Classification Study Period Department of Information Systems at STMIK Bandung Bali using Support Vector Machine (SVM) Method

Zulfachmi[1*], Aggry Saputra [2], and I Gusti Ngurah Janardana[3]
[1,2]Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University
[3]Department of Electrical and Computer Engineering, Udayana University
*Email: fahmi.alfahim.93@gmail.com

***Abstract*— Graduation is the final stage of learning process activities in universities. The undergraduate study period under STMIK Bandung Bali regulation is scheduled in 8 semesters (4 years) or less and a maximum of 14 semesters (7 years). Information Systems is one of the majors in STMIK Bandung Bali. The study period of this department can be influenced by many factors. These factors are the Kumulatid Credit Index (IPK), gender, scholarship, part-time work, Student Activity Unit (UKM). Purpose of this study was to determine the classification of accuracy factors. In this study using SVM (Support Vector Machine) method with accuracy of 99.07%.**

***Keyword*— Graduation, STMIK Bandung Bali, Support Vector Machines (SVM).**

I. INTRODUCTION

A college education is an education after secondary education includes educational programs diploma, bachelor's, master's, specialist and doctoral degrees. STMIK Bandung Bali is one of the universities located in Bali which has 2 courses namely Informatics Engineering and Information System with Bachelor degree program. Graduation is the end result of the process of teaching and learning activities during college lectures. Each college would want to create the best graduates in terms of quality and timeliness in completing the study.[1]

The duration study of Undergraduate Program according to Academic Regulations of STMIK Bandung Bali is scheduled in 8 semesters (4 years) or can be taken less than 8 semesters (3.5 years) and no later than 14 semesters (7 years). With the irrelevant comparison between students who enroll in college and students who graduated make database STMIK Bandung Bali become irregular. There are several factors that influence study period of a student being lectured. Factors that are expected to affect timely graduation include Grade Point Average (IPK), gender, employment, scholarship, Student Activity Unit (UKM).

Based on these problems, the researcher wanted to know the factors that influence length of study program of the students Information System by classifying the students' graduation into two categories: graduating on time less than 4 years (8 semesters) and graduation not on time for student education for more than 4 years (8 semesters) so that with data can help the campus in making a policy that improves students graduation STMIK Bandung Bali.

As for previous research by Siti Nur Asiyah and Kartika Fithriasari on classification study period of study program statistic with Support Vector Machine (SVM) and Iterative Dichotomiser3 (ID3) method by yielding accuracy of 90%[2]. In addition, research conducted by Pusphita Anna Octaviani about Classification of Support Vector Machine (SVM) On Primary School Accreditation Data (SD) in Magelang District get the accuracy level of 93.90% using kernel function Gaussian Radial Basic Function (RBF)[3]. Then the research Traffic Congestion On Twitter using Support Vector Machine (SVM) method conducted by Elly Susilowati produces an average accuracy of 90%[4]. While in the research Application of Support Vector Machine Method On Diagnosis of Hepatitis conducted by Raudlatul Munawarah yield truth percentage reach 70-96%[5]. In the research classification of final task theme using Support Vector Machine method to produce accuracy of 86.21%[6]. And then research on the classification of final project theme using Support Vector Machine method resulted in an accuracy of 86.21%[7]. Based on the level of accuracy generated by the method of Support Vector Machine above 85% then in this study using Support Vector Machine (SVM) method for classification of study period STMIK Bandung Bali students.

II. LITERATURE REVIEW

A. Data Mining

Data Mining is a process of searching patterns and hidden relationships in large amounts of data with the aim of classifying[8], prediction[9], clustering[10], estimation[11].

B. Support Vector Machine

Support Vector Machine (SVM) is a learning system that uses a hypothetical linear function in a high-dimensional space and is trained with algorithms based on optimization theory by applying learning bias derived from statistical theory. The main purpose of this method is to build OSH (Optimal Separating Hyperplane), which makes optimum separation function that can be used for classification.

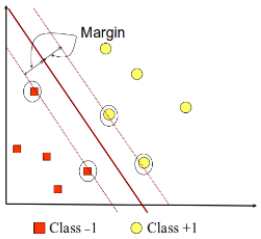

Fig 1. Hyperplane Concept on SVM

Data residing in boundary field is called the support vector. In Figure 1, two classes can be separated by a pair of parallel bounding plane. The first delimiter field limits first class while the second delimiter field limits second class, so that it is obtained:

$X_iW + b \geq +1, y_i = +1$ …….. (1)
$X_iW + b \leq -1, y_i = -1$

w is the normal field and b is the position of the alternate field to the coordinate center. The margin (distance) value between the bounding plane (based on the formula of spacing to the center) is $\frac{1-b-(-1-b)}{||\boldsymbol{w}||} = \frac{2}{||\boldsymbol{w}||}$ the value of this margin is maximized by still satisfying in equation (1). By multiplying b and w by a constant, a margin value multiplied by the same constellation will be generated. Therefore, the constraint in equation (2) is a scaling constraint that can be satisfied by rescaling b and w. Also because it maximizes 1 / || w || is equal to minimize || w || 2 and if the two boundary plots in equation (1) are represented in inequality (2).

$y_i (\mathbf{x_i w} + b) - 1 \geq 0$ ……..(2)

in the formula 3 explains the search for the best dividing field with the largest margin value being the problem of constraint optimization.:

$\min \frac{1}{2} ||w||^2$ …….. (3)
with $y_i (x_i w + b) - 1 \geq 0$

To classify data that can not be separated linearly SVM formulas must be modified as no solutions will be found. Therefore, the two boundary fields 1 must be changed so that they are more flexible with the addition of the variable εi ($\varepsilon i \geqslant 0$, $\forall i$: $\varepsilon i = 0$ if xi are correctly classified) to $\mathbf{XiW} + b \geqslant 1 - \varepsilon i$ for class 1 and $\mathbf{XiW} + b \geqslant -1 + \varepsilon i$ for class 2. The best separator segment with the addition of variable ξ i is often called soft margin hyperplane. Thus the best dividing field search formula is described in formula 4:

$\min \frac{1}{2} ||w||^2 + C(\sum_{i=1}^{n} \varepsilon i)$ …….. (4)
with $y_i (x_i w + b) \geq 1 - \varepsilon_i$
$\varepsilon_i \geq 0$

C is parameter that determines size of penalty due to an error in data classification and its value is determined by user. So the role of C is to minimize training errors and reduce model complexity.

Common kernel functions used in the SVM method are Linear Kernel, Polynomial Kernel, Radial Base Function Kernel (RBF), Sigmoid Kernel.

III. Research Methodology

A. Research variables

The research variables consist of dependent and independent variables. Dependent variables in this research is period study of student Information System which consist of 2 categories of study duration, that is categorized on time is student who graduated <= 4 year and not on time is student who pass > 4 years. While the independent variables are gender, IPK, scholarship (never or not receive scholarship), part time (never or never do part time), and Organization (active or not in organization). Research variables show in Table I.

TABLE I. RESEARCH VARIABLES

No.	Variabel	Category
1	Duration of Study (Y)	0 = ≤ 4 1 = > 4
2	Gender (X1)	0 = Woman 1 = Man
3	IPK (X3)	0 = < 3 1 = 3-3,5 2 = > 3,5
4	Work (X5)	0 = No 1 = Yes
5	Scholarship (X6)	0 = No 1 = Yes
6	UKM (X7)	0 = No 1 = Yes

B. Research Steps

In this study, the steps that go through data collection has been collected and then data sharing between sample data and test samples. After data is divided then next step is classification using the SVM method and then found the results of classification. The research steps can be seen in figure 2.

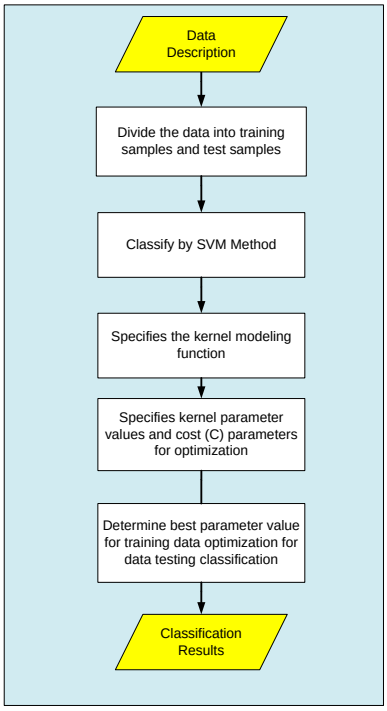

Fig 2. Flowchart Research

IV. Result And Discussion

A. Descriptive Analysis

Descriptive analysis is used to obtain a general description of data in this study. Based on 5 years of force that has completed study period, the data obtained by students who have passed as many as 107 data. From several periods it was found that 41% of percentage students were long study with less than 48 months (on time), while remaining 59% were students whose study period was more than 48 months (not on time). By sex, the number of students graduating on time with female sex is 45% and men 55%. As for timely sex of women is 22% and men 78%. Percentage based on IPK with length of study on time with IPK of less than three is 11%, IPK of three to three point five is 48%, and IPK of more than three point five is 41%. Meanwhile, students who graduate are not on time with a IPK of less than three is 51%, IPK of three to three point five is 44%, and IPK of more than three point five is 5%.

Based on the work variables, the data taken for students with a study duration of less than equal to 4 years (on time) and never doing part-time job is 27%, while those who never do part-time job is 73%. The number of students who graduated with a study duration of more than 4 years (not on time) and never do part time work is 65%, while those who never do part-time job is 35%. Based on the scholarship variables, the timely students who get scholarship amounted to 39% and the students on time did not get the scholarship of 61%, while the students who did not get the scholarship by 29% and the non-timely students did not get a scholarship of 71%. Based on the participation of students following UKM for a timely study period of 64% and students who do not follow UKM with 36% on-time study, while students who follow the UKM with a period of study is not on time 51% and 49% of students who do not timely follow UKM.

B. SVM Classification Using Linear Kernel

In SVM with linear kernel function, the value of C used is 0.1; 1; 10 and 100. These C values are then applied to training data to obtain eror values in each classification model. Here is the error value with training and testing data. From the data taken as much 107 divided for training data as much as 82 and testing 25. In Table II shows that there are many values of 0.00 on the value of classification error, so assumed from the 25 data testing tested no errors. From the table we use the C = 1 parameter value to be used in the SVM separator function by using linear kernel function because C = 1 is stable when tested with some other test data.

TABLE II. ERROR CLASSIFICATION VALUE BY USING LINEAR KERNEL FUNCTION

Parameter cost (C)	Error Classification
0,1	0,00
1	**0,00**
10	0,04
100	0,00

By using parameter C = 1, parameter value applied to data testing classification which evaluated by calculating accuracy of classification accuracy. Here is a confusion matrix using data testing as much as 25 data presented in table III.

TABLE III. CONFUSION MATRIX BY USING LINEAR KERNEL FUNCTION

Class	Prediction	
	On Time	Not On Time
On Time	7	0
Not On Time	0	18

$$\text{Accuracy} = \frac{7+18}{7+0+0+18} = 1\text{x}100\% = 100\%$$

C. SVM Classification Using Polynomial Kernel

SVM with polynomial kernel function, there are parameter d (degree) and C (cost). Determination of parameter d (degree) for separator function with polynomial kernel function is tested some parameter value with range 2 up to 5 and value of C used is 0,1; 1; 10 and 100. The values C and d are then applied to training data to obtain eror values in each classification model. The error values with training data and using different d and C parameters are presented in Table IV.

TABEL IV. ERROR CLASSIFICATION VALUE BY USING POLYNOMIAL KERNEL FUNCTION

Parameter C (cost)	Parameter d (degree)	Error Classification
0,1	**2**	**0,00**
	3	0,00
	4	0,04

	5	0,08
1	2	0,00
	3	0,08
	4	0,08
	5	0,08
10	2	0,00
	3	0,04
	4	0,08
	5	0,08
100	2	0,00
	3	0,04
	4	0,08
	5	0,08

From Table IV, the best parameters C (cost) and parameter d (degree) to be used in the SVM separator function model using the polynomial kernel function are C = 0.1 and d = 2. Using C = 1 and d = 2 parameters, parameter values are then applied to classification of training and testing data which is then evaluated by calculating accuracy of classification accuracy. Configuration matrix using data testing of 25 data can be seen in Table V.

TABLE V. CONFUSION MATRIX BY USING POLYNOMIAL KERNEL FUNCTION

Class	Prediction	
	On Time	Not On Time
On Time	7	0
Not On Time	0	18

$$\text{Accuracy} = \frac{7+18}{7+0+0+18} = 1\text{x}100\% = 100\%$$

D. SVM Classification Using the RBF Kernel

SVM with Radial Basis Function (RBF) kernel function, there are parameters (gamma) and C (cost). Determination of parameter (gamma) for separator function with RBF kernel function is tested some parameter value that is 0,003; 0.007; 0.015 and 0.031 and the value of C used is 0.1; 1; 10 and 100. The values of C and are then applied to training data to obtain eror values in each classification model. Table classification error value can be seen in Table VI. From table, the best parameters of C (cost) and parameter γ (gamma) to be used in SVM separator function model by using the RBF kernel function are C = 100 and γ = 0.003. Using parameters C = 100 and γ = 0.003, the parameter values are then applied to classification of data testing which is then evaluated by calculating the accuracy of the classification accuracy.

TABLE VI. ERROR CLASSIFICATION VALUE BY USING RBF KERNEL FUNCTION

Parameter C (cost)	**Parameter γ (gamma)**	**Error Classification**
0,1	0,003	0,28
	0,007	0,28
	0,015	0,28
	0,031	0,28
1	0,003	0,28
	0,007	0,28
	0,015	0,00
	0,031	0,00
10	0,003	0,00
	0,007	0,04
	0,015	0,00
	0,031	0,00
100	**0,003**	**0,00**
	0,007	0,00
	0,015	0,00
	0,031	0,04

TABLE VII. CONFUSION MATRIX BY USING RBF KERNEL FUNCTION

Class	Prediciton	
	On Time	Not On Time
On Time	7	0
Not On Time	0	18

$$\text{Accuracy} = \frac{7+18}{7+0+0+18} = 1\text{x}100\% = 100\%$$

TABLE VIII. RESULTS CLASSIFICATION BY USING SVM

Data	Fungsi Kernel		
	Linear	Polynomial	RBF
Training	99,07	99.07	99.07
Testing	100	100	100

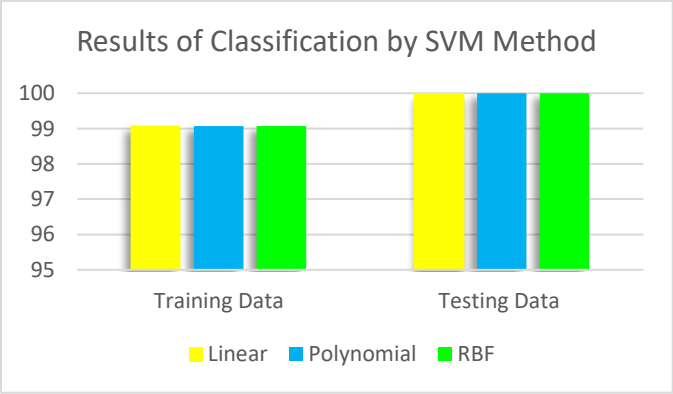

Fig 3. Results of SVM Classification

V. Conclusion

The results of descriptive analysis concluded that more students of male sex are passing on time than women. And for IPK variables, 89% of timely graduates are IPKs above 3. One

of the factors that caused the students not to finish the lecture time is because they work while studying.

The result classification student study period of STMIK Bandung Bali by using SVM method for data testing is 100%. And the result of classification in training data is 99.07%.

REFERENCE

[1] I. Bagus, A. Peling, I. N. Arnawan, I. P. A. Arthawan, and I. G. N. Janardana, "Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm," *Int. J. Eng. Emerg. Technol.*, vol. 2, no. 1, pp. 53–57, 2017.

[2] D. Ispriyanti and A. Hoyyi, "Analisis klasifikasi masa studi mahasiswa prodi statistika undip dengan metode support vector machine (svm) dan id3 (iterative dichotomiser 3) 1,2," vol. 9, no. 1, pp. 15–29, 2016.

[3] P. A. Octaviani, Yuciana Wilandari, and D. Ispriyanti, "Penerapan Metode Klasifikasi Support Vector Machine (SVM) pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang," *J. Gaussian*, vol. 3, no. 8, pp. 811–820, 2014.

[4] E. Susilowati, M. K. Sabariah, A. A. Gozali, J. T. Informatika, U. Telkom, and S. V. Machine, "Implementasi Metode Support Vector Machine Untuk Melakukan Klasifikasi Kemacetan Lalu Lintas Pada Twitter Implementation Support Vector Machine Method For Traffic Jam Classification On Twitter," vol. 2, no. 1, pp. 1478–1484, 2015.

[5] R. Faisal *et al.*, "Penerapan Metode Support Vector Machine Pada Diagnosa Hepatitis Penerapan Metode Support Vector Machine," no. February, 2016.

[6] O. Somantri and S. Wiyono, "Metode K-Means untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM)," vol. 3, no. 1, pp. 34–45, 2016.

[7] P. Agung, A. Wijaya, K. Budiarta, and M. Sudarma, "Bussines Intelligent in Telemarketing Using SVM," vol. 2, no. 1, pp. 62–66, 2017.

[8] S. Nur and K. Fithriasari, "Klasifikasi Berita Online Menggunakan Metode Support Vector Machine dan K- Nearest," vol. 5, no. 2, 2016.

[9] L. M. C. W, A. Daru, and D. Andrian, "Aplikasi Data Mining dengan Metode Support Vector Machine (SVM) untuk Prediksi Financial Distress pada Industri Jasa Go Public yang Terdaftar di Bursa Efek Indonesia," pp. 81–86, 2016.

[10] I. C. Dewi, B. Y. Gautama, and P. A. Mertasana, "Analysis of Clustering for Grouping of Productive Industry by K-Medoid Method," vol. 2, no. 1, pp. 26–31, 2017.

[11] L. Assaffat, "Analisis Akurasi Support Vector Machine Dengan Fungsi Kernel Gaussian Rbf Untuk Prakiraan Beban Listrik Harian Sektor Industri," *Momentum*, vol. 11, no. 2, pp. 64–68, 2014.