JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 7. NO. 1 AGUSTUS 2021
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v7i1.2001

THE IMPLEMENTATION OF NAÏVE BAYES AND SUPPORT VECTOR MACHINE (SVM) ALGORITHM, IN DETERMINING ACHIEVING STUDENTS IN SMP NEGERI 8 CIMAHI

Adhitiawarman [1*]; Dwi Hartanto[2]; Adjat Sudradjat[3]; Retno Sari[4]

Information System Study Program[1, 2], Informatic Study Program[3]
Universitas Nusa Mandiri
www.nusamandiri.ac.id
11190464@nusamandiri.ac.id [1*];11190568@nusamandiri.ac.id [2]; retno.rnr@nusamandiri.ac.id [4]

Information System Study Program
Universitas Bina Sarana Informatika
www.bsi .ac.id
adjat.ajt@bsi.ac.id [3]

(*) Corresponding Author

Abstract— In schools that excel, many students have the ability, sometimes they are still confused in determining student achievement. Many students make teachers observant in determining student achievement, students tend to have almost the same probability of being the same, making it difficult for teachers to make decisions. In selecting students who excel, teachers who find it difficult to make decisions because they are not supported by the system to predict the selection and when choosing still use manual calculations this causes a decrease in effectiveness and can be subjective in choosing students who qualify for scholarships, inefficient in terms of time and a human error occurs. This study aims to determine which accuracy is better than the Naïve Bayes Algorithm and Support Vector Machine (SVM) in determining outstanding students. This research, using 400 student data with 10 attributes and in processing the data using R Studio. After testing, was it found that the supporting vector algorithm was 93% accurate and 88% for the Naïve Bayes algorithm. The Support Vector Machine algorithm has better accuracy than the Naïve Bayes algorithm in determining outstanding students at SMP Negeri 8 Cimahi.

Keywords: Naïve Bayes, Support Vector Machine, Student achievement

Abstrak— *Di sekolah yang berprestasi sangatlah banyak siswa yang mempunyai kemampuan, guru terkadang masih merasa kebingungan dalam menentukan prestasi siswanya. Banyaknya siswa membuat guru harus jeli dalam menentukan prestasi siswa, siswa cenderung memiliki kemampuan hampir rata-rata sama sehingga menyulitkan guru dalam mengambil keputusan. Dalam pemilihan siswa yang berprestasi, guru terkadang kesulitan untuk mengambil keputusan dikarenakan belum didukung oleh sistem untuk memprediksi dalam seleksi dan saat memilih masih menggunakan perhitungan manual hal ini menyebabkan penurunan efektifitas dan bisa subjektif dalam memilih siswa yang memenuhi syarat untuk mendapatkan beasiswa, tidak efisien dari segi waktu dan terjadi human error[1][1](Kartika et al., 2017)(Kartika et al., 2017)(Kartika et al., 2017)(Kartika et al., 2017)(Kartika et al., 2017)(Kartika et al., 2017)(Kartika et al., 2017)(Kartika et al., 2017)(Kartika et al., 2017)(Kartika et al., 2017)(Kartika et al., 2017)(Kartika et al., 2017)(Kartika et al., 2017). Penelitian ini bertujuan untuk untuk mengetahui akurasi mana yang lebih baik dari Algoritma Naïve Bayes dan Support Vector Machine (SVM) dalam menentukan siswa berprestasi. Pada penelitian ini menggunakan sebanyak 400 data siswa dengan 10 atribut dan dalam pengolahan datanya mengunakan R Studio. Setelah dilakukan pengujian didapati hasil akurasi untuk algoritma support vector machine sebesar 93% dan untuk algoritma Naïve Bayes 88%. Algoritma Support Vector Machine nilai akurasinya lebih baik dari algoritma Naïve Bayes pada penentuan siswa berprestasi pada SMP Negeri 8 Cimahi.*

Kata Kunci:*, Naïve Bayes, Support Vector Machine, Siswa Berprestasi*

INTRODUCTION

Education is the most important thing in our lives, this means that every Indonesian human being has the right to get it and is expected to always develop therein. Education at the junior high school level is a foundation-laying in preparing the next generation to become humans who are ready to

compete in an increasingly advanced era. Quality and quality education is also in the heart of each student's parents, for example, a school that has many achievements is also a consideration.

In schools that excel some so many students have the ability, teachers sometimes still feel confused in determining student achievement. Learning achievement is defined as the achievement of students in completing the lesson stamp that is well-received [2]. The number of students makes teachers observant in determining student achievement, students tend to have almost the same ability, making it difficult for teachers to make decisions. This decision will certainly identify students who are likely to experience difficulties and obstacles in learning [3].

In selecting students who excel, teachers sometimes find it difficult to make decisions because they are not supported by the system to predict the selection and when choosing still use manual calculations this causes a decrease in effectiveness and can be subjective in choosing students who qualify for scholarships [4], inefficient in terms of time [5], human error occurs [1], large amounts of data cannot be handled [6].

Tabel 1 Research Literature

Literatur Supports	Kernel	Classifier	Accuracy
Application of Naïve Bayes Data Mining Method to Predict Learning Outcomes of Junior High School Students [7]	-	Naïve Bayes	56.79%
Analysis of the Performance of the C4.5 Algorithm and Naïve Bayes in Predicting the Success of Schools in Facing the National Examination [3].	-	Naïve Bayes	95.50%
Data Mining Method for Selection of Prospective Students for New Student Admissions at Pamulang University [8]	-	Support Vector Machine (SVM)	65%
Application of the Classification Algorithm to Support the Decision of Awarding Student Scholarships [9]	-	Naïve Bayes	80%
Application of the Support Vector Classification Method Machine (SVM) on Accreditation Data for Primary Schools (SD) in Magelang Regency [10]	Gaussian Radial Basis Function (RBF)	Support Vector Machine	93.902%

From table 1 it is explained that data mining can help in making decisions, data mining is very suitable for decision-making techniques [11], besides that data mining is a popular technique for determining student performance [12]. Due to data mining, the speed and accuracy of decision-making are effective and time-efficient, large data cannot be done quickly if it is still using manual methods [13]. Therefore, to assist Cimahi State Junior High School in selecting outstanding students, a method is needed that can assist in the data processing. This study aims to determine which accuracy is better than the Naïve Bayes Algorithm and Support Vector Machine (SVM) in determining outstanding students.

MATERIALS AND METHODS

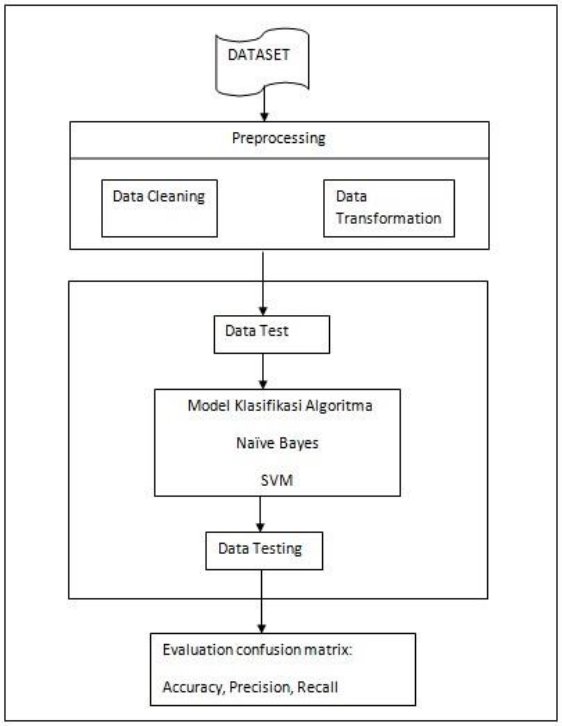
The stages in this research are:

a. Data collection
   In this study, the authors took data from SMP Negeri 8 Cimahi which consisted of 4 00 student data with the attribute names NIM, Name, Gender, Attendance, Skills Mid-Term Exam, Knowledge Final Exam, Skills Final Exam, Knowledge Final Exam, Predicate, Extracurricular.
b. Initial Processing of Data
   Preprocessing carried out on the dataset, namely:
   1) Data Cleaning
   2) The data cleaning stage includes the process of cleaning data by removing duplicate data, checking for inconsistent data, and correcting errors in data, such as typos.
   3) Data Transformation
      In the transformation data itself, it can be obtained from the unit test data sample, where the data is converted from a variable to a numeric by providing a label for each character to calculate the value of the matrix using the algorithm method that will be used.
c. Data Test
   Is part of the dataset that is tested to see its accuracy, or in other words, to see its performance.
d. Model Used
   The model used is Naïve Bayes and Support Vector Machine (SVM)
e. Data Testing
   Is a part of the dataset that is trained to make predictions or carry out the functions of an algorithm. We provide clues through algorithms so that the trained machine can look for correlations on its own or learn patterns from the given data.
f. Experiment and Model Testing
   Experiments in this study using R studio in the data processing.

VOL. 7. NO. 1 AGUSTUS 2021
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v7i1.2001

g. Evaluation and Validation of Results
To perform accuracy, precision, and recall measurements using a confusion matrix.

The following is a picture of the model from this research:



Source: [14]

Figure 1. Research Model

RESULTS AND DISCUSSION

Research data

This research data is a collection of value Rapo r, grade 9 student extracurricular activities, the amount of data that studied 400 students, with the translation of each data has attribute used in the calculation process that will be used for determining students who excel. Following the student, data attributes are presented in tabular form.

Table 2 Student data attributes

No	Attribute	Information
1	NIM	Attribute that informs the student's NIM
2	Nama	Attribute that informs the student's name
3	Pendidikan Agama	Attributes that inform the value of Religious Education lessons
4	PPKN	Attributes that inform the value of the PPKN lesson
5	Bahasa Indonesia	Attributes that inform the value of Indonesian lessons
6	Matematika	Attributes that inform the value of a Math lesson
7	IPA	Attributes that inform the value of a science lesson
8	IPS	Attributes that inform the value of social studies lessons
9	Bahasa Inggris	Attributes that inform the value of English lessons
1	Seni Budaya	Attributes that inform the value of Cultural Arts lessons
11	Penjaskes	Attributes that inform the value of Penjaskes lessons
12	Prakarya	Attributes that inform the value of Practice lessons
13	Bahasa Sunda	Attributes that inform the value of Sundanese lessons
14	Ketidak hadiran, Sakit, Ijin, Alpha	Attributes that inform attendance
15	Ektra Kulikuler 1	Attributes that indicate extracurricular value 1
16	Ektra Kulikuler 2	Attributes that indicate extracurricular value 2
17	Ektra Kulikuler 3	Attributes that indicate extracurricular value 3

Data Selection

In student data, attributes are selected and selected for use in the mining process. then the attributes that will be used in the study are presented in the table.

Table 3. Attribute selection

No	Attribute
1	NIM
2	Nama
3	Jenis Kelamin
4	Absensi
5	UTS Keterampilan
6	UTS Pengetahuan
7	UAS Keterampilan
8	UAS Pengetahuan
9	Predikat
10	Ekstrakulikuler

Model Testing with the Naïve Bayes Method

Naïve Bayes is a classification based on the Bayes theorem and is used to calculate the probability of each class with the assumption that one class is independent of one another. In this method, all attributes will contribute to decision making, with attribute weights that are equally important and each attribute is independent of one another (Saputra, 2018).

$$P(v_j) = \frac{N}{sum} \quad (1)$$

information:

$P(v_j)$ = Probability hypothesis v_j (prior)

N = Amount of training data where v = v_j
Sum = Amount of training data

Testing using the Rstudio application with the Naïve Bayes method. This process is a calculation to find the accuracy of the dataset. The following is the calculation table 4 from Naïve Bayes using R Studio.

Tabel 4. Calculation of Naïve Bayes eith R Studio

	accuracy	precision	recall
	<dbl>	<dbl>	<dbl>
A	0.8888889	NaN	0.0000000
B	0.8888889	0.8888889	0.9655172
C	0.8888889	0.8888889	0.8648649
D	0.8888889	NaN	0.0000000

Confusion Matrix and Statistics

Reference

Prediction	A	B	C	D
A	0	2	0	0
B	0	56	2	0
C	0	5	32	0
D	0	0	2	0

Overall Statistics

Accuracy	:	0.8889
95% CI	:	(0.8099, 0.9432)
No Information Rat	:	0.6364
P-Value [Acc > NIR]	:	1.249e-08
Kappa	:	0.7738
Mcnemar's Test P-Value	:	NA

Statistics by Class:

	Class: A	Class: B	Class: C	Class: D
Sensitivity	NA	0.8889	0.8889	NA
Specificity	0.9798	0.9444	0.9206	0.9798
Pos Pred Value	NA	0.9655	0.8649	NA
Neg Pred Value	NA	0.8293	0.9355	NA
Prevalence	0.0000	0.6364	0.3636	0.0000
Detection Rate	0.0000	0.5657	0.3232	0.0000
Detection Prevalence	0.0202	0.5859	0.3737	0.0202
Balanced Accuracy	NA	0.9167	0.9048	NA

R Console

From the calculation using the Naïve Bayes method with the results, the environment for the data set is 400 data, the overall test_set is 99 and the training_set is 301 with an accuracy of 88%.

Model Testing with Support Vector Machine

Support Vector Machine (SVM) is a selection method by assessing the standard parameters of the candidate set to draw the accuracy that has the best classification value [15] SVM has a basic principle of linear classifier, namely classification cases that can be separated linearly, however, SVM has been developed to work on non-linear problems by incorporating the kernel concept in a high-dimensional workspace. In high-dimensional space, a hyperplane (hyperplane) will be sought which can maximize the distance (margin) between data classes.

The measurement of classification performance on the original data and the resulting data from the classification model was carried out using cross-tabulation (confusion matrix) which contains information about the original data class represented on the matrix row and the predictive data class of an algorithm represented in the classification column.

Tabel 5. Three Class Confusion Matrix

Fgh		Prediction Class (h)		
		Class 1	Class 2	Class 3
Original Class (g)	Class 1	F11	F12	F13
	Class 2	F21	F22	F23
	Class 3	F31	F31	F33

Classification accuracy = $\frac{F11+F22+F33}{F12+F13+F21+F23+F31+F32+F11+F22+F33}$ ..... (2)

Classification error = $\frac{F12+F13+F21+F23+F31+F32}{F12+F13+F21+F23+F31+F32+F11+F22+F33}$ .....(3)

Classification accuracy shows the overall performance of the classification model, where the higher the classification accuracy, this means the better the classification model performance.

All classifications attempt to form models with high accuracy values and low error rates.

Table 6. Calculation of SVM with R Studio

	accuracy	precision	recall
	<dbl>	<dbl>	<dbl>
A	0.9393939	NaN	NaN
B	0.9393939	0.9682540	0.9384615
C	0.9393939	0.8888889	0.9411765
D	0.9393939	NaN	NaN

Confusion Matrix and Statistics

Reference

Prediction	A	B	C	D
A	0	2	0	0
B	0	62	4	0
C	0	2	32	0
D	0	0	0	0

Overall Statistics

Accuracy	:	0.9394

VOL. 7. NO. 1 AGUSTUS 2021
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v7i1.2001

95% CI	:	(0.8727, 0.9774)
No Information Rat	:	0.6364
P-Value [Acc > NIR]	:	1.616e-12
Kappa	:	0.8675
Mcnemar's Test P-Value	:	NA

Statistics by Class:

	Class: A	Class: B	Class: C	Class: D
Sensitivity	NA	0.9683	0.8889	NA
Specificity	1	0.8889	0.9683	1
Pos Pred Value	NA	0.9385	0.9412	NA
Neg Pred Value	NA	0.9412	0.9385	NA
Prevalence	0	0.6364	0.3636	0
Detection Rate	0	0.6162	0.3232	0
Detection Prevalence	0	0.6566	0.3434	0
Balanced Accuracy	NA	0.9286	0.9286	NA

R Console

From the calculation using the support vector machine (SVM) method with the results, for the data set that is 400 data, the overall test_set is 99 and the training_set is 301 with an accuracy of 93%.

Other attributes will be analyzed based on predetermined classes, namely gender class and class based on achievement. From visualize, data can be read and found new from that data. For more details, it will be explained below each visualize of the attributes used.

Class-based on gender

The results of visualizing observations on gender attributes are divided into two (2), namely (0) and (1), code (0) is for the male gender, while code (1) is for the female gender. Meanwhile, the predicate itself is divided into four (4), namely code (0), (1), (2), and (3) the translation of the code (0) = grade A, (1) = grade B, (2) = grade C, and (3) = grade D, from the graph below it can be seen that the predicate that is most abundant in the table below is female.

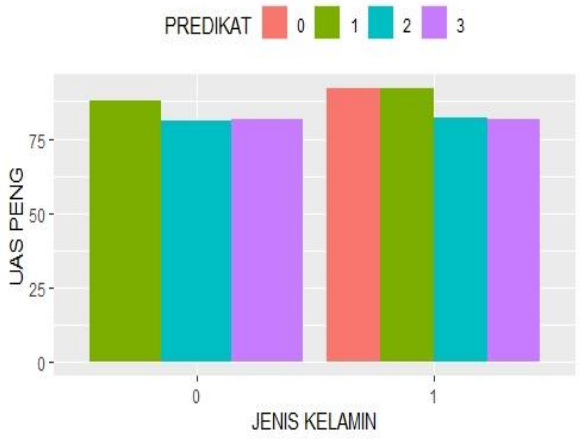


Figure 2 . Achievement chart seen from a gender

Class-based on student achievement

The results of visualizing observations on student achievement attributes are divided into five (5) namely UTS Knowledge, Final Exam Knowledge, Attendance, Extracurricular, and Predicate.

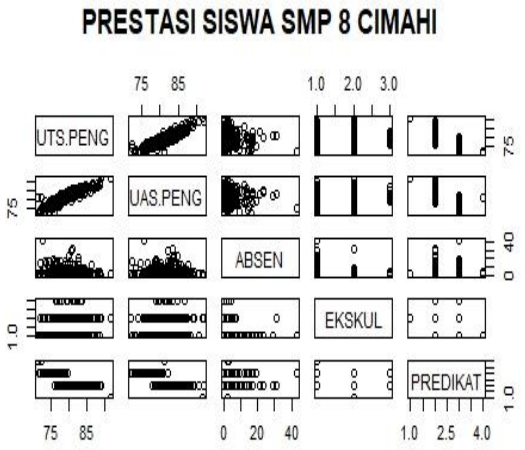


Figure 3. Prestasi Siswa SMP 8 Cimahi

From the results of the calculation process using R studio by displaying the results of the data that has been processed above. This column provides information about the 9th-grade student data taken in 2020. Next is to find out the accuracy level of the Naïve Bayes algorithm and support vector machine (SVM), in testing using the R studio application, each has a difference in accuracy results. This shows that in comparison to the use of the support vector machine (SVM) algorithm method with an accuracy of 93% the results are better than the use of the Naïve Bayes algorithm method with an accuracy of 88%.

CONCLUSION

Based on the results of testing, and data analysis it can be concluded that the implementation of the SVM algorithm and Naïve Bayes in the classification of student achievement at SMP Negeri 8 Cimahi, the accuracy results obtained from the calculation of the two methods using the Rstodi application, namely SVM produces an accuracy of 93% and the Naive Bayes method of 88%. Based on the comparison of methods obtained from the testing above, the SVM method is a good method with a total average accuracy of 0.93%.

REFERENCE

[1] J. I. Kartika, E. Santoso, and Sutrisno, "Penentuan Siswa Berprestasi Menggunakan Metode K-Nearest Neighbor dan Weighted

Product (Studi Kasus: SMP Negeri 3 Mejayan)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 5, pp. 352–360, 2017.

[2] N. Jannah and T. Yulianto, "Mengelompokkan Siswa Berprestasi Akademik dengan Menggunakan Metode K Means Kelas VII MT," *Zeta - Math J.*, vol. 2, no. 2, pp. 41–45, 2016.

[3] Y. Angraini, S. Fauziah, and J. L. Putra, "Analisis Kinerja Algoritma C4.5 Dan Naïve Bayes Dalam Memprediksi Keberhasilan Sekolah Menghadapi UN," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 5, no. 2, pp. 285–290, 2020.

[4] M. Marlina, W. Yusnaeni, and N. Indriyani, "Sistem Pendukung Keputusan Pemilihan Siswa Yang Berhak Mendapatkan Beasiswa Dengan Metode Topsis," *J. Techno Nusa Mandiri*, vol. 14, no. 2, pp. 147–152, 2017.

[5] A. Topadang and R. T. Tulili, "Sistem Pendukung Keputusan Pemilihan Siswa Berprestasi Di Jemaat Moria Samarinda Seberang Dengan Metode Simple Additive Weigthting," *J. Sains Terap. Teknol. Inf.*, vol. 10, no. 2, p. 122, 2018.

[6] M. L. Sibuea and A. Safta, "Pemetaan Siswa Berprestasi Menggunakan Metode K-Means Clustring," *Jurteksi*, vol. 4, no. 1, pp. 85–92, 2017.

[7] F. Rahman, D. Muhammad, and I. Firdaus, "Penerapan Data Mining Metode Naïve Bayes Untuk Prediksi Hasil Belajar Siswa Sekolah Menengah Pertama (Smp)," *Al Ulum Sains dan Teknol.*, vol. 1, no. 2, pp. 76–78, 2016.

[8] A. Saifudin, "Metode Data Mining Untuk Seleksi Calon Mahasiswa," *J. Teknol.*, vol. 10, no. 1, pp. 25–36, 2018.

[9] H. Sulistiani, "Penerapan Algoritma Klasifikasi Sebagai Pendukung Keputusan Pemberian Beasiswa Mahasiswa," pp. 300–305, 2018.

[10] P. A. Octaviani, Y. Wilandari, and D. ISpriyanti, "Penerapan Metode Klasifikasi Support Vector Machine (SVM) Pada Data Akreditasi Sekolah Dasar (SD) Di Kabupaten Magelang," *J. Gaussian*, vol. 3, pp. 811–820, 2014.

[11] M. A. Sembiring, M. F. L. Sibuea, and A. Sapta, "Analisa Kinerja Algoritma C.45 Dalam Memprediksi Hasil Belajar," *J. Sci. Soc. Res.*, vol. 1, no. 1, pp. 73–79, 2018.

[12] T. Setiyorini and R. T. Asmono, "Penerapan Metode K-Nearest Neighbor Dan Information Gain Pada Klasifikasi Kinerja Siswa," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 5, no. 1, pp. 7–14, 2019.

[13] A. Noviriandini and N. Nurajijah, "Analisis Kinerja Algoritma C4.5 Dan Naïve Bayes Untuk Memprediksi Prestasi Siswa Sekolah Menengah Kejuruan," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 5, no. 1, pp. 23–28, 2019.

[14] A. Adhitiawarman, D. Hartanto, A. Sudradjat, and R. Sari4, "Laporan Akhir Penelitian Mandiri 2021," Jakarta, 2021.

[15] Suhardjono, W. Ganda, and H. Abdul, "Prediksi Kellusan Menggunakan SVM Berbasis PSO," *Bianglala Inform.*, vol. 7, no. 2, pp. 97–101, 2019.