BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Data mining berfungsi sebagai teknik yang penting dalam analisis data besar, membantu mengekstrak informasi dan pengetahuan yang bermanfaat dari kumpulan data yang besar dan kompleks [1]. Dalam dunia medis, teknik ini telah menjadi alat penting untuk diagnosa dan analisis tren penyakit, termasuk untuk kanker payudara, yang merupakan salah satu penyebab kematian terbanyak di kalangan wanita di seluruh dunia. Penggunaan metode data mining, seperti algoritma *Naïve Bayes*, menawarkan potensi untuk meningkatkan akurasi dan efisiensi dalam diagnosa penyakit ini, yang pada akhirnya dapat mempengaruhi keputusan pengobatan dan hasil pasien.

Seiring kemajuan pesat teknologi digital, teknik analisis data berkembang menjadi semakin maju, terutama di bidang medis. Salah satu inovasi penting adalah penerapan *Machine Learning*, cabang kecerdasan buatan yang memungkinkan sistem belajar dari data dan membuat penilaian otomatis tanpa perlu diprogram ulang. Dalam dunia medis, *Machine Learning* sangat bermanfaat untuk menganalisis dan memprediksi data besar yang kompleks, seperti diagnosa kanker. Dengan mendeteksi pola tersembunyi dalam data medis, *Machine Learning* memungkinkan pengembangan model prediksi yang lebih akurat dan efisien. Penerapan algoritma *Naïve Bayes* yang didukung teknik seleksi fitur *Information Gain* berpotensi meningkatkan akurasi dan sensitivitas klasifikasi penyakit. Dalam

penelitian ini, *Machine Learning* berperan sebagai alat diagnostik terdepan yang mengoptimalkan pengolahan data klinis, memberikan diagnosa kanker payudara yang lebih cepat dan akurat, serta mendukung keputusan medis yang lebih baik [2].

Kanker payudara adalah masalah kesehatan global utama dengan tingkat morbiditas dan mortalitas yang tinggi [3]. Kemajuan dalam teknologi informasi dan data mining membuka peluang untuk metode diagnostik yang lebih maju. Algoritma *Naïve Bayes*, yang memanfaatkan teorema *Bayes* untuk prediksi berbasis probabilitas, terkenal dengan klasifikasi cepat dan efisien, sangat berguna untuk diagnosa kanker payudara. Dalam penelitian ini memanfaatkan dataset Kanker Payudara Wisconsin (Diagnostik) dari Kaggle dan UCI (*University of California*, Irvine) *Machine Learning Repository*. Dataset ini mencakup fitur penting untuk membedakan antara tumor jinak dan ganas, memungkinkan algoritma *Naïve Bayes* untuk mengidentifikasi metode diagnostik yang lebih cepat dan akurat [4].

Kanker payudara menyumbang 25,5% dari total kasus kanker baru di Indonesia pada tahun 2021, dengan peningkatan diagnosis yang signifikan sejak 2008 data ini diperoleh dari situs resmi Kementerian Kesehatan Republik Indonesia. Dataset Kanker Payudara Wisconsin (Diagnostik) yang digunakan dalam penelitian ini mencakup 569 sampel dengan 31 fitur dari gambar digital aspirasi jarum halus (FNA), yang membantu membedakan tumor ganas (M) atau jinak (B) [5]. Akurasi dalam diagnosa kanker payudara adalah krusial untuk perencanaan terapi yang efektif dan peningkatan hasil pasien, tetapi sering terhambat oleh kompleksitas data. Penelitian ini mengoptimalkan algoritma *Naïve*

Bayes dengan seleksi fitur *Information Gain* untuk meningkatkan klasifikasi kanker. Dalam menghadapi peningkatan kasus kanker di Indonesia, penggunaan metode diagnostik yang cepat dan akurat sangat krusial, dengan algoritma ini penanganan kanker bisa menjadi lebih efektif.

Fokus penelitian ini adalah penggunaan algoritma *Naïve Bayes* untuk klasifikasi diagnosa kanker payudara, dengan peningkatan melalui seleksi fitur *Information Gain*. Penelitian mengevaluasi model berdasarkan data dari sampel kanker payudara Wisconsin (Diagnostik) untuk akurasi, sensitivitas, dan spesifisitas. *Naïve Bayes*, yang mudah diimplementasikan dan efektif untuk dataset besar, membutuhkan pemilihan fitur yang cermat untuk menghindari penurunan akurasi akibat atribut berlebihan. Pemilihan fitur bisa membuat pengklasifikasi baik lebih efisien atau efektif dengan mengurangi jumlah data yang dianalisis. Beberapa metode pemilihan fitur yang sering digunakan antara lain *Document Frequency*, *Mutual Information*, *Information Gain*, dan *Chi-Square*, yang mana *Information Gain* diakui sebagai salah satu algoritma seleksi fitur terbaik.

Menurut Kumar dalam jurnal Lila Dini Utami [6] menyatakan bahwa : "*Information Gain* merupakan salah satu algoritma seleksi fitur yang digunakan untuk memilih fitur terbaik. Hasil dari proses *Information Gain* adalah kata penting yang bersifat informatif."

Sebagai tinjauan atas penelitian-penelitian terdahulu, pada penelitian yang dilakukan oleh Muhammad Ramanda Hasibuan dan Marji [7], metode *Information Gain* berhasil meningkatkan akurasi klasifikasi penyakit gagal ginjal hingga 96,8% menggunakan *Modified K-Nearest Neighbor*, signifikan lebih tinggi dari sistem

tanpa *Information Gain* yang mencapai 79,9%. Penelitian lanjutan oleh Avira Budianita [8] mengaplikasikan *Naive Bayes Classifier* dengan *Information Gain*, menghasilkan peningkatan akurasi prediksi waktu kelulusan mahasiswa dari 81,99% menjadi 83,60%, membuktikan efektivitas seleksi fitur dalam mengurangi kompleksitas data. Rahmanita [9] melanjutkan penggunaan kedua metode tersebut untuk klasifikasi penyakit dan hama tanaman jagung, mencapai akurasi 98,47% dan mempercepat proses diagnosa. Amelia Isnanda [10] meneliti penggunaan dalam analisis sentimen e-wallet selama pandemi, meningkatkan akurasi dari 84% menjadi 92% dengan *Information Gain*. Albet Dwi Pangestu [11] juga menerapkan *Naïve Bayes* dan *Information Gain* untuk analisis sentimen terhadap kebijakan PPKM Darurat, mencatat peningkatan performa dengan akurasi mencapai 81%. Keseluruhan penelitian ini menunjukkan signifikansi *Information Gain* dalam meningkatkan efektivitas algoritma *Naïve Bayes* di berbagai aplikasi data besar dan kompleks.

Berdasarkan kajian penelitian sebelumnya dan analisis permasalahan yang telah dijelaskan, penerapan metode data mining menggunakan algoritma *Naïve Bayes* dengan teknik seleksi fitur *Information Gain* dinilai tepat untuk klasifikasi diagnosa kanker payudara. *Naïve Bayes* dipilih karena efisiensinya dalam mengolah data besar dan kemudahan implementasinya, namun tetap memerlukan optimisasi fitur. *Information Gain* efektif dalam meningkatkan akurasi dengan memilih atribut yang paling relevan, sehingga mengurangi kompleksitas dan mempercepat proses. Penerapan *Information Gain* telah terbukti efektif di berbagai bidang klasifikasi, seperti pada penyakit gagal ginjal, analisis waktu

kelulusan mahasiswa, klasifikasi penyakit dan hama tanaman, serta analisis sentimen e-wallet dan kebijakan publik. Dengan teknik ini, penelitian diharapkan menghasilkan model klasifikasi yang lebih akurat dan efisien, serta mendukung pengambilan keputusan dalam diagnosis kanker payudara secara lebih tepat.

Dari uraian yang melatar belakangi masalah di atas maka yang penulis melakukan penelitian guna memberi solusi terhadap masalah yang terjadi dengan mengangkat judul **"Peningkatan Performa *Naïve Bayes* melalui Seleksi Fitur *Information Gain* Menggunakan *Machine Learning* untuk Klasifikasi Diagnosa Kanker Payudara"**.

1.2 PERUMUSAN MASALAH

Berdasarkan latar belakang di atas maka dapat dirumuskan permasalahan yang akan diteliti yaitu :

1. Bagaimana cara menerapkan algoritma *Naïve Bayes* yang dikombinasikan dengan metode seleksi fitur *Information Gain* untuk mengklasifikasikan diagnosa kanker payudara?
2. Seberapa tinggi tingkat akurasi model klasifikasi diagnosa kanker payudara yang dihasilkan dengan menggunakan kombinasi algoritma *Naïve Bayes* dan seleksi fitur *Information Gain*?

1.3 BATASAN MASALAH

Agar dapat lebih fokus dan pembahasan tidak menyimpang dari permasalahan yang ada, maka penelitian ini membatasi masalah sebagai berikut :

1. Penelitian ini hanya berfokus pada data yang diambil dari web kaggle.com.

2. Metode yang digunakan adalah *Naïve Bayes* dan dikombinasikan dengan metode *Information Gain*.
3. Alat bantu analisa menggunakan bahasa pemrograman *Python*.

1.4 TUJUAN PENELITIAN

Agar dapat mengacu pada permasalahan yang ada pada penelitian ini, penelitian ini menetapkan beberapa tujuan dari penelitian ini yaitu :

1. Untuk menghasilkan model klasifikasi yang efektif dalam mendiagnosa kanker payudara dengan menggunakan algoritma *Naïve Bayes* yang dikombinasikan dengan metode seleksi fitur *Information Gain*.
2. Untuk mengetahui tingkat akurasi model klasifikasi yang dikembangkan dalam mengidentifikasi kanker payudara berdasarkan dataset yang tersedia.
3. Untuk mengevaluasi kinerja algoritma *Naïve Bayes* dengan seleksi fitur *Information Gain* dibandingkan dengan metode klasifikasi lainnya dalam diagnosa kanker payudara.

1.5 MANFAAT PENELITIAN

Peneliti mengharapkan terdapat manfaat yang berguna dari hasil penelitian ini, diantaranya adalah :

1. Meningkatkan keefektifan diagnostik kanker payudara di fasilitas kesehatan dengan menyediakan model klasifikasi yang lebih akurat, menggunakan kombinasi algoritma *Naïve Bayes* dan seleksi fitur *Information Gain*, sehingga dapat membantu dalam pengambilan keputusan medis yang lebih tepat.

2. Memberikan wawasan mendalam mengenai tingkat akurasi dan keandalan model klasifikasi yang dikembangkan, memungkinkan praktisi medis untuk menilai dan mengadopsi teknologi ini.
3. Memberikan perbandingan yang signifikan antara algoritma *Naïve Bayes* yang disempurnakan dengan *Information Gain* dan metode klasifikasi lainnya, berkontribusi pada penelitian lebih lanjut dalam peningkatan metode diagnostik untuk kanker payudara.

1.6 SISTEMATIKA PENULISAN

Untuk memberikan gambaran umum mengenai keseluruhan penulisan ilmiah dapat dilihat melalui sistematika penelitian yang meliputi :

BAB I : PENDAHULUAN

Pendahuluan menjelaskan tentang latar belakang masalah, perumusan masalah, batasan masalah, tujuan dan manfaat penelitian, dan sistematika penulisan.

BAB II : LANDASAN TEORI DAN TINJAUAN PUSTAKA

Bab ini memaparkan landasan teori dan tinjauan pustaka yang relevan dengan penelitian ini, mencakup Data Mining, *Machine Learning*, Klasifikasi, algoritma *Naive Bayes*, seleksi fitur *Information Gain*, dan penggunaan *Python* sebagai alat implementasi. Tinjauan ini juga akan mengidentifikasi persamaan dan perbedaan dengan

penelitian sejenis untuk menonjolkan kontribusi unik dari penelitian ini.

BAB III : METODOLOGI PENELITIAN

Bab ini menjelaskan tentang kerangka kerja penelitian, metode pengumpulan data, dan penjelasan terkait metodologi yang digunakan untuk menyelesaikan masalah dengan memanfaatkan analisis sentimen menggunakan metode *Naïve Bayes* dan *Information Gain*, serta alat bantu yang digunakan semasa mengerjakan penelitian ini.

BAB IV : HASIL PENELITIAN DAN PEMBAHASAN

Bab ini menyajikan analisis metode klasifikasi *Naïve Bayes*, baik tanpa seleksi fitur maupun dengan seleksi fitur menggunakan *Information Gain*, serta hasil pengujian model. Selain itu, akan ditampilkan visualisasi data yang dilakukan menggunakan *Python* untuk menggambarkan performa model dalam klasifikasi diagnosis kanker payudara. Hasil penelitian ini juga akan mencakup perbandingan antara model tanpa seleksi fitur dan model yang menggunakan seleksi fitur.

BAB V : PENUTUP

Bab ini yang berisikan kesimpulan-kesimpulan yang diambil dari hasil analisis serta saran-saran yang mencakup keseluruhan dari hasil penelitian.