ABSTRACT

The dropout rate on Massive Open Online Courses (MOOCs) is still very high. One effort that can be done to reduce the dropout level is to interfere at-risk students. Hence a well performed prediction method is needed. Several methods have been applied to overcome this problem including statistic, machine learning and deep learning. Machine learning method is the most popular one. But the condition where the dropout class is dominating the samples makes the prediction result of any single machine learning method becomes not powerful and not stable.

In this study, two approaches are applied at the same time namely the approach at the data level and algorithm level. At the data level, the Synthetic Minority Over-sampling (SMOTE) technique is carried out at the preprocessing stage to balance the number of instances in the train-set. While at the algorithm level, the final prediction is done using the Ensemble Learning (EL) technique with three base-learners namely: Logistic Regression (LR), K-Nearest Neighbor (KNN) and Random Forest (RF). The algorithm was tested using Google Collaboratory on public dataset from KDDCUP2015. The results of the proposed method are evaluated and compared with the results of previous studies.

The result shows that SMOTE-Ensemble Learning (SEL) able to improve prediction performance with an average improvement of the harmonic mean of precision and recall (F1-score): 7.74% compared to previous work. SEL is more stable shown by the short range of F1-Score. This research is expected to be used as a reference for research in the field of dropout prediction in MOOCs.

Keywords: *Data Mining, Ensemble Learning, SMOTE, Dropout Prediction, Massive Open Online Courses*

INTISARI

Kasus *dropout* pada *Massive Open Online Courses* (MOOCs) masih sangat tinggi. Salah satu upaya yang dapat dilakukan untuk menekan tingginya angka *dropout* adalah dengan melakukan intervensi terhadap siswa yang berisiko. Oleh karena itu, diperlukan sistem prediksi yang handal. Berbagai metode mulai dari statistik, *machine learning* maupun *deep learning* sudah banyak diterapkan untuk mengatasi permasalahan tersebut. *Machine learning* menjadi yang paling populer dalam kasus ini. Namun, adanya dominasi pada kelas siswa *dropout* kerap menyebabkan hasil prediksi yang dihasilkan oleh berbagai metode *machine learning* tunggal menjadi kurang handal dan kurang stabil.

Pada penelitian ini, diterapkan dua pendekatan sekaligus yakni pendekatan pada *level* data dan *level* algoritme. Pada level data, *Synthetic Minority Over-sampling Technique* (SMOTE*)* dilakukan pada tahap *preprocessing* guna menyeimbangkan jumlah *instance* pada data latih. Sedangkan pada level algoritme, prediksi akhir dilakukan dengan teknik *Ensemble Learning* (EL) dengan tiga *base learner* yakni: *Logistic Regression* (LR), *K-Nearest Neighbor* (KNN) dan *Random Forest* (RF). Algoritme diujikan menggunakan Google Colaboratory pada lima kursus berbeda yang diambil dari data publik dari *platform* MOOC bernama XuetangX. Kinerja dari metode diukur dan dibandingkan dengan penelitian sebelumnya.

Hasil pengujian menunjukkan bahwa algoritme SMOTE-*Ensemble Learning* (SEL) terbukti lebih handal yang ditunjukkan dengan adanya kenaikan pada nilai rerata harmonis (F1-*Score*) antara *Precision* dan *Recall* yakni 7.74% apabila dibandingkan dengan penelitian sebelumnya. Metode SEL juga lebih stabil yang ditunjukkan dengan adanya *range* yang sempit untuk nilai *F1-Score.* Penelitian ini diharapkan dapat digunakan sebagai acuan untuk penelitian dibidang deteksi dini siswa yang beresiko pada pembelajaran berbasis MOOC.

Kata kunci -- *Data Mining, Ensemble Learning, SMOTE, Dropout Prediction, Massive Open Online Courses*