

ABSTRAK

Proses klasifikasi diawali dengan pembersihan data (*data cleaning*) untuk memastikan bahwa seluruh atribut numerik, seperti persentase penduduk miskin, pengeluaran per kapita, tingkat pengangguran terbuka, dan Indeks Pembangunan Manusia (IPM), berada dalam format yang sesuai agar dapat diproses optimal menggunakan aplikasi *Orange*. Setelah data dinyatakan bersih, algoritma Naive Bayes dipilih karena kemampuannya dalam mengolah data secara cepat, efisien, dan memberikan hasil klasifikasi yang kompetitif. Evaluasi model dilakukan menggunakan dua metode, yaitu *5-Fold Cross-Validation* dan *10-Fold Cross-Validation*, untuk memperoleh hasil yang stabil dan representatif. Berdasarkan evaluasi dari *widget Test & Score* pada *Orange*, model *Naive Bayes* dengan *5-Fold Cross-Validation* menghasilkan akurasi 89,8%, *AUC* 0,938, *precision* 0,952, *recall* 0,898, *F1-score* 0,909, dan *MCC* 0,652. Sedangkan dengan *10-Fold Cross-Validation*, diperoleh akurasi yang sama sebesar 89,8%, *AUC* 0,936, *precision* 0,926, *recall* 0,898, *F1-score* 0,908, dan *MCC* 0,612. Hasil ini menunjukkan bahwa performa model relatif stabil, dengan 5-fold memberikan hasil yang sedikit lebih baik pada *precision* dan *MCC*. Hasil Confusion Matrix dari pengujian terhadap 364 data uji menunjukkan bahwa model berhasil mengklasifikasikan 291 data “Tidak Miskin” dan 36 data “Miskin” secara benar. Namun, terdapat 32 data “Tidak Miskin” yang salah diklasifikasikan sebagai “Miskin” dan 5 data “Miskin” yang salah diklasifikasikan sebagai “Tidak Miskin”, dengan total kesalahan prediksi 37 data. Akurasi aktual dari pengujian ini adalah 89,8%, sesuai dengan hasil evaluasi *cross-validation*. Dari keseluruhan hasil tersebut, dapat disimpulkan bahwa algoritma *Naive Bayes* mampu mengklasifikasikan status kemiskinan dengan performa yang baik, ditunjukkan oleh nilai *AUC* yang tinggi, *precision* dan *recall* yang seimbang, serta akurasi yang konsisten. Oleh karena itu, model ini layak digunakan untuk analisis klasifikasi status kemiskinan pada data yang digunakan dalam penelitian ini.

Kata Kunci: Kemiskinan, Naïve Bayes, klasifikasi, data mining, evaluasi model, sosial ekonomi

ABSTRACT

The classification process begins with data cleaning to ensure that all numeric attributes, such as the percentage of poor people, per capita expenditure, open unemployment rate, and Human Development Index (HDI), are in the appropriate format so that they can be processed optimally using the Orange application. After the data is declared clean, the Naive Bayes algorithm is chosen because of its ability to process data quickly, efficiently, and provide competitive classification results. Model evaluation is carried out using two methods, namely 5-Fold Cross-Validation and 10-Fold Cross-Validation, to obtain stable and representative results. Based on the evaluation of the Test & Score widget on Orange, the Naive Bayes model with 5-Fold Cross-Validation produces an accuracy of 89.8%, AUC 0.938, precision 0.952, recall 0.898, F1-score 0.909, and MCC 0.652. Meanwhile, with 10-Fold Cross-Validation, the same accuracy was obtained at 89.8%, AUC 0.936, precision 0.926, recall 0.898, F1-score 0.908, and MCC 0.612. These results indicate that the model performance is relatively stable, with 5-fold giving slightly better results in precision and MCC. The Confusion Matrix results from testing 364 test data show that the model successfully classified 291 “Not Poor” data and 36 “Poor” data correctly. However, there were 32 “Not Poor” data that were misclassified as “Poor” and 5 “Poor” data that were misclassified as “Not Poor”, with a total prediction error of 37 data. The actual accuracy of this test was 89.8%, in accordance with the results of the cross-validation evaluation. From all the results, it can be concluded that the Naive Bayes algorithm is able to classify poverty status with good performance, indicated by high AUC values, balanced precision and recall, and consistent accuracy. Therefore, this model is suitable for use in the analysis of poverty status classification on the data used in this study.

Key word: Poverty, Classification, Naive Bayes, Orange, Data Mining