This study analyzes the comparison of data normalization techniques in the KNearest Neighbors (KNN) model for diabetes classification using the Pima Indians Diabetes dataset. The three normalization techniques evaluated are Min-Max Scaling, Z-Score Scaling, and Decimal Scaling. After preprocessing, such as handling missing values and removing duplicates, as well as feature selection using the Random Forest method, the features removed include SkinThickness, Insulin, Pregnancies, and BloodPressure. The evaluation was carried out using accuracy, precision, recall, F1-Score, specificity, and ROC AUC metrics. The results show that Min-Max Scaling provides a significant improvement in all metrics, with the highest accuracy of 0.8117 and ROC AUC of 0.8050. Z-Score Scaling provides good results, but not as good as Min-Max Scaling. Decimal Scaling shows the lowest performance. Statistical tests using Paired T-Test show significant differences between Min-Max Scaling and without normalization on all metrics (P-Value <0.05), while Z-Score Scaling and Decimal Scaling are only significant on some metrics, with P-Values of 0.08363 and 0.43839 respectively for accuracy and ROC AUC. Overall, Min-Max Scaling proved to be the best normalization method for improving KNN performance in diabetes classification.

Comparison of Data Normalization Techniques on KNN Classification Performance for Pima Indians Diabetes Dataset

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 meningkat dari 108 juta pada tahun 1980 menjadi 422 juta pada tahun 2014. Diperkirakan jumlah penderita diabetes akan mencapai 578 juta pada tahun 2030 dan 700 juta pada tahun 2045 [1]. Diabetes terjadi ketika tubuh tidak bisa mengatur kadar gula darah dengan baik, yang menyebabkan kadar gula darah terlalu tinggi [2][3]. Diabetes dapat menyebabkan berbagai komplikasi kesehatan serius, seperti kerusakan pada jantung, ginjal, mata, dan sistem saraf [3][4]. Oleh karena itu, deteksi dini dan klasifikasi risiko diabetes sangat penting untuk mencegah terjadinya komplikasi lebih lanjut. Dalam hal ini, teknologi informasi dan pembelajaran mesin dapat berperan penting dalam mempermudah dan mempercepat diagnosis serta klasifikasi penyakit diabetes. Salah satu algoritma pembelajaran mesin yang umum digunakan dalam 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 [5]. KNN mengandalkan perhitungan jarak seperti Euclidean Distance, yang mengukur jarak terpendek antara dua titik, dengan menghitung rumus akar kuadrat dari jumlah kuadrat selisih nilai pada setiap dimensi [6]. Meskipun sederhana dan mudah diimplementasikan, tantangan utama dalam penggunaan KNN adalah sensitivitasnya terhadap data yang belum dilakukan normalisasi. Jika data memiliki skala atau satuan yang berbeda, hal ini dapat menyebabkan ketidakseimbangan dalam perhitungan jarak dan menurunkan kinerja model [7]. Karena itu, normalisasi data menjadi langkah penting sebelum diterapkan pada algoritma KNN. Normalisasi bertujuan untuk mengubah nilai fitur ke dalam skala yang seragam, sehingga perhitungan jarak antar data dapat dilakukan secara akurat tanpa dipengaruhi perbedaan skala [8]. Terdapat tiga teknik normalisasi yang relevan dengan perhitungan Euclidean Distance, yaitu Min-Max Scaling, ZScore Scaling, dan Decimal Scaling [9]. Penelitian ini bertujuan untuk menganalisis perbandingan berbagai teknik normalisasi terhadap kinerja model KNN dalam klasifikasi diabetes pada dataset Pima Indians Diabetes. Teknik normalisasi diterapkan untuk mengatasi perbedaan rentang nilai antar fitur dan mengevaluasi dampaknya pada kinerja model [13]. Selain itu, pemilihan fitur juga dilakukan untuk memastikan hanya fitur relevan yang digunakan, sehingga dapat mengurangi overfitting [14]. Dengan membandingkan teknik normalisasi yang berbeda, diharapkan dapat dibangun model yang lebih akurat dan efisien, mendukung deteksi dini diabetes, serta membantu tenaga medis dalam pengambilan keputusan yang lebih tepat.

Beberapa penelitian sebelumnya telah membandingkan teknik normalisasi terhadap kinerja berbagai algoritma klasifikasi, yang berdampak signifikan pada akurasi klasifikasi di berbagai dataset. Sebagai contoh, penelitian Muasir Pagan et al. [10] membandingkan teknik normalisasi data yaitu Min-Max Scaling, Z-Score Scaling, dan Decimal Scaling terhadap kinerja algoritma KNN menggunakan sepuluh dataset. Hasil penelitian menunjukkan bahwa ZScore Scaling dan Decimal Scaling memberikan kinerja yang lebih baik dibandingkan Min-Max Scaling, dengan Z-Score Scaling secara konsisten menghasilkan akurasi, precision, recall, dan F1-score yang lebih tinggi di sebagian besar dataset. Temuan ini menyoroti pentingnya pemilihan teknik normalisasi yang tepat berdasarkan karakteristik dataset. Selanjutnya, penelitian Alshdaifat et al. [11] juga mengevaluasi dampak teknik normalisasi, yaitu Min-Max Scaling, Z-Score Scaling, dan Decimal Scaling terhadap kinerja algoritma klasifikasi seperti SVM dan ANN, dengan temuan yang juga relevan untuk KNN. Hasil penelitian pada 18 dataset menunjukkan bahwa Z-Score Scaling sering kali memberikan hasil terbaik karena kemampuannya menangani outlier, sementara Decimal Scaling kurang efektif dalam meningkatkan akurasi model secara keseluruhan. Selain itu, penelitian Saichon Sinsomboonthong [12] membandingkan kinerja delapan teknik normalisasi dalam klasifikasi menggunakan algoritma ANN pada enam dataset. Hasil penelitian menunjukkan bahwa Min-Max Scaling umumnya memberikan akurasi tertinggi dan MSE terendah. Namun, pada beberapa dataset, seperti White Wine Quality dan Pima Indians Diabetes, Adjusted-2 Min-Max Scaling memberikan hasil yang lebih baik. Teknik normalisasi Decimal Scaling, juga menunjukkan hasil kompetitif pada dataset tertentu, meskipun tidak mengungguli Min-Max Scaling