Application of a combination between Principal Component Analysis and Logistic Regression Based on Support Vector Machine on Educational Data Mining with Overlapping Data Problem

Siti Mutrofin1*, Maisarah Maisarah1, Slamet Widodo1, Raden Venantius Hari Ginardi2, Chastine Fatichah2

1Universitas Pesantren Tinggi Darul Ulum, Jombang, Indonesia
2Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

*sitimutrofin@ft.unipdu.ac.id

Abstract. In 2019, the government of the Republic of Indonesia issued a zoning-based policy for New Student Admissions (PPDB) from the level of elementary school (SD) to high school (SMA), especially for public schools. The policy is documented in Permendikbud No.51 / 2018. The government policy aims to ensure the equality of education and make prospective students not focus only on favorite schools. However, this policy raises new problems. One of them is that if the potential student has got a medium UN (National Examination) score or medium distance of the house to the destination school, then his potential to be accepted at the destination school is very small. It is even worse if the potential students do not know the lowest score and the farthest distance the destination school can accept. Thus, potential students will choose schools by only guessing without basing on valid data, so their chances of being accepted will be very small. This current research focused on Educational Data Mining at PPDB Public High School (SMA) in Jombang in the academic year 2019/2020 which aims to accommodate the needs of potential students to predict the destination schools based on their own grades and home distances using classification techniques of data mining. However, another problem emerged in this study. An overlapping data occurred where one data was also owned by more than one class. For example, a potential student of SMA Negeri 2 Jombang (SMAN 2 Jombang) has got a score of 80 in Bahasa Indonesia subject, which is the same as that of a student from SMA Negeri 3 Jombang (SMAN 3 Jombang). Data overlapping does not only occur in one data but almost all of the data. The data used in this study were 600 data, consisting of 308 from PPDB 2019 of SMAN 2 Jombang, and the rest were from SMAN 3 Jombang. The attributes used were the home distance from the destination school, overall UN scores, UN scores of Mathematics, Natural Sciences, Bahasa Indonesia, and English subjects.

The algorithm used was a combination of Principal Component Analysis (PCA) with Logistic Regression (LR)-based Support Vector Machine (SVM) with Anova kernel. The validation applied 10-fold cross-validation and the evaluation of algorithm performance used the aspects of accuracy, precision, and recall. The results of this current study showed an accuracy of 94.33%, a precision of 96.28%, and a recall of 92.53%. The results were better than those that did not apply PCA (70.83% accuracy, 69.62% precision, and 76.62% recall). By PCA, data could be seen from another angle that could separate or differentiate one class from the others. Even though there were 100% overlapping data, none of them, from all attributes, was 100% exactly the same.

1. Introduction
In 2019, the Republic of Indonesia government issued a zoning-based policy for New Student Admission (PPDB) process starting from elementary schools (SD) to senior high schools (SMA), especially for public schools. The policy is documented in Permendikbud No.51 / 2018 [1]. The policies are made to make the quality of education in Indonesia equitable and there will be no favorite school anymore which commonly relied heavily on the best scores of the students in the admission process. This policy raises new problems for the potential students whose scores are moderate and home distance from the destination school is not so close, especially in densely populated areas such as urban areas. Their potential to be accepted in the destination school is very small. If potential students’ homes are close to public schools, they will tend to underestimate the result of the National Examination (UN). An example of complaints came from one of students’ parents, Albert Mercelino, whose son registered in SMA Negeri 1 Kuta Utara, who said, “Keep using general zoning, not a radius. Use the score ranking too. That’s fair. The zoning policy is just like we teach our children not to struggle. The fools will relax. Learning or not, the most important thing is close to (school)” [2]. The second complaint came from Fitri Suhermin, one of the parents of a potential student of SMP Negeri 8 Surabaya, as quoted from the Indopos daily, “Disappointment because her child could not be accepted in SMP Negeri 8 which was only 700 meters from his home, which was due to the zoning system.” [3].

This current research was conducted to predict the potential students who were potentially accepted in the Public Senior High Schools in Jombang Regency, East Java Province, Indonesia by implementing Educational Data Mining with classification techniques. The general goal of this current research was to support the government policy and help provide recommendations to potential students in choosing the potential schools. However, in this current research overlapping data problems occurred, in which each datum overlaps one another among classes (note Table 1 printed in red), which were the data accepted in SMA Negeri 2 Jombang (SMADAJO) and those accepted in SMA Negeri 3 Jombang (SMAGAJO). Overlapping data problem is not only in one attribute of each data, but all the attributes of all data. Any algorithm undergoing overlapping data problems will not produce a good classification [4]. The solution is by removing dimensions or attributes. If not removed, the existing data may not be able to distinguish among classes, because there are the exact same data in both classes [4]. According to Gu and Cheng, overlapping data problems become one of the obstacles in data mining and pattern recognition because they affect the accuracy of the classification and the ability to generalize directly [5]. The specific purpose of this current research is to find out an algorithm that is suitable for overcoming the overlapping data problems. The solution to overlapping data problems is like Suyanto's statement, [4] “To well understand a problem, look at it from another perspective”. For example, by looking at data from a certain angle, they would look normal, not overlap.

To realize the objectives of this research, several algorithms were chosen according to the characteristics of the research data. One of the algorithms commonly used to overcome an overlapping data problem is the Principal Component Analysis (PCA) algorithm [4] [6] [7]. According to Suyanto [4]), "PCA is a mathematical method that transforms data into a new domain which results in many more important principle components". Whereas for overcoming overlapping data problems, the most
suitable one is Logistic Regression [8] based on Support Vector Machine (SVM) which refers to myKLR [9], because it has been proven to have good resistance [10]. The kernel used on both PCA and SVM is the Anova kernel, because according to Stitson et al. [11] ANOVA has proven to have good multi-dimensional performance.

2. Proposed method
The method proposed involved collecting datasets, dimensionality reduction, classification, validation, and evaluating the performance of classification algorithms. Illustrations related to the proposed method are presented in Figure 1.

2.1. Data
1.1. Data The data in this study were obtained directly from PPDB of the Government of the East Java Province of Indonesia in the academic year 2019/2020. The data sources were two public high schools in Jombang: SMADAJO and SMAGAJO. Both of them were used as classes in this study. The number of data was 600 data, consisting of 308 students of SMADAJO, and the rest were those accepted in SMAGAJO. Six attributes were applied: The home from the chosen school, the total scores of the National Exam (UN), the UN scores of Mathematics, Natural Sciences, English, and Bahasa Indonesia. All attributes for all data underwent overlapping data problems, but not all had exactly the same data between those in the SMADAJO and SMAGAJO classes. Data could only be accessed when the PPDB registration was opened. The East Java Provincial PPDB system could be accessed on https://ppdbjatim.net. All data were numeric both integer and real. The illustration of the dataset in this study can be seen in Table 1.

2.2. Dimensionality reduction
Overlapping data problem is one of the common problems in data mining and pattern recognition [5]. Adopting Suyanto [4], the illustration of data undergoing overlapping data problems were presented in Table 2 in the form of five objects with three attributes or dimensions (length, width, and height in meters) grouped into two classes (Tables and Chairs); Table 3 was in the form of five objects with two attributes or dimensions (length and width in meters) grouped into two classes (Tables and Chairs). The height dimension was removed because they could not distinguish between classes. In Table 2 and Table 3 there were overlapping data problems marked with red-printed scores. In figure 2 visualizing the data of five objects in a two-dimensional space (length and width), the position of a circle was close to two triangles; In figure 3 visualizing the data of five objects in one dimension of width (the dimension of length was removed), two overlapping objects occurred because both had the same width.
of 1.5; In figure 4 visualizing the data of five objects in one dimension of length (the dimension of width was removed), two overlapping objects occurred because they had the same length of 2.1; In figure 5 visualizing the data from another perspective. Tilting head to the right or left to see whether the data were randomly scattered or rather normal? The visualization of data could be handled by PCA, where the first principal component (PC1) and the second principal component (PC2) were the new dimensions in the PCA realm. The five objects could be separated into two different classes (triangles and circles) with one dividing line (dashed).

Table 2. Five objects with three attributes or dimensions (length, width, and height in meters) [4]

| Object | Length | Width | Height | Class |
|--------|--------|-------|--------|-------|
| Object 1 | 2.1    | 1.5   | 0.8    | Table |
| Object 1 | 2.3    | 1.7   | 0.8    | Table |
| Object 1 | 2.1    | 1.3   | 0.8    | Chair |
| Object 1 | 1.6    | 1.5   | 0.8    | Chair |
| Object 1 | 2.5    | 1.9   | 0.8    | Table |

Figure 2. Visualization of data of five objects in a two-dimensional space (length and width) [4]

Table 3. Five objects with two attributes or dimensions (length and width in meters) [4]

| Object | Length (L) | Width (W) | Class | Symbol |
|--------|------------|-----------|-------|--------|
| Object 1 | 2.1        | 1.5       | Table | ●      |
| Object 1 | 2.3        | 1.7       | Table | ●      |
| Object 1 | 2.1        | 1.3       | Chair | ▴      |
| Object 1 | 1.6        | 1.5       | Chair | ▴      |
| Object 1 | 2.5        | 1.9       | Table | ●      |
In this study, all selected algorithms were run by using RapidMiner with ANOVA kernel and the gamma score of 1 and a degree of 3. The scores were the default scores of RapidMiner. This ANOVA kernel was also used in the SVM kernel.

2.3. Classification
The classification algorithm involved three algorithms: Logistic Regression, Logistic Regression based on Support Vector Machine (SVM) (in RapidMiner known as Logistic Regression (SVM) referring to myKLR by Rueping [12]), and PCA with Logistic Regression (SVM). Three algorithms were chosen to confirm the expectation of this research that the combination of PCA, Logistic
Regression (SVM) and ANOVA kernel would result in better results compared to other algorithms. All algorithms in this study, from beginning to end of the process, applied RapidMiner.

![Figure 6. Illustration of the stratified 10-fold cross validation][13]

### 2.4. Validation

Validation in this study used the default from RapidMiner, namely 10-fold cross-validation with Stratified random sampling. The stratified random sampling was chosen because the distribution between training sets and testing sets can be spread evenly compared to other random samplings. The illustration of stratified 10-fold cross-validation is presented in Table 4 refers to Wahono [13].

### 2.5. Evaluation

The evaluation of algorithms applied confusion matrix which may produce accuracy, precision, and recall. These three evaluations were chosen because accuracy is not enough; another evaluation was needed to ensure that the performance of the algorithm in this study was really good. The illustration of the difference between accuracy, precision, and recall is presented in Figure 7 which refers to Data [14]. The illustration of the confusion matrix is presented in Figure 8 [15]. In True Positive (TP), a datum is predicted to be wrong because the actual datum is wrong. In False Positive (FP) a datum is predicted to be wrong but the actual datum is not wrong. In True Negative (TN), a datum is predicted not to be wrong because the actual datum is not wrong. In False Negative (FN), a datum is predicted not to be wrong but the actual datum is wrong. The formulas for accuracy, precision, and recall are shown in Table 5.

![Figure 7. Illustration of the difference between accuracy, precision, and recall according to Data][14]
Table 4. Confusion matrix [15]

| Object          | True SMADAJO | True SMAGAJO |
|-----------------|--------------|--------------|
| Prediction SMADAJO | TP           | FP           |
| Prediction SMAGAJO | FN           | TN           |

Table 5. Compound performance measures [15]

| Measures                                      | Defined as                                |
|-----------------------------------------------|-------------------------------------------|
| Accuracy / Correct Classification Rate (CCR)  | $\frac{TP + TN}{TP + TN + FP + FN}$      |
| Precision                                     | $\frac{TP}{TP + FP}$                      |
| Recall / True Positive Rate / Sensitivity / Probability of Detection (pd) | $\frac{TP}{TP + FN}$                      |

3. Experimental Results

The experiments in this study were carried out three times, using Logistic Regression, Logistic Regression (SVM), and PCA with Logistic Regression (SVM). The results of this study are presented in the form of a confusion matrix, as in Figure 8 which presents the results of Logistic Regression. Figure 9 presents the results of the Logistic Regression (SVM). Figure 10 presents the results of a combination of PCA with Logistic Regression (SVM). A more complete result is presented in Table 6.

![Figure 8](image1.png)

Figure 8. Illustration of the confusion matrix as a result of Logistic Regression

![Figure 9](image2.png)

Figure 9. Illustration of the confusion matrix as a result of Logistic Regression (SVM) with ANOVA kernel

![Figure 10](image3.png)

Figure 10. Illustration of the confusion matrix as a result of a combination of PCA and Logistic Regression (SVM) with the ANOVA kernel ANOVA

Table 6. Comparison of the three algorithms [15]

| Algorithms                        | Accuracy | Precision | Recall  |
|-----------------------------------|----------|-----------|---------|
| Logistic Regression               | 70.83%   | 69.62%    | 76.62%  |
| Logistic Regression (SVM)         | 72.33%   | 72.47%    | 74.35%  |
| PCA dan Logistic Regression (SVM) | 94.33%   | 96.28%    | 92.53%  |
Based on the experimental results shown in Figure 8, Figure 9, Figure 10, and Table 6 it was proven that the combination of PCA and Logistic Regression (SVM) with ANOVA kernel had better performance, as printed in blue; whereas the Logistic Regression had very poor performance, as printed in red. Therefore, it can be concluded that overlapping data problems can be overcome by PCA with Logistic Regression (SVM) with ANOVA kernel for predicting potential students to be potentially accepted in selected high schools. For further research, especially in the case of multi classes, it is better to know whether they can be solved by applying the same method, or whether the method proposed will only be suitable for binary classification cases. This is because there are more than two Public high schools in Jombang, and the potential students can choose not only SMADAJO and SMAGAJO.

4. Conclusion
The problem of this research is related to the overlapping data problems that occurred in the PPDB dataset of Public High Schools in Jombang which could be overcome by applying a combination of PCA algorithm and Logistic Regression (SVM) algorithm with ANOVA kernel. Based on this study, the proposed method successfully outperformed compared to not using PCA, with an accuracy of 94.33%, precision 96.28%, and 92.53% recall. For further research, it is better to prove whether the proposed algorithm can also be applied to multiclass cases, not just binary classes as in this study.

5. Acknowledgment
The researcher is grateful to the Directorate of Research and Community Service, Director General of Research and Development Reinforcement, Ministry of Research, Technology and Higher Education, for financing this research on Inter-University colaboration in 2018 with titled “Educational Data Mining Berbasis Klasifikasi untuk Menganalisis Mahasiswa Berpotensi Berhenti Kuliah pada Kasus Imbalanced Class Distribution”.

References

[1] L. P. Wardhana and K. Supriyoko, "Manajemen Penerimaan Peserta Didik Baru Secara Online Berbasis Zonasi, Prestasi dan Perpindahan," MMP (Media Manajemen Pendidikan), vol. 2, no. 2, pp. 228-235, 2019.
[2] P. Suyatra, "PPDB, Belajar Nggak Belajar, yang Penting Rumah Dekat Sekolah," Bali Express, 2 July 2019. [Online]. Available: https://baliexpress.jawapos.com/read/2019/07/02/144253/ppdb-belajar-nggak-belajar-yang-penting-rumah-dekat-sekolah. [Accessed 2 November 2019].
[3] J. Armanto, "Nah, Jokowi Akui Pelaksanaan PPDB Sistem Zonasi Banyak Masalah," Indopos, 21 June 2019. [Online]. Available: https://indopos.co.id/read/2019/06/21/178778/nah-jokowi-akui-pelaksanaan-ppdb-sistem-zonasi-banyak-masalah/. [Accessed 2 November 2019].
[4] S. Suyanto, Data Mining untuk Klasifikasi dan Klasterisasi Data, Bandung: Informatika, 2019.
[5] Y. Gu and L. Cheng, "Classification Of Class Overlapping Datasets By Kernel-MTS Method," International Journal of Innovative Computing, Information and Control, vol. 13, no. 5, pp. 1759-1767, 2017.
[6] N. Zhang, K. Leatham, J. Xiong and J. Zhong, "PCA-K-means Based Clustering Algorithm for High Dimensional and Overlapping Spectra Signals," in Ninth International Conference on Intelligent Control and Information Processing (ICICIP), Wanzhou, China, 2018.
[7] A. Füzy, R. Kovács, I. Cseresnyés, I. Parádi, T. Szili-Kovács, B. Kelemen, K. Rajkai and T. Takács, "Selection of plant physiological parameters to detect stress effects in pot experiments using principal component analysis," Acta Physiologiae Plantarum, vol. 41, no. 5, pp. 55-56, 2019.
[8] A. Nöu, Logistic Regression versus Support Vector Machines, Stockholm: Stockholms universitet, 2018.
[9] S. Rüping, "myKLR - kernel logistic regression," Technische Universität Dortmund, 2003.
[Online]. Available: https://www-ai.cs.uni-dortmund.de/SOFTWARE/MYKLR/index.html#Keerthi/etal/2003a. [Accessed 2 November 2019].

[10] A. Christmann, I. Steinwart and M. Hubert, "Robust learning from bites for data mining," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 347-361, 2017.

[11] T. Hofmann, B. Schölkopf and A. J. Smola, "Kernel Methods in Machine Learning," *The Annals of Statistics*, vol. 36, no. 3, pp. 1171-1220, 2008.

[12] V. V. Raghavan, V. N. Gudiwada, V. Govindaraju and C. R. Rao, Cognitive computing: Theory and applications, Elsevier, 2016.

[13] R. S. Wahono, N. S. Herman and S. Ahmad, "A comparison framework of classification models for software defect prediction," *Advanced Science Letters*, vol. 20, no. 10-12, pp. 1945-1950, 2014.

[14] S. Mutrofin, M. M. Machfud, D. H. Satyareni, R. V. H. Ginardi and C. Fatichah, "Komparasi Kinerja Algoritma C4.5, Gradient Boosting Trees, Random Forests, dan Deep Learning pada Kasus Educational Data Mining," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 7, no. 3, 2020.

[15] D. Bowes, T. Hall and D. Gray, "Comparing the performance of fault prediction models which report multiple performance measures: recomputing the confusion matrix," in *Proceedings of the 8th International Conference on Predictive Models in Software Engineering (PROMISE '12)*, New York, USA, 2012.

[16] B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad, “Classification of natural disaster prone areas in Indonesia using K-means,” *Int. J. Grid Distrib. Comput.*, vol. 11, no. 8, pp. 87–98, 2018.

[17] A. P. Windarto et al., “Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province,” *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019, doi: 10.1088/1742-6596/1255/1/012001.

[18] D. Hartama, A. Perdana Windarto, and A. Wanto, “The Application of Data Mining in Determining Patterns of Interest of High School Graduates,” *J. Phys. Conf. Ser.*, vol. 1339, no. 1, 2019, doi: 10.1088/1742-6596/1339/1/012042.

[19] Sudirman, A. P. Windarto, and A. Wanto, “Data mining tools | rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, no. 1, 2018, doi: 10.1088/1757-899X/420/1/012089.