

ABSTRAK

Diabetes melitus merupakan salah satu penyakit kronis yang jumlah penderitanya terus meningkat setiap tahun. Dalam bidang *data mining*, klasifikasi diabetes dapat dilakukan menggunakan algoritma Naive Bayes. Namun, kualitas data menjadi faktor penting dalam proses klasifikasi, khususnya adanya *missing value* yang dapat memengaruhi kinerja model. *Pima Indians Diabetes Dataset* memiliki beberapa atribut medis yang mengandung nilai nol yang dianggap sebagai *missing value* terselubung, sehingga diperlukan penanganan sebelum proses klasifikasi.

Penelitian ini bertujuan untuk mengevaluasi pengaruh penanganan *missing value* terhadap kinerja algoritma Gaussian Naive Bayes pada klasifikasi diabetes menggunakan *Pima Indians Diabetes Dataset*. Metode penanganan *missing value* yang digunakan adalah imputasi *mean* dan *median*. Penelitian dilakukan melalui tahapan identifikasi *missing value*, *preprocessing* data, pembagian data *training* dan *testing* dengan rasio 80:20, normalisasi menggunakan Min-Max Normalization, klasifikasi menggunakan Gaussian Naive Bayes, serta evaluasi menggunakan *accuracy*, *precision*, *recall*, dan *F1-score*. Selain itu, 5-Fold Cross Validation digunakan sebagai validasi tambahan untuk mengetahui konsistensi performa model.

Hasil penelitian menunjukkan bahwa skenario tanpa penanganan *missing value* menghasilkan performa tertinggi dengan *accuracy* sebesar 0,766, *precision* 0,661, *recall* 0,709, dan *F1-score* 0,684. Metode *median* menghasilkan *accuracy* sebesar 0,753, lebih tinggi dibandingkan metode *mean* sebesar 0,747, tetapi keduanya belum mampu melampaui performa *baseline*. Hasil 5-Fold Cross Validation juga menunjukkan *baseline* memperoleh rata-rata *accuracy* tertinggi sebesar 0,7552. Dengan demikian, penanganan *missing value* memberikan pengaruh terhadap kinerja Gaussian Naive Bayes, tetapi tidak selalu meningkatkan performa model dan efektivitasnya bergantung pada karakteristik dataset serta metode imputasi yang digunakan.

Kata kunci: *missing value*, Gaussian Naive Bayes, klasifikasi diabetes, *mean*, *median*

ABSTRACT

Diabetes mellitus is a chronic disease with a growing number of sufferers every year. In data mining, diabetes classification can be performed using the Naive Bayes algorithm. However, data quality is a critical factor in the classification process, particularly the presence of missing values, which can impact model performance. The Pima Indians Diabetes Dataset contains several medical attributes with zero values, considered hidden missing values, requiring treatment before the classification process.

This study aims to evaluate the effect of missing value treatment on the performance of the Gaussian Naive Bayes algorithm in diabetes classification using the Pima Indians Diabetes Dataset. The missing value treatment method used was mean and median imputation. The study included missing value identification, data preprocessing, splitting the training and testing data with an 80:20 ratio, normalization using Min-Max Normalization, classification using Gaussian Naive Bayes, and evaluation using accuracy, precision, recall, and F1-score. Furthermore, 5-Fold Cross Validation was used as additional validation to determine the consistency of model performance.

The results showed that the scenario without missing value handling produced the highest performance with an accuracy of 0.766, precision of 0.661, recall of 0.709, and F1-score of 0.684. The median method produced an accuracy of 0.753, higher than the mean method at 0.747, but neither method was able to surpass the baseline performance. The 5-Fold Cross Validation results also showed that the baseline achieved the highest average accuracy of 0.7552. Thus, missing value handling impacts Gaussian Naive Bayes performance, but it does not always improve model performance, and its effectiveness depends on the characteristics of the dataset and the imputation method used.

Keywords: missing value, Gaussian Naive Bayes, diabetes classification, mean, median