BAB III
METODE PENELITIAN

3.1 Alasan Penggunaan Metode

Penggunaan metode klasifikasi dengan algoritma *K-Nearest Neighbor* (K-NN) didasarkan pada efektivitas dan efisiensi algoritma ini dalam menangani data besar dan kompleks yang sering ditemui dalam penerimaan mahasiswa baru. K-NN tidak hanya efisien dalam mengklasifikasikan dan menganalisis data, tetapi juga memiliki keunggulan dalam memprediksi preferensi calon mahasiswa, yang sangat penting untuk pengembangan strategi pemasaran dan program studi di universitas. Fleksibilitas K-NN dalam pemilihan fitur dan penyesuaian parameter memungkinkan penelitian ini untuk disesuaikan dengan kebutuhan data dan pertanyaan penelitian. Selain itu, kemudahan interpretasi hasil yang ditawarkan oleh K-NN sangat penting untuk memastikan bahwa temuan penelitian dapat dengan mudah dipahami dan diterapkan oleh pihak universitas dalam pengambilan keputusan. Sifat K-NN yang sederhana namun efektif ini cocok untuk bekerja dengan data yang memiliki banyak variabel, seperti yang umum ditemui dalam data penerimaan mahasiswa, memberikan penjelasan yang lebih menyeluruh dan tepat mengenai elemen-elemen yang mempengaruhi keputusan penerimaan siswa.

Menurut (P.W Rahayu, dkk. 2024) menjelaskan bahwa tidak ada alasan utama yang mutlak untuk pemilihan algoritma tertentu dalam penelitian ini, pemilihan algoritma sebaiknya disesuaikan dengan karakteristik kasus yang ditemukan. Dalam kasus ini, penggunaan K-NN lebih cocok karena kesederhanaan dan efektivitasnya untuk data dengan banyak variabel yang tidak saling berkaitan. Meskipun K-NN memiliki kelebihan dalam menangani jenis data ini, penting untuk melakukan evaluasi kembali untuk memastikan bahwa algoritma yang dipilih benar-benar yang terbaik. Evaluasi ini melibatkan pengujian dan perbandingan dengan algoritma lain untuk memastikan akurasi dan efektivitas yang optimal dalam konteks data penerimaan mahasiswa baru. Dengan demikian, peneliti dapat memastikan bahwa metode yang digunakan benar-benar

memberikan hasil yang andal dan berguna untuk pengambilan keputusan di universitas.

3.2 Lokasi Penelitian

Penelitian ini difokuskan pada Universitas XYZ, sebuah institusi pendidikan tinggi terletak di salah satu kota terbesar di Indonesia. Universitas ini dipilih karena representatifnya dalam hal keragaman program studi dan populasi mahasiswa. Universitas XYZ dikenal karena sistem administrasinya yang terstruktur dengan baik, terutama dalam pengelolaan data penerimaan mahasiswa. Ini memungkinkan penelitian untuk mengakses data yang dibutuhkan dengan efektif dan efisien. Dengan populasi mahasiswa yang beragam dan meluasnya program studi yang ditawarkan, Universitas XYZ menjadi lokasi yang ideal untuk menganalisis bagaimana minat calon mahasiswa terhadap berbagai program studi berubah seiring waktu dan faktor apa saja yang mempengaruhinya.

3.3 Instrumen Penelitian

Dua komponen utama termasuk dalam instrumen penelitian yang digunakan. Pertama, *Python*, sebuah bahasa pemrograman tingkat tinggi, dijadikan alat analisis data utama karena kemampuannya yang luas dalam pemrograman dan analisis data. Kedua, dataset historis penerimaan mahasiswa baru di Universitas XYZ, yang mengandung informasi seperti data demografis dan pilihan program studi calon mahasiswa. Dataset ini akan diolah dengan *Python* untuk mengidentifikasi pola yang relevan, memberikan wawasan yang penting untuk memahami minat dan pilihan calon mahasiswa.

3.4 Sampel atau Sumber Data

Data untuk penelitian ini berasal dari divisi marketing Universitas XYZ. Sebelum mereka melakukan registrasi secara resmi, data yang digunakan adalah data lead atau target pendaftar. Data ini dianggap memberikan pemahaman yang berharga tentang calon mahasiswa yang telah menunjukkan minat atau niat untuk bergabung dengan universitas sebelum proses pendaftaran formal dimulai.

Variabel-variabel yang diambil, seperti jenis kelamin, jenis sekolah, asal sekolah, jenis seleksi, dan pilihan program studi, dipilih dengan hati-hati karena dianggap memberikan wawasan yang penting tentang preferensi calon mahasiswa dalam memilih program studi.

3.5 Teknik Pengumpulan Data

Teknik pengumpulan data dimaksudkan untuk menjamin kelengkapan, keakuratan, dan relevansi informasi yang digunakan. Langkah pertama dalam proses ini adalah memperoleh akses ke database Universitas XYZ, yang melibatkan prosedur izin yang ketat untuk mematuhi standar etika dan privasi. Setelah akses terjamin, data yang relevan untuk penelitian akan diekstrak, yaitu Data Penerimaan Mahasiswa Baru. Proses ekstraksi diikuti dengan langkah pembersihan dan pemrosesan data untuk menghilangkan kesalahan dan ketidakakuratan, memastikan bahwa data yang dianalisis berkualitas tinggi dan dapat diandalkan. Data ini kemudian dikodekan dan dikategorikan sesuai dengan kebutuhan analisis, seperti demografi dan pilihan program studi.

3.6 Teknik Analisis Data

Untuk memahami pola dan hubungan antar faktor dalam kumpulan data penerimaan siswa baru, Anda perlu menganalisis data tersebut. Teknik analisis data digunakan dalam penelitian ini untuk menemukan informasi yang relevan dan membantu dalam pengambilan keputusan.

Proses klasifikasi menggunakan algoritma *K-Nearest Neighbor*s (KNN) melibatkan beberapa langkah utama untuk memastikan model yang dihasilkan akurat dan andal. Berikut adalah langkah-langkah detail dari proses tersebut:

1. Mengumpulkan Data

   Langkah pertama dalam setiap proyek klasifikasi adalah mengumpulkan dataset yang relevan. Dataset ini harus mencakup fitur-fitur yang akan digunakan untuk prediksi dan label kelas yang

akan diprediksi. Dataset dapat berasal dari sumber publik atau dikumpulkan sendiri melalui eksperimen atau survei.

2. Pra-pemrosesan Data

Data mentah seringkali membutuhkan beberapa tahap pra-pemrosesan sebelum bisa digunakan untuk membangun model KNN.

- Normalisasi/Penskalaan Data: Melakukan normalisasi atau penskalaan pada data karena KNN sangat sensitif terhadap skala fitur. Ini memastikan bahwa semua fitur memiliki kontribusi yang seimbang dalam perhitungan jarak.
- Menghandle *Missing Values*: Menangani nilai yang hilang dalam dataset, misalnya dengan menggantinya menggunakan metode imputasi seperti mean atau median.

3. Memilih Nilai K

Untuk menentukan nilai K yang paling ideal, sejumlah nilai K yang berbeda dieksplorasi, dan validasi silang digunakan. *Overfitting* dapat terjadi pada nilai K yang kecil, sementara *underfitting* dapat terjadi pada nilai K yang besar. Dalam KNN, nilai K-jumlah tetangga terdekat-harus dipilih dengan hati-hati.

4. Mendefinisikan Data Latih dan Data Uji

Informasi dibagi menjadi dua bagian: data uji dan data pelatihan, sebelum model dibangun. Biasanya, pembagian ini dilakukan dengan perbandingan 70:30 atau 80:20. Data yang belum pernah dilihat model selama pelatihan digunakan untuk memastikan evaluasi yang objektif.

5. Menghitung Jarak

Dalam langkah ini, hitung jarak antara setiap titik data uji dengan semua titik data latih. Jarak *Euclidean* sering menjadi pilihan karena kesederhanaannya:

$$Jarak\ Euclidean\ = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Keterangan :

- $X = (x_1, x_2, \dots, x_n)$ adalah vektor fitur dari titik data pertama.
- $Y = (y_1, y_2, \dots, y_n)$ adalah vektor fitur dari titik data kedua.
- $x_1$dan y_1 adalah nilai fitur ke-i dari titik data X dan Y masing-masing.
- $\sum_{i=1}^{n}(x_i - y_i)^2$ adalah jumlah dari kuadrat perbedaan antara nilai-nilai fitur yang sesuai dari kedua titik data.

6. Menentukan K Tetangga Terdekat

K tetangga terdekat dari titik data uji diidentifikasi setelah jarak dihitung. Lokasi-lokasi dalam set pelatihan yang paling dekat dengan set pengujian disebut tetangga.

7. Voting

Setelah dilakukan voting untuk menentukan kelas dari titik data uji, prediksi untuk titik data uji tersebut adalah kelas yang paling sering terlihat antara K tetangga terdekat.

8. Evaluasi Model

Evaluasi performa model adalah langkah terakhir. Mengukur akurasi, presisi, *recall*, dan *F1-score* dilakukan dengan data uji. Ini memberikan gambaran yang jelas mengenai kinerja model. Teknik validasi silang digunakan untuk memastikan konsistensi hasil.