Diabetes merupakan salah satu penyakit tidak menular yang semakin meningkat prevalensinya di seluruh dunia, termasuk di Indonesia. Menurut data dari Organisasi Kesehatan Dunia (WHO), jumlah penderita diabetes diperkirakan akan terus meningkat dalam beberapa dekade mendatang. Diabetes terjadi ketika tubuh tidak bisa mengatur kadar gula darah dengan baik, yang menyebabkan kadar gula darah terlalu tinggi SITASI: Diabetes Mellitus Insights from Epidemiology, Biochemistry, Risk Factors, Diagnosis, Complications and Comprehensive Management. Diabetes dapat menyebabkan berbagai komplikasi kesehatan serius, seperti kerusakan pada jantung, ginjal, mata, dan sistem saraf SITASI: Preventing diabetes complications. Oleh karena itu, deteksi dini dan prediksi risiko diabetes sangat penting untuk mencegah terjadinya komplikasi lebih lanjut. Dalam hal ini, teknologi informasi dan pembelajaran mesin dapat memainkan peran penting dalam mempermudah dan mempercepat diagnosis serta prediksi penyakit diabetes.

Salah satu algoritma pembelajaran mesin yang umum digunakan dalam tugas klasifikasi adalah K-Nearest Neighbors (KNN). KNN merupakan metode klasifikasi yang bekerja berdasarkan kedekatan jarak antara titik data yang akan diklasifikasikan dengan data yang sudah terlabel. Meskipun KNN sederhana dan mudah diimplementasikan, salah satu tantangan utama dalam menggunakan algoritma ini adalah sensitifitasnya terhadap data yang belum dinormalisasi. KNN mengandalkan perhitungan jarak, seperti Euclidean, untuk mengukur kedekatan antar data. Jika data memiliki skala atau satuan yang berbeda-beda, hal ini dapat menyebabkan ketidakseimbangan dalam perhitungan jarak, yang pada gilirannya dapat menurunkan kinerja model SITASI: Challenges in KNN Classification.

Karena itu, normalisasi data menjadi langkah yang sangat penting dalam preprocessing data sebelum diterapkan pada algoritma KNN. Normalisasi data bertujuan untuk mengubah rentang atau distribusi nilai fitur agar memiliki skala yang seragam, sehingga perhitungan jarak antar data dapat dilakukan secara lebih adil. Terdapat tiga teknik normalisasi yang akan saya digunakan, di antaranya Min-Max Scaling, Z-Score, Decimal Scaling SITASI: A comprehensive review on data preprocessing techniques in data analysis. Setiap teknik memiliki karakteristik yang berbeda dalam menangani distribusi dan rentang data, yang dapat mempengaruhi hasil klasifikasi yang diperoleh. Oleh karena itu, penting untuk mengeksplorasi pengaruh masing-masing teknik normalisasi terhadap kinerja model KNN, guna menemukan teknik yang paling efektif.

Penelitian ini berfokus untuk menganalisis pengaruh berbagai teknik normalisasi terhadap kinerja klasifikasi KNN pada dataset Pima Indians Diabetes Database. Dataset ini merupakan dataset publik yang sudah digunakan dalam penelitian-penelitian sebelumnya mengenai prediksi diabetes. Dataset ini terdiri dari 8 fitur yang menggambarkan faktor-faktor risiko diabetes, seperti usia, kadar glukosa darah, tekanan darah, ketebalan kulit, insulin, massa tubuh, dan sebagainya, serta sebuah target variabel yang menunjukkan apakah seseorang menderita diabetes atau tidak SITASI LINK WEB: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Penelitian ini akan menguji berbagai teknik normalisasi yang telah disebutkan di atas dan mengevaluasi dampaknya terhadap akurasi klasifikasi KNN.

Selain teknik normalisasi, penelitian ini juga akan melibatkan tahap pemilihan fitur (feature selection) untuk memastikan bahwa hanya fitur-fitur yang relevan yang digunakan dalam model. Pemilihan fitur yang tepat dapat meningkatkan kinerja model dengan mengurangi overfitting dan meningkatkan interpretabilitas model SITASI: IGRF-RFE a hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset. Dalam penelitian ini, teknik normalisasi akan diterapkan sebelum / setelah tahap pemilihan fitur, untuk melihat bagaimana masing-masing teknik mempengaruhi kinerja model KNN dalam prediksi diabetes.

Evaluasi kinerja model akan dilakukan dengan menggunakan beberapa metrik evaluasi, seperti akurasi, precision, recall, dan F1-score. Metrik-metrik ini digunakan untuk mengukur seberapa baik model dalam melakukan klasifikasi data, terutama dalam konteks klasifikasi yang memiliki ketidakseimbangan kelas seperti pada dataset diabetes. Akurasi akan menjadi metrik utama, namun metrik tambahan akan digunakan untuk memberikan gambaran yang lebih lengkap mengenai kinerja model.

Penelitian ini diharapkan dapat berkontribusi dalam pengembangan sistem prediksi penyakit diabetes menggunakan algoritma KNN dengan mengeksplorasi pengaruh teknik normalisasi terhadap kinerja klasifikasi. Penelitian ini bertujuan mengidentifikasi teknik normalisasi paling efektif untuk meningkatkan akurasi model dalam memprediksi risiko diabetes serta menekankan pentingnya preprocessing data dan pemilihan fitur dalam membangun model yang akurat dan efisien. Selain itu, hasil penelitian ini dapat diterapkan dalam kesehatan masyarakat untuk mendukung deteksi dini diabetes, memungkinkan model KNN memberikan prediksi lebih akurat sehingga membantu tenaga medis dalam pengambilan keputusan yang lebih tepat. Dengan demikian, penelitian ini tidak hanya memiliki relevansi akademik, tetapi juga berpotensi memberikan dampak positif dalam bidang kesehatan dan pencegahan penyakit.