Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. Sec. 5. No. 1; e-ISSN: 2808-4519

Application of the K-Nearest Neighbor Algorithm to Analyze the Learning Ability of Grade VII Students in English Subjects

Dewi Ratna Sairo [1]*, Fajar Hariadi [2], Raynesta Mikaela Indri Malo [3]

[1,2,3] *Informatics Engineering Study Program, Wira Wacana Sumba Christian University,Indonesia*
dewiratnazairo@gmail.com[1*], fajar@unkriswina.ac.id[2],raynesta@unkriswina.ac.id[3]

Abstract

SMP Negeri 2 Pahungan Lodu currently lacks a system for classifying students based on their English language learning abilities. As a result, the learning process remains generalized, failing to account for the varying levels of student understanding. This situation poses challenges for teachers in adapting their instructional methods, leading to suboptimal academic interventions for students experiencing learning difficulties. To address these challenges, this study employs the K-Nearest Neighbor (KNN) algorithm as a classification method to categorize students into three groups: Able, Quite Able, and Underable. These categories are determined based on academic data, including assignment scores, practice assessments, midterm (UTS) and final exam (UAS) scores, report card grades, and student attendance levels. This research utilizes a data mining approach with the KNN algorithm, which operates by calculating the Euclidean distance between student data points and assigning categories based on the nearest neighbors. The dataset used in this study comprised 63 students after undergoing data cleaning. Subsequently, the data was divided into 50 training samples and 13 test samples. The results indicate that the KNN algorithm successfully classifies the test data with an accuracy rate of 84.62%. These findings demonstrate that the KNN algorithm is an effective tool for academic decision-making and for developing learning strategies tailored to students' ability levels.

Keywords: *K-Nearest Neighbor, Data Mining, Classification, English, SMP N 2 Pahunga Lodu, Class VII.*

1. Introduction

Students' academic ability is often the main benchmark in assessing the success of education in school. Students' academic grades are generally obtained from various evaluation components, such as assignment scores, practice, midterm exams, end-of-semester exams and report card scores. In addition, the level of student attendance or attendance in the classroom also plays an important role in determining student success. Each component of academic grades provides a different picture of a student's learning ability. The grades of assignments and practices, for example, reflect the extent to which students are able to apply the concepts they have learned in real situations. Meanwhile, the Mid-Semester Exam (UTS) and the Final Semester Exam (UAS) provide an evaluation of students' ability to understand the material on a wider scale and in more formal exam conditions. The report card, as a final recapitulation, summarizes all the components of the evaluation and provides an overall picture of the student's academic ability during one semester.

SMP Negeri 2 Pahunga Lodu is one of the educational institutions in East Sumba that does not have a system of classifying students based on the level of learning ability. The absence of this system has an impact on the learning process that is still general without considering the differences in individual students' understanding. As a result, teachers have difficulties in adjusting teaching methods, so the provision of materials and guidance is not optimal. This causes students who have learning difficulties to be at risk of falling behind, while more capable students do not get the appropriate challenges to develop their skills to the fullest.

Students' learning ability to understand subjects, especially English, is still a challenge in various schools, including SMP N 2 Pahunga Lodu. The learning process that is carried out in a general way without considering the differences in students' abilities causes ineffectiveness in the delivery of material, as well as makes it difficult for teachers to carry out academic interventions in a targeted manner. Thus, the KNN algorithm can help in classifying students based on their learning patterns and abilities, making it easier for educators to provide more targeted interventions.

Therefore, this study aims to analyze the learning ability of grade VII students in English subjects by focusing on data on assignment scores, practices, UTS, UAS, report card scores, and student attendance levels. By utilizing the KNN algorithm, the results of this analysis will make a practical contribution by grouping students into learning ability categories (Able, Moderate, and Underable), which allows teachers to adjust teaching methods according to individual needs. Schools can use the results of this research to improve the quality of education through student grouping, academic intervention development, and evaluation. Previous research, such as those conducted by [1], showing that KNN is effective in improving student understanding in online learning, while [2], proving that KNN can help teachers

design more effective learning strategies based on student performance analysis. Thus, the results of this study are expected to improve the quality of education at SMP N 2 Pahunga Lodu.

2. Theoretical Foundations

2.1. Learning Ability

Student learning ability is an important aspect in the world of education that reflects the individual's ability to understand, master, and apply knowledge or skills acquired during the learning process. This process is influenced by various indicators such as assignment scores, practice, mid-semester exams (UTS), final semester exams (UAS), report card scores, and student attendance levels. Assignment and practice scores provide an overview of students' ability to apply theory to real-world situations, while UTS and UAS measure students' mastery of material in formal situations. Student attendance or attendance rate is a significant indicator that indicates student involvement in learning, where a good attendance rate usually reflects motivation and consistency [3]. The factors that affect students' learning abilities are divided into internal and external. Internal factors include learning motivation, level of interest in the subject matter, and students' physical condition, such as health and concentration. Meanwhile, external factors include support from family, teaching methods applied by teachers, and a conducive learning environment. A supportive learning environment both at school and at home is able to increase students' engagement and understanding, enabling them to achieve more optimal outcomes. With the right analysis of learning ability, educators can provide a more personalized and strategic approach to help students reach their full potential[4].

2.2. Data Mining

Data mining is the process of extracting information from large data sets to find useful patterns, trends, or relationships. In the context of education, data mining is often used to analyze student learning outcomes by processing data such as assignment scores, exams, attendance, and demographics. This process involves several main stages, including *pre-processing* (cleaning and preparing data), pattern exploration (using algorithms such as classification and *clustering*), and evaluation (interpreting the results of the analysis). With this approach, educational institutions can identify the needs of students more specifically, such as providing additional support to those who have difficulty taking lessons

The factors that affect students' learning abilities are divided into internal and external. Internal factors include learning motivation, level of interest in the subject matter, and students' physical condition, such as health and concentration. Meanwhile, external factors include support from family, teaching methods applied by teachers, and a conducive learning environment. A supportive learning environment both at school and at home is able to increase students' engagement and understanding, enabling them to achieve more optimal outcomes. With the right analysis of learning ability, educators can provide a more personalized and strategic approach to help students reach their full potential[4].

2.3. K-Nearest Neighbor(KNN)

The K-Nearst Neighbor (KNN) algorithm is one of the simplest and effective methods of classification and regression in data analysis. The main principle of this algorithm is to find the nearest neighbors. K-Nearest Neighbor (KNN) works by comparing new data with a number of nearby data already in the dataset. Distance metrics such as Euclidean, Manhattan, or Minkowski are commonly used to measure the proximity between data. These algorithms are non-parametric, which means they don't make any assumptions about the distribution of data. As a result, K-Nearest Neighbor (KNN) has become flexible in addressing different types of datasets, including non-linear ones [5].
The formula for K-Nearest Neighbor (KNN) is as follows:

$$d = (p, q)\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

Information:

d(p, q)	: The distance between two points p and q of space with an n dimension.
p_i	: Value of feature i from point p (first data)
q_i	: Feature value i from point q (second data)
n	: The number of features or dimensions in the data
(p_i-q_i)	: The difference between the value of the feature of two points, then squared to avoid negative values
∑i=1	: Sum of all quadrant differences and features

2.4. Classification

Classification is one of the techniques in data mining that is used to group data into categories based on patterns that have been learned from historical data. Classification aims to develop a knowledge model based on similarities or differences in data groups [6].
Confusion matrix is a method of evaluating the performance of a classification algorithm, especially in supervised learning problems. This matrix is used to compare the results of the classification performed by the system with the actual labels of the data. The confusion matrix not only shows how accurate the model is, but also provides detailed information about the types of classification errors made [7].
The formula in the Classification is as follows:

$$\text{Akurasi}\,\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}100\% \quad (2)$$

Information:
HCMC : Correct prediction for positive class
TN : Correct prediction for negative class
FP : Wrong predictions are categorized as positive even though they are negative
FN : Wrong predictions are categorized as negative even though they are positive

3. Research Methods

The framework of thinking that will be described is the stages that will be carried out, namely in figure 1
.

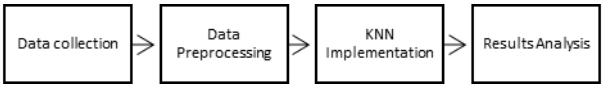

Fig. 1: Research Flow

3.1. Data Collection

Data collection has been carried out at SMP N 2 Pahunga Lodu related to the data on the scores of grade VII students in English subjects in the form of excel format. The research used score data in the 2024/2025 academic year semester 1 with student data of 66 students consisting of classes A and B. Data processing was carried out with a prepared dataset.

3.2. Preprocessing Data

Data preprocessing is very important because after collecting data, it is necessary to double-check the data that has been collected and data processing is carried out. The following is the raw data that has been prepared along with the script.

```
from google.colab import files
uploaded = files.upload()
```

Fig. 2: Script Upload File

	Nama	Presensi	Tugas	Praktek	UTS	UAS	Nilai raport	ANALISIS
0	Adi Putra Hamana Ndakulinya	100	85	100	43.75	50.0	70	Cukup Mampu
1	Alfian Umbu Nggau Behar	82	70	100	35.00	44.0	60	Kurang Mampu
2	Alfred Umbu Ndama Yilu	91	60	100	25.00	52.0	60	Kurang Mampu
3	Alief Prananda	82	65	100	25.00	54.0	62	Kurang Mampu
4	Alwi Rohi	91	100	100	100.00	99.0	98	Mampu
...	...	...	...	...	...	...	...	...
61	Valentino Markus Rohi	100	80	100	68.75	45.0	74	Cukup Mampu
62	Vanesia Aprilia Randa	100	80	100	100.00	80.0	89	Mampu
63	Veronika Triliska Dendo	100	80	100	75.00	70.0	76	Cukup Mampu
64	Vinsensius Marvel Umbu Hina	100	60	100	68.75	20.0	60	Kurang Mampu
65	Widya Rambu Tamu	100	85	100	100.00	100.0	96	Mampu

Fig. 3: Dataset Views

Based on the data in Figure 2 obtained from the results of data collection, the variables used in this study include Assignment scores, Practice, Final Semester Exams (UTS), Final Semester Exams (UAS), Report Card Scores, and Attendance Levels. Next, the Drop Missing values process is performed to check the empty rows in the table. For the script below:

```
[ ] # Cek nilai kosong
    print("Jumlah nilai kosong per kolom:")
    print(df.isnull().sum())

    # Hapus baris yang memiliki nilai kosong
    df_clean = df.dropna()
```

```
Jumlah nilai kosong per kolom:
Nama            0
Presensi        0
Tugas           0
Praktek         0
UTS             3
UAS             1
Nilai raport    0
ANALISIS        0
dtype: int64
```

Fig. 4: Script Drop Missing Values

After checking the data, the next process is to clean the data on the empty table because it does not meet the criteria for analysis. The results of this process are shown in Figure 5.
The data preprocessing script is as follows:

```
# Menghapus baris yang memiliki nilai kosong (NaN)
df_clean = df.dropna()

# Tampilkan data yang sudah dibersihkan saja, bukan seluruh data
print("Jumlah data setelah dibersihkan:", len(df_clean))
display(df_clean.reset_index(drop=True))
```

```
Jumlah data setelah dibersihkan: 63
```

Fig. 5: Script Displays Cleaned Data

	Nama	Presensi	Tugas	Praktek	UTS	UAS	Nilai raport	ANALISIS
0	Adi Putra Hamana Ndakulinya	100	85	100	43.75	50.0	70	Cukup Mampu
1	Alfian Umbu Nggau Behar	82	70	100	35.0	44.0	60	Kurang Mampu
2	Alfred Umbu Ndama Yilu	91	60	100	25.0	52.0	60	Kurang Mampu
3	Alief Prananda	82	65	100	25.0	54.0	62	Kurang Mampu
4	Alwi Rohi	91	100	100	100.0	99.0	98	Mampu
5	Ambu Utang Jua	100	95	100	75.0	40.0	73	Cukup Mampu
...	...	...	...	...	...	...	...	...
59	Umbu Rendra	55	35	100	40.0	57.0	60	Kurang Mampu
60	Umbu Saputra Wijaya	82	80	100	75.0	40.0	74	Cukup Mampu
61	Valentino Markus Rohi	100	80	100	68.75	45.0	74	Cukup Mampu
62	Vanesia Aprilia Randa	100	80	100	100.0	80.0	89	Mampu

Fig. 6: View cleaned data

Based on the table above, out of a total of 65 available student data, only 62 data were used after the data cleaning process was carried out. A total of 3 student data were deleted because they had incomplete values (missing values) in one or more columns. The remaining data is then used for the implementation process of the K-Nearest Neighbor (KNN) algorithm.

3.3. Implementation of KNN

The implementation of KNN (K-Nearest Neighbor) is the process of applying the KNN algorithm to classify or regress data. KNN works to measure the distance between the Train data and the test data using the Euclidean Distance method.
Here is the script:

```
from sklearn.model_selection import train_test_split

# Membagi data menjadi 80% latih dan 20% uji
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Tampilkan jumlah data latih dan data uji
print("Jumlah data latih:", len(X_train))
print("Jumlah data uji  :", len(X_test))
```

```
Jumlah data latih: 50
Jumlah data uji  : 13
```

Fig. 7: Splitting Test Data Training Data

Furthermore, this study will use the Euclidean Distance formula to calculate the distance between test data and training data. A total of 13 test data will be compared with 50 training data to determine the category of each test data.

Here is the script:

```
import numpy as np
from collections import Counter

def hitung_jarak_euclidean(x1, x2):
    return np.sqrt(np.sum((x1 - x2)**2))
# Jumlah tetangga terdekat
k = 3
y_pred = []
# Loop setiap data uji
for i in range(len(X_test)):
    distances = []
    for j in range(len(X_train)):
        dist = hitung_jarak_euclidean(X_test.iloc[i].values, X_train.iloc[j].values)
        distances.append((dist, y_train.iloc[j], j))  # Simpan juga index
    # Urutkan berdasarkan jarak terdekat
    distances.sort(key=lambda x: x[0])
    # Ambil k tetangga terdekat
    k_neighbors = distances[:k]
    print(f"\n🔍 Data Uji ke-{i+1} (Label Asli: {y_test.iloc[i]})")
    print("3 Tetangga Terdekat:")
    for idx, (jarak, label, train_index) in enumerate(k_neighbors, start=1):
        print(f"  {idx}. Index Latih: {train_index} | Label: {label} | Jarak: {jarak:.4f}")
    # Tentukan label prediksi berdasarkan mayoritas
    labels = [label for _, label, _ in k_neighbors]
    most_common = Counter(labels).most_common(1)[0][0]
    y_pred.append(most_common)
```

Fig. 8: Script Calculating Training Data and Test Data

	Nama	Kategori
0	Vinsensius Marvel Umbu Hina	Kurang Mampu
1	Umbu Saputra Wijaya	Cukup Mampu
2	Adi Putra Hamana Ndakulinya	Cukup Mampu
3	Renaldi Kahumbu Nggiku	Kurang Mampu
4	Ambu Utang Jua	Cukup Mampu
5	Melantonia Hunggu Djawa	Kurang Mampu
6	Desilva Rambu Loda	Cukup Mampu
7	Ayunda Rambu Ula	Kurang Mampu
8	Jezicka Auryn Rambu Hamu	Cukup Mampu
9	Veronika Triliska Dendo	Cukup Mampu
10	Umbu Lapu Kilingoru	Kurang Mampu
11	Ardian Saputra Wila Lay	Kurang Mampu
12	Rambu Ayu May Nggiri	Mampu

Fig. 9: Displaying Test Data Results

The figure above shows the prediction of 13 test data that has been analyzed using the KNN model that was previously trained with 50 data.

3.4. Classification

Furthermore, this study will make a prediction with the actual results and the transcript.

```python
# Ambil indeks data uji
test_indices = X_test.index

# Buat DataFrame perbandingan hasil
hasil_perbandingan = pd.DataFrame({
    'Nama': df_clean.loc[test_indices, 'Nama'].values,
    'Kategori Asli': y_test.values,
    'Hasil Prediksi': y_pred
})

# Tampilkan tabel hasil
print("Perbandingan Prediksi KNN dengan Kategori Asli:")
display(hasil_perbandingan)
```

Fig. 10: Original Label Prediction Script and KNN Prediction Results

```
Perbandingan Prediksi KNN dengan Kategori Asli:
```

	Nama	Kategori Asli	Hasil Prediksi
0	Vinsensius Marvel Umbu Hina	Kurang Mampu	Kurang Mampu
1	Umbu Saputra Wijaya	Cukup Mampu	Cukup Mampu
2	Adi Putra Hamana Ndakulinya	Cukup Mampu	Kurang Mampu
3	Renaldi Kahumbu Nggiku	Kurang Mampu	Kurang Mampu
4	Ambu Utang Jua	Cukup Mampu	Cukup Mampu
5	Melantonia Hunggu Djawa	Cukup Mampu	Cukup Mampu
6	Desilva Rambu Loda	Kurang Mampu	Cukup Mampu
7	Ayunda Rambu Ula	Kurang Mampu	Kurang Mampu
8	Jezicka Auryn Rambu Hamu	Cukup Mampu	Cukup Mampu
9	Veronika Triliska Dendo	Cukup Mampu	Cukup Mampu
10	Umbu Lapu Kilinggoru	Kurang Mampu	Kurang Mampu
11	Ardian Saputra Wila Lay	Kurang Mampu	Kurang Mampu
12	Rambu Ayu May Nggiri	Mampu	Mampu

Fig. 11: Displaying Original Labels and Predictions

The data shown above is a combination of the actual classification results and the prediction results obtained from the KNN algorithm. The actual data shows the original category of each test data, while the prediction data is the result of model classification based on the comparison of the distance between the training data, the comparison between these two data is used to measure how accurate the KNN model is in predicting the test data.

```python
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay

# Hitung confusion matrix
cm = confusion_matrix(y_test, y_pred, labels=np.unique(y_test))

# Tampilkan confusion matrix
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=np.unique(y_test))
disp.plot(cmap='Blues')

# Hitung akurasi
akurasi = accuracy_score(y_test, y_pred)

# Tampilkan hasil akurasi dalam persen
print(f"Akurasi Model KNN: {akurasi * 100:.2f}%")
```

Fig. 12: Script Confusion Matrix

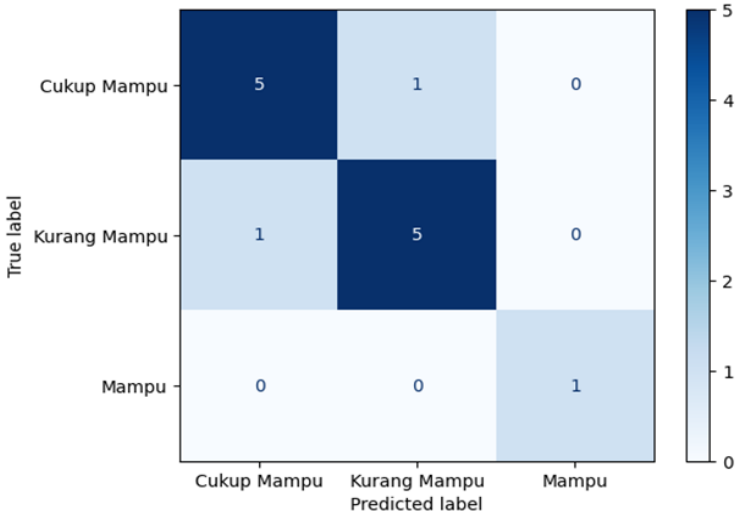

Fig. 13: Confusion Matrix

At the model evaluation stage using a confusion matrix, an analysis was carried out on the compatibility between the label predicted by the KNN algorithm and the true label. Based on the confusion matrix above, it can be explained that the model managed to correctly classify 5 'Quite Able' data, 5 'Underable' data, and 1 'Able' data correctly. However, there was a misclassification (false pediction) as many as 1 data 'Quite Able' which was predicted as 'Underable', and 1 'Underable' data predicted as 'Quite Able'. No classification was found in the 'Able' class to other classes, other than one data that was properly classified. Thus, cunfusion matrix in provides a visual overview of the performance of the KNN model in classifying test data into three predetermined categories, namely 'Quite Able', 'Underable', and 'Able'.

$$\text{Akurasi} = \frac{5+5+1}{13} = \frac{11}{13} \times 100 = 84{,}62\%$$

3.5. Results Analysis

From the classification results, an accuracy of 84.62% was obtained, which shows that the model has a moderate level of accuracy in classifying students based on their learning ability.

4. Conclusion

This study succeeded in classifying the learning ability of grade VII students in English subjects into three categories, namely Able, Quite Able, and Less Capable by applying the K-Nearest Neighbor (KNN) algorithm. Of the 63 valid student data, 13 were used as test data, and the classification results with a score of K=3 showed an accuracy of 84.62%. Most students are in the category of Quite Capable and Underprivileged. These findings show that the KNN algorithm can be used as an effective tool for educators in analyzing students' abilities and designing more appropriate and targeted learning strategies, so that teachers can adjust learning methods and approaches according to each student's abilities.

Acknowledgement

I would like to express my deepest gratitude to Supervisor I and Supervisor II for the guidance, direction, and support that has been provided during the process of preparing and completing this research.

References

[1] M. Raschintasofi, N. Khumairo, E. Rasywir, and A. Feranika, "Analysis of the Level of Understanding of Students of Universitas Dinamika Bangsa in Online Learning Using the K-Nearst Neighbor Algorithm," *J. Manaj. Technology. Then Sist. Inf.*, vol. 2, no. 1, pp. 69–77, 2022, doi: 10.33998/jms.2022.2.1.29.

[2] I. M. R. P. Dhita and G. A. V. M. Giri, "Implementation of KNN Algorithm to Predict School Student Performance," *J. Nas. Technology. Inf. and App.*, vol. 1, no. 3, pp. 819–826, 2023.

[3] M. R. Aziz and W. S. Devi, "Building the Language Pillar: An Effective Method of Expanding English Vocabulary in Junior High School Students," in *Proceedings of the 2024 FIP UMJ National Seminar and Scientific Publications*, F. I. Education, Ed., Jakarta, 2024, pp. 2349–2353.

[4] D. Saputra, W. Wargadinata, and S. Fikri, "Literature Study on Factors Affecting Student Learning Outcomes," *Pendas J. Ilm. Educators. basis*, vol. 10, no. 2, pp. 399–408, 2025.

[5] A. A. Akhbar and H. Dwi, "Analysis of the K-Nearest Neighbor Method Using Rapid Miner to Predict Rain in Surakarta City," in *Proceedings of the National Seminar on Information Technology and Business*, 2023, pp. 1–5. doi: 10.47701/senatib.v3i1.2996.

[6] T. Novika, P. Poningsih, H. Okprana, A. P. Windarto, and H. Siahaan, "The Application of Data Mining for the Classification of Students' Comprehension Level in Mathematics Lessons," *J. Media Inform. Budidarma*, vol. 5, no. 1, p. 9, 2021, doi: 10.30865/mib.v5i1.2498.

[7] G. Rininda, I. Hartami Santi, and S. Kirom, "Application of Svm in Sentiment Analysis on Edlink Using Confusion Matrix Testing," *JATI (Journal of Mhs. Tek. Inform.*, vol. 7, no. 5, pp. 3335–3342, 2024, doi: 10.36040/jati.v7i5.7420.