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

Siska Howay¹ ✉ and Suahirman²
¹²Master 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 students graduate, 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 enroll 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 SVM, 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
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>
<thead>
<tr>
<th>National Exam Score</th>
<th>Hobby</th>
<th>School Status</th>
<th>Parents’ Occupation</th>
<th>Parents’ Income</th>
<th>Number of Shelters</th>
<th>Student options</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>Swimming pool</td>
<td>Country</td>
<td>Civil servants</td>
<td>500.000</td>
<td>2</td>
<td>TKJ</td>
<td>P</td>
</tr>
<tr>
<td>80</td>
<td>Football</td>
<td>Private</td>
<td>private sector employee</td>
<td>1.000.000</td>
<td>3</td>
<td>TBSM</td>
<td>L</td>
</tr>
<tr>
<td>78</td>
<td>Dance</td>
<td>Country</td>
<td>Farmer</td>
<td>2.500.000</td>
<td>1</td>
<td>TPMG</td>
<td>P</td>
</tr>
</tbody>
</table>
### 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.

---

<table>
<thead>
<tr>
<th>No</th>
<th>Hobi</th>
<th>Status, Sekolah</th>
<th>P. Org. Tua</th>
<th>Jenis Kelamin</th>
<th>Pendapatan Orang Tua/Jumlah Tanggungan</th>
<th>Pilihan Siswa</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>Sing</td>
<td>Private</td>
<td>Civil servants</td>
<td>3.500.00</td>
<td>4</td>
<td>ATPH</td>
</tr>
<tr>
<td>75</td>
<td>Read</td>
<td>Private</td>
<td>private sector employee</td>
<td>4.000.000</td>
<td>3</td>
<td>APAT</td>
</tr>
<tr>
<td>85</td>
<td>Swimming pool</td>
<td>Private</td>
<td>Farmer</td>
<td>5.000.000</td>
<td>5</td>
<td>TKJ</td>
</tr>
<tr>
<td>70</td>
<td>Paint</td>
<td>Private</td>
<td>Civil servants</td>
<td>5.500.000</td>
<td>6</td>
<td>TBGM</td>
</tr>
<tr>
<td>80</td>
<td>Dive</td>
<td>Country</td>
<td>private sector employee</td>
<td>6.000.000</td>
<td>4</td>
<td>TPMG</td>
</tr>
<tr>
<td>85</td>
<td>Gardening</td>
<td>Private</td>
<td>Farmer</td>
<td>1.500.000</td>
<td>3</td>
<td>ATPH</td>
</tr>
</tbody>
</table>

*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 Next button contained in the application, it will appear like figure 4.2 then click next:

![Figure 4.2 Data Import Process](image)

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

![Figure 4.3 Data finished importing](image)

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

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 Training 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

![Figure 4.7 Accuracy Results of the KNN algorithm](image)

**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](image)

**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](image)

**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

![Figure 4.10 NBC algorithm accuracy results](image)

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](image)

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

![Figure 4.12 SVM method accuracy results](image)

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBC</td>
<td>74.78%</td>
</tr>
<tr>
<td>KNN</td>
<td>54.56%</td>
</tr>
<tr>
<td>SVM</td>
<td>58.70%</td>
</tr>
</tbody>
</table>
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