Comparative Study of KNN and SVM Methods for Analyzing College Major Consistency Based on High School Background

Anabel Fiorenza Rizkyllah[1], Allsela Meiriza[2] *, Dinna Yunika Hardiyanti[3]
Information System, Computer Science Faculty[1], [2], [3]
Sriwijaya University
Indralaya, South Sumatera, Indonesia
09031182227002@student.unsri.ac.id [1], allsela@unsri.ac.id[2], dinna.yunika@gmail.com[3]

***Abstract*—Selecting a college major that aligns with students' high school background is an essential factor in supporting academic achievement and career preparation. This study focuses on a comparative analysis of the Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) algorithm ms in evaluating the consistency of college major selection. A dataset of 636 students was collected and processed for analysis. Model evaluation was performed using 5-Fold Cross Validation, in which the dataset was repeatedly partitioned into training and testing sets to ensure reliable and unbiased performance assessment. The results suggest that SVM demonstrates higher effectiveness, achieving average scores across precision, recall, F1-score, and accuracy of 85%. Meanwhile, KNN obtained average performance scores of 78%. These findings highlight that SVM provides better performance in analyzing the consistency between students' high school majors and their chosen college majors. These findings also contribute to the development of decision support systems and counseling services to guide students in making more informed major choices.**

Keywords— K-Nearest Neighbor, Support Vector Machine, Data Mining, High School Background

I. INTRODUCTION

Today, education is no longer just a necessity, but a moral and social obligation that must be fulfilled to produce a generation that is ready to face the challenges of the future. Rapid technological advances have changed the way people live, learn, and work, requiring the education sector to prepare individuals who can adapt to these changes. To actively participate in this era, education must not only adapt to technological advances, but also guide students in choosing the right path for self-development.

Quality education requires careful career planning from an early age, including choosing a major that suits one's potential. Determining a major based on interests and abilities helps reduce the risk of mismatch between the desired career field and educational background. Choosing the right major not only determines the learning process during college, but also has a significant impact on future career opportunities because most professions have qualifications that correspond to the college major [1]. Thus, choosing a major is an essential foundation for producing competent individuals who are ready to face global competition and capable of making positive contributions in various sectors of life.

However, in practice, not all students are able to determine a college major that is fully aligned with their academic potential and long-term career objectives. In practice, many students face challenges when choosing a major in college. Some choose based on trends or the popularity of certain majors without considering their suitability in terms of ability, interest, or talent [2]. This situation often leads to the phenomenon of cross-major selection, where graduates from certain programs in high school or vocational school choose to continue their studies in fields that are not directly related to their previous majors. For example, students from science and mathematics programs who then select majors in the social sciences and humanities during the university admission process. This phenomenon can be triggered by various factors, including anxiety and uncertainty that often arise among senior students, which ultimately influence the academic decision-making process [3].

Data mining plays an essential role in processing and analyzing data to uncover previously unseen patterns and insights. This method has been broadly applied in various fields, including education, marketing, engineering, finance, sports, and healthcare [4]. Its rapid development in the education sector shows that Data Mining can provide new insights into learning behavior, academic achievement, and the factors that influence them. This trend has encouraged various studies that utilize Data Mining. Several previous studies have even proven its effectiveness in identifying correlation patterns between educational variables and student learning outcomes. Various algorithmic approaches have also begun to be compared to find the most appropriate method for generating data-based predictions and recommendations.

Through data mining, binary classification is often used to separate data into two mutually exclusive categories based on patterns from historical data. Various algorithms have been developed and used in classification modeling. The K-Nearest Neighbor (KNN) algorithm works by finding the proximity of new data to a number of its closest neighbors. Previous studies

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i4.2521, Copyright ©2025
Submitted: September 15, 2025, Revised: September 23, 2025, Accepted: September 28, 2025, Published: October 15, 2025

have proven the effectiveness of KNN. For example, research on water quality classification shows that KNN is capable of achieving accuracy comparable to Logistic Regression, which is around 89% [5]. Another study utilized KNN for sentiment analysis of online learning from social media, and achieved an accuracy rate of 84.65% [6]. These findings show that KNN can be used flexibly on different data, both in environmental and social contexts.

On the other hand, Support Vector Machine (SVM) has also been widely applied in various fields with promising results. Research on air quality prediction shows that SVM is capable of achieving an accuracy of 83.33% [7]. In addition, research related to the classification of outstanding students based on data mining recorded an accuracy rate of up to 89.80% [8]. This fact confirms that SVM can be relied upon to handle complex data, both in the environmental and academic fields.

Several comparative studies between SVM and KNN also yielded mixed results. Researchers who conducted a study using the enhanced SVM and KNN methods on Human Activity Recognition (HAR) systems reported that KNN achieved an accuracy of 97.08%, slightly higher than SVM with 95.88%, although SVM outperformed KNN in processing speed [9]. Meanwhile, another study that applied both SVM and KNN methods for senior high school selection found that SVM produced higher accuracy, reaching 97.1% compared to KNN's 88.5%, and also showed faster processing time [10].

Based on these findings, it may be inferred that the effectiveness of KNN and SVM is greatly influenced by the characteristics of the data and the context of the problem being analyzed. Therefore, this study focuses on these two algorithms, as KNN is known to be simple yet effective in proximity-based classification, while SVM excels at constructing optimal hyperplanes to produce accurate class separation. These different characteristics make it relevant to analyze both algorithms together for the purpose of obtain a more comprehensive comparison of their classification performance.

Based on previous studies using binary modeling, no specific research has been found that highlights the analysis of consistency between college major choices and majors taken in high school. However, understanding this consistency is important because inconsistencies can affect learning effectiveness, student adaptation to subject matter, and future career planning. In the Indonesian context, studies examining the consistency between high school majors and university major selection are still very limited, even though the problem of students choosing majors that are not in line with their background is often encountered in various universities. Therefore, this study focuses on the application of binary modeling to analyze the consistency between high school background and college majors using relevant data obtained from SMA Negeri 13 Palembang.

This study aims to analyze the consistency between the majors chosen by students in high school and the majors pursued in college, so that the results can be used to evaluate the major selection system, improve educational literacy, support the development of academic and career guidance in schools, and help students build confidence in making decisions in the future.

II. RESEARCH METHODOLOGY

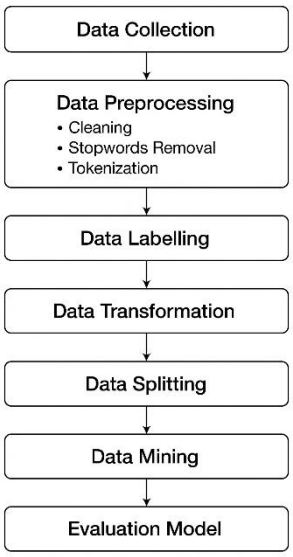

Fig. 1. Research Process Flow

The research process flow is shown in Fig. 1. The research process began with data collection, which involved collecting data using existing data. Data preprocessing includes cleaning, stopword removal, and tokenization to confirm that the data is ready and clean for processing. The data then undergoes data transformation to represent it in a format suitable for analysis, followed by data labeling to assign classes or categories to the data. The prepared data set is then divided into training and test data, and used in the data modeling phase to build a machine learning model. The final stage is model evaluation, which aims to measure the performance of the model using appropriate evaluation metrics to determine its accuracy and effectiveness.

The data in this study were analyzed using Google Colab as a cloud-based computing platform, which provides a Python programming environment with support for various libraries. Pandas was applied for data processing and management, while NumPy facilitated numerical computation. Matplotlib and Seaborn were used for data visualization to present the results more clearly. Scikit-learn, specifically KNeighborsClassifier, SVC, train_test_split, CountVectorizer, accuracy_score, classification_report, cross_val_score, KFold, and confusion_matrix, was employed for model construction and evaluation. In addition, the re and RapidFuzz libraries were utilized for text preprocessing and similarity measurement, thereby ensuring more accurate data representation and analysis.

A. Counseling guidance

Guidance and counseling are forms of assistance provided to students, either individually or in groups, with the aim of fostering independence and optimal development [11]. These

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i4.2521, Copyright ©2025
Submitted: September 15, 2025, Revised: September 23, 2025, Accepted: September 28, 2025, Published: October 15, 2025

services cover multiple domains such as personal, social, academic, and career development, enabling students to grow in a balanced and holistic manner. In particular, guidance and counseling play an important role in supporting students' career planning by helping them explore their interests, understand their abilities, and make informed decisions about future educational and professional pathways. Through a range of structured activities and supportive programs, students are assisted in recognizing their potential, overcoming challenges, and preparing for both academic success and career readiness.

B. The Role of High School Background in College Major Selection

In general, high school education in Indonesia is divided into two main streams, namely science and social sciences. This division aims to facilitate students' interests so that they are better prepared in determining their field of expertise in higher education and in developing specific skills relevant to the demands of the labor market [12]. However, in practice, not all students are able to choose a stream with well-considered judgment. As a result, many university students face various challenges during their higher education, including difficulties in understanding course materials due to a mismatch between the chosen major and their personal interests or abilities [13]. This condition indicates that the selection of a study stream at the high school level, plays an important role in the success of higher education. The misalignment between secondary education background and the major pursued at university not only affects academic achievement but can also influence learning motivation and career planning.

C. Data Collection

Data collection involves the process of gathering relevant and necessary data to answer research questions or solve a problem [14]. This stage is the main foundation, because the quality of the analysis results is highly dependent on the quality of the data collected.

D. Data Pre-processing

Data preparation is a fundamental stage in the data processing process, where unstructured original data is processed before entering the modeling phase. At this stage, various problems commonly found in data, such as missing values, outliers, and unbalanced data, need to be addressed systematically [15]. After all data had been successfully digitized, data checks were performed to address potential duplication, recording errors, or incomplete data. The data preprocessing steps applied in this study are presented as follows [16].

1) Cleaning

Cleaning is the stage of cleaning data by removing symbols, numbers, punctuation marks, and converting letters to lowercase. This stage ensures that the selected and integrated data is suitable for mining, including repairing damaged data, deleting unnecessary data, and standardizing values for consistency [17].

2) Stopwords Removal

Stopword Removal is done by removing common words that do not provide significant meaning in the analysis process.

3) Tokenization

The function of tokenization is to break text into pieces of words (tokens) to make it easier to process.

E. Data Transformation

Data transformation involves converting original data into a more structured and acceptable format to improve its quality and usefulness in analysis and modeling [18]. This stage covers various techniques, such as normalization, standardization, coding, and data aggregation. The goal is to ensure that the data used in the analysis process is comparable in scale, free from inconsistencies, and easily understood by modeling algorithms. With data transformation, hidden patterns in data sets can be more easily identified, model performance can be improved, and analysis results become more accurate and relevant for decision making.

F. Data Labeling

Data labeling tasks play a vital role in every machine learning application [19]. Data labeling is the stage where raw data is marked or categorized so that it can be recognized by machine learning systems. Through this labeling process, data that was originally unstructured becomes clearly meaningful in accordance with the purpose of the analysis. For example, in sentiment analysis research, each sentence can be labeled with the category "positive" or "negative" according to its meaning, while in image recognition, each photo is labeled to indicate the objects in it, such as 'cat' or "dog." With the labeling process, machine learning models can learn the patterns of relationships between input data and predetermined categories, enabling them to classify or predict new data more accurately.

G. Data Splitting

Data splitting is a technique that aims to divide a data set into several parts according to the needs of the analysis. With this separation, the evaluation process can be carried out more objectively so that model performance is not only focused on familiar data, but can also be generalized to new data. In practice, data splitting is often done with various proportions, such as 70:30 or 80:20. In a 70:30 ratio, most of the data is used for model learning, while the rest is used to test the extent to which the model is able to classify data that has not been studied before. Meanwhile, in an 80:20 ratio, a larger portion of the data is used for learning so that the model has a wider opportunity to recognize patterns, even though the portion of data for testing becomes smaller. The selection of this ratio usually considers the amount of data available and the desired level of accuracy, because the more data used for learning, the greater the chance for the model to recognize patterns well [20].

In addition to division by specific proportions, there is also the K-Fold Cross Validation technique, which divides the data into a number of folds. Each fold is used alternately as part of the test data, while the other folds are used for learning. This process is repeated until all folds have had a turn, so that the evaluation results are more comprehensive, representative, and

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i4.2521, Copyright ©2025
Submitted: September 15, 2025, Revised: September 23, 2025, Accepted: September 28, 2025, Published: October 15, 2025

not dependent on a single data division.

H. Data Mining

In this study, Data Mining is understood as an approach to processing large and diverse data sets to generate valuable information [21]. This process does not only rely on statistical calculations, but also utilizes specific algorithms and analysis techniques to identify previously unseen patterns. In this way, data that was initially just a collection of numbers can be interpreted into more meaningful knowledge and used as a basis to support the decision-making process.

In general, the benefits of Data Mining can be seen from its ability to uncover patterns, trends, and hidden relationships in large data sets. Through this process, new information can be obtained and then used to support more targeted decision making and strategic planning. The results are obtained through a combination of conventional data analysis and predictive techniques, thus providing more comprehensive insights [22].

Data mining techniques generally include various approaches such as classification, prediction, and clustering, each of which has a different purpose in processing data. One of the most commonly used techniques is classification. Classification is a method that serves to group data into specific classes by predicting previously unknown data categories [23]. Thus, classification can be understood as the process of placing data objects into predetermined classes based on predefined patterns or models.

1) Classification

Classification represents a data mining process that is used to group data based on specific characteristics [24]. This process aims to place each piece of data into a predetermined category or class so that patterns or relationships within the data can be more easily recognized. Thus, classification serves as an analysis method that helps the system make decisions or predictions based on the characteristics of the data.

In certain contexts, classification can be applied in the form of binary modeling, which is the process of grouping data into two mutually exclusive classes. This approach is often chosen because it provides a clear picture when the research problem focuses only on two main categories, allowing for more focused and simplified analysis.

a. K-Nearest Neighbor

The K-Nearest Neighbor (KNN) method is a classification technique used to determine the class of new data by comparing it with existing data [25]. The basic principle of this method is to calculate the distance between new data and data whose class is already known, then select the data with the closest distance as a reference. The concept of nearest neighbor itself refers to data that has the highest level of similarity or the smallest difference, so that new data can be classified into the appropriate group. In this study, KNN was configured using the parameter n_neighbors=5, which means the algorithm considers the 5 nearest neighbors when determining the class label.

b. Support Vector Machine

Support Vector Machine (SVM) is a machine learning method that focuses on finding the best hyperplane to act as a class boundary [26]. This concept aims to separate data between classes as much as possible, resulting in more accurate classification results. By utilizing high-dimensional feature space, SVM can transform complex data into a more organized form that can be clearly separated. This principle makes SVM effective in solving various classification problems, especially when data patterns are difficult to separate simply. In this study, the SVM algorithm was configured with kernel='linear', allowing the model to construct a linear decision boundary to separate classes. In addition, the parameter random_state=42 was set to ensure result consistency during repeated training.

I. Evaluation Model

The evaluation model stage is a crucial step in the Data Mining process, which aims to assess the accuracy of the developed model. The model evaluation stage generally uses a confusion matrix [27]. At this stage, the performance of classification methods such as Support Vector Machine (SVM) or K-Nearest Neighbor (KNN) is compared using various evaluation metrics, such as accuracy, precision, recall, and f1-score. The use of these metrics allows researchers to understand the strengths and weaknesses of each method in processing data. Therefore, the evaluation stage not only measures the success of the model but also serves as a basis for determining the most appropriate method to use in research.

1) Accuracy

Accuracy shows the ratio between the number of correct predictions and the total available data. This value is used to assess the extent to which the classification model is capable of producing accurate results. The accuracy equation is written as follows.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \times 100 \quad (1)$$

2) Prescision

Precision is the ratio between the number of correct positive predictions and the total data predicted as positive. The precision value is used to measure the accuracy of the model in generating positive predictions. The precision equation is shown as follows.

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (2)$$

3) Recall

Recall is the ratio between the number of correct positive predictions and the total actual data included in the positive class. The recall value indicates the extent to which the model is able to detect relevant data. The recall equation is shown below.

$$Recall = \frac{TP}{TP+FN)} \times 100 \quad (3)$$

4) F1-Score

The F1-score is a measure that combines precision and recall through a weighted harmonic mean. This value is used to

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i4.2521, Copyright ©2025
Submitted: September 15, 2025, Revised: September 23, 2025, Accepted: September 28, 2025, Published: October 15, 2025

provide a more balanced assessment of model performance. The F1-score equation is shown below.

$$F1 - Score = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \quad (4)$$

III. RESULT AND ANALYSIS

The data for this study was obtained from the Guidance and Counseling Department of SMA Negeri 13 Palembang, which stores documents related to students' majors and college choices. Initially available in print, this information was manually inputed into Microsoft Excel for digital processing. The data set used consists of 636 students who graduated from 2019 to 2024.

1) Name

This attribute serves as an identifier. Although not analyzed directly, this name assists in the data validation and tracking process to prevent duplication.

2) Class

This attribute shows the students' educational background while still in school. This information is essential for determining whether there is a correlation between educational background (e.g., science or social studies) and the consistency of college majors after graduation.

3) University

This attribute contains data about the universities that students are applying to. Analysis of this attribute can reveal students' tendencies in choosing certain universities and whether institutional preferences influence consistency in their major choices.

4) Major

This reflects the field of study students choose in higher education. This attribute is a key variable in this study, as the consistency analysis focuses on the correlation between high school majors and the college majors they choose. Table I below presents the original dataset used in this study.

TABLE I. Original data set

NO.	Name	Class	University	Major
1	Hastiliya	XII MIPA 1	Universitas Sriwijaya	Teknologi Hasil Perikanan
2	Isnaini Sahputri	XII MIPA 1	Universitas Sriwijaya	Ilmu Keperawatan
3	Natasya Odelia Indari	XII MIPA 1	Universitas Sriwijaya	FMIPA / Biologi
4.	Rizqia Meilika	XII MIPA 1	Universitas Sriwijaya	Manajemen
…			Universitas Sriwijaya	
634	Sharfina Khoirunnisa	XII MIPA 6	Poltekkes Palembang	Teknologi Pertanian
635	Yvonne Ansya Nabila	XII MIPA 5	Politeknik Sriwijaya	DIV Ilmu Keperawatan
636	Willdan Joddy Alviandra	XII IPS 1	Universitas Sriwijaya	Teknik Mesin

The data obtained is raw data, with variations in spelling, duplication, and additional irrelevant information. Therefore, data cleaning is necessary. The first stage of pre-processing is the cleaning stage. The cleaning stage involves cleaning the data by removing symbols, numbers, punctuation marks, and converting letters to lowercase. Table II below presents an example of text after the cleaning process.

TABLE II. Cleaned data set

Name	Class	University	Major
hastiliya	xii mipa	universitas sriwijaya	teknologi hasil perikanan
isnaini sahputri	xii mipa	universitas sriwijaya	ilmu keperawatan
natasya odelia indari	xii mipa	universitas sriwijaya	fmipa biologi
rizqia meilika	xii mipa	universitas sriwijaya	manajemen
sharfina khoirunnisa	xii mipa	universitas sriwijaya	teknologi pertanian
yvonne ansya nabila	xii mipa	poltekkes palembang	div ilmu keperawatan
willdan joddy alviandra	xii ips	politeknik sriwijaya	teknik mesin

After the data cleaning stage, the next step is to remove stopwords —common words that do not contribute significant meaning to the analysis, such as “and,” “which,” “in,” and “to.” The goal is to leave only necessary words. Table III below presents an example of text after the stopwords removal process.

TABLE III. Dataset after stopwords removal

Name	Class	University	Major
hastiliya	xii mipa	universitas sriwijaya	teknologi hasil perikanan
isnaini sahputri	xii mipa	universitas sriwijaya	ilmu keperawatan
natasya odelia indari	xii mipa	universitas sriwijaya	fmipa biologi
rizqia meilika	xii mipa	universitas sriwijaya	manajemen
sharfina khoirunnisa	xii mipa	universitas sriwijaya	teknologi pertanian
yvonne ansya nabila	xii mipa	poltekkes palembang	div ilmu keperawatan
willdan joddy alviandra	xii ips	politeknik sriwijaya	teknik mesin

After the stopwords removal stage, the next step is tokenization, which aims to break the text into pieces of words (tokens) to make it easier to process. Table IV below presents an example of text after tokenization.

TABLE IV. Data set after tokenization

Name	Class	University	Major
“hastiliya”	“xii”	“universitas”	“teknologi” “hasil”

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i4.2521, Copyright ©2025
Submitted: September 15, 2025, Revised: September 23, 2025, Accepted: September 28, 2025, Published: October 15, 2025

	“mipa”	“sriwijaya”	“perikanan”
“isnaini” “sahputri”	“xii” “mipa”	“universitas” “sriwijaya”	“ilmu” “keperawatan”
“natasya” “odelia” “indari”	“xii” “mipa”	“universitas” “sriwijaya”	“fmipa” “biologi”
“rizqia” “meilika”	“xii” “mipa”	“universitas” “sriwijaya”	“manajemen”
“sharfina” “khoirunnisa”	“xii” “mipa”	“universitas” “sriwijaya”	“teknologi” “pertanian”
“yvonne” “ansya” “nabila”	“xii” “mipa”	“poltekkes” “palembang”	“div” “ilmu” “keperawatan”
“willdan” “joddy” “alviandra”	“xii” “ips”	“politeknik” “sriwijaya”	“teknik” “mesin”

Next is data transformation to ensure consistency and prepare data for analysis. This transformation process includes several essential activities, such as:

1) Class column Standardization

There are only two categories: Mathematics and Natural Sciences (MIPA) and Social Sciences (IPS).

2) Standardization of words/spelling in the central column

Variations in the spelling of major columns have been standardized for consistency.

3) Removal of irrelevant attributes

Remove additional information such as faculty name, education level (DIII/DIV), and city name. Table V below presents an example of text after data transformation.

TABLE V. Data set after transformation

Name	Class	University	Major
hastiliya	mipa	universitas sriwijaya	teknologi hasil perikanan
isnaini sahputri	mipa	universitas sriwijaya	ilmu keperawatan
natasya odelia indari	mipa	universitas sriwijaya	biologi
rizqia meilika	mipa	universitas sriwijaya	manajemen
sharfina khoirunnisa	mipa	universitas sriwijaya	teknologi pertanian
yvonne ansya nabila	mipa	poltekkes palembang	ilmu keperawatan
willdan joddy alviandra	ips	politeknik sriwijaya	teknik mesin

After the Transformation stage, each data item was labeled to ensure consistent selection of majors. This labeling was based on a reference file containing lists of college majors that corresponded to students' high school majors. Table VI below presents a list of college majors by subject group.

TABLE VI. List of College majors based on subject groups

Natural Science	Social Science	Natural Science and Social Science
teknologi hasil perikanan	bahasa Inggris	seni kuliner
ilmu keperawatan	ilmu komunikasi	manajemen konvensi acara
biologi	sosiologi	psikologi
gizi	ilmu hukum	agroekoteknologi
perlindungan tanaman	hubungan internasional	agribisnis

Based on the references in the reference file, each student is categorized as “consistent” if their choice of major is appropriate, and “inconsistent” if it is not. Thus, the data set is ready for the data separation and modeling stage. Below is Fig. 2, which shows the number of consistent and inconsistent students.

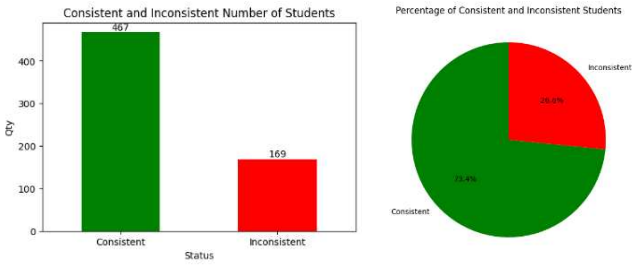

Fig. 2. Number of consistent and inconsistent students diagram

Based on Fig. 2 above, the results of data labeling for a total of 636 students show that 467 students (73.4%) have major choices consistent with their high school majors, while 169 students (26.6%) are classified as inconsistent. For modeling purposes, the labels were converted to numerical values: 1 for the consistent category and 0 for the inconsistent category. This distribution shows that the majority of students chose majors that matched their interests and academic abilities, making the dataset relatively balanced for classification analysis.

At the data-splitting stage, the study used K-fold cross-validation. This technique divides the entire dataset into five folds (5-fold), where in each iteration, one fold is used as test data and the other four as training data. This process is repeated 5 times, so that each fold serves as both training and test data. Thus, the model evaluation becomes more representative because it involves all data in the testing process and can reduce bias that may arise when relying on a single data division.

In the data modeling stage, the K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms are used to classify students as consistent or inconsistent in their major choices. The training data is used to “teach” the model to recognize patterns from existing attributes, while the test data is used to evaluate the model's accuracy and performance. This modeling process involves parameter selection, model training, and initial evaluation to ensure the model can distinguish between consistent and inconsistent categories. The results of

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i4.2521, Copyright ©2025
Submitted: September 15, 2025, Revised: September 23, 2025, Accepted: September 28, 2025, Published: October 15, 2025

the modeling stage will be the basis for comparing the effectiveness of each algorithm in predicting student consistency.

After going through the data modeling stage, the research process continues with an evaluation stage to measure model performance. The evaluation results are shown in Figure 3 below, which shows a comparison diagram of precision, recall, F1-Score, and accuracy values between the KNN and SVM algorithms.

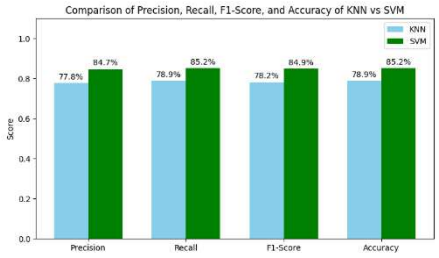

Fig. 3. Comparison of Precision, Recall, F-1 Score, and Accuracy of KNN vs SVM graph

Based on Fig. 3 above, it can be observed that the Support Vector Machine (SVM) method outperforms the K-Nearest Neighbors (KNN) method in terms of precision, recall, F1-score, and accuracy. SVM achieves higher results across all metrics, with precision of 84.7%, recall of 85.2%, F1-score of 84.9%, and accuracy of 85.2%, while KNN shows lower values with precision of 77.8%, recall of 78.9%, F1-score of 78.2%, and accuracy of 78.9%. These results indicate that SVM provides more consistent and accurate predictions, as well as a better balance between precision and recall when compared to KNN. Therefore, the figure highlights that SVM demonstrates superior classification performance in this comparison.

Although the overall results in Figure 3 clearly show the superiority of SVM over KNN, a more detailed evaluation is still needed to understand the specific strengths and weaknesses of each model at the class level. For this reason, the confusion matrix in Figure 4 is presented to illustrate how the KNN model performs in classifying individual classes, including both correct predictions and misclassifications. This analysis provides a deeper view of the types of errors that occur, complementing the general performance metrics shown previously. Furthermore, the section below presents the confusion matrices of both KNN and SVM.

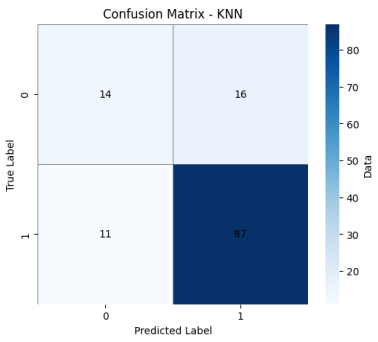

Fig. 4. Confusion Matrix - KNN

The confusion matrix in Figure 4 above shows the results of the KNN model evaluation on the test data. In class 0, 14 data points were predicted correctly, while 16 data points were incorrectly predicted as class 1. For class 1, the model successfully predicted 87 data points correctly, but there were still 11 data points that were incorrectly predicted as class 0. These results indicate that while KNN performs well in identifying class 1, it struggles with correctly classifying class 0, as reflected by the higher number of misclassifications in that category.

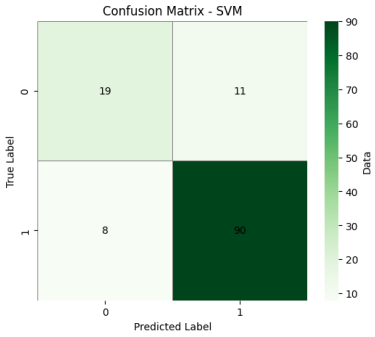

Fig. 5. Confusion Matrix - SVM

The confusion matrix in Figure 5 above shows the results of the SVM model evaluation on the test data. Of the total data, there were 19 class 0 data that were predicted correctly, while 11 class 0 data were incorrectly predicted as class 1. For class 1, the model successfully predicted 90 data correctly, but there were still 8 class 1 data that were incorrectly predicted as class 0. Overall, these results show that the SVM model has quite good performance with a dominant correct prediction rate, although there are still classification errors, especially in class 0.

Based on the confusion matrix comparison, it appears that SVM produces better predictions than KNN. SVM is able to classify data more accurately across both classes, while KNN shows a higher number of misclassifications. This visualization confirms the previous evaluation results that SVM outperforms KNN, both in terms of classification accuracy and model consistency in recognizing data patterns.

IV. CONCLUSION

From the results of this study, the majority of 467 students selected a college major that was consistent with their high school background, while 169 students chose a different field. When analyzed using classification models, the results indicate that SVM shows higher effectiveness than K-Nearest Neighbor (KNN), as it achieved average superior scores in precision, recall, f1-score, and accuracy. This indicates that SVM is more suitable for analyzing the consistency of college major selection based on high school background.

These outcomes are consistent with earlier research, such as the Comparison of K-Nearest Neighbor and Support Vector Machine for Choosing a Senior High School study, which showed that SVM has a higher performance of 97.1% compared

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i4.2521, Copyright ©2025
Submitted: September 15, 2025, Revised: September 23, 2025, Accepted: September 28, 2025, Published: October 15, 2025

to KNN at 88.5%, as well as a faster processing time. This study adds value by deepening the understanding of how machine learning algorithms, particularly SVM, can be applied to examine college major consistency. The findings also provide a foundation for developing decision support systems or enhancing counseling services, enabling schools to guide students in making more informed decisions that align with their potential and academic background.

REFERENCES

[1] A. Nurhartanto and T. D. Wengrum, "Edukasi Pemilihan Jurusan Kuliah Melalui Metode Pemetaan Bakat," *ANDASIH J. Pengabdi. Kpd. Masy.*, vol. 2, no. 1, pp. 33–39, 2021.

[2] A. Rizka, R. E. Putri, Y. Yusman, and M. Fajar, "Sistem Rekomendasi Jurusan Kuliah dalam Pengambilan Keputusan Menggunakan Metode MOORA," *J. Penerapan Sist. Inf. (Komputer Manajemen)*, vol. 4, no. 2, pp. 364–373, 2023.

[3] Listiowatty, "Cemas Dan Ragu Lintas Jurusan Studi Pada Siswa Kelas XII Mipa," *J. Pendidik.*, vol. 22, no. 2, pp. 102–113, 2021.

[4] W. Lastari and J. Jasmir, "Penerapan Data Mining Untuk Memprediksi Prestasi Siswa SMA Pada Dinas Pendidikan Provinsi Jambi," *J. Manaj. Sist. Inf.*, vol. 8, no. 2, pp. 310–321, 2023.

[5] M. S. Denny and D. E. Herwindiati, "Klasifikasi Kualitas Air Menggunakan K-Nearest Neighbors, Naïve Bayes, Dan Logistic Regression," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 6, no. 4, pp. 851–858, 2024.

[6] A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 2, p. 121, 2021.

[7] L. Widyarini and H. D. Purnomo, "Air Quality Prediction Using the Support Vector Machine Algorithm," *J. Inf. Syst. Informatics*, vol. 6, no. 2, pp. 652–661, 2024.

[8] R. Syahri and D. Puspita, "Classification of Outstanding Student using Support Vector Machine (SVM) Based on Data Mining," *JITE (Journal Informatics Telecommun. Eng.*, vol. 7, no. 1, pp. 13–23, 2025.

[9] A. Y. Shdefat, N. Mostafa, Z. Al-Arnaout, Y. Kotb, and S. Alabed, Optimizing HAR Systems: Comparative Analysis of Enhanced SVM and K-NN Classifiers, vol. 17, no. 1. Springer Netherlands, 2024.

[10] M. Irfan, A. R. Nurhidayat, A. Wahana, D. S. Maylawati, and M. A. Ramdhani, "Comparison of K-Nearest Neighbor and support vector machine for choosing senior high school," *J. Phys. Conf. Ser.*, vol. 1280, no. 2, 2019.

[11] Y. A. Batubara, J. Farhanah, M. Hasanahti, and A. Apriani, "Konseling Bagi Peserta Didik," *J. Buatan Alumni Bimbing. dan Konseling Islam (JKA BKI)*, vol. 4, no. 1, p. hlm 3, 2022.

[12] S. Azzahra, V. M. M. Rambitan, D. D. R. Turista, and S. P. Makkadafi, "Pengaruh latar belakang jurusan dan minat belajar terhadap hasil belajar siswa kelas XI pada mata pelajaran biologi di SMA IT Granada Samarinda," *Biocaster J. Kaji. Biol.*, vol. 5, no. 3, pp. 545–558, 2025.

[13] H. Aravik *et al.*, "Sosialisasi Pemilihan Jurusan Kuliah Pada Sekolah SMA di Kota Palembang," *AKM Aksi Kpd. Masy.*, vol. 6, no. 1, pp. 221–228, 2025.

[14] S. A. Mazhar, "Methods of Data Collection: A Fundamental Tool of Research," *J. Integr. Community Heal.*, vol. 10, no. 01, pp. 6–10, 2021.

[15] K. R. Putra, S. Umaroh, N. Fitrianti, and S. Nugraha, "Resultant: Data Preparation Techniques to Improve XGBoost Algorithm Performance," *MIND (Multimedia Artif. Intell. Netw. Database) J.*, vol. 8, no. 1, pp. 42–51, 2023.

[16] A. P. Mareta *et al.*, "Perbandingan Metode Naïve Bayes , Decision Tree , dan KNN Dalam Analisis Sentimen Aplikasi Gojek Di Playstore," vol. 7, no. 2, pp. 725–734, 2025.

[17] R. Rahmi, D. Antoni, H. Syaputra, F. Fatoni, and T. B. Kurniawan, "Metode Klasifikasi Gejala Penyakit Coronavirus Disease 19 (COVID-19) Menggunakan Algoritma Neural Network," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 12, no. 1, pp. 16–23, 2023.

[18] S. Borrohou, R. Fissoune, and H. Badir, "The role of data transformation in modern analytics: A comprehensive survey," *J. Comput. Lang.*, vol. 84, p. 101329, 2025.

[19] S. Zhang, O. Jafari, and P. Nagarkar, "A Survey on Machine Learning Techniques for Auto Labeling of Video, Audio, and Text Data," pp. 1–13, 2021.

[20] I. Muraina, Ideal Dataset Splitting Ratios In Machine Learning Algorithms: General Concerns For Data Scientists And Data Analysts. 2022.

[21] D. H. Khoiriyah and R. Ambarwati, "Dynamic Segmentation Analysis for Expedition Services: Integrating K-Means and Decision Tree," *J. Inf. Syst. Informatics*, vol. 6, no. 1, pp. 363–377, 2024.

[22] Sreejit Ramakrishnan, "The Importance of Data Mining & Predictive Analysis," *Int. J. Eng. Technol. Manag. Sci.*, vol. 7, no. 4, pp. 593–598, 2023.

[23] F. Alghifari and D. Juardi, "Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes," *J. Ilm. Inform.*, vol. 9, no. 02, pp. 75–81, 2021.

[24] A. Jananto, S. Sulastri, E. Nur Wahyudi, and S. Sunardi, "Data Induk Mahasiswa sebagai Prediktor Ketepatan Waktu Lulus Menggunakan Algoritma CART Klasifikasi Data Mining," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 10, no. 1, pp. 71–78, 2021.

[25] N. M. S. M. Maylita, H. Z. Zahro, and N. Vendyansyah, "Penerapan Metode K-Nearest Neighbor (KNN) untuk menentukan status gizi balita (Studi Kasus : Posyandu Ananda Kelurahan Langkai, Kota Palangka Raya, Kalimantan Tengah)," vol. 6, no. 2, pp. 1–5, 2022.

[26] A. Desiani *et al.*, "Penerapan Metode Support Vector Machine Dalam Klasifikasi Bunga Iris," *Indones. J. Appl. Informatics*, vol. 7, no. 1, p. 12, 2023.

[27] Y. Ansori and K. F. H. Holle, "Perbandingan Metode Machine Learning dalam Analisis Sentimen Twitter," *J. Sist. dan Teknol. Inf.*, vol. 10, no. 4, p. 429, 2022.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i4.2521, Copyright ©2025
Submitted: September 15, 2025, Revised: September 23, 2025, Accepted: September 28, 2025, Published: October 15, 2025