ABSTRACT

Online learning grows rapidly in recent years. In online learning environment, dropout rate is relatively higher than in traditional learning. One of several strategies to reduce the number of dropouts is analyzing cause-and-effect relationship between factors that affect student outcome. Bayesian Network is widely used as a machine learning method that can analyze causality. However, Bayesian Network performance is not optimal. The accuracy ranges from 56.57% to 82.14%.

This study used a public dataset, namely OULAD. The purpose of this study was to analyze changes in the causality graph of the Bayesian Network-based data handling method.. We used Hill Climber as part of structure learning and bayesian score in parameter learning. We did adjust with two max nodes of parents. We applied SMOTE *oversampling* and *feature selection* using *filter* method as part of preprocessing phase. There are four experimental schemes in eight datasets to compare and evaluate which schemes and datasets get the highest performance. There are six indicators to evaluate the schemes. The indicators are accuracy, precision, recall, ROC, AUROC and computation time.

The test results show the algorithm modification using SMOTE and *feature selection* can improve Bayesian Network performance. The highest accuracy and AUROC by using the combination of preprocessing step are 87.65 and 97.88. The average value of each AUROC is stable with a value above 80. For the cause-and-effect relationship, we applied using one course that get the highest performance. Factors that affect student performance for each scheme are different. However, the main factor in every schemes that always selected is the duration of study. The duration of study means student have a good retention in accessing resources until the end of learning.

Keywords: Causality, SMOTE, Feature Selection, Bayesian Network, Student Performance

INTISARI

Pembelajaran *online* berkembang pesat dalam beberapa tahun terakhir. Dalam lingkungan pembelajaran *online*, angka putus sekolah relatif lebih tinggi daripada pembelajaran tradisional. Salah satu strategi untuk mengurangi angka *dropout* yaitu dengan menganalisis hubungan sebab akibat antara faktor-faktor yang mempengaruhi hasil belajar siswa. Bayesian Network banyak digunakan sebagai metode *machine learning* yang dapat menganalisis kausalitas. Namun, kinerja Bayesian Network (BN) masih belum optimal dengan rentang nilai 56.37% hingga 82.14%.

Penelitian ini menggunakan dataset publik yaitu OULAD. Tujuan penelitian ini adalah untuk menganalisis perubahan grafik kausalitas terhadap metode *handling* data berbasis Bayesian Network. Kami menggunakan hill-climber sebagai *structure learning* dan *bayesian scoring* dalam pembobotan parameter *learning*. Kami melakukan dua metode *preprocessing* yaitu SMOTE *oversampling* dan pemilihan fitur. Ada empat skema eksperimental dan delapan jenis dataset yang digunakan dalam penelitian ini untuk mengevaluasi skema dan dataset mana yang mendapatkan kinerja algoritme tertinggi. Evaluasi kinerja algoritme menggunakan enam indikator yaitu akurasi, presisi, *recall*, ROC, AUROC dan waktu komputasi.

Hasil pengujian menunjukkan skema algoritme yang menggunakan SMOTE dan pemilihan fitur mampu meningkatkan kinerja Bayesian Network. Akurasi dan AUROC tertinggi dengan dua *preprocessing* data adalah 87.65 dan 97.88. Sementara untuk nilai rata-rata pada setiap AUROC yaitu stabil diatas 90. Untuk hubungan sebab-akibat, kami hanya menggunakan satu dataset yang mendapatkan kinerja tertinggi. Faktor yang mempengaruhi kinerja siswa untuk tiap skema berbeda-beda. Namun, faktor yang selalu mempengaruhi kinerja siswa di tiap skemanya adalah durasi studi. Variabel durasi studi ini menandakan mahasiswa memiliki retensi yang baik dalam mengakses sumber daya hingga akhir pembelajaran.

Kata kunci – Kausalitas, SMOTE, Seleksi Fitur, Bayesian Network, Kinerja Belajar Siswa