BAB 1

PENDAHULUAN

1.1 Latar Belakang

Penyakit tidak menular (Non-Communicable Diseases/NCDs) seperti hipertensi, diabetes, dan penyakit jantung merupakan penyebab kematian terbesar di dunia dan terus meningkat setiap tahun. Penyakit-penyakit ini bersifat kronis, berkembang dalam jangka panjang, dan sering kali tidak menunjukkan gejala yang signifikan pada tahap awal[1]. Pengenalan dini serta upaya berbasis identifikasi faktor risiko menjadi penting untuk menekan dampak penyakit dan peningkatan kualitas hidup masyarakat[2]. Salah satu metode yang paling umum digunakan untuk mengumpulkan data mengenai faktor risiko adalah melalui kuesioner kesehatan yang mencakup informasi mengenai gaya hidup, aktivitas fisik, pola makan, kebiasaan merokok, kualitas tidur, tingkat stres, dan riwayat keluarga. Meskipun mudah dan murah, data kuesioner menghadirkan berbagai tantangan yang signifikan, seperti nilai yang hilang, jawaban yang tidak konsisten, bias jawaban, dan format input yang beragam, sehingga data tersebut tidak dapat langsung digunakan dalam analisis atau prediksi [3].

Seiring dengan kemajuan teknologi informasi, Machine Learning telah menjadi pendekatan yang banyak digunakan dalam analisis kesehatan karena mampu mengidentifikasi pola dalam data dan menghasilkan prediksi otomatis. Machine learning merupakan sebuah tools yang dapat mengubah data menjadi sebuah informasi yang bermanfaat melalui pembuatan model dan analisis data [4]. Berbagai penelitian telah menunjukkan bahwa model-model seperti Naive Bayes, Random Forest, dan Support Vector Machine dapat memprediksi penyakit jantung, diabetes, dan hipertensi dengan tingkat akurasi yang cukup tinggi. Penelitian terkait mengklasifikasikan risiko diabetes dengan akurasi 87,88% dan tingkat kesalahan 12,12% [5]. Hasil klasifikasi juga didapatkan pada data penyakit jantung yang menunjukkan bahwa metode ini terbilang efektifk dengan akurasi mencapai 79,92 %[6]. Namun, keberhasilan model machine learning sangat bergantung pada kualitas data yang digunakan, sehingga diperlukan proses pengolahan data yang sistematis dan terstruktur dalam pengolahan data [7]. Pipeline pengolahan data berperan penting dalam meningkatkan kualitas data melalui tahapan pembersihan data, penanganan missing value, transformasi data, serta pembagian data latih dan data uji[8],[9]. Dengan adanya pipeline, proses pengolahan data menjadi lebih konsisten, otomatis, dan siap digunakan dalam pemodelan machine learning. Selain itu, integrasi pipeline

dengan sistem informasi berbasis web memungkinkan proses prediksi dilakukan secara cepat, real-time, dan mudah diakses oleh pengguna [10],[11],[12]

Berdasarkan urgensi tersebut dilakukan penelitian yang untuk memprediksi resiko penyakit yang mengintegrasikan pipeline dalam pengolahan data dengan model machine learning berbasis web. Sistem ini diharapkan mampu memberikan hasil prediksi yang cepat, akurat, dan mudah dipahami, sehingga dapat membantu dalam deteksi dini serta mendukung pengambilan keputusan di bidang kesehatan.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, penelitian ini memiliki rumusan masalah terkait pengolahan data bagaimana merancang pipeline pengolahan data kuesioner kesehatan yang mampu menangani missing value, inkonsistensi jawaban, dan noise guna menghasilkan dataset yang bersih untuk pemodelan machine learning Selain itu, penelitian ini juga mempertanyakan bagaimana algoritma Naive Bayes dapat dimanfaatkan untuk mengenali pola faktor risiko pada data kuesioner sehingga mampu memberikan prediksi tingkat risiko penyakit degeneratif dalam kategori rendah, sedang, atau tinggi. serta bagaimana mengembangkan sistem informasi berbasis web yang dapat mengintegrasikan pipeline dan model tersebut sehingga mampu memberikan hasil prediksi risiko penyakit secara otomatis, informatif, dan mudah dipahami oleh pengguna.

1.3 Batasan Masalah

Agar penelitian lebih terarah maka, batasan masalah dalam penelitian ini sebagai berikut:

1. Data yang digunakan adalah data kuesioner
2. Hanya membahas penyakit degeneratif secara umum seperti, hipertensi, jantung, dan diabetes.
3. Hasil prediksi risiko dikategorikan menjadi tiga kategori risiko rendah, sedang, tinggi.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut :

1. Mengembangkan pipeline pengolahan data kuesioner yang dapat meningkatkan kualitas data melalui pembersihan, penanganan missing value, transformasi data, dan standarisasi format.

2. Menerapkan algoritma Naive Bayes untuk memprediksi tingkat risiko penyakit berdasarkan faktor risiko yang diperoleh dari data kuesioner.
3. Membangun sistem informasi berbasis web yang mampu menampilkan hasil prediksi risiko secara otomatis dan informatif sebagai pendukung deteksi dini.

1.5 Manfaat Penelitian

1. Penelitian ini berkontribusi pada penerapan machine learning untuk memprediksi risiko penyakit tidak menular melalui integrasi pipeline pemrosesan data dalam data kuesioner kesehatan.
2. Penelitian ini membahas pentingnya kualitas data, khususnya penanganan missing value dan tahapan data pre-processing, terhadap peningkatan kinerja dan stabilitas model prediksi.
3. Hasil penelitian ini menghasilkan sistem prediksi risiko penyakit yang terstruktur dan otomatis untuk mendukung deteksi dini dan pengambilan keputusan berbasis data di sektor kesehatan.