Vector machine to predict student retention: A computerized approach

N W S Saraswati¹, K K Widiartha¹ and L P A Prapitasari²

¹ Dept of Informatics, STMIK STIKOM Indonesia, Denpasar, Indonesia
² Dept of Information System, STMIK STIKOM Bali, Denpasar, Indonesia

*sumartini.saraswati@stiki-indonesia.ac.id

Abstract. Student retention can be used as a measure of the university performance. Poor student retention would affect the university because it depicts a poor image of a university management. Student retention can be interpreted as the percentage of students who remain active until they are graduated and then we call these kind of students as active students. In this study, we conducted a research on students database to predict whether they are active or inactive students by using the Support Vector Machine and then at last, we as well measured the quality of the prediction model. We use first and second semester GPA, financial status, work status and student’s origin as predictor attribute. Prediction model that is developed from the student database of STMIK STIKOM Indonesia show a good result with an accuracy level of 97.46%. By excluding GPA from SVMs model, the performance of the model decrease. A model with all the attributes used in the classification provides a balance of positive and negative tuples of recognition compared to when one of attributes is omitted from the model.

1. Introduction

Student retention can be used as a measure of the university performance. Poor student retention would affect the university in general because it depicts a poor image of a university management [1]. It is applied vice versa, means that if the student retention is high, it will increase people’s trust about the quality of the university so that in the end it will bring benefits for the university.

Student retention can be interpreted as the percentage of students who remain active until they are graduated and then we call this kind of students as active students. Inactive students are those who stop attending lectures and do not register back in the next semesters. The inactive students tend to be dropped out of the university [2]. In Indonesia, the dropout rate is one of the indicators of the university quality assessment, in which it is used when determining the value of university accreditation by the government.

In this study, we conduct a prediction on the students whether they are active or inactive students, by using the Support Vector Machine and we also measure the quality of the prediction model. We have also experimented to find out what attributes are influencing the predictions. The inactive students referred to students who do not conduct academic registration for more than 2 semesters in a row. The data is taken from the academic database STMIK STIKOM Indonesia for the student year of 2014 until the year 2016. The prediction model is built by using the data mining method.

Data mining conducted in this case is to classify the students into two groups, either they belong to the active student's group or inactive students group. The prediction function in data mining is to assist...
the university in taking precautions before the student has to be dropped out. Possible actions for predicted inactive students, as such planning intensive lectures or suggestions for the students to transfer to a more controlled department. This precaution is expected to minimize dropout rates from the universities.

In Indonesia, there is an inter-regional education disparity. The differences occur in the quality of education of the areas that are closer to and areas that are far from the capital city. For example, the quality of education in the island of Java is different from the quality of education in the remote areas of Papua. This inter-regional educational gap has resulted in the possibility of adding one more retention factors, for an example, the student’s origin which tells where the student had previously studied. In order to prove this hypothesis, in this study, the origin of each student is added. It will show the influence of the student origin in determining the quality of the prediction model using the Support Vector Machine.

Support Vector Machine is a supervised machine learning method which analyzes data and recognizes patterns, used for classification and regression analysis. Support Vector Machine (SVM) was introduced by Vapnik in 1992 as an efficient classification technique for non-linear problems. SVM differs from classification techniques in the 1980s, such as the decision tree and the Artificial Neural Network, which in concept are not very clear and often trapped in local optimum solutions.

SVM seeks to find the hyperplane by maximizing the distance between classes (margins). In this way, SVM guarantees a high generalization capability for future data. In SVM, the outermost data objects closest to the hyperplane are called support vectors. Data objects, called support vectors, are the most difficult to classify because their positions are almost overlapping with other classes. Given its critical nature, only this support vector is taken by SVM to find the most optimal hyperplane. In this way, SVM can work more efficiently.

Ji Wu Jia’s research stated that SVM as the most popular approach method for solving problems of the student retention prediction model. This study uses modifications to the kernel function so that the prediction model accuracy increases to 94.29% using two significant attributes, the Total Credit Hours and GPA [4].

Anbuselvan et al. in his research found that SVM works well on data with a large number of dimensions and gives better results than Decision Tree and Rule Induction methods. SVM gives 89.84% accuracy compared to 86.32% of the decision tree and rule induction with 81.98% accuracy [5]. Support Vector Machine gives an accuracy of 98.06% on dropout prediction done by Siti Nurhayati in 2015 [6]. The paper of Gopalakrishnan tells that Naïve Bayesian is more suitable for predicting graduation, whereas AdaBoost and SVM are better at predicting persistence [7].

Support vector machine is a machine learning method that can complete supervised classification, such as in the field of text categorization, tone recognition, image classification, micro-gene array expression, protein structure predictions, other data classification related problems. Most supervised classification methods based on traditional statistical methods will provide improved performance with the number of samples approaching an infinite number. But in reality, the number of samples is practically limited. SVM is able to overcome the number of data dimensions and the limited number of samples. In the research by Srivasta and Bhambhu carried out a classification for (Diabetes data, Heart Data, Satellite Data and Shuttle data) which have two or multi-classes. His research’s result was kernel functions affecting the classification results [8]. Support vector machines can also be used to detect
fraudulent credit card transactions, character recognition in automated zip code reading, and predicting compound activity in drug discovery [9].

Support vector machines are a specific type of machine learning algorithm that is among the most widely used for many statistical learning problems, such as spam filtering, text classification, handwriting analysis, face and objects recognition, and countless others. Support vector machines have also come into widespread use in practically every area of bioinformatics within the last ten years, and their area of influence continues to expand today. The support vector machine has been developed as a robust tool for classification and regression in noisy, complex domains [10].

For classifying the proteins support vector machine (SVM) has been developed to extract a feature from the proteins sequences. Al-maghasbeh’s research describes a method for predicting and classifying secondary structure of proteins. Support vector machine (SVM) modules were developed using multi-agent system principle for predicting the proteins and its function, and achieved maximum accuracy, specificity, sensitivity, of 92%, 94.09%, and 91.59% respectively [11].

SVM in the research of Lavanya and Divya is used to predict aircraft accidents based on a collection of historical records. Prediction of an aircraft accident will save life and cost. The method used is proven to handle noisy, unrelated and lost data. Predictive results are tabulated and range from 85% to 90% [12].

SVM in text processing has also been used to do web mining. In the research of Wulandini and Nugroho conducted a classification for text data on Tropical diseases such as Dengue Fever, Malaria, and Bird Flu [13]. SVM also used by Baravati to filter spam emails [14].

Based on those researches, thus this research chose the method of Support Vector Machine to make the predictive model of undergraduate student retention.

SVM was chosen as the method of classification in this study because SVM has three advantages compared to other classification methods, such as:

- it has a high data generalization ability,
- able to produce a good prediction model despite being trained by relatively smaller data set,
- relatively easy to implement because the determination of support vector can be formulated in the problem of QP (Quadratic Programming) [3].

2.2. Data set
Several reasons that caused students not to continue their studies are as follows [5]:

- Personal difficulties reason which can be health issues, financial and socialization issues.
- Academic difficulties reason.
- Full-time students vs part-time students, part-time students are fewer in continuing their studies into the second year of their studies compared to the full-time students.
- Choosing the wrong major or college and university.
- Studying in the alternative choice universities, students tend to choose their main pick university if they have a choice.
- Lose interest in their subjects.
- Failing in managing time.
- Dissatisfaction on the university management.

From data owned by STMIK STIKOM Indonesia, in this research, we use the first three reason on the above list. Those factors described as finance factor, academic factor and work status of student factor. The origin of the student is an additional factor to be studied in this paper. The dimensions or attributes used in data mining are:
• The first and second semester GPA. This dimension represents the academic ability factor of the students. The type of data of this factor is numeric.
• The status that states the students have paid their study fees with installment payments method or not, which is stated in a binary type of data. This status represents the financial factor of the student.
• The status of whether the student is working or not. This status also states full time vs. part-time work of the student. The type of data of this dimension is binary.

The student’s origin is stated in nine-digit of binary series. In the pre-processing and transforming stages, there are inherent challenges as the support vector model only supports input or data in the form of numerical and binary and the region of student’s attribute is categorical based data. This attribute whose value are category based must be converted into numeric or binary forms before they can be analysed using support vector machine [5].

2.3. Methodology for model prediction
In this study, 2488 record data STMIK STIKOM Indonesia students are used and divided into the same amount of data between students who are active and inactive. Active students are students who have registered them self each semester or students who took a maximum of two semesters left. Meanwhile the inactive students are students who stop attending lectures and do not perform administrative registration more than 2 semesters in a row.

The testing technique used is k-Fold Cross-Validation with k = 10. So from 1244 data, 124 records are taken as test data and the rest is treated as training data. Each experiment was conducted 10 times where the test data and the train data were exchanged from the overall data contained in the dataset. SVM used in this research is SVM Light. SVM Light is a support vector machine program in C language written by Thorsten Joachims.

A method that can be used to do classification using SVM is by first transforming the data source into a vector. Based on the data format in SVM Light, the data vector is represented as the following:

• data of active student is represented with polarity value of 1,
• data of inactive student is represented with polarity value of -1.

An example of the vector representation of the data that is used in this research is as shown in Figure 1.

![Figure 1. Vector data.](image)

Polarity is the first number on the vector data shown by figure 1. that deciding whether the record is a positive tuple or negative tuple. The next number is the dimension of the vector that is representing the value of a classification attribute. The total number of the vector data dimension in this research is 13, which are listed as the following.

• Dimension no.1: GPA of 1st semester
• Dimension no.2: GPA of 2nd semester
• Dimension no.3: Financial status
- Dimension no. 4: Working status
- Dimension no. 5-13: Student’s origin

The classification process using SVM in this study is divided into several stages including data pre-processing, training and data testing. The complete classification process using SVM Light can be seen in figure 2. Data pre-processing is intended for changing the record in the database into vector data that will be the input for SVM. The training process aims to build model that have defined support vector and hyperplane. And in the end the testing process will determine the quality of the model that has been obtained through the training process.

![SVM classification process diagram]

**Figure 2.** SVM classification process.

### 3. Result
In this research, there are five kinds of tests that are conducted, they are:

- testing with 10 Fold Cross-Validation for all attributes,
- testing with 10 Fold Cross-Validation in which the GPA attribute is omitted,
- testing with 10 Fold Cross-Validation which eliminates the attribute of arrears status of the student payment,
- testing with 10 Fold Cross-Validation which eliminates the status attribute of student work,
- testing with 10 Fold Cross-Validation which eliminates attributes of student origin.
The result of each experiment is stated in Table 1 through Table 10. The correct of active student field refers to the number of the correct classification records of the active student data. As well as Incorrect of active Student field refers to the number of the incorrect classification records of the active student data.

**Table 1.** Experiment result when GPA was omitted.

| Experiment | Correct of Active Student | Incorrect of Active Student | Correct of Non-Active Student | Incorrect of Non-Active Student |
|------------|---------------------------|-----------------------------|-------------------------------|--------------------------------|
| folding 1  | 123                       | 1                           | 82                            | 42                             |
| folding 2  | 124                       | 0                           | 102                           | 22                             |
| folding 3  | 124                       | 0                           | 89                            | 35                             |
| folding 4  | 124                       | 0                           | 82                            | 42                             |
| folding 5  | 124                       | 0                           | 96                            | 28                             |
| folding 6  | 124                       | 0                           | 87                            | 37                             |
| folding 7  | 124                       | 0                           | 70                            | 54                             |
| folding 8  | 124                       | 0                           | 49                            | 75                             |
| folding 9  | 124                       | 0                           | 56                            | 68                             |
| folding 10 | 124                       | 0                           | 44                            | 80                             |
| **AVERAGE**| **123.9**                 | **0.1**                     | **75.7**                      | **48.3**                       |

**Table 2.** Confusion Matrix of Experiment when GPA was omitted.

|          | POSITIVE | NEGATIVE |
|----------|----------|----------|
| POSITIVE | 123.9    | 0.1      |
| NEGATIVE | 48.3     | 75.7     |

**Table 3.** Experiment result when worked attribute was omitted.

| Experiment | Correct of Active Student | Incorrect of Active Student | Correct of Non-Active Student | Incorrect of Non-Active Student |
|------------|---------------------------|-----------------------------|-------------------------------|--------------------------------|
| folding 1  | 119                       | 5                           | 119                           | 5                              |
| folding 2  | 122                       | 2                           | 123                           | 1                              |
| folding 3  | 121                       | 3                           | 122                           | 2                              |
| folding 4  | 117                       | 7                           | 119                           | 5                              |
| folding 5  | 122                       | 2                           | 120                           | 4                              |
| folding 6  | 123                       | 1                           | 121                           | 3                              |
| folding 7  | 121                       | 3                           | 121                           | 3                              |
| folding 8  | 122                       | 2                           | 118                           | 6                              |
| folding 9  | 120                       | 4                           | 119                           | 5                              |
| folding 10 | 119                       | 5                           | 122                           | 2                              |
| **AVERAGE**| **120.6**                | **3.4**                     | **120.4**                     | **3.6**                        |

**Table 4.** Confusion matrix of experiment when worked attribute was omitted.

|          | POSITIVE | NEGATIVE |
|----------|----------|----------|
| POSITIVE | 120.6    | 3.4      |
| NEGATIVE | 3.6      | 120.4    |
Table 5. Experiment result when financial attribute was omitted.

| Experiment | Correct of Active Student | Incorrect of Active Student | Correct of Non-Active Student | Incorrect of Non-Active Student |
|------------|---------------------------|----------------------------|-------------------------------|---------------------------------|
| folding 1  | 121                       | 3                          | 120                           | 4                               |
| folding 2  | 122                       | 2                          | 124                           | 0                               |
| folding 3  | 122                       | 2                          | 121                           | 3                               |
| folding 4  | 117                       | 7                          | 115                           | 9                               |
| folding 5  | 122                       | 2                          | 122                           | 2                               |
| folding 6  | 123                       | 1                          | 121                           | 3                               |
| folding 7  | 121                       | 3                          | 120                           | 4                               |
| folding 8  | 122                       | 2                          | 119                           | 5                               |
| folding 9  | 120                       | 4                          | 121                           | 3                               |
| folding 10 | 119                       | 5                          | 122                           | 2                               |
| AVERAGE    | 120.9                     | 3.1                        | 120.5                         | 3.5                             |

Table 6. Confusion Matrix of Experiment when financial attribute was omitted.

|       | POSITIVE | NEGATIVE |
|-------|----------|----------|
| POSITIVE | 120.9   | 3.1      |
| NEGATIVE | 3.5     | 120.5    |

Table 7. Experiment result when region attribute was omitted.

| Experiment | Correct of Active Student | Incorrect of Active Student | Correct of Non-Active Student | Incorrect of Non-Active Student |
|------------|---------------------------|----------------------------|-------------------------------|---------------------------------|
| folding 1  | 120                       | 4                          | 121                           | 3                               |
| folding 2  | 122                       | 2                          | 123                           | 1                               |
| folding 3  | 121                       | 3                          | 123                           | 1                               |
| folding 4  | 117                       | 7                          | 119                           | 5                               |
| folding 5  | 122                       | 2                          | 122                           | 2                               |
| folding 6  | 123                       | 1                          | 121                           | 3                               |
| folding 7  | 121                       | 3                          | 120                           | 4                               |
| folding 8  | 122                       | 2                          | 120                           | 4                               |
| folding 9  | 120                       | 4                          | 121                           | 3                               |
| folding 10 | 119                       | 5                          | 123                           | 1                               |
| AVERAGE    | 120.7                     | 3.3                        | 121.3                         | 2.7                             |

Table 8. Confusion Matrix of Experiment when region attribute was omitted.

|       | POSITIVE | NEGATIVE |
|-------|----------|----------|
| POSITIVE | 120.7   | 3.3      |
| NEGATIVE | 2.7     | 121.3    |
Table 9. Experiment result with complete attribute.

| Experiment | Correct of Active Student | Incorrect of Active Student | Correct of Non-Active Student | Incorrect of Non-Active Student |
|------------|---------------------------|-----------------------------|-------------------------------|---------------------------------|
| folding 1  | 119                       | 5                           | 121                           | 3                               |
| folding 2  | 124                       | 0                           | 122                           | 2                               |
| folding 3  | 121                       | 3                           | 120                           | 4                               |
| folding 4  | 117                       | 7                           | 119                           | 5                               |
| folding 5  | 122                       | 2                           | 122                           | 2                               |
| folding 6  | 123                       | 1                           | 121                           | 3                               |
| folding 7  | 121                       | 3                           | 120                           | 4                               |
| folding 8  | 122                       | 2                           | 120                           | 4                               |
| folding 9  | 120                       | 4                           | 121                           | 3                               |
| folding 10 | 119                       | 5                           | 123                           | 1                               |
| AVERAGE    | 120,8                     | 3,2                         | 120,9                         | 3,1                             |

Table 10. Confusion Matrix of Experiment with complete attribute.

|          | POSITIVE | NEGATIVE |
|----------|----------|----------|
| POSITIVE | 120,8    | 3,2      |
| NEGATIVE | 3,1      | 120,9    |

Measurements used to evaluate the prediction model in this research are accuracy, recall, precision, and specificity. Accuracy states the level of recognition of the prediction model where the value is a percentage of the number of tuples in the test data that is correctly classified. The recall or measure of completeness is known as a true positive rate, whereas the specificity known as the true negative rate is the portion of the correctly classified tuples. The last measure of precision is a measure of certainty in which what percentage of positive tuples labeled positive is in fact true.

Table 11. Prediction model quality.

|                           | Without GPA | Without Working Attribute | Without Financial Attribute | Without Student’s origin | Complete Attribute |
|---------------------------|-------------|---------------------------|----------------------------|--------------------------|-------------------|
| Accuracy                  | 80,48%      | 97,18%                    | 97,34%                     | 97,58%                   | 97,46%            |
| Recall/Sensitivity        | 99,92%      | 97,26%                    | 97,50%                     | 97,34%                   | 97,42%            |
| Precision                 | 71,95%      | 97,10%                    | 97,19%                     | 97,81%                   | 97,50%            |
| Specify                   | 61,05%      | 97,10%                    | 97,18%                     | 97,82%                   | 97,50%            |

Testing results show that the prediction model gives a good result (decent accuracy) with the percentage accuracy of 97.46%. The next experiment was done by eliminating the GPA attribute of the prediction model in which the accuracy decreased to 80.48%. This decrement also caused the model to be much declining in quality while recognizing the inactive students (negative tuple), which in this case is represented by the specificity measure where previously 97.50% then decreased to 61.05%. The quality of the model also decreases if the financial and working status attributes are omitted from the prediction model whose value is shown in table 11.

Surprisingly, the prediction model shows an increase in accuracy when the Student’s origin factor or attribute is omitted from the model. The initial accuracy is 97.46% then increased to 97.58% or in other words there is an increase of 0.12%. But for the size of recall/sensitivity, the experiment resulted in a decrease of 0.08%, which means the model's ability to recognize the positive tuples (active students) decreases.
4. Conclusion
The conclusion of this research is that the prediction model that is developed from the database of STMIK STIKOM Indonesia’s students shown good quality with recognition ability or accuracy level of 97.46%. The quality of the prediction model will decrease when the GPA attribute is omitted from the model. By excluding GPA from SV'Ms model, the performance decrease by 17%-46%. We can conclude that GPA is an important factor in determining the quality of drop out student prediction. It means that this predictor has to be considered as the most relevant input among others. For experiments with omitting attributes of student’s origin, although it shows an increase in accuracy value but does not indicate an improvement in model quality as the recall value decreases. Thus it can be said that in the case study using dataset STMIK STIKOM Indonesia students, the origin of the student gives the effect of increasing the value of sensitivity or the recognition of positive tuples (active students) on the prediction model. With the use of all attributes in the classification provides a balance of positive tuple recognition (active students) and negative tuples (non-active students) recognition rather than when one of the attributes is omitted from the model.

References
[1] Zhang Y, Oussena S, Clark T and Hyensook K 2010 Using data mining to improve student retention in Higher Education: A case study Proc. 12th Int. Conf. Enterp. Inf. Syst. 1 190–197, (2010).
[2] Hastuti K 2012 Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif Semin. Nas. Teknol. Inf. Komun. Terap 2(1) 241–249
[3] Suyanto 2017 Data mining Untuk Klasifikasi dan Klasterisasi Data, 1 st ed (Penerbit Informatika)
[4] Jia J, Laeng M and Mareboyana M 2013 Machine Learning Algorithms and Predictive Models for Undergraduate Student Retention Proc. World Congr. Eng. Comput. Sci. 1 23–25
[5] Sangodiah A, Beleya P, Muniandy M, Heng L E and Spr C R 2015 Minimizing Student Attrition in Higher Learning Institutions in Malaysia Using Support Vector Machine J. Theor. Appl. Inf. Technol. 71(3) 377–385.
[6] Nurhayati S, Luthfi E T and Papua U Y 2015 Prediksi Mahasiswa Drop Out Menggunakan Metode Support Vector SISFOTENIKA 3(6) 82–93
[7] Gopalakrishnan A, Kased R, Yang H, Love M B, Graterol C and Shada A 2018 A Multifaceted Data Mining Approach to Understanding What Factors Lead College Students to Persist and Graduate Computing Conference
[8] Srivastava D K and Bhambhu L 2009 Data Classification Using Support Vector Machine JATIT
[9] Burbidge R and Buxton B 2001 An introduction to support vector machines for data mining Keynote Pap. young OR12 2–14
[10] H Bhavsar and M H Panchal 2012 A Review on Support Vector Machine for Data Classification Int. J. Adv. Res. Comput. Eng. Technol. 1(10) 2278–1323
[11] Al-maghasbeh M K A and Al-khraisat H M M 2015 Towards A Multi Agent System Based Data Mining For Proteins Prediction And Classification Int. J. Sci. Technol. Res. 4(12) 157–160
[12] Lavanya B and Divya B 2017 Big Data Analysis Using Svm And K-Nn Data Mining Techniques Int. J. Comput. Sci. Mob. Comput. 6(1) 84–91
[13] Wulandini F and Nugroho A S 2009 Text Classification Using Support Vector Machine for Webmining Based Spatio Temporal Analysis of the Spread of Tropical Diseases in International Conference on Rural Information and Communication Technology 189–192
[14] Baravati H B, Hosseinkhani J, Keikhaee S and Shahsavari A 2014 Structured Stream Data Mining Using SVM Method as Basic Classifier Int. J. Adv. Comput. Sci. Inf. Technol. 3(3) 306–316