Integral : ***Jurnal Penelitian Pendidikan Matematika***
p - ISSN 2654-4539
e – ISSN 2654-8720
Vol. 6 No. 1, November 2023
Page 12 of 17

Optimization Of Support Vector Machine (Svm) Based Forward Selection For Prediction Of Incoming Students Continue To Private College

Iqbal Fahmi[1] , Pujiono[2] ,Much Arif Soeleman[3]
[1]Informatics Engineering, TEDC Bandung Polytechnic,[2 3]Magister Informatics Engineering, Computer Science, Dian Nuswantoro University
[1]fahmiiq96@gmail.com, [2]pujiono[a]dsn.dinus.ac.id,[3]
arief22208@gmail.com

Abstract.

The large volume of society can cause problems if it is not commensurate with improving the quality of human resources. A factor that can support human resource capacity is improving the quality of education. High school student data has quite diverse data. With a case study at a high school in Brebes Regency, this experiment is used as a basis for predicting the distribution of high school graduates in the following year. The data mining process is assisted by the WEKA application. The classification used is a support vector machine classification based on forward selection to determine the attributes that are most influential in prediction. The highest results from the SVM experiment were obtained by kernel anova with an accuracy value of 96.17%. Then the FS-SVM algorithm with anova kernel parameter C of 0.5 with an accuracy level of 99.71%.

Keywords: High School, Course, Data, Classification

I. INTRODUCTION

Population count in 2020, Brebes Regency is a region in Central Java with a population of 1.98 million people, and Indonesia 270,200,000 people. The Central Statistics Agency (BPS) and the Ministry of Home Affairs released the results of the 2020 Census data at the Borobudur Hotel, Jakarta.

The Deputy Regent of Narjo and his staff and the Head of BPS Brebes Prita Rextiana were actually present at the event from the Secretariat meeting. According to the 2020 SP results, Cilacap Regency (1.9 million) ranked second, Banyumas ranked third (1.78 million), and Semarang City ranked fourth (1.65 million).[1] If there are still residents who lack education, this will have an impact on the quality of human resources.

Research that has been carried out by Khoirunnisa, Lia Susanti, Ira Tasfiyyutu and Rokhmah conducted research on the predictions of Al-Hidayah Vocational School students entering college in 2021 from the results testing the highest level of accuracy among the three algorithms that have been tested include Decision Tree with an accuracy value of 93.60%, while Naive Bayes with an accuracy value of 92.40% and KNN an accuracy value of 94.96%.[4] The problem underlying this research is finding the best subset of attributes by selecting features on a particular dataset to achieve optimal performance of the classification model.

II. Related Research Journal

In 2019, Rahmat et al. [9] SVM in data mining supports PMB advertising strategies that require precise data processing to find patterns in the data to extract hidden information from the data.

In this research, the SVM method was chosen for data mining analysis. Six attributes are available, next research in 2019. Suharjono. S et al. [10] did research on the application of support vector machine (SVM) taxonomy to student data. When tested with 796 test data, the classification accuracy obtained using the support vector machine method based on particle swarm optimization can increase prediction accuracy from 85.81% to 86.3%.
00.62% increase.

III. RESEARCH METHODS

Support vector machines (SVM) aim to develop efficient methods for learning computing to separate*hyperplanes* in dimensional feature space. The SVM computing process can be done easily. described as an attempt to obtain an optimal hyperplane acting as a boundary between one class and another. For space dimension, Enter the evidence xi

(i=1...k) which belongs to class 1 or class 2 and the corresponding label is -1 for class 1 and + 1 for class 2 Data indicated by x∈ R d while each label is given yi represents {-1 1} for i = 1,2,. ,1, where l is the amount of data. Suppose
two classes - 1 and + 1 can be completely separated from each other by a hyperplane of dimension d. SVM divided by 2

a. *Linear*SVM

Used for linearly separable data, which means that if a data set can be classified into two classes with a single row, the data is called separable data.

b. *Non linear* SVM

Used for data that can be discretized nonlinearly meaning that the dataset cannot classified linearly or straight line, then the data is called nonlinear data and the classifier used is called a nonlinear classifier.

Forward Selection is a gradual rule that aims to add the variables used one by one to an equation that is based on a certain Alpha for input. The input alpha is a value that determines whether one of the predictors is not currently in the model. Therefore, the forward selection technique is one of the methods for selecting the best model in regression by gradually eliminating the independent variables that build the model.

Data processing begins by searching for data in high school schools in Brebes. The first process after finding the data, then the data will be shared. The goal is to divide the dataset into training data and testing data. The division can be done by dividing 60% training data and then 40% testing data or 50% training data and 50% testing data.

Dpreviously collected and processed data cannot be used for testing at this time. Not all of this information is included in the model, so the distributed information is first used in the data separation process so that the data is distributed evenly. Separate data has different purposes, test data is data with smaller parts (test data).

So test data is used. The next process that takes place, the SVM model is run simultaneously with the kernel to determine the accuracy value for each fold and the average accuracy value for all k folds. To compare machine support vectors and support vectors using forward selection.

The next step in the initial stage is data deletion*missing value* as many as 50, so the data becomes 2095. After that determine the variables that will be used as predictor variables and target variables. The variables are student's major type, gender, average report card score, father's professional school exam score, mother's profession, father's education, mother's education, parents' income as predictor variables while the status variable is the target variable.

Table 2. Student dataset

NO	Jurusan Siswa	Jenis Kelamin	Nilai rata-rata Rapot	Nilai Ujian Sekolah	Profesi Ayah	Profesi Ibu	Pendidikan Ayah	Pendidikan Ibu	Penghasilan Orang Tua
1	IPS	L	Cukup	Baik	Petani	Pns	SMP	S1	Sedang
2	IPA	P	Cukup	Baik	Pedagang	Ibu Rumah Tangga	SMA	SMA	Sedang
3	IPA	P	Cukup	Baik	PNS	Karyawan	S1	SMA	Sedang
4	IPA	P	Baik	Baik	Nelayan	Pedagang	SMP	SMA	Sedang
5	IPS	L	Cukup	Baik	Pedagang	PNS	SMA	S1	Besar
6	IPA	P	Cukup	Baik	Pedagang	Petani	SMA	SMP	Besar
7	IPS	P	Baik	Baik	Petani	Ibu Rumah Tangga	SD	S1	Besar
8	IPA	L	Cukup Baik	Baik	Pns	Pedagang	S1	SMA	Besar
9	IPA	P	Baik	Cukup Baik	Nelayan	Petani	SMA	SMA	Sedang
10	IPA	P	Cukup	Cukup Baik	PNS	Pedagang	SMA	S1	Besar
11	IPA	L	Baik	Cukup	Nelayan	PNS	SMP	SMA	Sedang
	...								
	...								
2194	IPA	L	Cukup Baik	Cukup Baik	Karyawan Swasta	Pedagang	SMA	S1	Sedang

In this stage the support vector machine is tested without additional optimization, we change the data set from the first year, two years and the last three years to k-validation using 10 by changing cost=01, cost=0.3 and cost=0.5 to The kernel type uses polynomials and gets the RMSE value.

Linear SVM manual patterns and calculations to find a hyperplane by taking 2 inputs, as follows

Table 3 dataset 2 input

x_1	x_2	Kelas(y)
1	1	1
1	-1	-1
-1	1	-1
-1	-1	-1

There are two features (x1 and x2), then w will also have 2 features (w1 and w2).
The formulation used is as follows:

$$\frac{1}{2}\|w\|^2 = \frac{1}{2}({w_1}^2 + {w_2}^2)$$

Requirement

$y_i(W.X_i+b) \geq 1, \quad i=1,2,3,\ldots\ldots,N$

$y_i(W1.x1+W2.x2) \geq 1$

The following equation is obtained

$(w1+w2+b) \geq 1$, for $y1=1, x1=1, x2=1$

$(-w1+w2-b) \geq 1$, for $y2=-1, x1=1, x2=-1$

$(w1-w2-b) \geq 1$, for $y3=-1, x1=-1, x2=1$

$(w1+w2-b) \geq 1$, for $y4=-1, x1=-1, x2=-1$

Add equations 1 and 2

$(w1+w2+b) \geq 1$

$(-w1+w2-b) \geq 1$ +

$2W2 = 2$

So $W2 = 1$

Adding equations 1 and 3

$(w1 + w2 + b) \geq 1$

$(w1 - w2 - b) \geq 1$ +

$2W_1 = 2$

So $W_1 = 1$

Adding equations 2 and 3

$(-w1 + w2 - b) \geq 1$

$(w1 - w2 - b) \geq 1$ +

$-2b = 2$

So $b = -1$

So we get the hyperplane equation :

$w_1x_1 + w_2x_2 + b = 0$

$x_1 + x_2 - 1 = 0$

$x_2 = 1 - x_1$

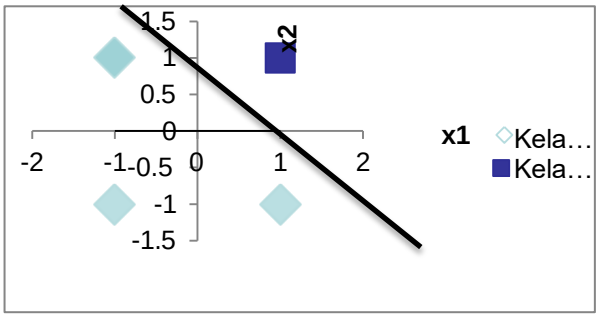


Figure 2 Line visualization*hyperplane*

IV. RESEARCH RESULTS AND DISCUSSION

In the table above. It can be seen that the accuracy of SVM with the linear kernel is lower than the other three kernels. In this experiment, SVM showed a fairly high accuracy score. This prediction is not optimal because SVM has internal weaknesses

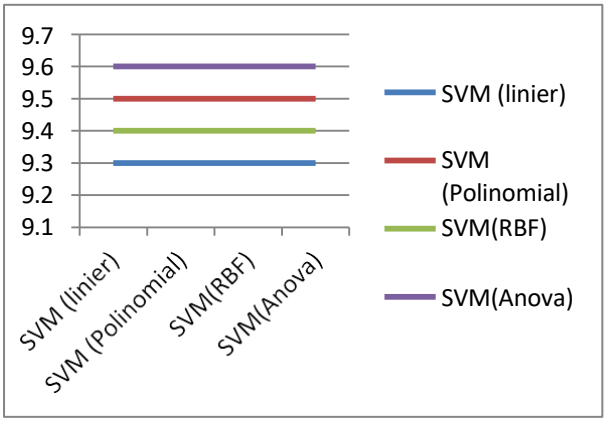
1. Create an empty: $Yk = \{\emptyset\}$, $k = 0$
2. Choose the best features : $X+ = \arg maxx +\in Yk\ [J(Yk + X+)]$
3. If $((Yk + x +) > ((Yk)$
a. Update $Yk+1 = Yk + x +$
b. $k = k + 1$
c. Go back to step -2

The dataset is trained (Training) and tested (Testing), the first dataset is divided by applying 10 k fold validation and divided into two parts, 90% training and 10% testing. Shows a comparison of accuracy values data classification using SVM.

Table 3 SVM accuracy results

Algoritma	TP	TN	FP	FN	Accuary
SVM (Linier)	40	45	20	20	93,12%
SVM (Polinomial)	13	65	45	6	95,71%
SVM (RBF)	25	57	27	7	95,98%
SVM (Anova)	25	57	27	7	96,17%

In the table above. It can be seen that the accuracy of SVM with a linear kernel is lower than the other three kernels. In this experiment, SVM showed a fairly high accuracy score. This prediction is not optimal because SVM has weaknesses in determining optimal parameter values.



Graphic image of 3 SVM

It can be seen in the image above that the highest accuracy value was obtained by the anova kernel with an accuracy value of 96.17%. The graph of the lowest accuracy value was for the linear kernel.

Table 4 Accuracy of SVM with Forward Selection

Parameter		Akurasi	
C	€	SVM	SVM + FS
0.1	0.01	92,15%	93,17%
0.3	0.001	**93,12%**	94,12%
0.5	0.0001	93,15%	**95,59%**

It can be seen that SVM accuracy increases with the False Positive and False Negative values in each SVM kernel. So when compared, there is a striking difference in accuracy before and after selecting the property select function shown in the table above. Apart from that, the settings on the SVM will be set with 3 tests and the accuracy of the SVM will be compared with the accuracy of SVM+FS as can be seen in the table below.

Table 5 compares the accuracy of SVM with SVM & FS

Algoritma	TP	TN	FP	FN	Accuary
SVM (Linier)	70	65	60	60	95,59%
SVM (Polinomial)	53	75	65	36	98,48%
SVM (RBF)	65	77	47	17	98,78%
SVM (Anova)	29	65	32	15	99,04%

It can be seen that SVM accuracy increases with the False Positive and False Negative values in each SVM kernel. So when compared, there is a striking difference in accuracy before and after selecting the property select function shown in the table above.

V. CONCLUSIONS AND SUGGESTIONS

From the results of experiments using data classification of high school student graduates using the Support vector machine method which is optimized using forward selection, it can be concluded as follows:

a. In the experimental results above, the highest accuracy was obtained in the FS-SVM algorithm with anova kernel parameter C of 0.5 with an accuracy level of 99.044%.
b. From the results of testing high school student data, it can be concluded that using the Forward selection method can increase accuracy and the attributes that most influence performance, namely parent income, parent profession, report card grades, school exam scores, parent education, major and gender.

BIBLIOGRAPHY

[1] B. P. S. P. J. Tengah, "*Economic Census*," vol. 1999, no. December, pp. 101-1 1–6,

[2] N. H. Mf, *The Influence of Family Socioeconomics and Future Orientation*, through Self-Motivation on Interest in Continuing to Higher Education in Class XII State High School Students in Brebes Regency. 2018.

[3] *Ministry of National Education*, Department of National Education Policy. 2004.

[4] K. Khoirunnisa, L. Susanti, I. T. Rokhmah, and L. Stianingsih, "*Prediction of Al-Hidayah Vocational School Students Entering Higher Education Using the Classification Method*," J. Inform., vol. 8, no. 1, pp. 26–33, 2021, doi: 10.31294/ji.v8i1.9163.

[5] Anggada Maulana, "Basic Concepts of Data Mining," Data Mining Concepts., vol. 1, pp. 1–16, 2018.

[6] I. C. R. Drajana, "Support Vector Machine and Forward Selection Methods for Predicting Payments for Purchases of Copra Raw Materials," Ilk. J. Ilm., vol. 9, no. 2, pp. 116–123, 2017, doi: 10.33096/ilkom.v9i2.134.116-123.

[7] K. R. Sulaeman, C. Setianingsih, and R. E. Saputra, "Analysis of the Support Vector Machine Algorithm in Stroke Classification," eProceedings Eng., vol. 9, no. 3, pp. 922–928, 2022, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/17909/17544%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/ 17909.

[8] S. Puspita, H. Sakti, and M. Adami, "SVM Algorithm in Data Mining Course Understanding Level (Case Study in Software Engineering Course)," vol. 2020, no. Semnation, pp. 291–300, 2020.

[9] A. Handayanto, K. Latifa, N. D. Saputro, and R. R. Waliyansyah, "*Analysis and Application of the Support Vector Machine (SVM) Algorithm in Data Mining to Support Promotion Strategies (Analysis and Application of the Support Vector Machine (SVM) Algorithm in Data Mining to Support Promotional Strategies*)," vol. 7, no. November, pp. 71–79, 2019.

[10] P. Time, K. Student, and M. Svm, *"Pso-Based,"* vol. 7, no. 2, pp. 97–101, 2019.

[11] A. S. Ritonga and E. S. Purwaningsih, "*APPLICATION OF THE SUPPORT VECTOR MACHINE (SVM) METHOD IN CLASSIFICATION OF SHIELD METAL ARC WELDING QUALITY*," vol. 5, no. 1, pp. 17–25, 2018.

[12] A. Darmawan, N. Kustian, W. Rahayu, T. Tabebuya, and K.visitor, "*IMPLEMENTATION OF DATA MINING USING SVM MODELS,"* vol. 2, no. 3, pp. 299–307, 2018.

[13] D. I. Pusphita Anna Octaviani1, Yuciana Wilandari2, "*APPLICATION OF THE SUPPORT VECTOR MACHINE (SVM) CLASSIFICATION METHOD ON PRIMARY SCHOOL (SD) ACCREDITATION DATA IN MAGELANG DISTRICT Pusphita*," vol. 3, pp. 811–820, 2014.

[14] D. et al Purnamasari, "*Get Easy Using Weka."*

[15] G. W, "No Tit.צוכצle," *Metod. Penelit., vol. 113, 2010, [Online]. Available: https://www.ptonline.com/articles/how-to-get-better-mfi-results.*

[16] D. Sugiyono Prof., "*Prof. Dr. Sugiyono, quantitative qualitative research methods and R&D. intro (PDFDrive).pdf*," Bandung Alf. p. 143, 2011.

[17] S. Arikunto, *"Research Procedures for a* Practice Approach-Xth Revision." 2010.