Emerging Information Science and Technology
Vol. 1, No. 3, (2020), pp. 93-103
DOI: 10.18196/eist.v1i3.10627

Implementation of Multiclass Support Vector Machine for Classification of New Students Receiving Achievement Scholarships at Universitas Muhammadiyah Yogyakarta

Nurfahmi[1], Slamet Riyadi[1*], and Asroni[1]

[1]*Universitas Muhammadiyah Yogyakarta, Jln.Brawijaya, Tamantirto, Kasihan, Bantul, Yogyakarta 55183, Indonesia*

**Corresponding author: riyadi @umy.ac.id*

Abstract

The selection process for scholarship grantees at Universitas Muhammadiyah Yogyakarta (UMY) still utilizes the conventional method, namely Microsoft Excel. It is conducted by inputting all student data, then sorted from the highest to the lowest. Scholarships for prospective new students must be right on the target, meaning that they meet the criteria for students eligible for scholarships. It is intended not to harm other prospective students who should get scholarships. In previous studies, the classification process used many algorithms such as Naive Bayes, C.45, Decision Tree, k-Nearest Neighbor, and Support Vector Machine. The use of the Support Vector Machine algorithm employed a two-class classification. Support Vector Machine and Decision Tree algorithms are two classification methods that can obtain precise and accurate results. This study aims to use the Multiclass Support Vector Machine (LibSVM) algorithm to classify the achievement scholarship rankings for new students. The minimum amount of data used affected the classification results. From the 2015 to 2019 data, the highest amount of data was 2015, obtaining the highest accuracy result of 84.34% using the sigmoid kernel type and the k-fold value of 3. The classification was based on the entry system stages. PMDK stage 2 obtained an accuracy of 81.38%, with the most data amounting to 268 from stage 3.

Keywords: *Scholarship Selection, Classification, Multiclass Support Vector Machine (LibSVM)*

1. Introduction

Scholarships are assistance in the form of educational support funds usually given to students, both for those excelling and those economically disadvantaged. Several institutions usually award scholarships. The university is one of the educational institutions routinely providing scholarships for selected students through several selection stages it determined.

In data mining, classification techniques are placing or selecting the right attributes and parameters based on the dataset to be classified in determining the accuracy of the calculation process. The classification process in previous studies used many algorithms, such as Naive Bayes, C.45, Decision Tree, k-Nearest Neighbor and Support Vector Machine. The Support Vector Machine algorithm only used a two-class classification. The Support Vector Machine and Decision Tree algorithm are two classification techniques obtaining precise and accurate results (Somvanshi, Chavan, & et al., 2016).

The selection process for scholarship grantees at the University of Muhammadiyah Yogyakarta (UMY) uses a conventional method, namely Microsoft Excel. It is performed by inputting all student data, then sorted from the highest to the lowest. Prospective students must meet the criteria for scholarship grantees to be right on target, thereby not harming others more eligible to receive the scholarships. This study aims to utilize the multiclass Support Vector Machine (LibSVM) algorithm to classify new students' achievement scholarship rankings.

2. Method

This research was conducted through several stages, as presented in Figure 1. In this study, literature study were conducted well before focus on specific issue. The rest of other step were the continuing activity after the issue about New Students Receiving analysis become the target of this research study.

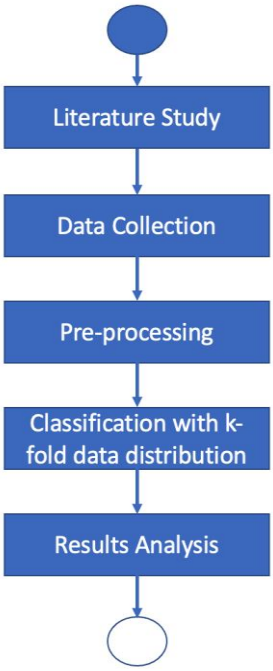

Figure 1. Research Stages

2.1 Literature Study

A literature study is the first stage of this research. This stage aims to collect some information about previous studies, particularly on the methods and algorithms used.

2.2 Data Collection

The second stage is data collection. There are two types of data sources, primary and secondary. This study used secondary data sources obtained from second parties. The data were obtained from the Bureau of Information Facilities (BSI) of Universitas Muhammadiyah Yogyakarta. The data obtained had 16 attributes, including study program, student number, entry number, name, gender, entry

system, entry system name, school origin, city, province, high school major, the reason for retreat, retreat information, rank, choice 1, and choice 2.

Original data obtained from BSI of UMY amounted to 1,401 records. Of the 1,401 data obtained, only 730 could be used in the data mining process because many were nulls. There was also report card value data. There were six attributes in the report card value data comprising the entry number, type of report card, semester, subject, value and k13 value. Report card value data are presented in Figure 3.3.

2.3 Pre-Processing

Pre-processing in data mining was carried out through four stages: data cleaning, data integration, data selection and data transformation. The authors did not use data integration at this stage since no data requiring it. Each stage of data processing is explained as follows:

a. Data Cleaning

Data cleaning is a step to eliminate null, noise, and inconsistent data. The data obtained had many deficiencies, such as missing or null data. Data cleaning is crucial in the data mining stage. Thus, the data would not affect algorithm performance.

b. Data Selection

Data selection is a step to determine the attributes used in the data mining process. The selected attributes were data relevant to the analysis to be carried out. The data used were school origin, gender, entry system, regional origin, mathematics, chemistry, physics, biology, and rank. The data chosen for the entry system attributes were PMDK 1, PMDK 2, and PMDK 3 because this research was conducted on the entry system of interests and talents.

c. Data Transformation

Data Transformation is a process of changing attributes and is consolidated into a form suitable for the data mining process. The attributes changed included school origin, entry system, regional origin and average score. Table 1 shows the attributes and its category.

From Table 1, The target attribute of rank consists of four categories:

1. Non-Ranked is for students who do not get scholarships or discount tuition fees
2. PBUD 1 is for students ranked first by obtaining free education development funds (DPP) in the first year and free from fixed and variable school administration contributions (SPP) in the first year
3. PBUD 2 is for students ranked second by obtaining free education development funds (DPP) in the first semester and free from fixed and variable school administration contributions (SPP) in the first semester
4. PBUD 3 is for students ranked third by obtaining free education development funds (DPP) in the first semester

Table 1. Dataset Attributes

Attribute	Attribute Name	Category
Target	Rank	Non-Ranked
		PBUD 1
		PBUD 2
		PBUD 3
Input	School Origin	SMAN
		MAN
		OTHERS
	Gender	M
		F
	Entry System	PMDK 1
		PMDK 2
		PMDK 3
	Origin	Java
		Outside Java
	Average Mathematics Score	A
		B
		C
	Average Chemical Score	A
		B
		C
	Average Physics Score	A
		B
		C
	Average Biology Score	A
		B
		C
	Average English Score	A
		B
		C

The input of School Origin attribute of rank consists of three categories:

1. SMAN is The name of the school starting with SMAN
2. MAN The name of the school starting with MAN
3. OTHERS School names starting with other than SMAN and MAN

The input of Entry System attribute of rank consists of three categories:

1. PMDK 1 is PMDK stage 1 and PMDK batch 1
2. PMDK 2 is PMDK stage 2 and PMDK batch 2
3. PMDK 3 is PMDK stage 3 and PMDK batch 3

The input of Registrant Origin attribute of rank consists of three categories:

1. Java is registrant who comes from Java
2. Outside Java is registrant who comes from other places beside Java

The input of Average Score attribute of rank consists of three categories:

1. A is average score that range from 85 to 100
2. B is average score that range from 75 to 84
3. C is average score that under 75

2.4 Classification

This stage is the primary and most important process in the preparation of the final project. At this stage, the authors began modeling to implement the Support Vector Machine (LibSVM) algorithm using the RapidMiner software. This implementation used testing and training data, resulting in an information.

3. Results and Discussion

This study used a dataset of new students in the PMDK entry system, totaling 730 samples. The dataset came from the last five years of data, batch 2015 to 2019. This dataset had four categories of rankings: Non-Ranked, PBUD 1, PBUD 2 and PBUD 3.

3.1 Data Mining Capitalization

At this stage, the authors built data mining classification models. There were two classification models created, using data based on years and entry stages.

Support Vector Machine (LibSVM) can only test datasets with numeric properties. Meanwhile, the dataset obtained from the BSI was categorical. Then, the data set was converted into numeric form. Data were converted using a blending type, namely nominal to numerical, as shown in Figure 2. The dataset operator was connected with the nominal to numerical operator through the port out and port exa.

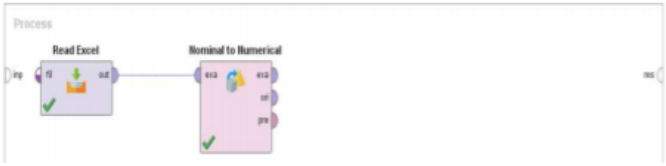

Figure 2. Convert Dataset

3.2 Data Classification by Year

In this test, all datasets were used, namely 730 data divided by batch. Training and testing data were shared automatically by the Cross-Validation operator. The first step taken in this test is demonstrated in Figure 3, namely adding the data to be tested, the nominal to the numerical operator and the Cross-Validation operator on the process sheet and connecting the operators with the appropriate ports. Hence, operators can connect to one another.

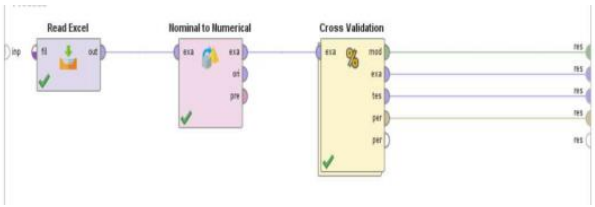

Figure 3. Add Cross-Validation

Next, the authors set the Cross-Validation operator by double-clicking it, and then it appeared, as depicted in Figure 4. The authors then added the SVM operator (LibSVM) in the training section, connected the port tra and port mod in the training section with the port tra and port mod on the SVM operator.

In the testing section, the researchers added the apply model and performance operators, then connected the port mod and port test in the testing section with the port mod and port uni on the apply model operator, then connected the port lab on the apply model operator with the port lab on the operator performance, then connected the port per on the operator performance with the port per in the testing section and connected the port exa with the port test.

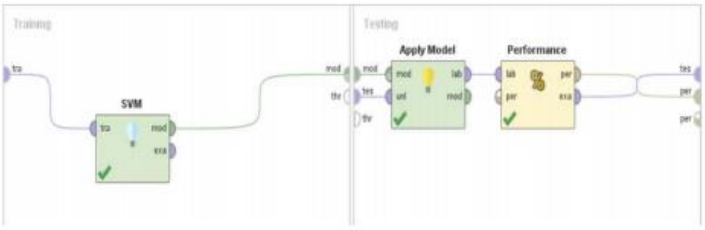

Figure 4. Closs Validation Sheet

The next step was to set the parameters on the Cross-Validation operator, as displayed in Figure 5. The operator set is the number of folds for three repetitions using the shuffled sampling type.

Figure 5. Cross-Validation Parameters

The SVM operator (LibSVM) has used the SVM C_SVC parameter type and kernel type poly with a gamma value of 0.0, C of 0.0, cache size of 80, and epsilon of 0.001, as displayed in Figure 6, according to the previous test results obtaining the highest accuracy. At this stage, the classification process was tested using three kernel types, including poly, rfb and sigmoid using three k-fold values of 3, 6, and 9.

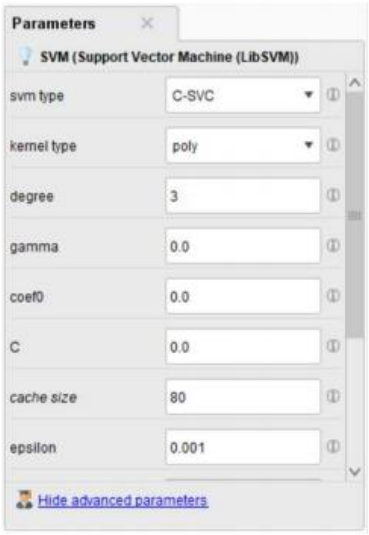

Figure 6. LibSVM Parameters

The classification of batch 2015 consists of 172 data, specifically Non-Ranked, three PBUD 1, nine PBUD 2, and 21 PBUD 3. Table 5 depicts that classification results using the sigmoid kernel type and the k-fold value of 3 obtain an accuracy of 84.34%, higher than using the kernel type and other k-fold values. The best accuracy results in the 2015 batch of data reveal that the number of classification accuracy of Non-Ranked = 137 data, PBUD 1 = 0, PBUD 2 = 0, PBUD 3 = 8 and the number of incorrect classifications of Non-Ranked = 2, PBUD 1 = 3, PBUD 2 = 9, PBUD 3 = 13.

Table 2. Classification Results of Batch 2015

k-fold	*Accuracy*		
	poly	*rbf*	*sigmoid*
k=3	83.76%	80.85%	**84.34%**
k=6	83.85%	80.87%	83.23%
k=9	83.74%	80.85%	81.40%

Table 6 shows the results of batch 2016 data classification with a total of 130 data, comprising Non-Ranked = 90 data, PBUD 1 = 3 data, PBUD 2 = 10 data, and PBUD 3 = 27 data. In the classification process using 2016 data, the highest accuracy is 71.57% using kernel type poly with k-fold value of 6. The classification process using 2016 data obtain the highest accuracy results, with the number of classification accuracy of Non-Ranked = 87 data, PBUD 1 = 0 , PBUD 2 = 0, PBUD 3 = 6 and the number of incorrect classifications of Non-Ranked = 3 data, PBUD 1 = 3, PBUD 2 = 10, PBUD 3 = 21.

Table 3. Classification Results of Batch 2016

k-fold	*Accuracy*		
	poly	*rbf*	*sigmoid*
k=3	70.82%	69.27%	70.03%
k=6	**71.57%**	69.34%	69.30%
k=9	70.79%	69.26%	69.21%

The classification process using data batch 2017 obtain the highest accuracy results using kernel type poly with k-fold values of 6 and 9, obtaining the same accuracy results. Data for batch 2017 consists of 144 data, with Non-Ranked = 108 data, PBUD 1 = 2 data, PBUD 2 = 6 data, and PBUD 3 = 28 data. The highest classification result is 78.47%. The overall classification results are depicted in Table 7. The best accuracy results of batch 2017 data obtain the number of classification accuracy of Non-Ranked = 105 data, PBUD 1 = 0, PBUD 2 = 1, PBUD 3 = 7 data and the number of incorrect classifications of Non-Ranked = 3 data, PBUD 1 = 2, PBUD 2 = 5, PBUD 3 = 21.

Table 4. Classification Results of Batch 2017

k-fold	*Accuracy*		
	poly	*rbf*	*sigmoid*
k=3	76.39%	75.00%	75.00%
k=6	**78.47%**	75.00%	77.08%
k=9	**78.47%**	75.00%	77.78%

Data for batch 2018 comprise 165 data, with Non-Ranked = 136 data, PBUD 1 = 2 data, PBUD 2 = 8 data, and PBUD 3 = 19 data. Table 8 displays the comparison of the accuracy results of the classification process for each kernel and the k-fold value. The accuracy result using the sigmoid kernel type with the k-fold = 3 value obtains the highest accuracy than the others with an accuracy value of 82.42%. The classification process using data from batch 2018 obtain classification accuracy for the number of correct classifications, namely Non-Ranked = 136, PBUD 1 = 0, PBUD 2 = 0, PBUD 3 = 0 and for the number of incorrect classifications, Non-ranked = 0 data, PBUD 1 = 2, PBUD 2 = 8, and PBUD 3 = 19.

Table 5. Classification Results of Batch 2018

k-fold	*Accuracy*		
	poly	*rbf*	*sigmoid*
k=3	81.82%	81.42%	**82.42%**
k=6	81.81%	82.41%	82.41%
k=9	82.39%	82.39%	82.39%

The 2019 batch data classification results obtain the highest accuracy result of 68.62% using kernel type poly with k-fold = 9. This classification process used data totaling 120 data, with Non-Ranked = 81 data, PBUD 1 = 3 data, PBUD 2 = 9 data, and PBUD 3 = 27 data. Table 9 is the result of the classification process of several kernel types and k-fold values. The classification process using 2019 data obtain the

highest accuracy result with the number of classification accuracy of Non-Ranked = 75 data, PBUD 1 = 0, PBUD 2 = 0, PBUD 3 = 7 and the number of incorrect classifications comprises No-ranked = 6 data, PBUD 1 = 3, PBUD 2 = 9, and PBUD 3 = 20.

Table 6. Classification Results of Batch 2019

k-fold	*Accuracy*		
	poly	*rbf*	*sigmoid*
k=3	65.00%	67.50%	65.00%
k=6	65.83%	67.50%	65.83%
k=9	**68.62%**	67.70%	65.26%

3.3 Data Classification Based on Path Stages

The next classification process was based on the entry system stages. In this stage classification, the attributes were omitted because the entry system stages have the same values. According to the existing entry system stages, this classification process was divided into PMDK 1, PMDK 2 and PMDK 3.

The classification process using PMDK 1 utilized 253 data, in which Non-Ranked data = 187 data, PBUD 1 = 6 data, PBUD 2 = 15 data, and PBUD 3 = 45 data. Classification using this first stage obtain the highest accuracy result of 75.49% using kernel type poly and k-fold value of 3. Table 10 presents the classification results using PMDK 1 data. The classification results obtained from the highest accuracy value are the number of classification accuracy of Non-Ranked = 181 data, PBUD 1 = 0 data, PBUD 2 = 2 data, PBUD 3 = 8 data and the number of incorrect classifications of Non-Ranked = 6 data, PBUD 1 = 6 data, PBUD 2 = 13 data, PBUD 3 = 37 data.

Table 7. Classification Results of PMDK Stage 1

k-fold	*Accuracy*		
	poly	*rbf*	*sigmoid*
k=3	**75.49%**	73.91%	74.30%
k=6	73.51%	73.90%	72.71%
k=9	73.51%	73.92%	71.54%

PMDK 2 data classification used 268 data, where Non-Ranked data = 218, PBUD 1 = 4, PBUD 2 = 14, and PBUD 3 = 32. This classification process can be seen in table 11. This classification obtains the same four results, namely using the rbf and sigmoid type kernels using the k-fold values of 3 and 9.The highest accuracy result obtained is 81.38%, with the number of classification accuracy of Non-Ranked = 218, PBUD 1 = 0, PBUD 2 = 0, PBUD 3 = 0 and for the number of incorrect classifications of Non-Ranked = 0 data, PBUD 1 = 4, PBUD 2 = 14, and PBUD 3 = 32.

Table 8. Classification Results of PMDK Stage 2

k-fold	*Accuracy*		
	poly	*rbf*	*sigmoid*
k=3	81.00%	**81.38%**	**81.38%**
k=6	80.24%	81.37%	81.37%
k=9	80.26%	**81.38%**	**81.38%**

The PMDK 3 data classification process used 209 data consisting of Non-Ranked = 149 data, PBUD 1 = 3 data, PBUD 2 = 13 data, and PBUD 3 = 44 data. Table 12 shows the classification results of PMDK 1 data, which obtained the highest accuracy using kernel type poly with k-fold value of 9. This classification obtain the highest accuracy result of 78.99% with the number of classification accuracy of Non-Ranked = 143 data, PBUD 1 = 0, PBUD 2 = 0, PBUD 3 = 22 and the number of incorrect classifications of Non-Ranked = 6, PBUD 1 = 3, PBUD 2 = 13, and PBUD 3 = 22.

Table 9. Classification Results of PMDK Stage 3

k-fold	*Accuracy*		
	poly	*rbf*	*sigmoid*
k=3	74.13%	71.26%	68.88%
k=6	78.95%	71.25%	72.20%
k=9	**78.99%**	71.30%	69.87%

3.4 Process Analysis and Classification Results

The Cross-Validation operator carried out the process of sharing testing data and training data in this classification process. Cross-validation folded the data according to the k-fold value and repeated the classification process as much as the k-fold value. The distribution of testing and training data was different in each classification process according to the k-fold value. If the k-fold value is 3, then the amount of data classified is divided into three folds. In the first classification process, the testing data used was the data in the first fold. For the second repetition process, the testing data used was the data in the second fold, and so on until the repetition of the process was complete to as many as k-fold values.

Based on the research conducted, many data affected the accuracy results obtained from the classification process. The more data used during the classification process, the higher the average accuracy result obtained. Likewise, the fewer data used, the lower the accuracy results obtained.

Apart from many data, kernel and k-fold value choice also significantly affected accuracy results. Each kernel and k-fold value obtained different results. From this research, kernel type poly provided the highest accuracy result than other kernel types.

4. Conclusions

After carrying out a series of model developments for testing and analyzing the Classification of New Students' Achievement Scholarship Grantees of the Faculty of Engineering, Universitas Muhammadiyah Yogyakarta, the conclusion is that the Support Vector Machine algorithm could be implemented to classify new students awarded achievement scholarships, where the amount or the minimum amount of data used affected the classification results. From the 2015 to 2019 data, batch 2015 had the highest number of 172 and the highest accuracy result of 84.34% using the sigmoid kernel type and the k-fold value of 3 and the classification based on the entry system stage, PMDK stage 2 obtained an accuracy of 81.38% with the most data from stage 3 as many as 268.

References

[1] Dillak, R. Y., Pangestuty, D. M., dan Bintiri, M. G. "Klasifikasi Jenis Musik Berdasarkan File Audio Menggunakan Jaringan Syaraf Tiruan Learning Vector Quantization". *in Seminar Nasional Informatika 2012 (semnasIF 2012) UPN "Veteran" Yogyakarta*, Yogyakarta, 2012.

[2] E. P. Wigner, "On a modification of the Rayleigh–Schrodinger perturbation theory," (in German), *Math. Naturwiss. Anz. Ungar. Akad. Wiss.*, vol. 53, p. 475, 1935.

[3] Fakhriyani, Widodo, and Adhi, B. P. "Perbandingan Algoritma Naive Bayes Dan Support Vector Machine Dalam Seleksi Kelulusan Pemberkasan Beasiswa BPPPPA Fakultas Teknik Universitas Negeri Jakarta", Jurnal PINTER, 108-115. 2018.

[4] Han, J., Kamber, M., and Pei, J. "Data mining: concepts and techniques". 3rd Edition. San Francisco: Morgan Kaufmann publications, 2012.

[5] P.A. Octaviani, Y. Wilandari, and D. Ispriyanti, "PENERAPAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE (SVM) PADA DATA AKREDITASI SEKOLAH DASAR (SD) DI KABUPATEN MAGELANG, *Jurnal Gaussian*, Volume 3, Nomor 4, pp. 811 – 820, 2014.

[6] Perdana, N. G., and Widodo, T. "Sistem Pendukung Keputusan Pemberian Beasiswa Kepada Peserta Didik Baru Menggunakan Metode TOPSIS". *in SEMINAR NASIONAL TEKNOLOGI INFORMASI & KOMUNIKASI TERAPAN 2013 (SEMANTIK 2013)*, Semarang, 2013. 265-272.

[7] Prasetyo, E. "DATA MINING- Mengolah data menjadi Informasi Menggunakan Matlab". Yogyakarta : CV. ANDI OFFSET (Penerbit ANDI), 2014.

[8] Rachman, F., and Purnami, S. W. "Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM)". JURNAL SAINS DAN SENI ITS Vol. 1, No. 1, (Sept. 2012) ISSN: 2301-928X, D-130 - D-135. 2012.

[9] Somvanshi, M., Chavan, P., et al. "A review of machine learning techniques using decision tree and support vector machine. International Conference on Computing Communication Control and Automation (ICCUBEA)" (pp. 1-7). 2014. IEEE Conference Publications.

[10] Sudarajat, A., and Budi, I. "Analisis Kinerja Algoritma Support Vector Machine (Svm) Pada Data Seleksi Penerima Beasiswa Menggunakan Particle Swarm Optimization (Pso) (Studi Kasus: Politeknik Tedc Bandung)". TEDC Vol. 13 No. 1, Jan. 2019.