Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Comparison of SVM, NBC, and KNN Classification Methods in Determining Students' Majors at SMK N02 Manokwari

Siska Howay[1] ✉ **and Suhirman**[2]
[12]*Master Program Information Technology, Universitas Teknologi Yogyakarta, Indonesia*
Corresponding Author: Siska Howay, **E-mail**: Siska.howay@student.uty.ac.id

| ABSTRACT

The stages of choosing a major for prospective SMK students are rarely the beginning of the next career determination. The determination of the major aims to make students more directed in receiving lessons based on the abilities and talents of the students, and, of course, when the student graduates, they already have the *skills* to get a job if they do not continue their education to college. The method used in this study is data mining techniques. But not all data mining algorithms perform well in classifying the selection of interest paths at the SMK level. Therefore, this study will discuss the comparative analysis of the performance level of the *Support Vector Machine* (SVM) classification algorithm and *the Naïve Bayes Classifier* (NBC) and *K-Nearest Neighbors* (KNN*)*. Comparison of NBC, KNN and SVM methods was measured using feeding accuracy for the KNN method to get an accuracy of 54.56%, then for the NBC method to get an accuracy of 74.78%, and the SVM method to get an accuracy of 58.70%. Then it can be concluded that the three methods, based on the attributes used by the NBC method, got high accuracy, which is 74.78%.

| KEYWORDS

Comparison; SVM; NBC; KNN; classification methods

| ARTICLE INFORMATION

ACCEPTED: 01 February 2023 **PUBLISHED:** 13 February 2023 **DOI:** 10.32996/jcsts.2023.5.1.3

1. Introduction

Selecting majors for prospective SMK students is the first step in determining the next career. Students are more inclined to choose a major influenced by friends and many parents' choices. The determination of the major aims to make students focused on receiving lessons based on the abilities and talents of the students, and of course, when the students graduate, they already have *the skills* to get a job if they do not continue to lectures. Choosing the right major can improve achievement and give students a sense of comfort in learning. The lack of interest in learning caused by choosing the wrong major causes the passion for learning to disappear, which triggers students to skip school often, and classes become rowdy, which results in decreased achievement.

Vocational high school (SMK) is a formal education unit that organizes vocational education. Vocational education is a part of education that prepares students to be able to work in a certain field and be able to work in a group of jobs. So in each new school year, every student who wants to enrol in SMK will choose a major while studying at SMK. Currently, the selection test for new students at SMK Negeri 02 Manokwari uses the following components: UN scores, hobbies, school status, parental occupation, gender, parental income, number of benefits and student choices. Furthermore, these components are processed using MS Excel, and the decision is made subjectively in the student admissions meeting. This results in the student's major not matching the student's abilities and talents.

Several previous studies have predicted majors in high schools and colleges using data mining techniques. These studies include: Howay & Rianto (2021) using K-means. Vista & Pasaribu (2021) using the C4.5 method. Sinaga et al., (2021) using the Decision

Tree, KNN, and SVM Methods. All three studies were conducted in high schools. Application of the *Naïve Bayes* Method in Predicting the Determination of the Department of IT Students firdaus, K. L (2021). This research was conducted at the university.

Based on previous studies, the authors proposed a classification study using the SVM, NBC, and KNN methods in determining the majors of students in SMK and comparing the mottoes of SVM, NBC, and KNN.

To guide participants to choose the right specialization path is very important in the type of learning that exists in schools so that it is a basis for continuing high teacher training. There are many methods that can be done to choose the path of interest. One of the methods used is data mining techniques. But not all data mining algorithms have good performance in classifying the selection of interest paths at the SMK level. Therefore, this study will discuss the comparative analysis of the performance level of the *Support Vector Machine* (SVM) classification algorithm and *the Naïve Bayes Classifier* (NBC) and *K-Nearest Neighbors* (KNN*)*. The comparison analysis in question is a comparison of the accuracy rate of the three algorithms. SVM is a *machine learning* method that works on the principle of Structural Risk Minimization (SRM) with the aim of finding the best *hyperplane* that separates two *classes* in the *input space,* While the Naïve Bayes Classifier is a statistical classification that can be used to predict the probability of membership of a class and KNN is a method that uses a supervised algorithm, in which the results of a newly classified test sample based on the majority of the categories on the k-NN aim to classify new objects based on attributes and training samples.

Based on the description above, the author is interested in conducting a study entitled Comparison of classification methods S VM, NBC, dan KNN in determining the majors of students di SMK N 02 Manokwari.

2. Research Methodology

2.1 Stages carried out in this study:

1. *Selection*

This stage selects major data consisting of prediction variables and one target variable. The target variable is the Department. Meanwhile, the variables predict UN scores, hobbies, school status, parents' occupation, gender, parents' income, number of benefits and student choices.

2. *Preprocessing*

The amount of data taken corresponds to the number of new students who apply. From the existing data, data cleaning is carried out if there is missing data or double or outlier data.

3. *Transformation*

After the data cleansing process from errors, further transformations are carried out on the data according to the type of data. The transformation stage is where the types will be grouped into data categories for each prediction variable and data categories for the target variable.

4. *Data Mining*

At this stage, the selection of majors for the supervised learning classification function is used by SVM, NBC, and KNN algorithms.

5. *Evaluation*

At this stage, the evolution of the predicted results obtained from the three methods and obtained the best method is close to the actual data classification. Evaluation Criteria with Confusion Matrix method and ROC (Receiver Operating Characteristic) curve. Accuracy and error are used to get the Performance value. It can further be depicted in figure 3.1

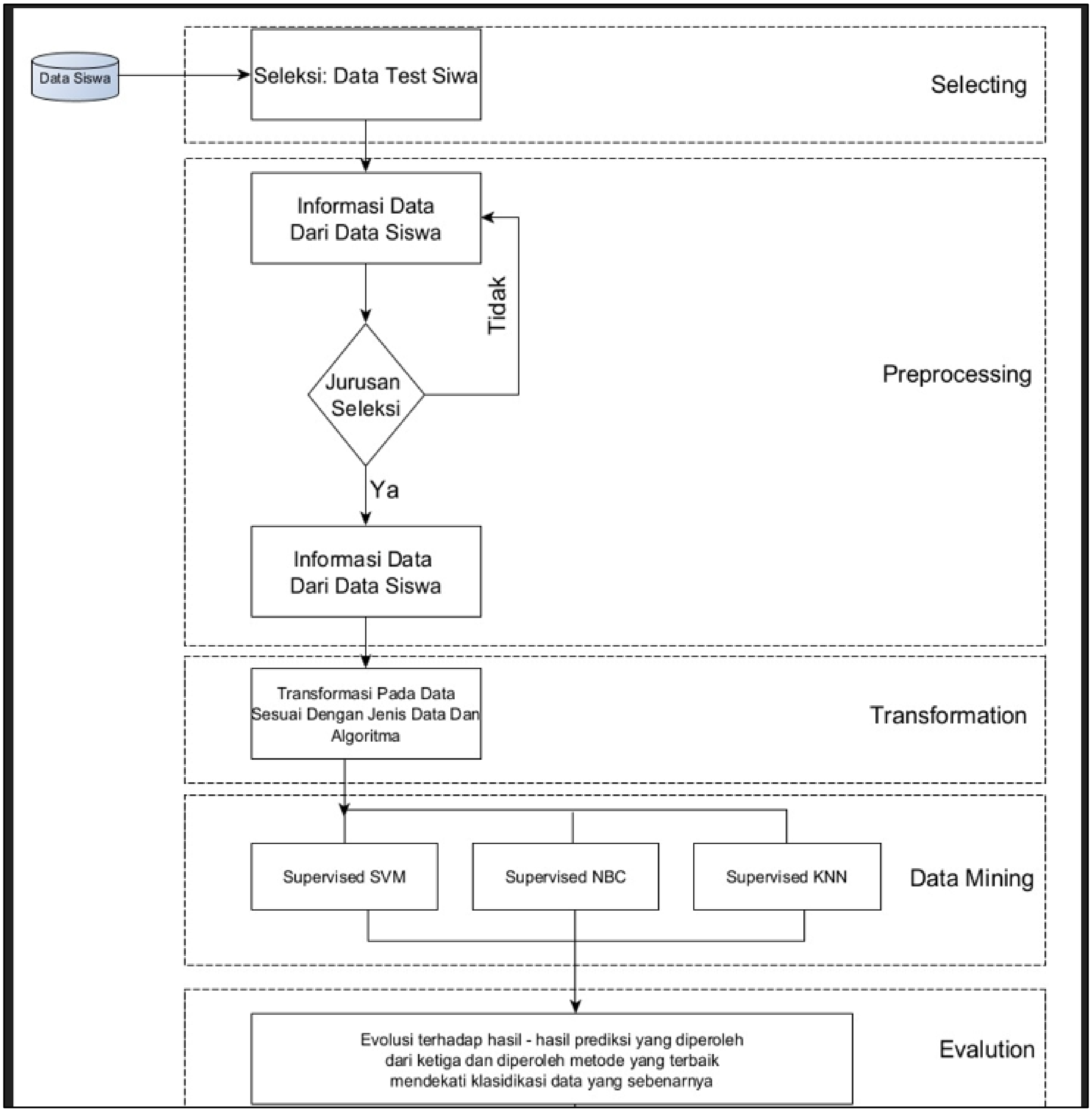

Data Siswa
Seleksi: Data Test Siwa
Selecting
Informasi Data Dari Data Siswa
Tidak
Jurusan Seleksi
Ya
Preprocessing
Infomasi Data Dari Data Siswa
Transformasi Pada Data Sesuai Dengan Jenis Data Dan Algoritma
Transformation
Supervised SVM
Supervised NBC
Supervised KNN
Data Mining
Evolusi terhadap hasil - hasil prediksi yang diperoleh dari ketiga dan diperoleh metode yang terbaik mendekati klasidikasi data yang sebenarnya
Evalution

Figure 3.1 Research Flow

2.2 Consumed Attributes

Data on prospective SMK students is obtained from the admission section of new students; the initial attributes of this data are Name, National Exam Score, Hobbies, School Status, Parents' Occupation, Gender, Parents' Income, Number of Shelters and Student Choices. Attributes can be seen in Table 3.1

Table 3.1 Attributes

National Exam Score	Hobby	School Status	Parents' Occupation	Parents' Income	Number of Shelters	Student options	Gender
70	Swimming pool	Country	Civil servants	500.000	2	TKJ	P
80	Football	Private	private sector employee	1.000.000	3	TBSM	L
78	Dance	Country	Farmer	2.500.000	1	TPMG	P

85	Sing	Country	Civil servants	3.500.00	4	ATPH	L
75	Read	Private	private sector employee	4.000.000	3	APAT	P
85	Swimming pool	Private	Farmer	5.000.000	5	TKJ	L
70	Paint	Private	Civil servants	5.500.000	6	TBSM	P
80	Dive	Country	private sector employee	6.000.000	4	TPMG	L
85	Gardening	Private	Farmer	1.500.000	3	ATPH	P

3. Results and Discussion

3.1 Implementation and Results

In this chapter, we will discuss the implementation of the applications used in this study. The purpose of the implementation is to find out whether the application used in this study can provide precise or accurate results. To carry out the implementation, it is necessary to prepare several things as follows:

1. hardware requirements (Hardware)
   Hardware analysis is the main need of a physical computer system, which consists of interrelated components in the form of inputs, processes and outputs. The required hardware is as follows:

a. . Intel Celeron CPU B815 A43E processor.
b. RAM (*Random Access Memory*) 4 GB.
c. Storage media *(hard drive*) with a capacity of 500 GB.

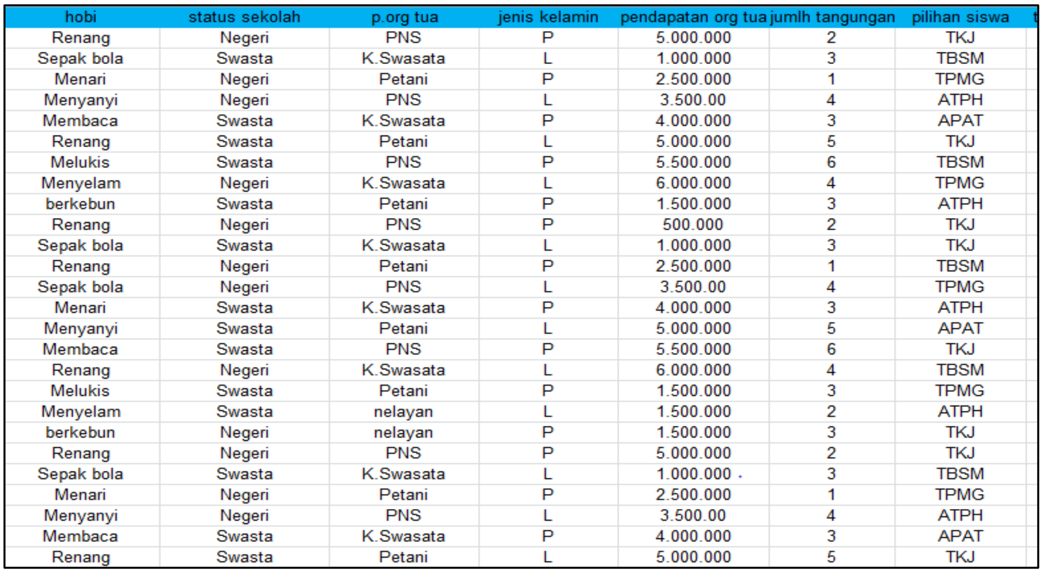
2. *Software* Needs
   Software needs are one of the most important factors in this study. Software required:

a. Windows 7 Operating System (32/64 bit)
b. RapidMiner Studio App

3.2 RapidMiner Application Implementation Results

The results of the implementation of accuracy from the KNN, NBC, and SVM methods using the RapidMiner Application are based on new student admission data at SMK N02 Manokwari. The following below is an explanation of the accuracy prediction process of the KNN, NBC and SVM methods that are processed using the RapidMiner application.

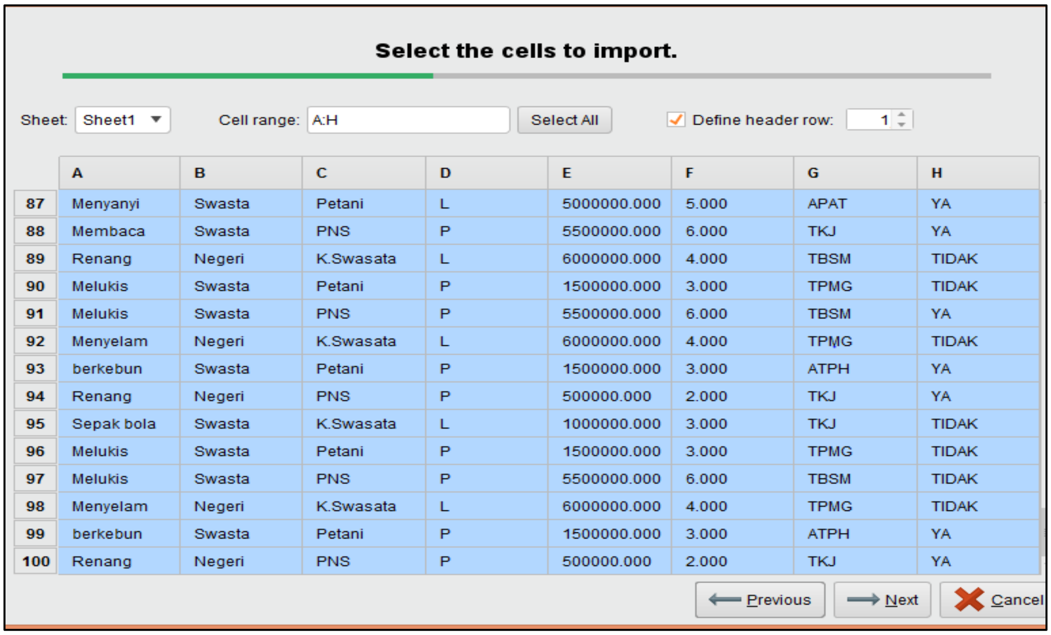
1. The first thing we need to prepare in this process is data, where the data used in this research is student data. Furthermore, it can be seen in figure 4.1



hobi	status sekolah	p.org tua	jenis kelamin	pendapatan org tua	jumlh tangungan	pilihan siswa
Renang	Negeri	PNS	P	5.000.000	2	TKJ
Sepak bola	Swasta	K.Swasata	L	1.000.000	3	TBSM
Menari	Negeri	Petani	P	2.500.000	1	TPMG
Menyanyi	Negeri	PNS	L	3.500.00	4	ATPH
Membaca	Swasta	K.Swasata	P	4.000.000	3	APAT
Renang	Swasta	Petani	L	5.000.000	5	TKJ
Melukis	Swasta	PNS	P	5.500.000	6	TBSM
Menyelam	Negeri	K.Swasata	L	6.000.000	4	TPMG
berkebun	Swasta	Petani	P	1.500.000	3	ATPH
Renang	Negeri	PNS	P	500.000	2	TKJ
Sepak bola	Swasta	K.Swasata	L	1.000.000	3	TKJ
Renang	Negeri	Petani	P	2.500.000	1	TBSM
Sepak bola	Negeri	PNS	L	3.500.00	4	TPMG
Menari	Swasta	K.Swasata	P	4.000.000	3	ATPH
Menyanyi	Swasta	Petani	L	5.000.000	5	APAT
Membaca	Swasta	PNS	P	5.500.000	6	TKJ
Renang	Negeri	K.Swasata	L	6.000.000	4	TBSM
Melukis	Swasta	Petani	P	1.500.000	3	TPMG
Menyelam	Swasta	nelayan	L	1.500.000	2	ATPH
berkebun	Negeri	nelayan	P	1.500.000	3	TKJ
Renang	Negeri	PNS	P	5.000.000	2	TKJ
Sepak bola	Swasta	K.Swasata	L	1.000.000	3	TBSM
Menari	Negeri	Petani	P	2.500.000	1	TPMG
Menyanyi	Negeri	PNS	L	3.500.00	4	ATPH
Membaca	Swasta	K.Swasata	P	4.000.000	3	APAT
Renang	Swasta	Petani	L	5.000.000	5	TKJ

Figure 4.1 student data

2. In this process, researchers use the RapidMiner application, then download and install the application on the top or computer used, then open the application and import the prepared data into the RapidMiner application; if it is imported, click the Netx button contained in the application, it will appear like figure 4.2 then click next:

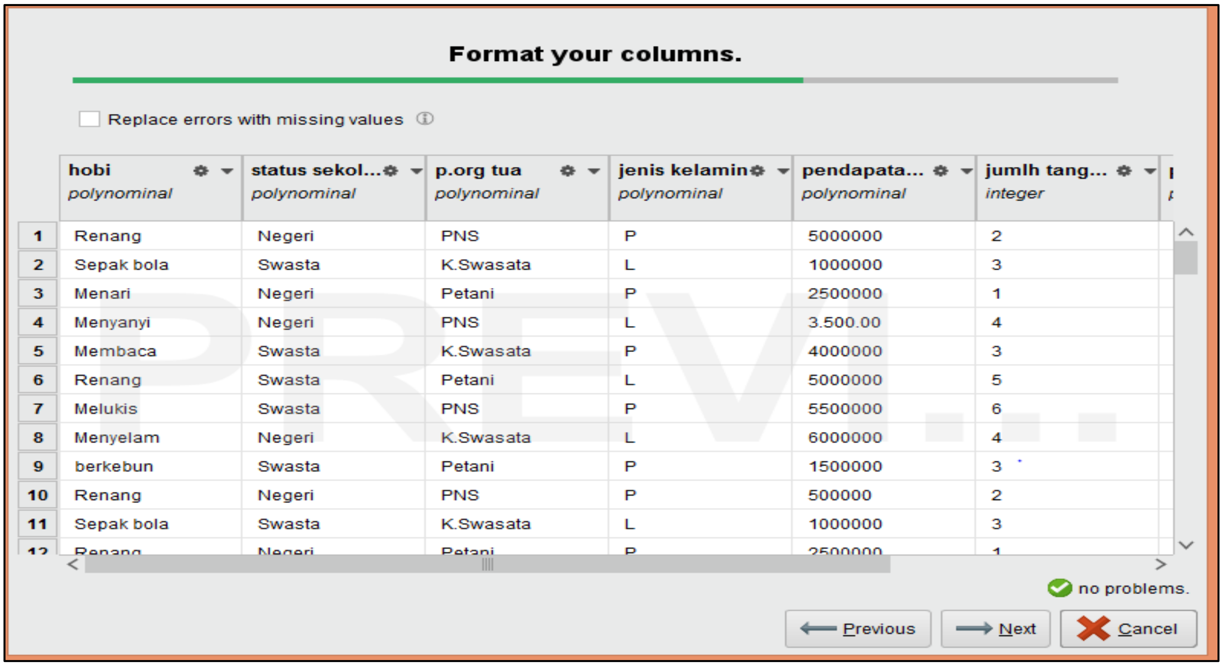


Select the cells to import.

	A	B	C	D	E	F	G	H
87	Menyanyi	Swasta	Petani	L	5000000.000	5.000	APAT	YA
88	Membaca	Swasta	PNS	P	5500000.000	6.000	TKJ	YA
89	Renang	Negeri	K.Swasata	L	6000000.000	4.000	TBSM	TIDAK
90	Melukis	Swasta	Petani	P	1500000.000	3.000	TPMG	TIDAK
91	Melukis	Swasta	PNS	P	5500000.000	6.000	TBSM	YA
92	Menyelam	Negeri	K.Swasata	L	6000000.000	4.000	TPMG	TIDAK
93	berkebun	Swasta	Petani	P	1500000.000	3.000	ATPH	YA
94	Renang	Negeri	PNS	P	500000.000	2.000	TKJ	YA
95	Sepak bola	Swasta	K.Swasata	L	1000000.000	3.000	TKJ	TIDAK
96	Melukis	Swasta	Petani	P	1500000.000	3.000	TPMG	TIDAK
97	Melukis	Swasta	PNS	P	5500000.000	6.000	TBSM	TIDAK
98	Menyelam	Negeri	K.Swasata	L	6000000.000	4.000	TPMG	TIDAK
99	berkebun	Swasta	Petani	P	1500000.000	3.000	ATPH	YA
100	Renang	Negeri	PNS	P	500000.000	2.000	TKJ	YA

Figure 4.2 Data Import Process

3. If it appears as shown above, then click Next, and it will appear like figure 4.3

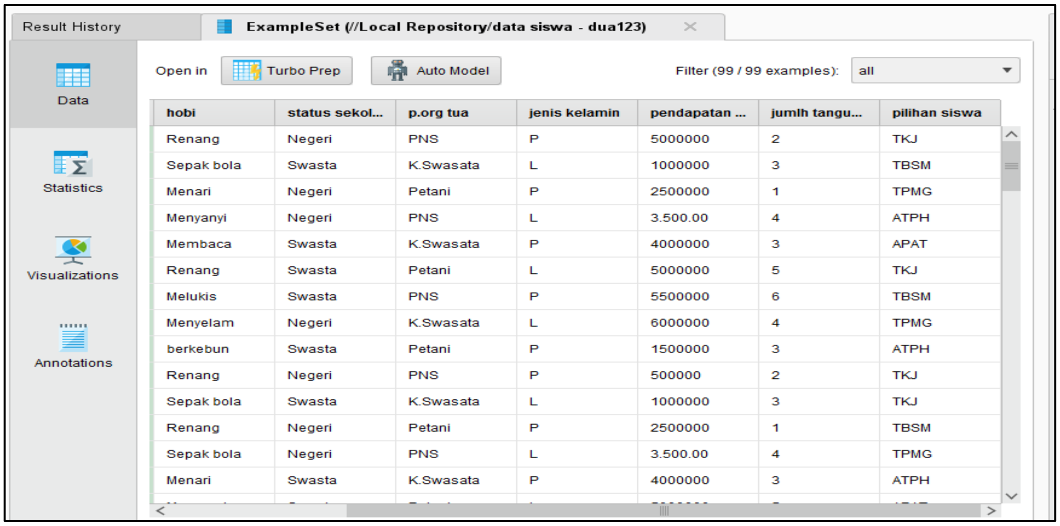


Format your columns.

	hobi *polynominal*	status sekol... *polynominal*	p.org tua *polynominal*	jenis kelamin *polynominal*	pendapata... *polynominal*	jumlh tang... *integer*
1	Renang	Negeri	PNS	P	5000000	2
2	Sepak bola	Swasta	K.Swasata	L	1000000	3
3	Menari	Negeri	Petani	P	2500000	1
4	Menyanyi	Negeri	PNS	L	3.500.00	4
5	Membaca	Swasta	K.Swasata	P	4000000	3
6	Renang	Swasta	Petani	L	5000000	5
7	Melukis	Swasta	PNS	P	5500000	6
8	Menyelam	Negeri	K.Swasata	L	6000000	4
9	berkebun	Swasta	Petani	P	1500000	3
10	Renang	Negeri	PNS	P	500000	2
11	Sepak bola	Swasta	K.Swasata	L	1000000	3
12	Renang	Negeri	Petani	P	2500000	1

Figure 4.3 Data finished importing

4. If it has appeared as shown above eating, click Next, and it will appear where the data is saved; then click Finish eating will appear as figure 4.4



ExampleSet (//Local Repository/data siswa - dua123)

Filter (99 / 99 examples): all

hobi	status sekol...	p.org tua	jenis kelamin	pendapatan ...	jumlh tangu...	pilihan siswa
Renang	Negeri	PNS	P	5000000	2	TKJ
Sepak bola	Swasta	K.Swasata	L	1000000	3	TBSM
Menari	Negeri	Petani	P	2500000	1	TPMG
Menyanyi	Negeri	PNS	L	3.500.00	4	ATPH
Membaca	Swasta	K.Swasata	P	4000000	3	APAT
Renang	Swasta	Petani	L	5000000	5	TKJ
Melukis	Swasta	PNS	P	5500000	6	TBSM
Menyelam	Negeri	K.Swasata	L	6000000	4	TPMG
berkebun	Swasta	Petani	P	1500000	3	ATPH
Renang	Negeri	PNS	P	500000	2	TKJ
Sepak bola	Swasta	K.Swasata	L	1000000	3	TKJ
Renang	Negeri	Petani	P	2500000	1	TBSM
Sepak bola	Negeri	PNS	L	3.500.00	4	TPMG
Menari	Swasta	K.Swasata	P	4000000	3	ATPH

Figure 4.4 data successfully stored in RapidMiner

5. If it appears as shown above, the data has been successfully imported. After the data has been successfully imported, click the desing menu in the application and continue with the drop of student data that has been saved; then use the Cross Validation operator where the data will be divided into two divisions, namely testing data and Trening data automatically based on composers and then connected the data; the following can be seen in figure 4.5

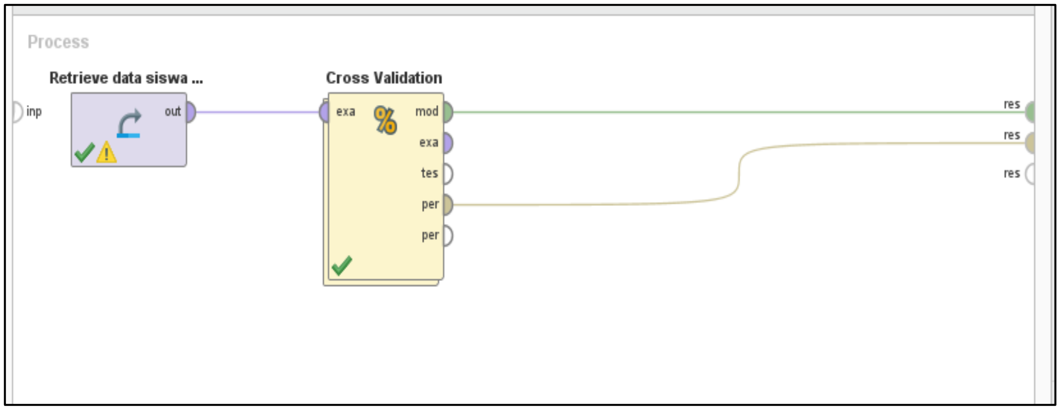


Figure 4.5 drop student data and cross validity

6. If you have finished the process of dropping student data and cross validation, then click validation and select the Algorithm used; where later, there are three algorithms that will be processed eating the first process, namely the KNN Algorithm process for the trending part, Then for testing choose the operator Apply model, then for measurement using the performance operator then connected then in the settings section click accuracy because the result of this process is accuracy can then be seen in the picture Viewed in Figure 4.6

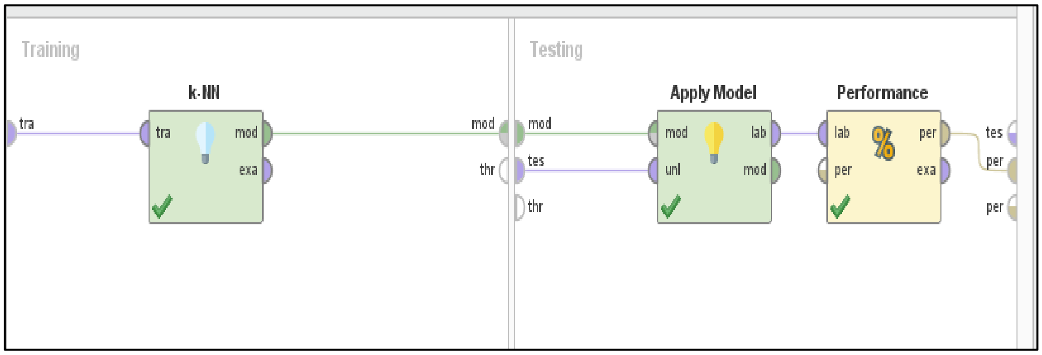
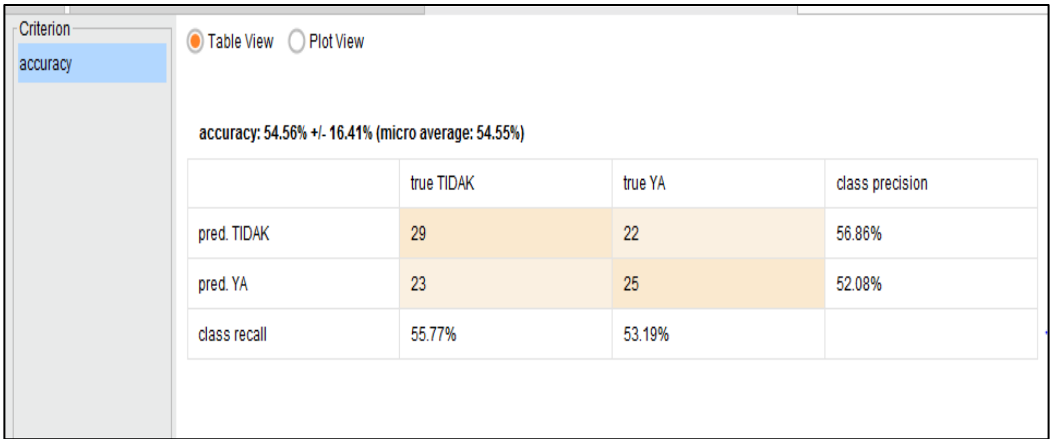


Figure 4.6 KNN algorithm process

7. If the algorithm process has been completed, then check again if it is bernar, then click Run or run, and it will appear as shown in 4.7



Criterion

accuracy

● Table View ○ Plot View

accuracy: 54.56% +/- 16.41% (micro average: 54.55%)

	true TIDAK	true YA	class precision
pred. TIDAK	29	22	56.86%
pred. YA	23	25	52.08%
class recall	55.77%	53.19%	

Figure 4.7 Accuracy Results of the KNN algorithm

8. Because the KNN algorithm process has been completed, the NBC algorithm process, where the initial process is the same as the KNN in the second process, is created again for NBC. Furthermore, it can be seen in figure 4.8

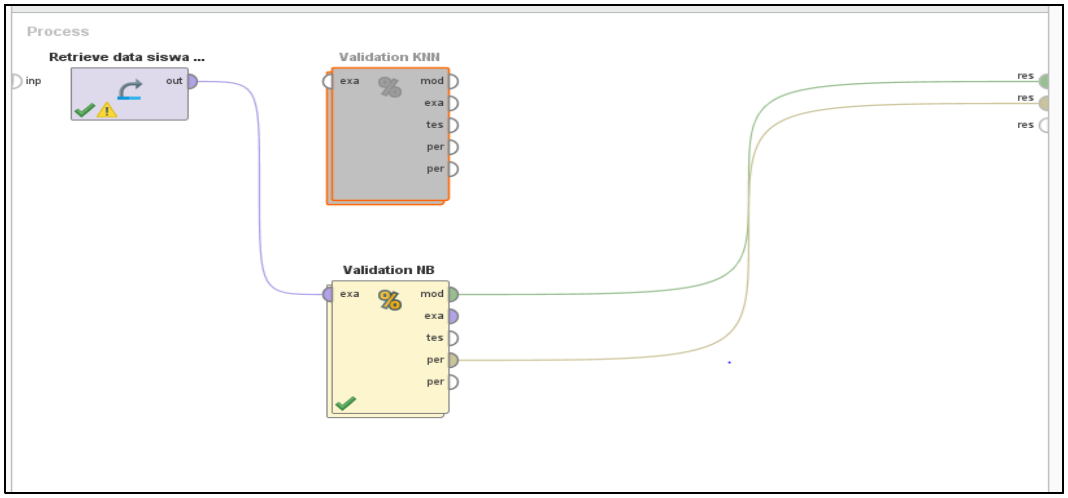


Figure 4.8 validation of the NBC algorithm

9. If the validation process is complete, as shown above, click on NBC validation and for the testing section, select the NBC algorithm, and for testing it remains the same then, it can be seen in figure 4.9

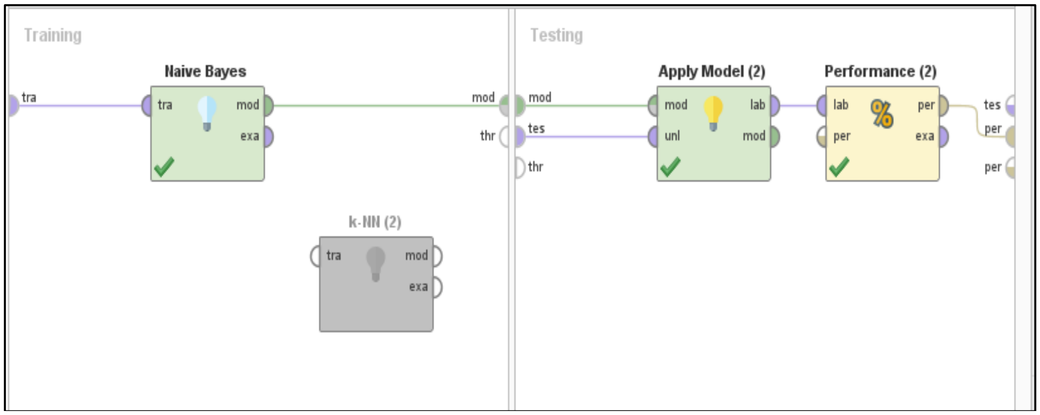
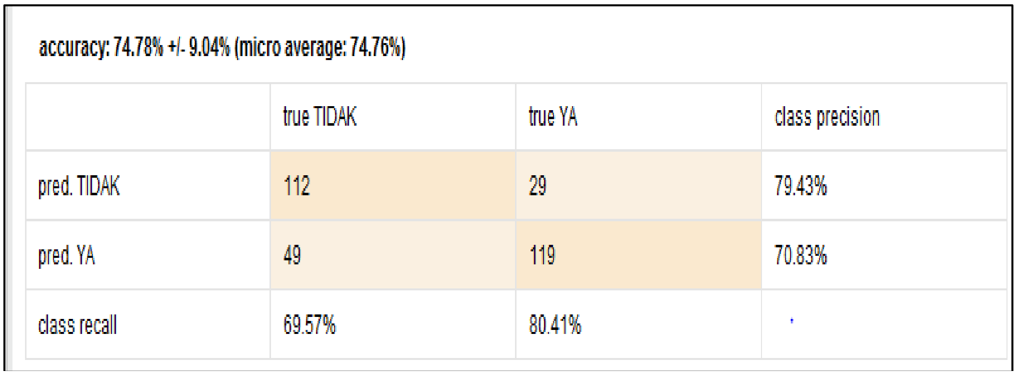


Figure 4.9 NBC testing and trending

10. After checking if all the processes are connected, then run or run it will appear as shown in Figure 4.10



accuracy: 74.78% +/- 9.04% (micro average: 74.76%)

	true TIDAK	true YA	class precision
pred. TIDAK	112	29	79.43%
pred. YA	49	119	70.83%
class recall	69.57%	80.41%	

Figure 4.10 NBC algorithm accuracy results

11. The following is the process of classifying the accuracy of the SVM method; in this part of the method, we get the difference from the NBC and KNN methods because the SVM method can only process data in the form of nominals or numbers, then the nominal to the numerical operator is added, then it can be seen in figure 4.11

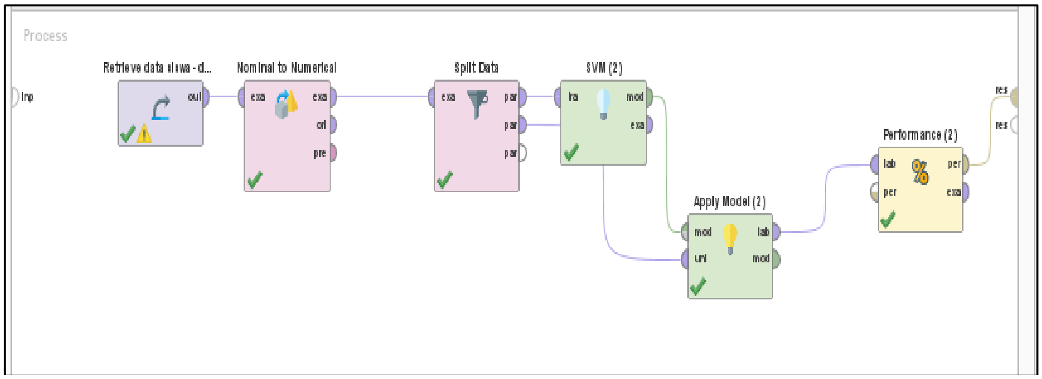
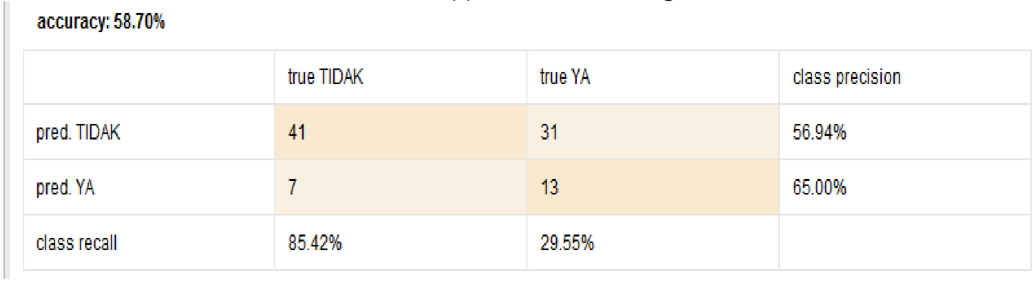


Figure 4.11 of the SVM process

12. If all processes have been connected, then run or run it will appear as shown in figure 4.12



accuracy: 58.70%

	true TIDAK	true YA	class precision
pred. TIDAK	41	31	56.94%
pred. YA	7	13	65.00%
class recall	85.42%	29.55%	

Figure 4.12 SVM method accuracy results

13. The table of accuracy comparison results of SVM, KNN, and NBC methods can be seen in Table 4.1

Table 4.1 accuracy comparison results

Method	Accuracy
NBC	74,78%
KNN	54,56%,
SVM	58,70 %

4. Conclusion

Based on the results of the implementation, the conclusions regarding the comparison of NBC, KNN and SVM methods are:

1. Comparison of NBC, KNN, and SVM methods using student data attributes in SMK N02 Manokwari schools batch 2018-2020 and the amount of data used 100 data can then be run with the rapidminer application.
2. Comparison of NBC, KNN and SVM methods measured using feeding accuracy for the KNN method to get 54.56% accuracy, the NBC method got 74.78% accuracy, and the SVM method got 58.70% accuracy. Then it can be concluded that from the three methods, based on the attributes used by the NBC method, which gets high accuracy, which is 74.78%.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Ardi, I. (2021). Implementasi *Artificial Neural Network* Dalam Memprediksi Jumlah Peserta Les Bahasa Inggris Menggunakan Metode *Back Propagation, Jteki Jurnal Teknik Komputer Dan Informatika 1*(1) 45 – 51. Http://Jteki.Ppj.Unp.Ac.Id

[2] Arifin, S (2019). Analisa Perbandingan Tingkat Performansi Metode

[3] Support Vector Machine Dan *Naïve Bayes Classifier* Untuk Klasifikasi Jalur Minat Sma, *ISSN : 2302-3805.*

[4] Firdaus, (2020). Penerapan Metode *Naïve Bayes* Dalam Prediksi Penentuan Jurusan Mahasiswa TI, *Journal Of Telecommunication, Electronics, And Control Engineering Jtece. 02*(02).82-88

[5] Howay, R (2021). Sistem Rekomendasi Jurusan Pada Sekolah Menengah Kejuruan (SMK) Dengan Algoritma K-Means, *Syntax Idea, 3*(10),V3i10.1443 E-ISSN: 2684-883X Published By: Ridwan Institute, Https://Doi.Org/10.36418/Syntax-Idea.

[6] Hidayanti, T, A. (2021).Perbandingan Dan Analisis Metode Klasifikasi Untuk Menentukan Konsentrasi Jurusan,*Jurnal Ilmiah Informatika Global Volume 11 Issn Online : 2477-3786.*

[7] Mariati, L W (2020). Model Klasifikasi Kepuasan Mahasiswa Teknik Terhadap Sarana Pembelajaran Menggunakan Data Mining, *Jurnal Teknologi Informasi*, E-ISSN 2656-14(2), Https://Doi.Org/10.47111/JTI.

[8] Marisa, (2021). Educational Data Mining (Konsep Dan Penerapan), *Jurnal Teknologi Informasi* 4(2)

[9] Siregar, R (2021).Mengukur Tingkat Kepuasan Mahasiswa Dalam Pembelajaran Dengan Naïve Bayes, E-ISSN : 2798-2580.

[10] Sinaga, R and Dalimunthe, L R (2021). Komparasi Metode Decision Tree, KNN, Dan SVM Untuk Menentukan Jurusan Di SMK, *Jurnal Sistem Komputer Dan Informatika (JSON), 3(2)* e-ISSN 2685-998XDOI 10.30865/Json.V3i2.3598