalam membantu pemilik memberikan bonus, berdasarkan penelitian ini, penulis berhasil mengklasifikasikan 23 data dari 25 data memprediksi pelanggannya dengan nilai precission mencapai 100%, nilai recall mencapai 91%, nilai accuracy mencapai 92%.