BAB I

PENDAHULUAN

1.1 LATAR BELAKANG MASALAH

Machine Learning merupakan cabang dari *Artificial Intelligence* yang berdasarkan pada algoritma dan mampu beroperasi tanpa perlu diprogram secara eksplisit[1]. Dalam konteks ini, kemampuan *Machine Learning* untuk belajar dari data dan meningkatkan kinerja seiring waktu menjadi kunci utama dalam pengembangan aplikasi cerdas yang adaptif . Seluruh pemahaman tentang *Machine Learning* selalu melibatkan pengolahan data[2]. Data merupakan bagian penting dari proses dalam bidang *Artificial Intelligence*, terutama *Machine Learning*. Tanpa data, tahap pelatihan untuk proses pengujian model tidak dapat dilakukan[3]. Penting untuk diingat bahwa kualitas dan kelengkapan data memiliki dampak signifikan terhadap keberhasilan pelatihan model *Machine Learning.*

Pemanfaatan *Machine Learning* sangat luas dan terus berkembang. Pemanfaatan *Machine Learning* telah menjadi elemen lengkap dalam berbagai aspek kehidupan, baik di dunia bisnis, ilmu pengetahuan, teknologi maupun kesehatan[4]. Dalam *Machine Learning*, terdapat empat metode pembelajaran, yakni *Supervised Learning*, *Unsupervised Learning, Semi-Supervised Learning* dan *Reinforcement Learning*[5]. Keempat metode pembelajaran dalam *Machine Learning* ini memberikan kerangka kerja yang beragam untuk mengatasi berbagai jenis masalah.

Supervised Learning merujuk pada metode *Learning* yang digunakan ketika sudah mengetahui hasil yang sedang dicari atau diperkirakan[6]. *Supervised Learning* memiliki beberapa algoritma populer seperti *Back-propagation*, *Linear regression*, *Random Forest, Support Vector Machines*, *Naïve Bayesian*, *Rocchio Method, Decision Tree*, *k-Nearest Neighbor, Neural Network* dan *Logistic Regression*[4]. Salah satu subkategori dalam *Supervised Learning* adalah klasifikasi[7]. Dimana tujuannya untuk mengelompokkan data ke dalam kategori atau kelas yang sudah ditentukan berdasarkan pola atau fitur yang dapat dipelajari oleh algoritma.

Klasifikasi merupakan langkah mengelompokkan objek berdasarkan karakteristik atau ciri yang serupa ke dalam beberapa kelas[8]. klasifikasi memainkan peran kunci dengan mengelompokkan objek berdasarkan ciri-ciri serupa ke dalam kelas-kelas tertentu. Terdapat beberapa jenis pengklasifikasian dalam *Machine Learning*, dan salah satunya adalah metode *Naïve Bayes classifier* [9]. *Naïve Bayes classifier* merupakan sebuah pendekatan statistik yang memungkinkan prediksi probabilitas kepemilikan suatu kelas tertentu[10]. Metode ini didasarkan pada asumsi sederhana yang menyatakan bahwa setiap fitur dalam data bersifat independen, sehingga mempermudah perhitungan probabilitas secara efisien. *Naïve Bayes classifier* memerlukan dataset sebagai elemen kunci dalam melatih dan menguji model. Pemilihan dataset yang representatif dan diversifikasi merupakan faktor kritis dalam memastikan performa yang optimal untuk algoritma ini.

Salah satu teknik *Supervised Learning* dalam *Machine Learning* yang digunakan untuk mengklasifikasikan atau melakukan regresi adalah pendekatan *Decision Tree*[11]. Yang bekerja dengan membagi dataset menjadi subset yang semakin kecil berdasarkan serangkaian keputusan, membentuk struktur berhirarki yang dapat dengan mudah diinterpretasikan. Teknik tersebut mampu melakukan klasifikasi pada data baru dengan memanipulasi data yang telah diklasifikasikan sebelumnya, kemudian menggunakan informasi tersebut untuk menetapkan sejumlah aturan[12]. Yang memungkinkan sistem untuk mengambil keputusan dengan lebih tepat berdasarkan pola yang ditemukan dalam data.

Untuk keberhasilan dalam menerapkan *Naïve Bayes classifier* dan *Decision Tree*, pemilihan dataset yang sesuai menjadi hal yang sangat penting. Dataset merupakan kumpulan data yang dapat digunakan sebagai materi eksperimen penelitian[13]. Dalam penelitian ini menggunakan dataset yang bersumber dari *Repository Kaggle* bernama "*Stroke Prediction Dataset"* yang terdiri dari 5110 data, 11 atribut dan 1 label. Di dalam label terdapat 2 kelas yaitu "0" dan "1" yang berarti "tidak" dan "iya". Pada analisis kesehatan, dataset menjadi instrumen yang tak ternilai dalam mengeksplorasi gejala dan penyakit yang mungkin berkaitan dengan faktor risiko tertentu. Salah satu penyakit yang sangat diperdalam melalui penggunaan dataset adalah penyakit *Stroke*. Untuk mengolah dataset *Stroke Prediction* diperlukan sebuah teknik klasifikasi seperti *Naïve Bayes classifier* dan *Decision Tree*.

Memilih dataset *Stroke Prediction* dari *Repository Kaggle* memberikan keunggulan karena dataset tersebut telah melalui berbagai tahap validasi sehingga

dapat dipercaya dan siap digunakan untuk analisis yang lebih mendalam. Selain itu, *Kaggle* juga menyediakan beragam dataset yang kaya akan informasi dan mencakup berbagai domain, sehingga memungkinkan untuk menemukan dataset yang sesuai dengan kebutuhan spesifik penelitian. Dengan menggunakan dataset yang kredibel dan relevan seperti dataset *Stroke Prediction, Naïve Bayes classifier* dan *Decision Tree* dapat diterapkan dengan lebih efektif untuk menganalisis faktor-faktor risiko yang berkaitan dengan penyakit *Stroke*, yang pada gilirannya dapat memberikan kontribusi besar dalam pengembangan strategi pencegahan dan pengobatan lebih lanjut.

Penelitian ini penting karena penyakit *Stroke* merupakan salah satu penyebab utama kematian dan kecacatan di seluruh dunia[14]. Dengan populasi yang semakin menua dan gaya hidup yang tidak sehat, jumlah kasus *Stroke* cenderung meningkat. Oleh karena itu, perbandingan kinerja antara algoritma *Naïve Bayes classifier* dan *Decision Tree* dalam mengklasifikasi penyakit *Stroke* dapat memberikan wawasan yang berharga untuk pengembangan sistem klasifikasi yang lebih efektif dan cepat dalam diagnosis dini, penanganan, serta pencegahan penyakit *Stroke*.

Selain itu, penerapan teknik *Naïve Bayes classifier* dan *Decision Tree* telah terbukti dalam penelitian sebelumnya, penerapan teknik *Naïve Bayes classifier* dijumpai pada penelitian yang dilakukan Srinivas pada tahun 2023 untuk deteksi penyakit *Stroke* menggunakan metode *Naïve Bayes classifier.* Akurasi yang dihasilkan dari metode *Naïve Bayes Classifier* ini sebesar 77,40%[15]. Elias Dritas juga melakukan penelitian pada tahun 2022 dengan menggunakan beberapa

algoritma dalam *Machine Learning* salah satunya *Naïve Bayes classifier* untuk *Classification Stroke Risk Prediction* menunjukan akurasi sebesar 84%[16]. Penelitian dengan metode yang sama dilakukan Dinda pada tahun 2022 yang membandingkan *k-Nearest Neighbor* dan *Naïve Bayes Classifier classifier* untuk klasifikasi penyakit *Stroke.* Hasil dari penelitian ini *Naïve Bayes classifier* mendapatkan akurasi sebesar 74,45%[17]. Hal ini menegaskan bahwa kemampuan model *Naïve Bayes classifier* dalam mengklasifikasikan data dapat menghasilkan presentase dengan baik.

Penggunaan algoritma *Decision Tree* untuk klasifikasi penyakit *Stroke* ditemukan juga pada penelitian yang dilakukan oleh Harshitha K V pada tahun 2021 menunjukan akurasi pada algoritma *Decision Tree* sebesar 91%[18]. Akurasi sebesar 93% dihasilkan dalam penelitian yang dilakukan oleh MD. Monirul islam pada tahun 2021 penelitian tersebut bertujuan untuk prediksi *Stroke* menggunaan tekik klasifikasi *Machine Learning*[19]. Kelvin Leonardi Kohsasih juga melakukan klasifikasi penyakit *stroke* dengan mengkomparasi algoritma *Naïve Bayes classifier* dan *Decision Tree,* algoritma *Decision Tree* mencapai akurasi sebesar 95%[20]

Kemajuan yang signifikan dalam performa klasifikasi *Naïve Bayes classifier* dan *Decision Tree* menjadi dasar mengapa penelitian ini akan mengkomparasi kedua algoritma tersebut untuk pengklasifikasian penyakit *stroke*. Penelitian ini berjudul "**Komparasi Algoritma *Naïve Bayes* Dan *Decision Tree* Dalam Melakukan Klasifikasi Penyakit *Stroke* Pada Dataset *Stroke Prediction***".

1.2 RUMUSAN MASALAH

Dengan mempertimbangkan uraian latar belakang yang telah disampaikan, masalah dapat dirumuskan sebagai berikut :

1. Bagaimana penerapan serta komparasi algoritma *Naïve Bayes Classifier* dan *Decision Tree* untuk mengklasifikasi penyakit *Stroke* pada dataset *Stroke Prediction*?
2. Seberapa akurat algoritma *Naïve Bayes Classifier* dan *Decision Tree* dalam mengklasifikasi penyakit *Stroke* berdasarkan dataset *Stroke Prediction*?

1.3 BATASAN MASALAH

Dalam penelitian ini, tujuan dari masalah adalah untuk memastikan bahwa pembahasan tetap terfokus pada aspek inti dari permasalahan yang telah dirumuskan. Berikut adalah batasan masalah yang telah ditetapkan dalam penelitian ini :

1. Data yang digunakan dalam penelitian ini merupakan data *Public* yang tersedia di *Repository Kaggle* .
2. Variabel dependen dalam penelitian ini adalah hasil klasifikasi penyakit *Stroke*, yang dapat berupa "*Stroke*" atau "*non-Stroke*".
3. Pada penelitian ini menggunakan algoritma *Naïve Bayes classifier* dan *Decision Tree* sebagai pengklasifikasian terhadap dataset *Stroke Prediction* sebanyak 5110 data.

4. Evaluasi performa algoritma *Naïve Bayes slassifier* dan *Decision Tree* dalam penelitian ini didasarkan pada akurasi, presisi, *recall*, dan *f1-score.*
5. Alat bantu analisa dengan menggunakan bahasa pemrograman *Python.*

1.4 TUJUAN PENELITIAN

Tujuan penelitian adalah memecahkan permasalahan yang tergambar dalam latar belakang dan rumusan masalah. Adapun tujuan yang hendak dicapai sebagai berikut :

1. Untuk melakukan penerapan dan komparasi algoritma *Naïve Bayes classifier* dan *Decision Tree* untuk mengklasifikasi penyakit *Stroke* pada dataset *Stroke Prediction.*
2. Untuk mengetahui tingkat akurasi yang dihasilkan oleh algoritma *Naïve Bayes classifier* dan *Decision Tree* dalam melakukan klasifikasi penyakit *Stroke* berdasarkan dataset *Stroke Prediction*.

1.5 MANFAAT PENELITIAN

Berikut adalah beberapa manfaat penelitian ini :

1. Meningkatkan pemahaman tentang penggunaan algoritma *Naïve Bayes classifier* dan *Decision Tree* dalam klasifikasi penyakit *Stroke*.
2. Menilai algoritma yang terbaik antara *Naïve Bayes classifier* dan *Decision Tree* dalam mengklasifikasi penyakit *Stroke* berdasarkan dataset *Stroke Prediction.*

1.6 SISTEMATIKA PENULISAN

Untuk memudahkan pemahaman dalam menyusun laporan ini, penulis telah menyediakan sistematika penulisan berikut agar dapat diikuti:

BAB I : PENDAHUlUAN

Dalam bab ini, akan menguraikan informasi mengenai asal mula permasalahan, perumusan masalah, batasan masalah, tujuan, dan manfaat penelitian, yang menjadi dasar pelaksanaan penelitian ini.

BAB II :LANDASAN TEORI

Bagian landasan teori ini mencakup konsep-konsep yang dikemukakan oleh para pakar dan relevan dengan topik yang akan dibicarakan. Isi landasan teori ini melibatkan teori-teori *Machine Learning*, proses klasifikasi, metode *Naïve Bayes classifier*, *Decision Tree,* dataset, penyakit *Stroke*, bahasa pemrograman *Python*, dan penelitian sebelumnya yang serupa.

BAB III :METODOLOGI PENELITIAN

Dalam bab ini akan menggambarkan setiap tahap prosesnya. Bab ini akan membahas pendekatan penelitian, pengumpulan data dan sumber data, *Data Preparation*, pembagian *Training* dan *Testing* menggunakan *Cross Validation* dan tahapan implementasi algoritma *Naïve Bayes classifier* dan *Decision Tree*.

BAB IV : ANALISIS DAN HASIL

Dalam bab ini, akan dilakukan penerapan dan pengujian akurasi dan komparasi menggunakan teknik *Naïve Bayes classifier* dan *Decision Tree* pada dataset "*Stroke Prediction*" serta penghitungan menggunakan *confusion matrix* selain itu, pengujian akan dilakukan dengan memanfaatkan bahasa pemrograman *python*.

BAB V : PENUTUP

Dalam bab ini, akan memasuki tahap akhir dari penelitian. Bab penutup ini akan menguraikan rangkuman dari hasil penelitian dan memberikan rekomendasi yang tepat dan dapat memberikan manfaat bagi pihak-pihak yang terlibat dalam proses penelitian ini.