SYSTEMIC: Information System and Informatics Journal

ISSN: 2460-8092, 2548-6551 (e)

Vol 5 No 1 - Agustus 2018

Feature Selection With the RandomForest Packages to Predict Student Performance

Slamet Wiyono[1], Dyah Apriliani[2], Taufiq Abidin[3], Dairoh[4]

[1,2,3] Teknik Informatika Politeknik Harapan Bersama, Tegal, Indonesia

slamet2wiyono@gmail.com[1]

Kata Kunci

Seleksi Fitur
Random Forest
R Studio
Prediksi
Kinerja Mahasiswa

Abstrak

Setiap program studi berupaya meningkatkan kualitas pendidikan dan akreditasi. Salah satu elemen yang menjadi nilai akreditasi adalah siswa yang lulus tepat waktu. Semakin banyak mahasiswa aktif, semakin banyak mahasiswa akan lulus tepat waktu. Dengan demikian, kepala program studi perlu membuat prediksi mahasiswa yang akan tidak aktif di semester berikutnya. Untuk membuat prediksi, kita harus menentukan fitur apa yang dibutuhkan. Artikel ini adalah hasil dari riset pemilihan fitur untuk memprediksi status aktif mahasiswa. Pilihan fitur menggunakan tujuh fitur menggunakan paket RandomForest dari R Studio. Satu fitur sebagai output adalah status aktif mahasiswa dan enam fitur sebagai input i.e; grade point (GP), grade point average (IPK), pekerjaan orang tua, jurusan sekolah, kategori sekolah, dan kota asal mahasiswa. Hasil pemilihan fitur menunjukkan fitur terkuat hingga terlemah adalah; nilai poin (GP), nilai poin rata-rata (IPK), pekerjaan orang tua, jurusan asal, sekolah asal, dan kota asal mahasiswa.

Keywords

Feature selection
Random forest
R Studio
Prediction
Student performance

Abstract

Each study program seeks to improve the quality of education and accreditation. One element that becomes the value of accreditation is students who graduate on time. The more active students, the more students will graduate on time. Thus, the head of the study program needs to make predictions of students who will be inactive in the next semester. To make predictions, we must determine what features are needed. This article is the result of feature selection research to predict the active status of students. The selection of features using seven features using the RandomForest package from R Studio. One feature as output is the active status of students and six features as input i.e; grade point (GP), grade point average (GPA), parent work, school majors, school category, and student hometown. The results of the selection of features show the strongest features to the weakest are; grade points (GP), grade point average (GPA), work of parents, majors of origin, schools of origin, and student hometown.

1. INTRODUCTION

Each study program seeks to improve the quality of education and accreditation. One element that becomes the value of accreditation is students who graduate on time [1]. The more students who graduate on time, the better the value of accreditation. Non-active students will influence graduate on time. Handling of potentially non-active students is needed to prevent non-active students. With this prevention, it is expected to reduce the number of non-active students, so that the graduation rate on time has increased. With the increase in graduation rates on time, it is expected to further enhance the accreditation of study programs.

Research on the predictions of student activity has been done. Among them are researched to predict the students' performance of the Faculty of Computer Science, Dian Nuswantoro University using Decision Tree Algorithm [2]. Other than, the same research had also carried out using the KNN algorithm [3]. The research that had carried out using one algorithm. Thus there is no comparison, so it is possible to have another algorithm that is better for making predictions.

DOI:10.29080/systemic.v5i1.618

Research to predict student activity by comparing several algorithms were many done. Some research i.e: research by comparing Support Vector Machine (SVM) and Decision Tree algorithms [4], comparing of J48, Random Forest, Multilayer Perceptron, IB1, and Decision Table algorithm [5], comparing of Logistic Regression, Decision Tree, Naïve Bayes, dan Neural Network algorithm [6], comparing of K-Nearest Neighbor (KNN), Support Vector Machine (SVM) , and Random Forest algorithm [7], perbandingan algoritme Decision Tree, Support Vector Machine (SVM), dan Naïve Bayes [8]. Comparing of Support Vector Machine (SVM), Neural Network, Naïve Bayes, and K-Nearest Neighbor (KNN) algorithm [9], and comparing of K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree algorithm [10]. From some of the research, the researcher only focuses on the selection of algorithms to prediction. The selection of the right algorithm will certainly produce accurate predictions. However, the right algorithm will produce a prediction that is less accurate if the attributes used for predictions are incorrect.

Feature selection is needed to identify what factors influence student performance. feature selection is one important means to attack problems with various aspects of data, and to enable existing tools to apply, otherwise not possible [11]. One of the algorithms commonly used to features selection is Random Forest. A random forest (RF) classifier is an ensemble classifier that produces multiple decision trees, using a randomly selected subset of training samples and variables [12]. Some research on feature selection using these algorithms i.e; research for the classification and features selection for diagnosis and prediction of breast cancer [13], airborne lidar feature selection for urban classification [14], and feature selection for protein division [15]. Research had shown that using feature selection will improve performance [16]. The purpose of this research was doing features selection to look for features that most influence student performance.

2. RESEARCH METHOD

The data used in the research were from data of Informatics Engineering student of Politeknik Harapan Bersama from 2014 to 2017 (1530 observation). Data used included: grade point (GP), grade point average (GPA), parent work, school majors, school category, student hometown, and active status of students. The tool used in this study was R Studio software. This tool was used to features selection using RandomForest packages. The RandomForest package is an implementation of Breiman's random forest algorithm (based on Breiman and Cutler's Fortran code) which is used for classification and regression. Thus, it is possible to calculate estimates between data points [17]. The research procedure is shown in Figure 1.

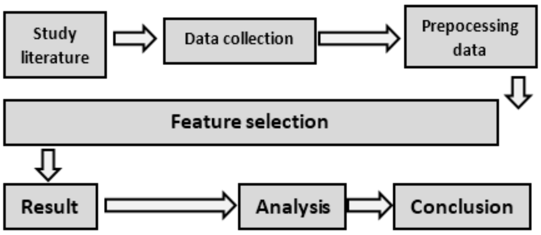

Fig.1 Research Procedure

Study Literature

Sources of literature come from journals, proceedings, and books. Study literature continues to be carried out in tandem with the other research stages until the end of the study. This has done so that in the next stage other sources of reference that support the research were found, the reference sources can be used as literature to help complete the research that was conducted.

Data Collection

The data used in the study are data from Informatics Engineering students of Harapan Bersama Polytechnic from 2014 to 2016. The data used include GPA, credits taken, hometown, origin school, parent work, and student activity every semester.

Preprocessing Data

At this stage, the input and output data or target data are determined. In addition, data normalization is also carried out, namely by converting character data into numerical data.

Feature Selection

Feature selection is done to determine what features are most influential. In addition, feature selection is done to get fewer features so that it will facilitate the computing process.

Result

At this stage, scores are obtained for each feature that affects student activity

Analysis

The analysis is done by analyzing the score of each feature.

Conclusion

The final results are the most influential features and which features are less influential.

The randomForest () function owned by R Studio (the randomForest package) is an interesting part of rpart (), has sufficient complexity, and often provides accurate prediction results.. “For each of a large number of bootstrap samples (by default, 500) trees are independently grown”. “In addition, a new random sample of variables is chosen for use with each new tree”. “The out-of-bag (OOB) prediction for each observation is determined by a simple majority vote across trees whose bootstrap sample did not include that observation”. “Trees are grown to their maximum extent, limited however by nodesize (minimum number of trees at a node)”. “Additionally, maxnodes can be used to limit the number of nodes”. “There is no equivalent to the parameter cp”. “The main tuning parameter is the number mtry of variables that are randomly sampled at each split”. “The default is the square root of the total number of variables; this is often satisfactory”. “It may seem surprising that it is (usually) beneficial to take a random sample of variables”. Essentially, mtry controls the trade-off between the amount of information in each individual tree, and the correlation between trees. A very high correlation limits the ability of an individual tree to convey information that is specific to that tree” [18].

Gain Ratio

“The gain ratio is an extension of the information gain measure, which attempts to overcome the bias that the information gain measure is prone to selecting features with a large number of values” [13]. “Thereby, the information gain measure is used as an attribute selection measure of the decision tree and is obtained by computing the difference between the expected information requirement, classifying a tuple in tuples, and the new information requirement for attribute A after the partitioning. The measure of the expected information requirement is given by” [19]

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

“where *m* is the number of distinct classes; pi indicates the probability by calculating the proportion of belonging to class Ci in tuples *D*. The new information requirement for attribute *A* is measured by”

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \tag{2}$$

“where v indicates that D was divided into v partitions or subsets, {D1,D2,⋯,Dv}. Thus, the information gain measure Gain(A) for attribute A can be calculated by the formula”.

$$\mathrm{Gain}(A) = Info(D) - Info_A(D) \tag{3}$$

“Then, a ‘split information’ function was used to normalize the information gain measure Gain(A). The split information function was defined by”

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \tag{4}$$

Finally, to calculate the size of the information gain used a profit ratio with the calculation of Gain (A) divided by SplitInfo (A) which is a measure of information split.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \tag{5}$$

The best feature is seen from the size of the gain ratio. The greater the value of the gain ratio, the more important the feature.

Random Forest

“The feature evaluation approach based on random forest is known as an embedded method” [20] “and provides a variable importance criterion for each feature by computing the mean decrease in the classification accuracy for the out of bag (OOB) data from bootstrap sampling” [21]. “Assuming bootstrap samples b = 1, …, B, the mean decrease in classification accuracy D¯¯j for variable xj as the importance measure is given by”

$$\overline{D}_j = \frac{1}{B}\sum_{b=1}^{B}(R_b^{oob} - R_{bj}^{oob}) \tag{6}$$

where Roobb denotes the classification accuracy for OOB data ℓoobb using the classification model Tb; and Roobbj is the classification accuracy for OOB data ℓoobbj permuted the values of variable xj in ℓoobb (j = 1, …, N). Last, the z-score of the xj variable representing the most important variable can be found by using the calculation with the formula zj = D¯¯jsj / B√, after the standard deviation sj from the decrease in classification accuracy is calculated. In this study, the feature evaluation procedure is performed automatically using the 'RandomForest' R package.

Correlation-based feature selection

“Unlike the feature evaluation methods mentioned above, a feature subset was evaluated simply by using the filter algorithm Correlation-based Feature Selection (CFS). The CFS assessed the worth of a set of features using a heuristic evaluation function based on the correlation of features, and Hall and Holmes” [22] claiming that most features must be correlated with classes that are highly uncorrelated with each other. Thereby, the formula below is used to evaluate the criteria for a subset.

$$merits = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \quad (7)$$

The symbol f is a feature where; c is class, rcc is the average feature correlation with class, rfff is the average feature between correlations, and k is the number of attributes in a subset. To explore the feature space the first best search is used, and the five subsets that do not develop sequentially have been set to stop criteria to avoid searching for the entire subset of feature space.

3. RESULTS AND DISCUSSION

The most influential feature on student-active status i.e; grade point (GP), grade point average (GPA), parent's work, school majors, school category, and student's hometown. The correlation value score is shown in Figure 2. Figure 2 shows the Mean Decrease in Gini correlation score. The sequence of feature correlation values with student-active status is shown in Table 1. Table 1 shows the sequence of correlation values biggest to the smallest.

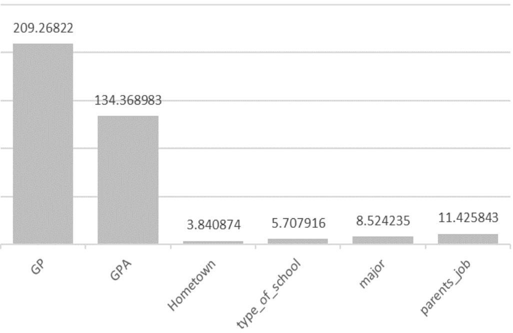

Fig 2. Correlation Value Score

The feature selection was done using the RandomForest packages R Studio using seven features. One feature as output is student-active status and six features as input i.e: grade point (GP), grade point average (GPA), parent work, school majors, school category, student hometown. The results of feature selection as shown in Table 4.1 show that the strongest feature that influences is grade point (GP) with score 209.27 and grade point average with score 134.37. This shows that the activity of students in the next semester is strongly influenced by GP and GPA. The lower the GP and GPA, the more potential it is not active in the next semester. The lowest score that influences the student-active in the next semester is student's hometown with score 3.8 and school category with score 5.7. This shows that students from Tegal and its surroundings, as well as those from outside Tegal, do not have a strong influence on predicting student-active in the next semester. Likewise, the school category both from public and private schools do not have a strong influence on predicting student-active in the next semester. Parent work and school majors have a middle influence. This shows that the income of parents and majors from students has enough influence on student academic status. The lower the income of parents of students, the more potentially in-active in the next semester, and students from IT majors have a higher potential for active-status than students who come from science majors or even others.

Table 1. Sequence of Correlation Value Score

Seq	Feature	Score
1	Grade point (GP)	209.27
2	Grade point average (GPA)	134.37
3	Parent’s work	11.43
4	School majors	8.52
5	School category	5.71
6	Student’s hometown	3.84

Based on the results of the study, efforts are needed from the management of study programs to increase GP and GPA score so that students who are potentially in-active will decrease. In addition, efforts are also needed from the head of the study program to provide information relating to tuition funding assistance such as scholarships or the assistance of student side jobs. Thus students who have economic problems caused by a lack of parents' income can be handled. In addition, the head of the study program must also pay special attention to students who come from other than IT and or Science. Thus students who have difficulty in following academic activities can be handled.

4. CONCLUSION

The features that most influence the activity of students in the next semester i.e: grade points (GP), grade point average (GPA), work of parents, majors of origin, schools of origin, and student hometown. Then, to increase the number of students who are active in the next semester, the head of the study program needs to make efforts to improve the academic score of students. In addition, The head of the study program also needs to provide information relating to funding assistance for tuition such as scholarships or side job assistance for students.

Acknowledgments

Thank you for P3M (Pusat Penelitian dan Pengabdian Kepada Masyarakat) Politeknik Harapan Bersama which had funded this research

REFERENCES

[1] BAN-PT, *Buku I Naskah Akademik Akreditasi Institusi Perguruan Tinggi*. Jakarta: Badan Akreditasi Nasional Perguruan Tinggi, 2011.

[2] D. Untari, "Data Mining Untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Metode Decision Tree C4.5," Universitas Dian Nuswantoro, 2014.

[3] S. Wiyono and T. Abidin, "IMPLEMENTATION OF K-NEAREST NEIGHBOUR (KNN) ALGORITHM TO PREDICT STUDENT ' S PERFORMANCE," *SIMETRIS*, vol. 9, no. 2, pp. 873–878, 2018.

[4] S. Wiyono, "Perbandingan Algoritma Machine Learning SVM dan Decision Tree untuk Prediksi Keaktifan Mahasiswa," *SINKRON*, vol. 3, no. 1, pp. 105–108, 2018.

[5] M. S. Mythili and A. R. M. Shanavas, "An Analysis of students ' performance using classification algorithms," *IOSR-JCE*, vol. 16, no. 1, pp. 63–69, 2014.

[6] K. Hastuti, "Analisis Komparasi Algoritma Klasifikasi Data Mining untuk Prediksi Mahasiswa Non Aktif," vol. 2012, no. Semantik, pp. 241–249, 2012.

[7] R. A. Nugraheni and K. Mutijarsa, "Comparative Analysis of Machine Learning KNN, SVM, and Random Forests Algorithm for Facial Expression Classification," in *ISEMANTIC*, 2016, pp. 163–168.

[8] P. Suryachandra and P. V. S. Reddy, "Comparison of Machine Learning Algorithms for Breast Cancer," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, 2016, vol. 3, pp. 1–6.

[9] N. Zerrouki, F. Harrou, A. Houacine, and Y. Sun, "Fall Detection Using Supervised Machine Learning Algorithms: A comparative study," in *2016 8th International Conference on Modelling, Identification and Control (ICMIC)*, 2016, pp. 665–670.

[10] S. Wiyono and T. Abidin, "COMPARATIVE STUDY OF MACHINE LEARNING KNN , SVM , AND DECISION TREE ALGORITHM TO PREDICT STUDENT ' S PERFORMANCE," *IJRG*, vol. 7, no. January, pp. 190–196, 2019.

[11] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Singapore: Springer Science+Business Media, 1998.

[12] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 24–31, Apr. 2016.

[13] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random Forest Classifier Combined with Feature Selection for Breast Cancer Diagnosis and Prognostic," vol. 2013, no. May, pp. 551–560, 2013.

[14] N. Chehata, L. Guo, and C. Mallet, "AIRBORNE LIDAR FEATURE SELECTION FOR URBAN CLASSIFICATION USING RANDOM FOREST," *Int. Arch. Photogramm. - Citeseer*, vol. XXXVIII, no. c, pp. 207–212, 2009.

[15] B. Li, Y. Cai, K. Feng, and G. Zhao, "Prediction of Protein Cleavage Site with Feature Selection by Random Forest," vol. 7, no. 9, pp. 1–9, 2012.

[16] Q. Zhou, H. Zhou, and T. Li, "Cost-sensitive Feature Selection Using Random Forest: Selecting Low-cost Subsets of Informative Features," *Knowledge-Based Syst.*, vol. 95, pp. 1–11, 2016.

[17] L. Breiman, A. Cutler, A. Liaw, and M. Wiener, "Breiman and Cutler's Random Forest for Classification and Regression," 2018, p. 17.

[18] J. Maindonald and J. Braun, *Data Analysis and Graphics Using R*. Cambridge, 2010.

[19] M. Alves, A. Roberto, C. Daleles, C. Atzberger, D. Alves, and M. Pupin, "Remote Sensing of Environment Object Based Image Analysis and Data Mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas," *Remote Sens. Environ.*, vol. 123, pp. 553–562, 2012.

[20] M. Pal and G. M. Foody, "Feature Selection for Classification of Hyperspectral Data by SVM," vol. 48, no. 5, pp. 2297–2307, 2010.

[21] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests : A survey and results of new tests," *Pattern Recognit.*, vol. 44, no. 2, pp. 330–349, 2011.

[22] M. A. Hall and G. Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining," vol. 15, no. 6, pp. 1437–1447, 2003.