

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pertumbuhan eksponensial data medis dalam era digital saat ini telah mengubah lanskap pemrosesan dan analisis data di bidang kesehatan. Data medis mencakup berbagai atribut yang sangat beragam, termasuk data citra medis, data laboratorium, data catatan medis elektronik, dan banyak lagi. Dalam konteks ini, teknologi Machine Learning telah muncul sebagai alat penting dalam memahami, mengelola, dan mengambil keputusan berdasarkan data medis. Machine Learning memiliki potensi untuk meningkatkan diagnosis penyakit, meramalkan perkembangan kesehatan pasien, serta mendukung pengambilan keputusan klinis yang lebih baik.

Sejauh ini puskesmas memanfaatkan data rekam medis untuk membuat kebijakan berupa penyuluhan kesehatan kepada masyarakat, dengan memberikan informasi jumlah pasien yang berobat, jenis penyakit yang diderita, serta laporan keluarnya pasien. Namun, terkait pola kecenderungan penyakit yang diderita oleh masyarakat belum digali lebih dalam, agar dapat digunakan sebagai acuan dalam penyuluhan kesehatan, sehingga lebih tepat sasaran.[1]

Dalam konteks algoritma Machine Learning, Regresi, Random Forest, dan Naive Bayes telah terbukti efektif dalam berbagai aplikasi. Namun, dalam konteks klasifikasi data medis, belum ada kajian komprehensif yang membandingkan kinerja ketiga algoritma tersebut. Oleh karena itu, penelitian ini bertujuan untuk memberikan pemahaman mendalam tentang keefektifan Regresi, Random Forest, dan Naive Bayes dalam mengatasi kompleksitas data medis.

Regresi logistik ialah metode analisis data yang memanfaatkan konsep matematika untuk mengidentifikasi keterkaitan antara dua faktor data. Dengan menggunakan keterkaitan ini, metode ini dapat memprediksi nilai salah satu faktor berdasarkan faktor yang lain.[2] Dalam konteks klasifikasi data medis, regresi logistik sering digunakan untuk memprediksi probabilitas suatu kondisi medis

berdasarkan data pasien. Misalnya, dokter mungkin ingin memprediksi apakah seorang pasien memiliki risiko tinggi atau rendah terkena penyakit jantung berdasarkan berbagai faktor seperti usia, tekanan darah, kolesterol, dan gaya hidup. Dengan menggunakan regresi logistik, algoritma dapat menghasilkan model yang memetakan faktor-faktor ini ke dalam probabilitas risiko, yang kemudian diklasifikasikan sebagai risiko tinggi atau rendah. Hal ini membantu dalam pengambilan keputusan klinis dan penanganan pasien yang lebih baik, memungkinkan intervensi dini dan perawatan yang lebih personal. Algoritma ini menjadi sangat berguna dalam pengolahan data medis karena kemampuannya untuk menangani data yang kompleks dan memberikan prediksi yang dapat diandalkan.

Random Forest adalah salah satu algoritma pembelajaran mesin yang terkenal karena kemampuannya menghasilkan prediksi yang akurat dengan menggabungkan banyak pohon keputusan.[3] Algoritma ini termasuk dalam metode bagging, di mana beberapa model decision tree dibangun dengan menggunakan subset data yang diambil secara acak dari dataset yang sama. Proses ini bertujuan untuk menciptakan variasi antar pohon keputusan sehingga meningkatkan keandalan dan mengurangi risiko overfitting yang sering terjadi pada decision tree tunggal.[4] Setelah semua pohon keputusan terbentuk, hasil prediksi dari setiap pohon digabungkan menggunakan metode voting untuk klasifikasi atau rata-rata untuk regresi.

Algoritma Naive Bayes (NBC) adalah salah satu metode klasik dalam klasifikasi berbasis Bayesian yang memiliki keunggulan berupa struktur sederhana dan efisiensi komputasi yang tinggi. Keistimewaan algoritma ini terletak pada kemampuannya untuk melakukan estimasi parameter, seperti rata-rata dan varians variabel, dengan hanya membutuhkan data pelatihan dalam jumlah kecil[5]. Dengan asumsi bahwa setiap variabel bersifat independen, algoritma ini cukup menghitung estimasi untuk masing-masing variabel secara individu tanpa memerlukan matriks kovarians secara keseluruhan[6]. Naive Bayes memprediksi

kejadian di masa depan berdasarkan pola yang telah terbentuk dari pengalaman di masa lalu.

Kondisi kesehatan pasien sering kali tercermin dalam beragam atribut dan pola yang kompleks, termasuk data citra medis, hasil tes laboratorium, dan rekam medis elektronik. Oleh karena itu, keberhasilan algoritma klasifikasi dalam mengenali dan menginterpretasi pola ini dapat menghasilkan dampak positif terhadap diagnosis penyakit, pemantauan kondisi pasien, dan pengambilan keputusan klinis.

1.2 Rumusan Masalah

1. Bagaimana perbedaan kinerja algoritma regresi linear, Random Forest, dan Naive Bayes dalam klasifikasi data medis?
2. Bagaimana perbedaan recall antara Regresi, Random Forest, dan Naive Bayes dalam mengklasifikasikan data medis?
3. Bagaimana perbedaan F1-score antara Regresi, Random Forest, dan Naive Bayes dalam mengklasifikasikan data medis?
4. Manakah algoritma yang memiliki nilai akurasi paling tinggi antara Regresi, Random Forest, dan Naive Bayes dalam mengklasifikasikan data medis?

1.3 Batasan Masalah

1. Penelitian ini akan membatasi jenis data medis tertentu. Jenis data medis yang digunakan adalah data laboratorium, Data pasien dan demografis
2. Analisis akan difokuskan pada algoritma Regresi Linear, Random Forest, dan Naive Bayes.
3. Metrik evaluasi yang digunakan adalah akurasi, sensitivitas, spesififikasi.
4. Fokus analisis kinerja algoritma ini akan terbatas pada tugas klasifikasi data medis.

5. Implementasi dan pengujian algoritma akan dilakukan menggunakan bahasa pemrograman Python

1.4 Tujuan Penelitian

Tujuan utama dari penelitian ini adalah:

1. Menganalisis kinerja algoritma Regresi Logistic, Random Forest dan Naive Bayes untuk klasifikasi data medis penyakit STROKE
2. Membandingkan kinerja ketiga algoritma ini untuk mengidentifikasi yang paling efektif dalam berbagai konteks klasifikasi data medis penyakit STROKE.
3. Menganalisis keunggulan dan kelemahan masing-masing algoritma dalam konteks klasifikasi data medis, sehingga dapat diambil keputusan yang lebih baik dalam pemilihan algoritma.
4. Menguji coba matriks evaluasi terhadap diagnosis penyakit berdasarkan data medis penyakit STROKE dengan menggunakan algoritma Regresi Logistic, Random Forest dan Naive Bayes.

1.5 Manfaat Penelitian

Manfaat dari penulisan proposal penelitian ini antara lain :

1. Menyediakan analisis mendalam terhadap kinerja algoritma Regresi Logistic, Random Forest dan Naive Bayes dalam klasifikasi data medis penyakit STROKE.
2. Membantu pengambilan keputusan dalam penanganan pasien.
3. Meminimalkan kesalahan dalam klasifikasi data medis.
4. Menjadi referensi untuk penelitian selanjutnya dalam bidang klasifikasi data medis.
5. Membantu puskesmas dan instansi kesehatan dalam meningkatkan efisiensi proses skrining dan deteksi dini penyakit.

6. Mendukung pengambilan keputusan berbasis data di puskesmas dan instansi kesehatan untuk perencanaan kebijakan kesehatan yang lebih tepat.
7. Mempermudah tenaga medis dalam mengelompokkan pasien berdasarkan tingkat risiko penyakit, sehingga penanganan dapat lebih terstruktur.

1.6 Sistematika Penulisan

Penulisan pada penelitian ini mencakup 5 bab pokok pembahasan

BAB I PENDAHULUAN, berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan

BAB II TINJAUAN PUSTAKA, berisi tentang teori-teori dasar tentang Machine Learning, algoritma Regresi Linear, Random Forest, Naive Bayes, serta klasifikasi data medis.

BAB III METODE PENELITIAN, Bab ini menjelaskan metodologi yang digunakan dalam penelitian, termasuk pengumpulan data, pra-pemrosesan data, serta pelatihan dan pengujian model.

BAB IV HASIL DAN PEMBAHASAN, berisi analisis kinerja ketiga algoritma dalam klasifikasi data medis dan presentasi hasil penelitian.

BAB V PENUTUP, berisi kesimpulan dari penelitian dan saran untuk penelitian selanjutnya.