BAB I
PENDAHULUAN

I.1 Latar Belakang

Beberapa tahun ini datamining banyak menarik perhatian di bidang ilmu computer, hal ini dikarenakan datamining dapat mengubah data dalam jumlah besar menjadi suatu informasi atau pengetahuan yang berguna. Salah satu metode pada datamining adalah metode klasifikasi. Klasifikasi digunakan untuk mendeskripsikan atau meramalkan kecenderungan dari suatu data pada waktu yang akan datang. Klasifikasi merupakan suatu metode yang mencoba menemukan hubungan antara atribut masukan dan atribut target (Riany & Testiana, 2023).

Salah satu metode klasifikasi pada datamining adalah *Naive Bayes Classifier*, dimana metode ini bertujuan untuk membangun sebuah model klasifikasi menggunakan dataset training yang diproses untuk menentukan suatu kelas dari dataset. *Naive Bayes Classifier* merupakan suatu metode yang mudah dalam pengimplementasiannya dan sederhana (Ramadandi, S, 2020).

(Agung Prabowo, 2021) mengatakan bahwa *Naive Bayes Classifier* adalah suatu metode yang digunakan untuk melakukan klasifikasi dengan statistic dan probabilitas. *Naive Bayes Classifier* memiliki asumsi yang kuat dengan independensi dari kejadian awal atau kondisi, metode ini hanya membutuhkan jumlah data pelatihan atau data training yang cukup kecil untuk menentukan estimasi dan parameter yang diperlukan pada proses klasifikasi. Sedangkan (Purba & Syahputra, 2021) mengatakan bahwa algoritma *Naive Bayes Classifier* adalah suatu metode penghitungan secara statistik untuk memprediksi peluang yang akan datang dengan berdasar kepada pengalaman atau permasalahan yang dihadapi sebelumnya sehingga dikenal sebagai Teorema *Bayes*.

Naive Bayes Classifier lebih tepat digunakan pada dataset yang besar dan dapat menangani data yang tidak lengkap atau *missing value*, serta kuat menangani atribut yang tidak relevan dan noise data, akan tetapi metode ini memiliki kelemahan dimana suatu probabilitas tidak dapat mengukur seberapa besar tingkat akurasi sebuah prediksi (Arfan Haqiqi et al., 2021).

Beberapa metode yang digunakan untuk melakukan seleksi atribut adalah *relief-f* dan *gain ratio*. *Relief-F* adalah suatu metode seleksi atribut yang dikembangkan oleh Kononenko pada tahun 1994 dan merupakan pengembangan dari metode *relief* (Yusra et al., 2021). Metode *Relief-F* ini memberikan estimasi atribut yang efisien dengan menghitung nilai perbedaan jarak antar instan, dimana perbedaan jarak ini digunakan untuk menghitung nilai bobot (Freda et al., 2024).

Selain *Relief-F* metode yang dapat digunakan untuk melakukan seleksi atribut adalah *gain ratio*. *Gain ratio* adalah metode untuk mencari atribut yang relevan dengan menghitung seluruh hubungan atribut terhadap suatu dataset (Edusainstek et al., 2018). *Gain ratio* digunakan sebagai metode seleksi atribut untuk meningkatkan akurasi dari metode KNN, dimana *gain ratio* dapat meningkatkan akurasi dari algoritma KNN dengan mengurangi dimensi fitur dari dataset (Setiyorini & Asmono, 2020).

Sedangkan (Purnamasari, 2021) melakukan penelitian dengan judul penerapan *Naive bayes classifier* untuk pemilihan konsentrasi mata kuliah. Data yang digunakan pada penelitian ini adalah data mahasiswa sebanyak 40 orang, dengan proses pengambilan data menggunakan sample pada sebuah kelas mahasiswa angkatan 2013. Hasil dari penelitian ini dapat menjadi rekomendasi bagi dosen wali untuk membantu mahasiswa dalam memilij konsentrasi mata kuliah berdasarkan dataset yang ada.

(Saelan et al., 2020) dalam penelitiannya tentang komparasi algoritma klasifikasi untuk prediksi minat sekolah tinggi pelajar pada *student alcohol consumption*. Algoritma yang digunakan pada penelitian ini adalah *naive bayes classifier* dan *decision tree*. Dengan menggunakan teknik *cross validation* diperoleh statistic yang menunjukkan bahwa algoritma *decision tree* memiliki kinerja yang lebih baik daripada algoritma *naive bayes*.

(Fitriani et al., 2020) melakukan penelitian tentang seleksi fitur relief pada klasifikasi malware android menggunakan support vector machine. Akurasi klasifikasi Seleksi Fitur Relief-F menghasilkan 33.33333%, hasil akurasi Seleksi Fitur pembanding lainnya juga memberikan hasil sama dengan Seleksi Fitur Relief-F. Sedangkan hasil klasifikasi tanpa Seleksi Fitur memberikan hasil yang cukup

tinggi yaitu 95%. Hasil pengujian menunjukkan bahwa Seleksi Fitur tidak cocok digunakan dengan data yang sedikit karna memberikan hasil yang jauh lebih rendah dari tanpa menggunakan Seleksi Fitur.

Berdasarkan penelitian – penelitian tersebut, peneliti mengusulkan metode seleksi atribut menggunakan r*elief-f* dan *gain ratio* untuk meningkatkan akurasi dari algoritma klasifikasi *Naive bayes*.

I.2 Rumusan Masalah

Rumusan masalah pada penelitian ini adalah sebagai berikut:

1. Bagaimana pengaruh seleksi atribut Relief-F terhadap peningkatan akurasi Naïve Bayes Classifier pada dataset besar.
2. Bagaimana perbandingan tingkat akurasi Naïve Bayes Classifier setelah diterapkan seleksi atribut Relief-F dan Gain Ratio.
3. Sejauh mana penggunaan seleksi atribut Relief-F dan Gain Ratio dapat meningkatkan performa klasifikasi pada dataset House Vote dan Bank Marketing.

I.3 Tujuan

Tujuan dari penelitian ini adalah sebagai berikut:

1. Untuk menganalisis pengaruh seleksi atribut menggunakan metode Relief-F dan Gain Ratio terhadap akurasi algoritma Naïve Bayes Classifier pada dataset dengan karakteristik berbeda, yaitu dataset simbolik (House Vote) dan dataset numerik-kategorikal berskala besar (Bank Marketing).
2. Untuk membandingkan tingkat akurasi Naïve Bayes Classifier setelah diterapkan seleksi atribut menggunakan metode Relief-F dan Gain Ratio.
3. Untuk mengevaluasi sejauh mana metode seleksi atribut Relief-F dan Gain Ratio mampu meningkatkan performa klasifikasi pada dataset House Vote dan Bank Marketing.

I.4 Manfaat

Manfaat penilitian ini diharapkan bisa digunakan sebagai:

1. Memberikan informasi tentang efektivitas metode seleksi atribut Relief-F dan Gain Ratio dalam meningkatkan akurasi klasifikasi menggunakan Naïve Bayes Classifier.

2. Menjadi referensi bagi peneliti dan praktisi dalam memilih metode seleksi fitur yang tepat untuk meningkatkan performa model klasifikasi.
3. Berkontribusi dalam pengembangan penelitian di bidang data mining dan machine learning, khususnya dalam penerapan algoritma klasifikasi dengan optimasi fitur.

I.5 Batasan Masalah

Penelitian ini mempertimbangkan beberapa batasan masalah sebagai berikut:

1. Penelitian ini hanya menggunakan algoritma Naïve Bayes Classifier sebagai metode klasifikasi utama.
2. Seleksi atribut dalam penelitian ini hanya terbatas pada dua metode, yaitu Relief-F dan Gain Ratio.
3. Dataset yang digunakan dalam penelitian ini adalah House Vote (435 data) dan Bank Marketing (45.211 data) dari UCI Machine Learning Repository.

I.6 Sistematika Pembahasan

Struktur dari proposal tesis ini adalah sebagai berikut:

1. Bab 1 Pendahuluan menjelaskan latar belakang masalah yang melandasi penelitian, rumusan masalah, tujuan penelitian, manfaat yang diharapkan, serta ruang lingkup atau batasan masalah. Tujuannya adalah untuk memberikan gambaran awal mengapa penelitian dilakukan dan arah penelitian secara umum.
2. Bab 2 Tinjauan Pustaka berisi kajian literatur yang relevan dengan topik penelitian, termasuk teori-teori, hasil penelitian terdahulu, dan konsep-konsep penting yang menjadi dasar penyusunan kerangka teori dan kerangka pemikiran. Tujuannya adalah untuk menunjukkan pemahaman peneliti terhadap bidang yang diteliti serta posisi penelitian dalam konteks ilmu yang lebih luas.
3. Bab 3 Metodologi Penelitian menjelaskan secara rinci metode yang digunakan dalam penelitian, termasuk jenis penelitian, sumber dan teknik pengumpulan data, populasi dan sampel, variabel penelitian, serta

teknik analisis data. Tujuannya adalah agar penelitian dapat direplikasi dan hasilnya dapat dipertanggungjawabkan secara ilmiah.