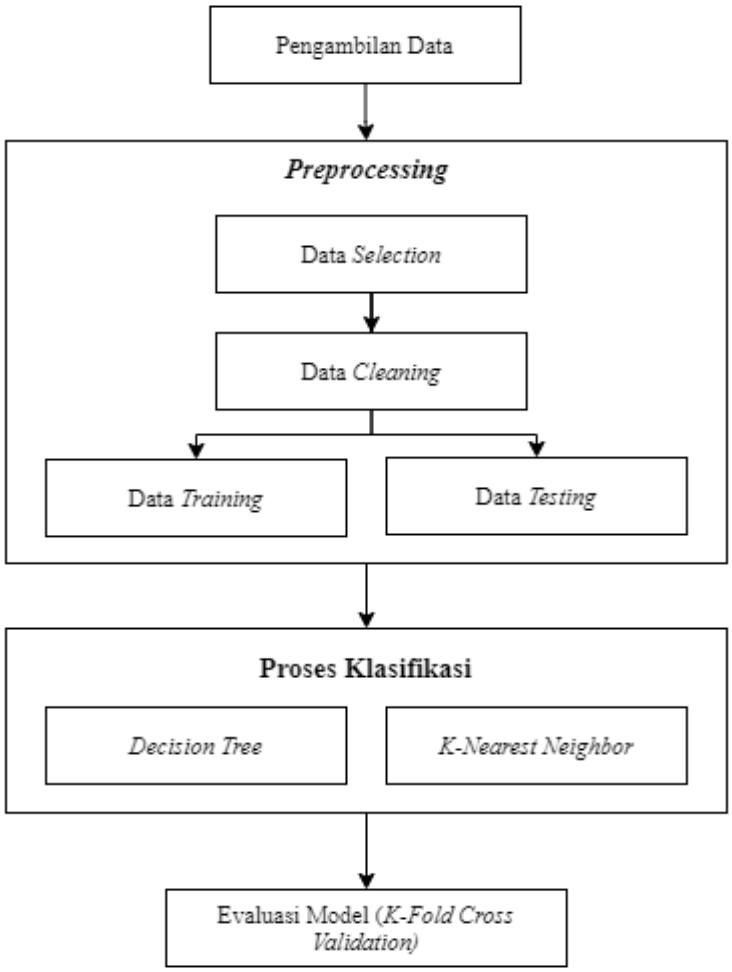
BAB III

METODOLOGI PENELITIAN

Metodologi penelitian menjelaskan tahapan-tahapan yang digunakan dalam penelitian perbandingan akurasi algoritma klasifikasi *Decision Tree* dan *K-Nearest Neighbor* (K-NN). Beberapa tahapan dalam metodologi penelitian yang digunakan yaitu melakukan pengumpulan data, *preprocessing*, proses klasifikasi dan hasil klasifikasi. Pada Gambar 3.1 menunjukkan tahapan penelitian yang akan dilakukan.



Gambar 3.1 Tahapan Penelitian

3.1 Pengumpulan Data

Tahap pengumpulan data ini mencari bahan dasar, yaitu mengumpulkan data pencemar udara ISPU di DKI Jakarta yang diambil dari website Portal Satu Data Indonesia *www.data.go.id*, data Indeks Standar Pencemar Udara (ISPU) wilayah DKI Jakarta tahun 2020 yang terhitung antara tanggal 1 Januari 2017 masa sebelum pandemi COVID-19 hingga 28 Februari 2021 saat pandemi COVID-19 berlangsung.

3.2 *Preprocessing*

Sebelum tahap implementasi diterapkan, tahap *preprocessing* terlebih dahulu dilakukan. Jumlah data awal yang dapat diperoleh dari pengumpulan data, tetapi tidak semua data digunakan dan tidak semua atribut dipergunakan karena data tersebut harus melalui tahap pengolahan awal data atau disebut dengan preparation data.

3.2.1 Data *Selection*

Pada proses seleksi data dari data yang dikumpulkan, dilakukan penyeleksian dengan memilih dan memisahkan data berdasarkan kriteria-kriteria yang ditentukan. Kemudian mengurangi jumlah atribut dan *record* yang ada sehingga didapat data yang tetap informatif.

3.2.2 Data *Cleaning*

Pada tahap ini akan dilakukan pembuangan data yang tidak diperlukan pada data karakteristik kualitas udara yang tidak akan memberikan pengaruh terhadap hasil klasifikasi. Dalam langkah ini, data yang bernilai kosong (*null*) akan dibersihkan dengan cara dihapus secara manual dan mengisi nilai yang telah hilang pada data yang tidak lengkap (*missing value*), dilakukan penghapusan atribut atau mengganti data tersebut, mengidentifikasi atau menghilangkan

outliers dan memperhalus data *noise*, dan memperbaiki ketidak konsistenan data. Pembersihan data dilakukan kemudian dilanjutkan proses transformasi.

3.2.3 Data *Training* dan Data *Testing*

Data yang sudah melalui tahap *preprocessing*, dilakukan proses klasifikasi menggunakan data *training* pada algoritma *Decision Tree* dan *K-Nearest Neighbor*. Sementara pada tahap data *testing* mengalami tahap pembelajaran digunakan untuk melihat performa dan hasil akhir dari model.

3.3 Proses Klasifikasi *Decision Tree* dan *K-Nearest Neighbor*

3.5.1 *Decision Tree*

Pada tahap ini dilakukan penerapan algoritma *Decision Tree* yang di visualisasikan pada pohon keputusan menggunakan rumus Indeks gini yang menghitung menggunakan data titik-titik split. Penerapan indeks gini akan membuat proses klasifikasi pada data semakin detail. Hasil algoritma *Decision Tree* berupa pohon keputusan nantinya dapat digunakan untuk mengeksplorasi data dengan menemukan hubungan tersembunyi antara jumlah calon variable input dengan sebuah variabel target.

3.5.2 *K-Nearest Neighbor*

Tahapan ini akan dilakukan analisa bagaimana penerapan algoritma *K-Nearest Neighbor* menggunakan rumus *euclidean distance* untuk menyelesaikan masalah berupa membagi data *training* dan data testing, kemudian data testing diklasifikasi menggunakan *K-Nearest Neighbor*. Adapun langkah-langkah klasifikasi *K-Nearest Neighbor* bekerja sebagai berikut :

1. Menginisialisasi nilai k dan mengumpulkan semua data yang dibagi menjadi data *training* dan data *testing*.
2. Menghitung jarak antara sampel menggunakan rumus *eucledian distance.* Kemudian mengitung total jarak keseluruhan data.
3. Urutkan jarak dari hasil perhitungan jarak sehingga diperoleh jarak dari terkecil sampai terbesar.
4. Ambil contoh k- tetangga terdekat berdasarkan data sampel yang memiliki jarak terdekat dengan data uji.
5. Klasifikasi label kelas dengan lebih banyak tetangga untuk sampel input.

3.4 Evaluasi Model

Tahapan ini mengevaluasi hasil akurasi dari algortima *Decision Tree* dan *K-Nearest Neighbor* dengan menggunakan perhitungan data *training* data *testing*. Kemudian divalidasi menggunakan *K-Fold Cross Validation.*

Untuk membuktikan kinerja algoritma yang digunakan menggunakan persamaan sebagai berikut:

$$Accuracy = = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \mathrm{x}100\%$$

Akurasi yang dihasilkan dihitung berdasarkan *Confusion Matrix*. Perhitungan pada confusion matrix dihitung sesuai dengan prediksi positif yang benar (*True Positif*), prediksi positif yang salah (*False Positif*), prediksi negatif yang benar (*True Negatif*) dan prediksi negatif yang salah (*False Negatif*). Semakin tinggi nilai akurasi yang didapat maka semakin baik pula metode yang dihasilkan (Maulidah et al., 2020).

Hasil akurasi akan dilakukan validasi menggunakan persamaan sebagai berikut:

$$\text{Akurasi} = \frac{Jumlah\ Klasifikasi\ Benar}{Jumlah\ Data\ Uji}\ x\ 100\%$$

3.5 Kesimpulan dan Saran

Tahapan ini dilakukan guna memberikan kesimpulan yang didapat dari hasil pengujian dan saran untuk pengembangan dari penelitan lebih lanjut.