ABSTRACT

COMPARISON OF NAÏVE BAYES AND RANDOM FOREST MODELS IN PREDICTING THE STUDY DURATION CLASSIFICATION OF MATHEMATICS UNDERGRADUATES AT THE UNIVERSITY OF LAMPUNG

By

Shelvira Hestina Putri

The advancement of information technology has encouraged the utilization of data mining across various fields, including education. Data mining enables the analysis of large-scale data to discover meaningful patterns and generate predictions based on those patterns. In the context of higher education, the timeliness of student graduation is a crucial indicator related to study program accreditation and institutional efficiency. This study aims to compare the performance of two machine learning-based classification algorithms, namely Naïve Bayes and Random Forest, in predicting the study duration of undergraduate students in the Mathematics Study Program at the University of Lampung. The data used consists of academic records that have been labeled based on whether the students graduated on time or not. Model evaluation was conducted using a confusion matrix and two validation techniques: data splitting and k-fold cross-validation. The test results indicate that the Random Forest algorithm achieved the highest accuracy of 94.44%, outperforming Naïve Bayes. Additionally, both models attained their highest accuracy when using the data splitting method. These findings suggest that Random Forest is a more effective method for classifying the study duration of undergraduate mathematics students at the University of Lampung and provides a strong foundation for predictive efforts to support improvements in on-time graduation rates.

Keywords: Data Mining, Classification, Naïve Bayes, Random Forest, Comparison, Study Duration, On-time Graduation.

ABSTRAK

PERBANDINGAN MODEL *NAÏVE BAYES* DAN *RANDOM FOREST* DALAM PREDIKSI KLASIFIKASI MASA STUDI SARJANA MATEMATIKA UNIVERSITAS LAMPUNG

Oleh

Shelvira Hestina Putri

Perkembangan teknologi informasi telah mendorong pemanfaatan data mining dalam berbagai bidang, termasuk dunia pendidikan. *Data mining* memungkinkan analisis terhadap data berukuran besar untuk menemukan pola yang bermakna dan menghasilkan prediksi berdasarkan pola tersebut. Dalam konteks pendidikan tinggi, ketepatan waktu kelulusan mahasiswa merupakan indikator penting yang berkaitan dengan akreditasi program studi dan efisiensi institusi. Penelitian ini bertujuan untuk membandingkan performa dua algoritma klasifikasi berbasis *machine learning*, yaitu *Naïve Bayes* dan *Random Forest*, dalam memprediksi masa studi mahasiswa Program Studi Sarjana Matematika Universitas Lampung. Data yang digunakan adalah data akademik mahasiswa yang telah diberi label berdasarkan status kelulusan tepat waktu atau tidak. Evaluasi model dilakukan dengan menggunakan *confusion matrix* dan dua teknik validasi, yaitu metode data *splitting* dan *k-fold cross validation*. Hasil pengujian menunjukkan bahwa algoritma *Random Forest* memiliki akurasi tertinggi sebesar 94,44%, mengungguli *Naïve Bayes*. Selain itu, kedua model mencapai hasil akurasi tertinggi saat menggunakan metode data *splitting*. Hasil penelitian ini menunjukkan bahwa *Random Forest* merupakan metode yang lebih baik dalam klasifikasi masa studi sarjana Matematika Universitas Lampung.

Kata-kata kunci: Data Mining, klasifikasi, Naïve Bayes, Random Forest, Perbandingan, Masa Studi, Kelulusan Tepat Waktu.