Intisari

Berdasarkan survei *Programme for International Student Assessment*, Indonesia berada di posisi lebih rendah dibanding semua negara yang berpartisipasi, kecuali Peru dalam hal matematika dan sains, serta negara ke-lima dari bawah dalam hal membaca. Hal ini membuktikan bahwa kualitas dan performa belajar siswa di negara tersebut masih rendah. Penelitian ini bertujuan untuk mengetahui berbagai faktor pendukung performansi siswa Indonesia, dengan mengambil sampel data di salah satu Sekolah Menengah Pertama Muhammadiyah 2 Depok, Sleman, Yogyakarta. Untuk mengimplementasikan hal tersebut, maka pada penelitian ini memanfaaatkan Educational Data mining (EDM) yang merupakan suatu proses untuk mendeteksi dan menggali berbagai data yang berkaitan dengan dunia pendidikan.

Jumlah siswa yang diuji dalam penelitian ini hanya terdiri atas 63 siswa, sehingga dataset yang terbentuk adalah jenis dataset kecil. Dataset yang kecil memiliki kecenderungan nilai yang tidak representatif dikarenakan jumlah sampel untuk *training* maupun *testing* tidak ideal. Sehingga dalam penelitian ini dilakukan berbagai pendekatan dari pelaksanaan *preprocessing,* variasi jenis klasifikasi, jenis evaluator seleksi fitur serta sistem validasi. Untuk jenis klasifikasi, dimanfaatkan Naive Bayes serta REPTree, yang telah terbukti bekerja baik pada dataset kecil berdasarkan penelitian sebelumnya. Sedangkan evaluator seleksi fitur, dimanfaatkan penyeleksi atribut (*Gain Ratio),* penyeleksi subset (*Classifier Subset Evaluator)* serta reduksi dimensi (*Principal Component).* Selain itu, untuk jenis validasi yang dipakai dalam penelitian ini memanfaatkan 10-*fold cross* serta *holdout validation.*

Hasil dari penelitian ini menunjukkan parameter akurasi dan variansi yang berbeda pada tiap validasi yang diperoleh. Validasi *10-fold cross* menghasilkan akurasi dan vartiansi erbaik pada klasifikasi REPTree dengan evaluator *Principal Component.* Sedangkan, validasi *holdout* menghasilkan variansi terbaik pada klasifikasi Naive Bayes dengan evaluator *Gain Ratio.*Walaupun akurasi pada validasi *holdout* cenderung lebih tinggi dibandingkan dengan validasi 10-*fold cross,* namun pada validasi *holdout* mengandung *overfitting* dengan indikasi nilai variansi yang diperoleh lebih besar. Selain itu, dari penelitian ini juga memperoleh berbagai faktor yang mempengaruhi performa belajar siswa SMP Muhammadiyah 2 Depok Sleman, seperti faktor dengan siapa siswa tinggal bersama, faktor Pendidikan Ayah serta faktor lingkungan kelas.

Kata kunci: Educational Data Mining (EDM), Dataset kecil, Klasifikasi, Seleksi fitur, Validasi.

Abstract

Based on Programme for International Student Assessment, Indonesia is on the lower position than participated countries, except Peru in Mathematic and Sains field, and also belongs to five lowest in reading field. It proves that quality in education and student study performance of this country is still low. This research has purpose to know various Indonesian students' performance supporter factors, by taking data sampel in one of Junior High School, SMP Muhammadiyah 2 Depok, Sleman, Yogyakarta. For implementing it, this research will use Educational Data mining (EDM) which is a process to detect and dig out the relational data in educational world.

Tested students' amount in this research only consists of 63 students, so that it will form small dataset. This small dataset has higher probability to gain unrepresentative result because of unideal amounts between training and even testing dataset. So that, in this research does many approaches by using preprocessing, variation of classifier, evaluator to feature selection and also validation system. Classifiers which is used here are Naive Bayes and REPTree. Two of them are already proved work well in small dataset based on previous researchs. While for selection feature evaluators which is used here are atribut selection (Gain Ratio), subset selection (Classifier Subset Evaluator) and dimensional reduction (Principal Component). For validation systems are 10-fold cross and holdout validation.

Result of this research shows different accuration and variance as parameters in each validation. 10-fold cross validation result highest accuration in procedure combination of REPTree with Principal Component evaluator selection. While for holdout validation results best variance in combination of Naive Bayes with Gain Ratio evaluator selection. Although the accuration results of holdout validation tend to higher than results of 10-fold cross validation, but the holdout results have higher probability of overfitting due to higher variance indication. Besides that, this research also results various factors which influencing study performance of SMP Muhammadiyah 2 Depok, Sleman's students are student's living with factor, dad's educational degree factor and also class's environment factor.

Keywords : *Educational Data Mining (EDM), Small Dataset, Classification, Featue Selection, Validation.*