Classification of Student Grade Based on Academic Records Using Support Vector Machine

Amalia Dwi Nurfadzilah[1,*], Ayundyah Kesumawati[1]

[1]*Departement of Statistics, Faculty of Mathematics And Natural Sciences, Universitas Islam Indonesia*
[*]*Corresponding author. Email: 15611135@students.uii.ac.id*

ABSTRACT

Academic records have meaning “used” or “impressions” related to science in the form of notes. In this research, we use academic records of Statistics Students of Islamic University of Indonesia years 2015, which is like the percentage of late attendance, types of courses taken, schedule of days and hours of courses, and the number of SKS (Semester Credit System). This academic record is important because it has a pattern that affects the grade of the course. Therefore to find out the pattern of student academic records, it is necessary to classify the grade of the courses based on academic records, the Support Vector Machine (SVM) classification method is used because this method is reliable for classification with high dimensions and multiclass. In the academic record’s data it is known that there is an imbalance of data, so to overcome it, The Synthetic Minority Oversampling Technique (SMOTE) method are use SVM so the performance of classification would be better. We can conclude that by using the SVM and SMOTE method known that the classification has accuracy 58% with the Cost 10 and gamma 100 so that students who go in to "excellent" class are 363, "very good" class are 102, "good" class are 4, "fair" class are 2, "poor" class is only one, and "failed" class are 66.

Keywords: *Classification, Student’s Grade, Support Vector Machine (SVM), Synthetic Minority Oversampling Technique (SMOTE).*

1. INTRODUCTION

This study aims to classify the Statistics student of Islamic University of Indonesia based on academic records in to six class, with the study case the academic records of Statistics Students of Islamic University of Indonesia years 2015, which is like the percentage of late attendance, types of courses taken, schedule of days and hours of courses, and the number of SKS (Semester Credit System). This academic records is important because it has a pattern that affects the grade of the course. Therefore to find out the pattern of student academic records, it is necessary to classify the grade of the courses based on academic records, the Support Vector Machine (SVM) classification method is used because this method is reliable for classification with high dimensions and multiclass. To overcome the imbalance of data, The Synthetic Minority Oversampling Technique or SMOTE are used in SVM so that the result in performance of classification would be better.

The dataset uses six variables with the categorical variable: 1. types of the courses divided into four types, which are university compulsory courses, core courses, free choice subjects and compulsory elective courses; 2. Class taken: first class, second class, third class, and fourth class; 3. Day of the courses taken: monday class, tuesday class, wednesday class, thursday class, friday class, saturday class; 4. SKS (Semester Credit System): one SKS, two SKS, tree SKS; 5. the percentage of late attendances ; and 6. The Grades : Excellent, Very Good, Good, Fair, Poor, and Failed. All the variables will be changed into numerical variables except variable 5 and 6, and we will import it to the R software.

Support Vector Machine (SVM) shows the classification using the dot from hyperplane that would be maximized the margin line between one to another class of classification. The variable class would be defined as a hyperplane called support vectors. There is two kinds of SVM: Linear support vector machines and Nonlinear support vector machines (Hyeran & Seong-Han, 2003). The multiclass variables would be classified using Support Vector Machine and using kernel to implement the data. Kernel has function to optimize the data so the data would behave better accuracy in classification. (Rosalez, Perez et al., 2013; Han and Michelin, 2006; Renukadevi and Thangraj, 2013). Therefore to conduct the kernel and minimize the error on the testing data used cross validation and grid search method (Hsu et al., 2010). The best recommendation kernel is Radial Based Function (RBF) which has same performance with the linear kernel. The Radial Basis Function are used because it has efficient, simple, and easy calculation and best adapt with parameter optimization (Gaspar et al., 2012 and Shamshirband et al., 2016)

The maximum accuracy to the classification using mixed good parameters form C and γ (gamma) value then using Support Vector Machine and Radial Basis Function from Gaussian kernel are used to decrease the error. The gamma parameter are used to define the how the single training data reaches and influence on the testing data, if gamma value is low it means that the dots is far and the otherwise if gamma value is high so the dots is close to another. Because of that so if gamma value is too huge so the support vector are only influence just around the support vector area itself and there is no amount of C which can prevent the overfitting through the classification. Therefore if the gamma value too small then the model of the classification would be very constrained and could not handle the complexity of the data. The area of the influence by gamma that selected so it include the whole of training data also. The algorithm of the gamma is the inverse from standard deviation of radial basis kernel by gaussian function. A small gamma would define a large variance of the gaussian function, so even two dots are far each other can be considered similar. Then a small variance of gaussian function means that it has large gamma value. (Nitesh V, et al., 2002)

In this study we also uses Synthetic Minority Oversampling Technique (SMOTE) to overcome the imbalance data. Therefore the over sampling in the minor class approach by creating the synthetic training data than rather replace the data, this technique approached successfully on recent research from handwritten character recognition (Ha and Bunke, 1997). In this research would be created an extra training data based on the real data. Then the synthetic training data would be generated the testing data in a less application specific manner by operate in the feature space. The minor class are over sampled by taking each class and introducing synthetic testing data in along the line segment joining any or all of the *k* minor class of nearest neighbour by chosen randomly from the real data. (Nitesh V, et al., 2002)

To see the performances of the classification, we are use confusion matrix. The confusion matrix is a matrix that contains summary of the prediction result from classification. In the matrix there are amount of correct and incorrect prediction with the count values and broken values from each of the class. therefore in the matrix there are four measurement of the confusion matrix: accuracy, error, recall, and precision. (Prasetyo, 2012)

2. MATERIALS AND METHODS

In this study before doing the classification SVM, we also examine descriptive data to see whether the data is balanced or imbalance. If the dataset is imbalance, so we need to balancing the dataset to make it have good classification later. Balancing the dataset using SMOTE method by determining the oversampling and under sampling values on the variables that have the highest and smallest frequencies.

The over sampling in the minor class approach by creating the synthetic training data than rather replace the data, this technique approached successfully on recent research from handwritten character recognition (Ha and Bunke, 1997). In this research would be created an extra training data based on the real data. Then the synthetic training data would be generated the testing data in a less application-specific manner by operate in the feature space. The minor class are oversampled by taking each class and introducing synthetic testing data in along the line segment joining any or all of the *k* minor class of nearest neighbour by chosen randomly from the real data. (Nitesh V, et al., 2002)

In this case, the under sampled from major class would be randomly removed until found the best combination of percentage under-sampling and also the minor class. So the training data is forces to learn or train to experience the varying degrees od under sampling and the minor class has larger percentage at under sampling with the higher degree. In the recent experiment from Nitesh journal, in Nitesh journal if the under sampling at 200% then the data modified into twice as many elements from minor class at the major class. therefore it has 50 samples in minor class and had 200 samples in major class and we under sampling the majority at 200%, the major class would has 25 samples. By doing the combination of under and over sampling to the major and minor class the bias of the learn or training data toward the major class is reversed in the favour of the minor class. The Classifiers are learned on the data by SMOTING the minor class and the major class, the illustration is showed in **Figure 1**.

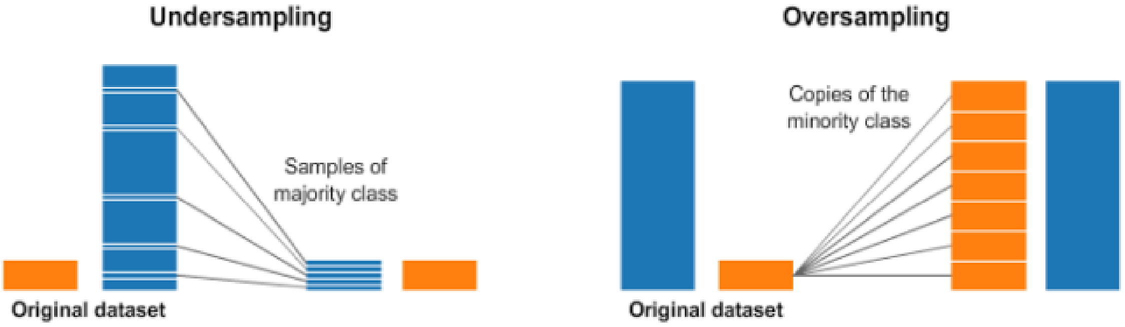

Figure 1. (a) Under Sampling and **(b)** Over Sampling On SMOTE

All of the phase in this journal is showed in **Figure 2** in the first phase (step 1 to 2) the after input the dataset and then determined the training data and the testing data to predict the class. The data in this study were taken from the amount of data set for training data as much as 66.67% of the total data and were taken randomly, therefore training data was taken randomly so a running set of seeds was carried out before being released at the time before random data so that only random done once.

After the training data is determined, the data testing is then determined. In this study, the remainder of the training data which is 33.33% will be used in data testing. Testing data is used to measure the improvement where the classification successfully did the classification correctly. Therefore, the data in the test data should not be in the training set that can be known whether the classifier is "smart" in doing the classification.

Separating data into training and testing so that the model obtained requires good generalization skills in doing data classification. Not a bad classification model can classify data very well in data training, but it is very bad in classifying new and unprecedented data. This is called overfitting. In this study the number of samples used was 1497 and the data samples were taken randomly, this sample will also be used as data testing. While the rest is used as training data as much as 2995 data.

After determined the training data and testing data, classification is carried out. But in this study the data is not imbalance so that the next step is SMOTE the dataset. In the imbalance dataset there are very significant class differences, therefore the data resampling was performed using the SMOTE method. To find the best value of the oversampling and under sampling, in the algorithm over sampling use “Perc Over” and under sampling use “Perc Under”. We try several values of the oversampling and under sampling and see the accuracy by their confusion matrix.

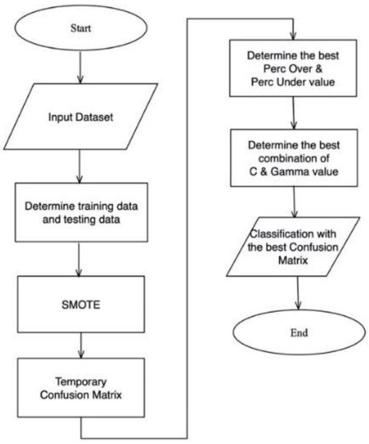
Start
Input Dataset
Determine training data and testing data
SMOTE
Temporary Confusion Matrix
Determine the best Perc Over & Perc Under value
Determine the best combination of C & Gamma value
Classification with the best Confusion Matrix
End

Figure 2. Experiment Phases

The next step is determined the best combination of parameter C and gamma. The good combination of the parameters C and gamma (γ) has maximum accuracy for the classification. In this study we also try several values of C and gamma to find the best accuracy. After find the best accuracy we also calculate the Precision, Recall, and error to show the performance the classification.

Support Vector Machine there are two types: Linear support vector machines and Nonlinear support vector machines. In this study we use the nonlinear support vector machines because the dataset are not linear and there are six class to classify. Therefore the main idea of the SVM is the hyperplane that used to make a decision of classification, which define in the positive and negative class that separates with the largest margin line and minimizing the dimension area from the SVM. The training data will has label $x_i \in \mathrm{R}^d$ and the label $y_i \in \{-1, +1\}$, for all $i = 1,\cdots,l$. Where *I* is the number of data and d is dimension that each class separated linearly in R^d. In this study we wish have the hyperplane which has smallest error among all of the possible hyperplane. Like an optimal hyperplane is the one which has the maximum margin line separating between two classes, and these closest dots are called Support Vector. In the Figure 1(a) showed the solid line represents optimal separating hyperplane and approach completely by d dimensional hyperplane. (Hyeran and Seong Han, 2003)

$$w.x_i + b = 0 \tag{1}$$

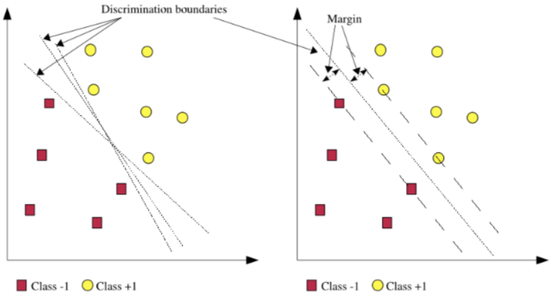

Figure 3. SVM Linear.

The hyperplane defines in $w \cdot x_i + b \geq +1$ defines for the positive dots or positive class and $w \cdot x_i + b \leq -1$ for the negative class. Then in the SVM the largest margin is define to finds the hyperplane that showed in Fig 3, by maximizing using equation 1/||w||. The best separate hyperplane the classes using minimizing equation (2) under the equation (3) method to find the best separate training dataset. (Hyeran and Seong Han, 2003)

$$min_w = \tau(w) = \frac{1}{2}\|w\|^2 \tag{2}$$

$$y_i(w.w_i + b) - 1 \geq 0, \forall_i \tag{3}$$

Optimization can be used with the Lagrange Multiplier, Lagrange Multiplier is a method of finding the minimum or maximum values (Muñoz, et al., 2019).

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i y_i\,(w.w_i + b - 1) \tag{4}$$

In the data technically between the two classes are not perfectly separable, but hyperplane that maximize the margin line while minimizing the proportion of the error misclassification can be determined. In this problem using positive slack variable ξ_i equation (3), and then becomes:

$$y_i(x_i.w \quad + b) \geq 1 - \xi_{i,} \forall_i \tag{5}$$

Optimization the error of the classification the slack variable ξi must run over unity, later $\sum_i \xi i$ is the upper boundaries for the misclassification errors. Afterwards the function τ(.) in equation (2) must be minimized and then becomes equation (6) : (Hyeran and Seong Han, 2003)

$$\frac{1}{2}\|w\|^2 + C \ \sum_{i=1}^{N} \xi_i \tag{6}$$

In the C parameters chosen using control from tradeoff between margin and error $\xi = (\xi_1,\cdots,\xi_l)$. the larger C values means that the classification has higher penalty dots to the error are assigned. By minimizing equation (6) under equation (5) would have generalize separating hyperplane.

In purpose to maximize the pattern of the classification, the non linear SVM took over fitting using soft margin that conduct replacing all of the dots of testing data with non linear kernel function matrix (Boser et al., 1992). The kernel function to defined with focusing in the dot product between two features space of the two classes could be replaced using kernel function then classification result (hyperplane) slightly than not knowing Φ mapping of what is used for each datum. The kernel is used to mapped more the data than the previous training data to the new best training data which has higher dimension without making the new feature space. (Hsu et al., 2010)

In the aim to accomplish the non linear classification decision function, the mapping Φ from the data and transform in to Euclidean space of H which performs as $\Phi : R^n \rightarrow H$, therefore the linear classification is calculated in the new space with the dimension as d. Subsequently the training data algorithm only depends on the dot product in H od the form $\Phi(x_i) \cdot \Phi(x_j)$. Now that calculation of the dot products is prohibitive if the dimension change the training vectors $\Phi(x_i)$ is very large, and since Φ is unknown as prior, the Mercer's algorithm for the positive class define the functions that allows to replace $\Phi(x_i).\Phi(x_j)$ by positive class definite symmetric kernel $k(x_i,x_j)$, that is $k(x_i,x_j) = \Phi(x_i).\Phi(x_j)$. Therefore the training phase of the classification, in this study needs kernel function and Φ is no needs since the implicitly defined by the choice of kernel. Then to the different kernel $k(x_i,x_j)$ these constructs the different hyperplane also in the feature space. These **Table 1** showed the three typical of kernel function. (Hyeran and Seong Han, 2003)

Table 1. Kernel Function

Kernel Function	Inner Product Kernel
Polynomial	$K(x_i, x_j) = (x_i.x_j + 1)^p$
Gaussian	$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$
Sigmoid	$K(x_i, x_j) = \tanh(\alpha\, x_i.x_j + \beta)$

3. RESULTS AND DISCUSSION

The data in this study were taken from the amount of data set for training data as much as 66.67% of the total data and were taken randomly, therefore training data was taken randomly so a running set of seeds was carried out before being released at the time before random data so that only random done once. After the training data is determined, the data testing is then determined. In this study the remainder of the training data which is 33.33% will be used in data testing. Testing data is used to measure the improvement where the classification successfully did the classification correctly. Therefore, the data in the test data should not be in the training set that can be known whether the classifier is "smart" in doing the classification. Separating data into training and testing so that the model obtained requires good generalization skills in doing data classification. Not a bad classification model can classify data very well in data training, but it is very bad in classifying new and unprecedented data. This is called overfitting.

In this study the number of samples used was 1497 and the data samples were taken randomly, this sample will also be used as data testing. While the rest is used as training data as much as 2995 data. Using SMOTE After the distribution of training data and testing data, classification is carried out. In the training data, there are very significant class differences and the data are imbalance see the **Fig 4**, So to overcome to the balance data we are doing resampling data using SMOTE method. The following are the results of the SMOTE method. In the fig 4 we can see that class 1 has the biggest frequency meanwhile the class 6 has very small frequency. So here we will under sampling the class 1 and oversampling the class 6 so all the class would be balance.

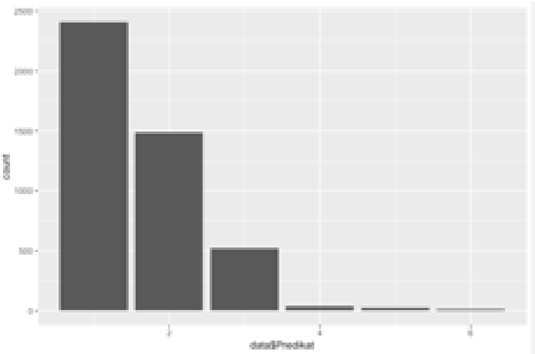

Figure 4. Plot the Dataset

We will try with several values of under sampling and oversampling to find the best balance training dataset. Look at the table 2 below:

Table 2. Experiment of under sampling and over sampling

Perc Over	Perc Under	Accuracy	Total data
2%	2%	0.44	63
10%	10%	0.29	189
7.5%	7%	0.56	513
7.5%	7.8%	0.57	564
7.5%	8%	0.57	576
7.5%	9%	0.59	639
7.5%	10%	0.56	702
8%	10%	0.59	801
8%	9%	0.57	729
8%	12%	0.57	945
8%	11%	0.58	873
8%	5%	0.56	441
8.3%	5%	0.51	441

from table 2 we choose the values of perc over and perc under by the highest accuracy, with the accuracy 0.59 and the highest total data training 801. In this study it can be concluded that the range between perc over and under must not be too large, the best range is 2% and we use perc over 8% and perc under 10%. Then after SMOTE the plot is in Fig 5.

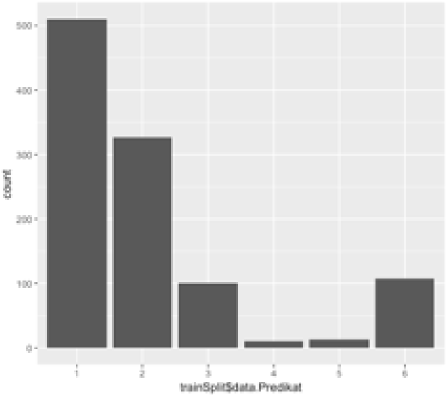

Figure 4. Plot the Dataset After SMOTE

In purpose to determining SVM modeling, first determine the value of C or Cost and determine the value of Gamma. C values commonly used in research are between 0.01, 0.1, 1, 100, and Gamma values commonly used in research that are between 0.01, 0.1, 1, 100. then in this study we will try the 20 combination of the values of each parameter. The best C value and gamma value is with the highest accuracy, the table 3 below is the comparison of 20 combination C and gamma values:

Table 3. Combination C and Gamma values

Cost	Gamma	accuracy	Error Rate
0.01	0.01	0.49	0.51
	0.1	0.49	0.51
	1	0.49	0.51
	100	0.49	0.51
0.1	0.01	0.494	0.506
	0.1	0.477	0.523
	1	0.477	0.523
	100	0.477	0.523
100	0.01	0.493	0.507
	0.1	0.554	0.446
	1	0.575	0.425
	100	0.580	0.42
1	0.01	0.494	0.506
	0.1	0.509	0.491
	1	0.563	0.437
	100	0.578	0.422
10	0.01	0.480	0.52
	0.1	0.530	0.47
	1	0.580	0.42
	100	0.580	0.42

Because of random when running the algorithm in R, there can be differences inaccuracy, but the results of several running in R are closed. In table 3, there is three highest accuracy values whit the accuracy is 0.58 in the combination cost and gamma: 100 and 100, 10 and 1, and 10 and 100. Then to find the best C and gamma we are see the confusion matrix table and compare all of it.

Table 4. *Confusion Matrix* SMOTE cost 10 and gamma 1

Prediction	Excellent	Very Good	Good	Fair	Poor	Failed
Excellent	353	93	45	2	2	11
Very Good	39	112	23	2	4	4
Good	4	10	29	0	0	1

Fair	0	0	0	0	0	0
Poor	0	0	0	0	0	0
Failed	1	1	0	0	0	65

Table 5. *Confusion Matrix* SMOTE cost 10 & gamma100 and cost 100 & gamma 100

Prediction	Excellent	Very Good	Good	Fair	Poor	Failed
Excellent	363	117	50	5	3	11
Very Good	30	102	12	2	0	3
Good	2	4	26	0	0	1
Fair	0	0	0	2	0	0
Poor	0	0	0	0	1	0
Failed	0	0	0	0	0	66

Based on table 4 and 5, the result of confusion matrix cost 10, gamma 100 and cost 100, gamma 100 are both same. In Table 4, the confusion matrix cost 10 and gamma 1 we can see that there are no members in the Poor and Failed class, so we do not choose this. The best combination are cost 10, gamma 100 and cost 100, gamma 100 but we are just use one of them, it is cost 10, gamma 100. Table 6 are the result of Recall and Precision, we can see from it that the precision and recall values are good. With the result of the classification the students who go in to "excellent" class are 363, "very good" class are 102, "good" class are 4, "fair" class are 2, "poor" class is only one, and "failed" class are 66.

Table 6. Precision and Recall

Class Variables	SMOTE	
	Precision	*Recall*
Excellent	0.661	0.918
Very Good	0.684	0.457
Good	0.787	0.295
Fair	1	0.222
Poor	1	0.25
Failed	1	0.814

4. CONCLUSION

We can conclude that by using Support Vector machine with Synthetic Minority Oversampling Technique (SMOTE) method known that the classification has accuracy 58% with the Cost 10 and gamma 100 so that students of Statistics Universitas Islam Indonesia years 2015 who got in to "excellent" class are 363, "very good" class are 102, "good" class are 4, "fair" class are 2, "poor" class is only one, and "failed" class are 66.

In this study it's very hard to make the accuracy higher more than 80% because of many classes each variables, and may only apply to students of Statistics UII years 2015 who get the transition from old curriculum to the new curriculum, so that's the effect the grades.

REFERENCES

[1] Boser, B.E., Guyon, I. M., Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the 5th Annual Workshop on Computational Learning Theory, Jul. 27-29, ACM, New York: 144-152. DOI: 10.1145/130385.130401.

[2] Gaspar, P., J. Carbonell and J.L. Oliveira, 2012. On the parameter optimization of support vector machines for binary classification. J. Integrative Bioinformat. 9: 201-201

[3] Ha, T. M., Bunke, H. (1997). Off-line, Handwritten Numeral Recognition by Perturbation Method. IEEE Transactions on Pattern Analysis and Machine Intelligence. 19(5): 535-539.

[4] Ha, T. M., Bunke, H. (1997). Off-line, Handwritten Numeral Recognition by Perturbation Method.

IEEE Transactions on Pattern Analysis and Machine Intelligence. 19(5): 535-539.

[5] Hsu, C. W., Chang, C. C., Lin, C. J. (2010). A Practical Guide to Support Vector Classification. Taipei: Department of Computer Science. National Taiwan University.

[6] Hyeran, B. Seong-Han, L., (2003). A Survey On Pattern Recognition Applications Of Support Vector Machine. International Journal of Pattern Recognition and Artificial Intelligence, XXII(3): 459-486.

[7] Muñoz, A., Moguerza, J. M., Martos, G., (2019). Support Vector Machines, Madrid: researchgate.

[8] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. Journal of artificial intelligence research. 16. 321-357.

[9] Prasetyo, E., (2012) . Data Mining: Konsep dan Aplikasi menggunakan Matlab. Yogyakarta: Andi Offset.

[10] Renukadevi, N.T., P. Thangaraj. (2013). Performance Evaluation of SVM - RBF Kernel for Medical Image Classification. Global J. Comput. Sci. Technol. Graph. Vis. 13. 1-7.

[11] Rosales-Perez, A., Escalante, H. J., Gonzalez, J. A., & Reyes-Garcia, C. A. (2013). Bias and Variance Optimization for SVMs Model Selection. In The Twenty-Sixth International FLAIRS Conference. Florida. USA: 136-141

[12] Shamshirband, S., Mohammadi, K,. Khorasanizadeh, H. Yee, P.L,. Lee M, (2016). Estimating The Diffuse Solar Radiation Using a Coupled Support Vector Machine-Wavelet Transform Model. Renewable Sustainable Energy Rev. 56. 428-435.DOI: 10.1016/j.rser.2015.11.055.