Abstrak 1

Penelitian ini menganalisis perbandingan berbagai teknik normalisasi data terhadap performa model K-Nearest Neighbors (KNN) dalam klasifikasi diabetes pada dataset Pima Indians Diabetes. Tiga teknik normalisasi yang dievaluasi adalah Min-Max Scaling, Z-Score Scaling, dan Decimal Scaling. Data yang telah melalui tahapan preprocessing, seperti penanganan missing values, penghapusan duplikat, dan pemeriksaan tipe data, kemudian melalui tahap seleksi fitur menggunakan metode Random Forest untuk menentukan atribut yang paling relevan. Fitur yang dihapus dalam tahap seleksi fitur meliputi SkinThickness, Insulin, Pregnancies, dan BloodPressure, karena kontribusinya yang rendah terhadap model. Data yang telah diproses kemudian diterapkan pada model KNN untuk memprediksi kemungkinan seseorang mengidap diabetes. Evaluasi dilakukan menggunakan beberapa metrik kinerja, seperti akurasi, presisi, recall, F1-Score, spesifisitas, dan ROC AUC. Hasil menunjukkan bahwa Min-Max Scaling memberikan peningkatan signifikan pada seluruh metrik, dengan akurasi tertinggi tercatat 0.8117 dan ROC AUC mencapai 0.8050. Z-Score Scaling juga menunjukkan hasil yang baik, namun tidak sebaik Min-Max Scaling secara statistik. Sementara itu, Decimal Scaling menunjukkan performa yang lebih rendah dibandingkan kedua metode lainnya. Secara keseluruhan, Min-Max Scaling terbukti sebagai metode normalisasi yang paling efektif untuk meningkatkan performa model KNN dalam klasifikasi diabetes.

This study analyzes the comparison of various data normalization techniques against the performance of the K-Nearest Neighbors (KNN) model in diabetes classification in the Pima Indians Diabetes dataset. The three normalization techniques evaluated are Min-Max Scaling, Z-Score Scaling, and Decimal Scaling. The data that has gone through the preprocessing stages, such as handling missing values, removing duplicates, and checking data types, then goes through the feature selection stage using the Random Forest method to determine the most relevant attributes. Features removed in the feature selection stage include SkinThickness, Insulin, Pregnancies, and BloodPressure, due to their low contribution to the model. The processed data is then applied to the KNN model to predict the likelihood of a person having diabetes. The evaluation is carried out using several performance metrics, such as accuracy, precision, recall, F1-Score, specificity, and ROC AUC. The results show that Min-Max Scaling provides a significant improvement in all metrics, with the highest recorded accuracy of 0.8117 and ROC AUC reaching 0.8050. Z-Score Scaling also shows good results, but not as good as Min-Max Scaling statistically. Meanwhile, Decimal Scaling shows lower performance than the other two methods. Overall, Min-Max Scaling proves to be the most effective normalization method for improving the performance of the KNN model in diabetes classification.

Abstrak 2

Penelitian ini menganalisis perbandingan teknik normalisasi data pada model K-Nearest Neighbors (KNN) untuk klasifikasi diabetes menggunakan dataset Pima Indians Diabetes. Tiga teknik normalisasi yang dievaluasi adalah Min-Max Scaling, Z-Score Scaling, dan Decimal Scaling. Setelah preprocessing, seperti penanganan missing values dan penghapusan duplikat, serta seleksi fitur menggunakan metode Random Forest, fitur yang dihapus meliputi SkinThickness, Insulin, Pregnancies, dan BloodPressure. Evaluasi dilakukan dengan metrik akurasi, presisi, recall, F1-Score, spesifisitas, dan ROC AUC. Hasil menunjukkan Min-Max Scaling memberikan peningkatan signifikan pada seluruh metrik, dengan akurasi tertinggi 0.8117 dan ROC AUC 0.8050. Z-Score Scaling memberikan hasil yang baik, tetapi tidak sebaik Min-Max Scaling. Decimal Scaling menunjukkan performa terendah. Uji statistik menggunakan Paired T-Test menunjukkan perbedaan signifikan antara Min-Max Scaling dan tanpa normalisasi pada semua metrik (P-Value < 0.05), sedangkan Z-Score dan Decimal Scaling hanya signifikan pada beberapa metrik, dengan P-Value masing-masing 0.08363 dan 0.43839 untuk akurasi dan ROC AUC. Secara keseluruhan, Min-Max Scaling terbukti sebagai metode normalisasi terbaik untuk meningkatkan performa KNN dalam klasifikasi diabetes.

This study analyzes the comparison of data normalization techniques in the K-Nearest 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 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.