Optimalisasi Kinerja Pemelajaran Mesin di Bidang Pendidikan dengan Contoh Kasus Prediksi Mahasiswa Putus Studi di Beberapa Perguruan Tinggi Indonesia = Optimizing Machine Learning Performance in the Field of Education with Case Example of Predicting Student Drop-out at Several Indonesian Universities

Darian Texanditama, author

Abstrak

Pemelajaran mesin dikenal sangat berguna dalam menyelesaikan permasalahan prediksi dan klasifikasi melalui pembelajaran pola dan perilaku data yang tersedia. Oleh karena itu, pemelajaran mesin dapat dimanfaatkan di berbagai bidang kehidupan dan industri modern. Namun, kinerja pemelajaran mesin sangat tergantung dari model pemelajaran mesin yang digunakan maupun dari kualitas data yang digunakan untuk pemelajaran. Data yang tidak bersih, tidak representatif, dan ketersediaannya terbatas akan mengurangi kualitas hasil prediksinya.
Penelitian ini bertujuan untuk menguji kombinasi beberapa metode pemrosesan data (yaitu MissForest, GAIN, ENN, dan TabGAN oversampling) dengan model pembelajaran mesin (yaitu model CatBoost dan model klasifikasi biner berbasis neural network) untuk memprediksi kasus mahasiswa putus studi di beberapa universitas di Indonesia menggunakan data dari PDDikti. Penambahan fitur dilakukan untuk memberi label bidang studi terhadap dataset tersebut. Selain penambahan fitur seleksi fitur relevan menggunakan korelasi Pearson serta feature importances juga dilakukan setelah pelatihan model awal. Google Colab dengan bahasa pemrograman Python digunakan untuk menjalankan algoritma pemrosesan data dan pelatihan model.
Hasil penelitian menunjukkan bahwa model CatBoost dengan kombinasi metode imputasi GAIN, undersampling ENN, dan tanpa fitur kelompok bidang studi memberikan F1-score tertinggi yaitu 66,38% dengan nilai precision 71,75% dan nilai recall 61,76%. Apabila digunakan model klasifikasi biner pemelajaran dalam akan didapatkan metrik terbaik F1-score 62,32%. Hasil terbaik penelitian ini menunjukkan peningkatan F1-score sebesar 2,15% dibandingkan dengan F1-score pada penelitian sebelumnya yang menggunakan model CatBoost bersama kombinasi Missforest dan ENN tanpa fitur kelompok
bidang studi. Penelitian ini menunjukkan bahwa oversampling dan undersampling memberikan dampak yang berlawanan terhadap metrik precision dan recall. Penelitian juga menemukan seleksi fitur dapat meningkatkan kinerja model namun tidak berdampak besar dibandingkan teknik-teknik lain misalnya balancing dan optimisasi hyperparameter.
.....Machine learning is known to be very useful in solving prediction and classification problems by learning the patterns and behavior of available data. Therefore, machine learning can be utilized in various areas of modern life and industry. However, the performance of machine learning is highly dependent on the machine learning model used as well as on the quality of the data used for learning. Data that is not clean, not representative, and scarce will reduce the quality of the prediction results.
This study aims to test the combination of several data processing methods (namely MissForest, GAIN, ENN, and TabGAN oversampling) with machine learning models (CatBoost and binary classification

models based on neural networks) to predict dropout cases at several Indonesian universities using data from PDDikti. The addition of features is done to label data with their respective fields of study. Other than adding features, selection of relevant features using Pearson's correlation as well as feature importances is also carried out after initial model training. Google Colab with the Python programming language is used to run data processing algorithms and train models.

This study shows that CatBoost with the combination of GAIN imputation, ENN undersampling, and no field of study feature results in the highest F1-score of 66.38%, which are composed of 71.75% in precision and 61.76% in recall. If a deep learning binary classification model is used instead, the best F1-score result is 62.32%. The best result from this study shows an increase in F1-score of 2.15% compared to the F1-score of the previous study (64.23%) which used CatBoost along with a combination of Missforest, ENN and no field of study features. This research shows oversampling and undersampling produce opposite effects on precision and recall scores. Research has also

found that feature selection can improve model performance but does not have a large impact compared to other techniques such as balancing and hyperparameter optimization