Data mining using filtering approaches and ensemble methods

Syifa Faradilla Fabrianne¹, Agung Triayudi², Ira Diana Sholihat³
Information Systems Department, Universitas Nasional, Jakarta, Indonesia

*agungtriayudi@civitas.unas.ac.id

Abstract. To develop a prediction paradigm, ensemble methods such as boosting based on the heuristic system can be used. Compared to using individual classifiers, the prediction results using ensemble learning techniques are usually more accurate. In this study, several ensemble techniques were discussed to obtain comprehensive knowledge about key methods. Among the various ensemble methods, a boosting mechanism has been implemented to predict student achievement. Because the ensemble method is considered an actual event in a prediction and classification, the boosting technique is used to develop an accurate predictive pedagogical model. The utilization of this boosting technique is based on the nature of each method proposed in educational data mining. By using the ensemble method and filtering approach, the predictions of student performance showed a substantial increase.

1. Introduction

Among researchers, the topic of developing pedagogical data to find productive information is currently being discussed. In this case, many data mining techniques are applied to achieve better insights from each academic data warehouse [1][2]. Knowledge extraction plays a significant role in driving the education wheel through various data mining techniques [3][4]. In educational data mining, predicting a student's performance is an important task [5][6][7]. On another occasion, a research study conducted suggested a prediction system using the Adaboost algorithm to provide the right solution in minimizing student failure [8][9]. However, some researchers [10][11] are known to have predicted student achievement using the ensemble approach because this method has enormous power in the accuracy of classifier predictions after the combined output is formed. Apart from the complexity of the data, this method also has the ability to achieve greater accuracy, thereby reducing overfitting of individual classifiers and generalization errors [12][13][14].

Diversity among classifiers is one of the key aspects with a major role in the application of this ensemble mechanism, which can be achieved through various subsamples of input data, as in the case of improvement [15]. There are of course some important and fundamental factors that differentiate the categories of ensemble methods [16]. In addition, classifiers in an ensemble can be sequential or concurrent. Based on the learning mechanism during the training phase and testing a sub-sample of the dataset, the ensemble is divided into two categories, namely ensemble base learning and Meta [17]. In this study, researchers used the Waikato Environment for Knowledge Analysis (WEKA) as a tool for data mining techniques. WEKA is a software that can perform various data mining processes by applying algorithms from machine learning.
2. Methodology

2.1 Dataset

In this study, the dataset's pedagogical data taken from a private university in Jakarta, Indonesia, the informatics engineering faculty. The data processed is data from 4 different classes: Programming class (Desktop-based), Business, Accounting, and Web Programming. In total, there are 60 rows of data taken from each class. In table 1, a description of the attributes and values used in the dataset has been presented.

| No | Fields                                | Description                        |
|----|---------------------------------------|------------------------------------|
| 1  | Demography                            | Immigrant (0), Indigenous (1)      |
| 2  | PR_Task A (Programming Task A)        | 0-25 (Possible Values)             |
| 3  | PR_Task B (Programming Task B)        | 0-25 (Possible Values)             |
| 4  | PR_Practice (Programming Practice)    | 0-50 (Possible Values)             |
| 5  | PR_Total (Programming Total)          | 0-100 (Possible Values)            |
| 6  | BS_Task A (Business Task A)           | 0-50 (Possible Values)             |
| 7  | BS_Task B (Business Task B)           | 0-50 (Possible Values)             |
| 8  | BS_Total (Business Total)             | 0-100 (Possible Values)            |
| 9  | AC_Task A (Accounting Task A)         | 0-25 (Possible Values)             |
| 10 | AC_Task B (Accounting Task B)         | 0-25 (Possible Values)             |
| 11 | AC_Practice (Accounting Practice)     | 0-50 (Possible Values)             |
| 12 | AC_Total (Accounting Total)           | 0-100 (Possible Values)            |
| 13 | WP_Task A (Web Programming Task A)    | 0-25 (Possible Values)             |
| 14 | WP_Task B (Web Programming Task B)    | 0-25 (Possible Values)             |
| 15 | WP_Practice (Web Programming Practice)| 0-50 (Possible Values)             |
| 16 | WP_Total (Web Programming Total)      | 0-100 (Possible Values)            |
| 17 | Total_Marks (Total Marks)             | 400                                 |
| 18 | Total_Obt (Total Obtained)            | 0-400 (Possible Values)            |
| 19 | Overall Grade                         | Cluster 1, Cluster 2, Cluster 3    |

Base ensemble learning or the ensemble method is an algorithm that is usually used in machine learning to find the best predictive solution. In Base Ensemble Learning, four methods can be used. The four methods are: 1) Voting is an essential technique in distributed learning algorithms that involves calculating all learning algorithms related to average class allocation; 2) Bagging is an ensemble method that simultaneously spreads out to process sub-samples to improve the classification accuracy of combining several models; 3) Boosting is a method used to improve classification accuracy performance; 4) X-Validation or the extensive ensemble procedures in ordinary machine learning or selection via cross validation are carried out through a 10-fold cross validation system.

![Model of Meta Ensemble Learning](image-url)
In Meta Ensemble Learning approach, a prediction is more focused on the meta-level rather than the base level. A single classifier known as the meta classifier is fully responsible for classifying the existing examples, then combining their predictions using the voting method.

In Meta Ensemble Learning, two methods can be used. Two methods are: 1) Stacking is a method that aims to achieve the most precise generalization accuracy; 2) Grading is applying a learning algorithm based on a meta classifier as a criterion in classifying instances at the meta-level. In this method, a weighted voting approach is still carried out and is responsible until the final class prediction.

The Arbiter Tree method is a method developed through a bottom-up method. Generally, the dataset is randomly divided into 'n' disjoint sub-partitions and stimulated by a combination of two learning algorithms.

![Figure 2. Schematic of two base learning algorithms and sole arbitrator](image)

### 3. Result and Discussion

In this Individual Performance Learning Classifier subsection, the data will process to predict student achievement using basic algorithms and meta-learning such as J48, k-NN (k-Nearest Neighbor), Naïve Bayes, Random Tree, and Boosting. The classification was carried out using filtering procedures, namely Synthetic Minority Oversampling Technique (SMOTE) and Spread Subsampling. Both methods are used throughout the education dataset to achieve significant results. Besides, if the dataset to be processed is irrelevant, the level of prediction accuracy will be low.

| CN            | CC      | TPR   | FPR   | Precision | Recall | F-Measure | ROC   | RAE    |
|---------------|---------|-------|-------|-----------|--------|-----------|-------|--------|
| Random Tree   | 75%     | 0.75  | 0.144 | 0.754     | 0.75   | 0.751     | 0.803 | 40.36% |
| Naïve Bayes   | 91.67%  | 0.917 | 0.046 | 0.915     | 0.917  | 0.914     | 0.982 | 15.59% |
| KNN           | 78.33%  | 0.783 | 0.145 | 0.78      | 0.783  | 0.78      | 0.819 | 38.18% |
| J48           | 85%     | 0.85  | 0.102 | 0.85      | 0.85   | 0.85      | 0.885 | 27.61% |

Table 2 has presented the results of some of the attributes of prior use of filtering approaches. Nine parameters have an essential role in this research, such as Correctly Classified (CC), True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F-Measure, Receiving Operating Characteristics (ROC), and Relative Absolute. Error (RAE). In the table above, it can also be seen that the Naïve Bayes classifier has the highest CC value, which is 91.67%.

In this Base Classification Results using the SMOTE Method subsection, the base classifier results from the entire dataset discussed using one of the filtering approaches, namely the SMOTE method.

| CN            | CC      | TPR   | FPR   | Precision | Recall | F-Measure | ROC   | RAE    |
|---------------|---------|-------|-------|-----------|--------|-----------|-------|--------|
| Random Tree   | 82.86%  | 0.829 | 0.084 | 0.829     | 0.829  | 0.827     | 0.872 | 26.06% |
| Naïve Bayes   | 94.28%  | 0.943 | 0.028 | 0.943     | 0.943  | 0.942     | 0.993 | 11.41% |
| KNN           | 85.71%  | 0.857 | 0.084 | 0.867     | 0.857  | 0.85      | 0.887 | 25.56% |
| J48           | 87.14%  | 0.871 | 0.066 | 0.871     | 0.871  | 0.87      | 0.89  | 21.22% |

Table 3 shows the results of a base classifier with the SMOTE method. As a result, there was an increase in CC values in the four classifiers. However, there is also a decrease in the RAE value, which indicates that the SMOTE method’s application has gone well.
In this Base Classification Results using the Spread Subsampling Method subsection, the results of the base classifier from the entire dataset discussed using one of the filtering approaches, namely the Spread Subsampling method.

Table 4. Results of the base classifier with the Spread Subsampling method

| CN          | CC      | TPR | FPR  | Precision | Recall | F-Measure | ROC   | RAE   |
|-------------|---------|-----|------|-----------|--------|-----------|-------|-------|
| Random Tree | 82.86%  | 0.829| 0.088| 0.825     | 0.829  | 0.825     | 0.87  | 26.07%|
| Naïve Bayes | 95.71%  | 0.957| 0.022| 0.959     | 0.957  | 0.957     | 0.991 | 10.21%|
| KNN         | 84.29%  | 0.843| 0.09 | 0.853     | 0.843  | 0.836     | 0.877 | 27.41%|
| J48         | 82.86%  | 0.829| 0.085| 0.836     | 0.829  | 0.831     | 0.881 | 28.29%|

In Table 4 has presented the results of a base classifier using Subsampling Spread. The table above shows that there are not many significant changes. There was a decrease in the CC value of the J48 classifier and the KNN classifier. Conversely, there was an increase in the CC value of the Naïve Bayes classifier, and the Random Tree classifier, the CC value was still unchanged at 82.86%.

Figure 3. The degree of difference in the results of classification by various methods

Figure 3 shows the level of differences in the results of classifications before using filtering approaches, after using the SMOTE method, and after using the Spread Subsampling method. Overall, the Naïve Bayes classifier's CC value has the highest increase in value among other classifiers. The most significant decrease was found for the RAE value in the Random Tree classifier and the kNN classifier. The Boosting approach is a learning classifier that can be applied to educational datasets to increase predictive accuracy. By using the boosting approach, the classifier is expected to get a higher value than before. Table 5 is the result of the classifier using the boosting approach. As a result, it can be seen that there is an increase in two classifiers, there are J48 classifier and Random Tree classifier.

Table 5. Results classifier by boosting approach

| CN          | CC      | TPR | FPR  | Precision | Recall | F-Measure | ROC   | RAE   |
|-------------|---------|-----|------|-----------|--------|-----------|-------|-------|
| Boost with Random Tree | 83.33%  | 0.833| 0.074| 0.83      | 0.833  | 0.831     | 0.88  | 26.91%|
| Boost with Naïve Bayes | 88.33%  | 0.883| 0.064| 0.887     | 0.883  | 0.88      | 0.959 | 18.11%|
| Boost with KNN       | 78.33%  | 0.783| 0.145| 0.78      | 0.783  | 0.78      | 0.819 | 38.81%|
| Boost with J48       | 90%     | 0.9  | 0.067| 0.905     | 0.9    | 0.901     | 0.942 | 15.61%|

Table 6 shows the results of the classifier boosting approach with one of the oversampling approaches, SMOTE. As a result, the CC values in the four classifiers all increased. Besides, the RAE values for the four classifiers have also increased.
Table 6. Results classifier boosting approach by SMOTE method

| CN       | CC         | TPR | FPR | Precision | Recall | F-Measure | ROC    | RAE   |
|----------|------------|-----|-----|-----------|--------|-----------|--------|-------|
| Boost with Random Tree | 85.56% | 0.856 | 0.066 | 0.857 | 0.856 | 0.855 | 0.895 | 22.38% |
| Boost with Naïve Bayes   | 96.67% | 0.967 | 0.012 | 0.968 | 0.967 | 0.967 | 0.998 | 5.88%  |
| Boost with KNN           | 90.00% | 0.9  | 0.056 | 0.903 | 0.9  | 0.895 | 0.922 | 18.63% |
| Boost with J48           | 94.44% | 0.944 | 0.026 | 0.944 | 0.944 | 0.944 | 0.961 | 8.69%  |

Table 7 shows the results of the boosting approach classifier with the Spread Subsampling method. As a result, the CC value of the three classifiers has increased, but one other classifier, the Naïve Bayes classifier, has decreased. For RAE values, all classifiers have an increase.

Table 7. Results classifier boosting approach with Spread Sub-sampling method

| CN       | CC         | TPR | FPR | Precision | Recall | F-Measure | ROC    | RAE   |
|----------|------------|-----|-----|-----------|--------|-----------|--------|-------|
| Boost with Random Tree | 84.29% | 0.843 | 0.074 | 0.846 | 0.843 | 0.843 | 0.885 | 23.90% |
| Boost with Naïve Bayes   | 95.71% | 0.957 | 0.018 | 0.958 | 0.957 | 0.957 | 0.978 | 8.25%  |
| Boost with KNN           | 85.71% | 0.857 | 0.084 | 0.867 | 0.857 | 0.857 | 0.887 | 25.35% |
| Boost with J48           | 91.43% | 0.914 | 0.044 | 0.914 | 0.914 | 0.913 | 0.937 | 13.06% |

In Figure 4, a histogram is presented that interprets the CC and RAE values of the four classifiers that existed before using the filtering approach, after using the SMOTE method, and after using the Spread Subsample method. The CC value of the four classifiers increases after using the SMOTE method. Based on the existing histogram, it can be seen that combining the use of the SMOTE method and the boosting approach has the highest success rate for increasing the CC value.

Figure 4. Histogram for comparison of CC and RAE values with the boosting approach.

Figure 5. Histogram of comparison of the ROC values of the three clusters.

In Figure 5, a histogram visualization of the comparison of the Receiving Operating Characteristics (ROC) values of the three existing clusters presented. It can be seen that cluster 2 has a value that tends to be higher in all classifiers, followed by cluster 1 then the last cluster 3. Besides, in Figure 6, the screenshot results of the dataset in which cluster predictions are presented. There are several differences between the predictions that have been made previously and the predictions obtained after doing this research.

Figure 6. Screenshot of cluster prediction dataset.
4. Conclusion
In this study, several classifiers have been evaluated and analyzed using the SMOTE method and the Spread Subsample method in the WEKA application. This study's main objective is to determine the methods that can be used to predict the results of student performance in a classroom in college. The combination of the Ensemble method with the Filtering Approaches has a substantial impact on the predictive accuracy of learning classifiers. The application of the ensemble method is considered an essential event in the prediction and classification procedure. The boosting technique itself is widely used to develop accurate predictive pedagogical models. After comparisons between several stand-alone and combined methods, the researcher found a method that has a reasonably high predictive accuracy value, namely one of the ensemble methods, namely boosting with the SMOTE method. Employing a combination of the ensemble method with the filtering approach, the prediction accuracy value will be higher, and the percentage of Relative Absolute Error (RAE) will decrease.

References

[1] Ashraf, M., Zaman, M., & Ahmed, M. (2020). An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches. Procedia Computer Science, 167, 1471-1483.

[2] Yulianto, L., Triyudi, A. and Sholihati, I. (2020) “Implementation Educational Data Mining For Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4.5”, Jurnal Mantik, 4(1, May), pp. 441-451.

[3] Nakayama, M., Mutsuura, K., & Yamamoto, H. (2018). Using note taking instructions to reform student’s note taking activities and improve learning performance in a blended learning course. International Conference Information Visualization, 326-331.

[4] Romero, C., Romero, J. R., & Ventura, S. (2014). A survey on pre-processing educational data. In Educational data mining (pp. 29-64). Springer, Cham.

[5] Ashraf, M., Zaman, M., & Ahmed, M. (2020). An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches. Procedia Computer Science, 167, 1471-1483.

[6] Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., & Pritchard, D. E. (2012). Model-Based Collaborative Filtering Analysis of Student Response Data: Machine-Learning Item Response Theory. International Educational Data Mining Society.

[7] Lu, O. H., Huang, A. Y., Huang, J. C., Lin, A. J., Ogata, H., & Yang, S. J. (2018). Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning. Educational Technology & Society, 220-232.

[8] Park, Y., Yu, J.H., & Jo, I. (2016). Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute☆. Internet and Higher Education, 29, 1-11.

[9] Rahman, M. H., & Islam, M. R. (2017, December). Predict Student's Academic Performance and Evaluate the Impact of Different Attributes on the Performance Using Data Mining Techniques. In 2017 2nd International Conference on Electrical & Electronic Engineering (ICEEE) (pp. 1-4). IEEE.

[10] Sutoyo, E., & Almaarif, A. (2020). Educational Data Mining for Predicting Student Graduation Using the Naïve Bayes Classifier Algorithm. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 4(1), 95-101.

[11] Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4), 49-64.

[12] U. Ali, K. S. Arif and D. Usman Qamar, "A Hybrid Scheme for Feature Selection of High Dimensional Educational Data," 2019 International Conference on Communication
Technologies (ComTech), Rawalpindi, Pakistan, 2019, pp. 71-75, doi: 10.1109/COMTECH.2019.8737829.

[13] Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students’ academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447-459.

[14] Triayudi, Agung, and Iskandar Fitri. "A new agglomerative hierarchical clustering to model student activity in online learning." Telkomnika 17, no. 3 (2019): 1226-35.

[15] Triayudi, A., and I. Fitri. "Alg Clustering To Analyze The Behavioural Patterns Of Online Learning Students." Journal of Theoretical & Applied Information Technology 96, no. 16 (2018): 5327-5337.

[16] Triayudi, Agung, Sumiati Sumiati, Thoha Nurhadiyan, and Vidila Rosalina. "Data Mining Implementation to Predict Sales Using Time Series Method." Proceeding of the Electrical Engineering Computer Science and Informatics 7 (2020): 1-6.

[17] Triayudi, Agung, Widyarto, O.W, Rosalina, Vidila. "CLG Clustering for Mapping Pattern Analysis of Student Academic Achievement." ICIC Express Letters 14, no. 12 (2020): 1225-1234.