BAB III: METODOLOGI PENELITIAN

3.1 Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode KDD (*Knowledge Discovery in Databases*). Fokus utama penelitian terletak pada pemanfaatan teknik *Data Mining* dan *Machine Learning* untuk mengekstraksi pengetahuan dari data mentah infrastruktur jaringan. Penggunaan paradigma *supervised learning* melalui algoritma *Decision Tree* memungkinkan transformasi data performa teknis yang kompleks menjadi model klasifikasi yang objektif. Pendekatan ini dipilih karena kemampuannya dalam menangani variabel numerik dalam jumlah besar dan menghasilkan output berupa skema keputusan yang logis serta transparan.

3.2 Dataset Penelitian

Dataset merupakan fondasi dari model yang dibangun. Detail mengenai data yang digunakan adalah sebagai berikut:

- Sumber Data: Data diperoleh secara sekunder dari platform data saintifik, Kaggle, melalui dataset *"Internet Speed Prediction Dataset"*. Data ini mensimulasikan berbagai kondisi trafik pada infrastruktur penyedia layanan internet (ISP) seperti Telkom.
- Volume Data: Memiliki total 5.000 entri, yang memberikan ruang lingkup yang cukup bagi algoritma untuk mempelajari variasi pola gangguan dan stabilitas sinyal.
- Dimensi Data (Fitur): Fitur-fitur yang dianalisis mencakup parameter *Quality of Service* (QoS) sebagai berikut:
    1. Ping Latency (ms): Mengukur durasi perjalanan paket data bolak-balik; indikator utama responsivitas jaringan.
    2. Download Speed (Mbps): Kapasitas maksimum penerimaan data dari server ke pengguna.
    3. Upload Speed (Mbps): Kapasitas maksimum pengiriman data dari pengguna ke server.
    4. Packet Loss Rate (%): Persentase kegagalan pengiriman data yang merepresentasikan stabilitas jalur transmisi.
    5. Signal Strength (dBm): Intensitas daya sinyal nirkabel yang diterima oleh terminal pengguna.
    6. Network Congestion: Indeks kepadatan lalu lintas pada segmen jaringan tertentu.
    7. ISP Quality: Penilaian kualitatif awal yang menjadi fitur pendukung dalam klasifikasi.

3.3 Prosedur Pengolahan Data

Untuk menjamin akurasi model, data diproses melalui alur kerja *pipeline* yang sistematis:

3.3.1 Pra-pemrosesan Data (*Data Pre-processing*)

Tahap ini bertujuan untuk menormalisasi data dan menghilangkan gangguan (*noise*):

- Identifikasi Nilai Kosong: Menggunakan pustaka *Pandas*, dilakukan audit data untuk mendeteksi *missing values* melalui fungsi isnull().sum(). Keberadaan nilai kosong dapat menyebabkan bias pada perhitungan *entropy*.
- Pembersihan Data: Baris yang tidak memiliki informasi lengkap dihapus (*listwise deletion*). Mengingat volume data yang cukup besar, penghapusan ini tidak mengganggu distribusi statistik secara keseluruhan namun meningkatkan integritas model.
- Feature Scaling (Opsional): Meskipun *Decision Tree* tidak terlalu sensitif terhadap skala data, dilakukan verifikasi rentang nilai untuk memastikan tidak ada pencilan (*outliers*) yang ekstrim.

3.3.2 Transformasi Variabel Target (Labelling)

Permasalahan ini merupakan klasifikasi, namun variabel Internet_speed pada dataset awal bersifat kontinu. Oleh karena itu, diterapkan teknik Binning berbasis Quantile (Kuantil). Teknik ini memastikan distribusi kelas yang seimbang, menghindari masalah *class imbalance* yang sering merusak akurasi model.

Rentang Kecepatan Internet (Mbps)	Kategori Kualitas	Dasar Statistik (Persentil)
$< 430,81$	Bad (Buruk)	Persentil ke-0 hingga ke-25 (Q1)
$430,81 – 1951,30$	Enough (Cukup)	Persentil ke-25 hingga ke-75 (Interkuartil)
$\ge 1951,30$	Good (Baik)	Persentil ke-75 ke atas (Q3)

3.3.3 Strategi Pembagian Data (*Data Splitting*)

Untuk menguji validitas generalisasi model, data dibagi menjadi:

- Training Set (70%): Sebanyak 3.500 data digunakan sebagai referensi bagi algoritma untuk membangun struktur pohon dan menghitung *Information Gain*.
- Testing Set (30%): Sebanyak 1.500 data digunakan sebagai simulasi data baru untuk menguji performa model.
- Stratified Sampling: Diterapkan untuk memastikan bahwa setiap "kantong" data (latih dan uji) memiliki persentase kelas Baik, Cukup, dan Buruk yang identik dengan dataset asli.

3.4 Pengembangan Model dan Evaluasi

Proses pemodelan dilakukan dengan langkah-langkah teknis berikut:

1. Konstruksi Algoritma: Menggunakan DecisionTreeClassifier dengan parameter Entropy. Rumus $Entropy(S) = \sum -p_i \log_2 p_i$ digunakan untuk menentukan fitur mana yang paling efektif mengurangi ketidakpastian data di setiap percabangan.
2. Pruning (Pemangkasan): Untuk mencegah pohon tumbuh terlalu spesifik (*overfitting*), dilakukan pembatasan max_depth = 4. Hal ini memastikan model menghasilkan aturan yang dapat dinalar manusia (*interpretable rules*).
3. Metrik Evaluasi: Performa model diukur secara komprehensif menggunakan:

- Accuracy: Persentase total prediksi yang benar.
- Precision: Ketepatan model dalam memprediksi kelas tertentu.
- Recall: Kemampuan model dalam menemukan kembali seluruh anggota kelas tertentu.
- F1-Score: Rata-rata harmonik antara presisi dan recall untuk melihat keseimbangan performa.
- Confusion Matrix: Tabel kontingensi untuk memetakan kesalahan klasifikasi antar kelas secara detail.