

ABSTRAK

Gerry Chandra (00000024282)

IMPLEMENTASI MACHINE LEARNING DENGAN ALGORITMA LOGISTIC REGRESSION DAN RANDOM FOREST UNTUK PREDIKSI PERFORMA CALON MAHASISWA BARU

Skripsi, Fakultas Sains dan Teknologi (2020).

(xiv + 73 halaman; 51 gambar; 7 tabel; 5 lampiran)

Berkembangnya internet dalam aplikasi industri memberikan dampak kepada institusi pendidikan. Salah satu dampak yang bisa dirasakan adalah terjadinya digitalisasi pada sebagian besar proses akademik universitas. Hal ini membuat proses pengolahan data dalam jumlah besar menjadi krusial untuk membantu pihak yang bersangkutan dalam menghasilkan keputusan yang tepat dan lebih baik. Salah satu masalah yang bisa dijawab dengan mengolah data pendidikan yang tersedia adalah rendahnya persentase mahasiswa lulus tepat waktu yang mempengaruhi akreditasi perguruan tinggi di Indonesia. Penelitian ini bertujuan untuk membentuk suatu model yang mampu melakukan prediksi terhadap performa calon mahasiswa baru. Model utama yang dikembangkan adalah klasifikasi *machine learning* dengan algoritma *Logistic Regression* dan *Random Forest* menggunakan informasi sekolah dan nilai-nilai SMA serta informasi orang tua sebagai variabel. Model *machine learning* dalam penelitian ini dibangun dengan menggunakan data mahasiswa Universitas Pelita Harapan tahun akademik 2018/2019. Performa model kemudian diukur dengan menggunakan metrik-metrik evaluasi model klasifikasi, yaitu *accuracy*, *precision*, *recall*, *F1-score* dan AUC. Kedua model utama yang dibangun dalam penelitian ini menghasilkan nilai *accuracy* dan *recall* sebesar 0,74, nilai *precision* sebesar 0,22 dan nilai *F1-score* sebesar 0,34. Nilai AUC untuk model *Logistic Regression* adalah 0,82 dan 0,79 untuk model *Random Forest*. Eksplorasi model dilakukan dalam penelitian ini untuk membuat model yang lebih seimbang antara label kelas dan juga model sederhana dengan hanya menggunakan dua variabel, yaitu nilai Bahasa Inggris dan nilai Matematika saat SMA sebagai *predictor*. Model tambahan yang dibuat tidak menghasilkan nilai-nilai metrik sebaik model utama. Analisis *feature importances* dari algoritma *Random Forest* menunjukkan variabel nilai Bahasa Inggris dan nilai Matematika sebagai dua variabel yang menjadi faktor terpenting untuk menentukan performa calon mahasiswa.

Kata Kunci : *Educational Data Mining*, Klasifikasi, *Logistic Regression*, *Machine Learning*, *Random Forest*.

Referensi : 22 (2008-2019)

ABSTRACT

Gerry Chandra (00000024282)

IMPLEMENTATION OF MACHINE LEARNING WITH LOGISTIC REGRESSION AND RANDOM FOREST ALGORITHM TO PREDICT PERFORMANCE OF PROSPECTIVE STUDENTS

Thesis, Faculty of Science and Technology (2020).

(xiv + 73 pages, 7 tables, 51 figures, 5 appendices)

The development of internet in industrial application affects educational institutions. One common example of the revolution is the digitalization of university's academic processes. It makes the process of handling massive amount of data becomes crucial in order to assist the concerning parties to make better decisions. One problem that can be solved by exploring the available educational data is the low percentage of university students who graduated on time that can affect university's accreditation in Indonesia. This research focuses on the development of a model that can be utilized to predict the performance of prospective students. The main model uses the idea of machine learning classification with Logistic Regression and Random Forest algorithm using school information, high school scores and parents/guardian data as input variables. Model built in this research uses students' data of Universitas Pelita Harapan of academic year 2018/2019 as sample. Then, performance of the model will be measured using classification model evaluation metrics, which are accuracy, precision, recall, F1-score and AUC. The main models built in this research gives 0.74 accuracy and recall score, 0.22 precision score and 0.34 F1-score. Logistic Regression model gives 0.82 AUC score, while Random Forest model results in 0.79 AUC score. Model exploration is done in this research to build a model with more balanced class labels and also a simpler model that only uses two input variables, English and Mathematics scores at high school. The performance of the additional models is not as good as the main model. Feature importances analysis of Random Forest algorithm shows that English and Mathematics score at high school are the two most deciding variables to predict the performance of prospective students.

Keywords : Classification, Educational Data Mining, Logistic Regression, Machine Learning, Random Forest.

Reference : 22 (2008-2019)