P-ISSN: 2656-1743 | E-ISSN: 2656-1735 **JURNAL RISET INFORMATIKA**
DOI: https://doi.org/10.34288/jri.v8i3.520 Vol. 8, No. 3. Juni 2026
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

A COMPARATIVE STUDY OF DISTANCE METRICS AND NEIGHBOR SELECTION IN K-NEAREST NEIGHBOR FOR VOCATIONAL STUDENT PERFORMANCE CLASSIFICATION

Ida Wahyuningrum[-1], Nita Novita[-2], Denny Alfian[-3], Muhammad Aris Ganiardi[-4]

[1,2,3,4]Department of Informatics Management
Politeknik Negeri Sriwijaya
ida_wahyuningrum@yahoo.com[-1], nitanovita_polsri@yahoo.com[-2], denny_alfian_mi@polsri.ac.id[-3],
marisg2010@gmail.com[-4]

Abstract

This study aims to evaluate parameter sensitivity in the K-Nearest Neighbor (KNN) algorithm, particularly the selection of distance metrics and k-values, for classifying academic performance in vocational education with heterogeneous and imbalanced data characteristics. The dataset consists of 750 first-year students from the Informatics Management program, including academic attributes (GPA, attendance, and core course grades) and demographic attributes (age, gender, educational background, and economic status). Data preprocessing involves data cleaning, one-hot encoding, Z-score normalization, and handling class imbalance using SMOTE. Model evaluation is conducted using K-Fold Cross Validation with accuracy, precision, recall, and macro-average F1-score as performance metrics. The results show that KNN performance is highly influenced by the combination of distance metrics and k-values. All metrics achieve accuracy above 84%, but differ in handling class imbalance. The Chebyshev metric (k = 10) provides the best balance with an F1-score of 0.6468, while the Minkowski metric (p = 3) achieves the highest recall of 0.7334. The Euclidean metric attains the highest accuracy of 0.8504 (k = 11), but tends to be biased toward the majority class. These findings indicate that optimizing KNN parameters should not rely solely on accuracy, but also consider balanced performance across classes. This study provides a practical evaluation framework for selecting KNN parameters to support more robust and fair academic prediction systems in vocational education data.

Keywords : KNN; Distance Metrics; Student Performance; Vocational Education; Classification

Abstrak

Penelitian ini bertujuan untuk mengevaluasi sensitivitas parameter pada algoritma K-Nearest Neighbor (KNN), khususnya pemilihan metrik jarak dan nilai k, dalam klasifikasi prestasi akademik mahasiswa pendidikan vokasi dengan karakteristik data heterogen dan tidak seimbang. Dataset terdiri dari 750 mahasiswa tahun pertama Program Studi Manajemen Informatika, mencakup atribut akademik (IPK, kehadiran, nilai mata kuliah inti) dan demografis (usia, jenis kelamin, latar belakang pendidikan, kondisi ekonomi). Tahapan prapemrosesan meliputi pembersihan data, one-hot encoding, normalisasi Z-score, serta penanganan ketidakseimbangan kelas menggunakan SMOTE. Evaluasi model dilakukan dengan K-Fold Cross Validation menggunakan metrik akurasi, presisi, recall, dan F1-score (macro average). Hasil menunjukkan bahwa performa KNN sangat dipengaruhi oleh kombinasi metrik jarak dan nilai k. Seluruh metrik menghasilkan akurasi di atas 84%, namun berbeda dalam menangani ketidakseimbangan kelas. Metrik Chebyshev (k = 10) memberikan keseimbangan terbaik dengan F1-score 0,6468, sedangkan Minkowski (p = 3) memiliki recall tertinggi sebesar 0,7334. Euclidean mencapai akurasi tertinggi sebesar 0,8504 (k = 11), namun cenderung bias terhadap kelas mayoritas. Temuan ini menegaskan bahwa optimasi parameter KNN tidak cukup hanya berdasarkan akurasi, tetapi juga harus mempertimbangkan keseimbangan performa antar kelas. Penelitian ini memberikan kerangka evaluasi praktis dalam pemilihan parameter KNN untuk mendukung sistem prediksi akademik yang lebih adil dan adaptif pada data pendidikan vokasi.

Kata Kunci : KNN; metrik jarak; kinerja mahasiswa; pendidikan vokasi; klasifikasi

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.520
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

INTRODUCTION

Vocational education in higher education institutions possesses unique characteristics compared to general academic education, as it emphasizes mastery of practical skills relevant to industry and workforce needs. Learning processes in vocational education are predominantly conducted in laboratories, workshops, and real-world work environments rather than traditional classrooms (Khamdun et al., 2021),(Ali & Koehler, 2020). This condition renders student performance assessment more complex, as it requires the evaluation of cognitive, affective, and psychomotor aspects, including attendance, behavior, and practical skills (Pritasari et al., 2026). Such complexity directly impacts faculty and management efforts to comprehensively monitor student progress. Consequently, data-driven approaches for supporting academic decision-making become increasingly urgent.

Recent advancements in Educational Data Mining (EDM) offer systematic solutions for analyzing student data and predicting academic performance and dropout risks (Shoaib et al., 2024). Various studies demonstrate that EDM techniques enhance institutions' ability to identify at-risk students early through predictive modeling based on historical academic, demographic, and learning behavior data (Abou Naaj et al., 2023),(Yusof et al., 2022),(Anadi et al., 2023). Review research affirms significant growth in machine learning applications in education, focusing on predicting grade point averages, timely graduation, and early risk detection(Yağcı, 2022),(Astu et al., 2024). EDM implementations have been tested across contexts from general higher education to large-scale online learning, employing algorithms such as Decision Tree, Naive Bayes, Support Vector Machine (SVM), Artificial Neural Network (ANN), and K-Nearest Neighbor (KNN)(Staneviciene et al., 2024),(Mohamed Nafuri et al., 2022).

The KNN algorithm is a widely used classification method in EDM for analyzing academic performance and graduation status(Wati et al., 2023). Studies report that KNN competitively predicts learning achievements, GPAs, and graduation status, either standalone or combined with techniques like k-means clustering. Comparative studies benchmark KNN against SVM(Johora et al., 2025), Random Forest, and Naive Bayes for academic performance prediction, program recommendations, or course suggestions in higher education(Feng et al., 2022). These findings position KNN as an effective model candidate, particularly with features encompassing academic, demographic, and learning participation attributes (Staneviciene et al., 2024).

The popularity of the KNN algorithm stems from its simplicity, non-parametric nature, and interpretability for educational practitioners (Xiao et al., 2022). However, performance hinges on distance metric and k-value selection, with suboptimal choices reducing prediction accuracy (Wati et al., 2023). Recent research indicates KNN matches or surpasses other algorithms on educational data under proper preprocessing, feature distribution handling, and metric selection. Empirical reviews show no universally superior algorithm; ANN or SVM may outperform KNN on some datasets, while KNN excels on others. This underscores contextual KNN configuration studies in vocational education domains.

Current studies confirm varying performances of distance metrics like Euclidean, Manhattan, Minkowski, and Chebyshev based on data characteristics and feature distribution (Setiawan, 2022). Comparative analyses on student performance datasets reveal specific metric-k combinations significantly boost KNN accuracy and stability, especially with mixed features and high scale variance. Challenges intensify with large-scale datasets integrating academic scores, demographics, and e-learning behavior indicators (Manurung et al., 2025)(Lakhdar et al., 2024). Arbitrary parameter selection risks degrading classification utility for policymakers, emphasizing systematic sensitivity analyses in education.

Literature on dropout and timely graduation prediction predominantly targets general education, underrepresenting vocational contexts (Yamin, 2026)(Mohamed Nafuri et al., 2022). Higher education studies succeed with machine learning for graduation and dropout prediction but inadequately address vocational emphases on practice-based learning and skills assessment(Ali & Koehler, 2020). Vocational research favors classical statistics for risk factors without deep exploration of distance-based algorithms like KNN. Thus, interactions between distance metrics and k-values on hybrid vocational student data remain underexplored, creating research opportunities.

To address this research gap, this study aims to evaluate the effectiveness of various

distance metrics and k-value parameters in the KNN algorithm for predicting academic performance within complex vocational environments. Specifically, this study analyzes the impact of distance metrics (Euclidean, Manhattan, Minkowski, and Chebyshev) and k-values (1–27) on KNN performance in classifying vocational student academic outcomes. The dataset integrates academic and demographic features, mirroring complex real-world vocational environments. Key contributions include: (1) a systematic KNN parameter sensitivity analysis in vocational education, (2) the identification of optimal configurations for hybrid vocational datasets, and (3) practical recommendations for KNN-based classification systems in vocational institutions. These outcomes aim to strengthen Educational Data Mining (EDM) for early at-risk detection and targeted academic decision-making.

RESEARCH METHODS

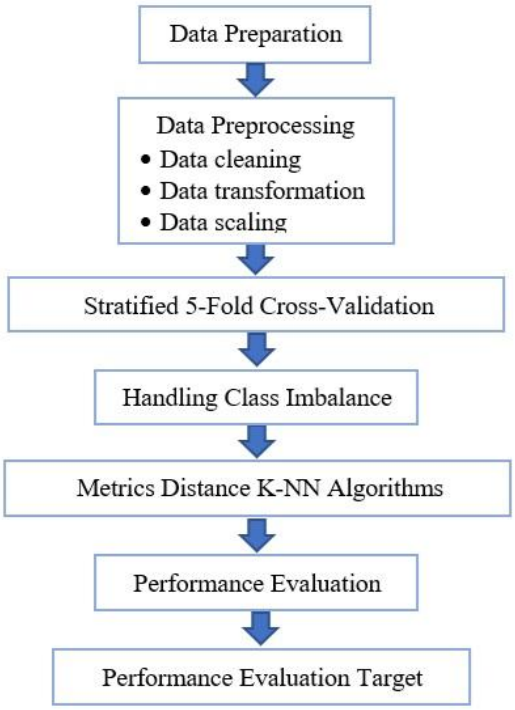

Figure 1. Research stages

Figure 1 illustrates the research workflow, beginning with the Data Preparation stage to set up the dataset, followed by the Data Preprocessing phase, which encompasses data cleaning and data transformation. To ensure the objectivity of the resulting model, a Handling Class Imbalance step is performed to address disproportionate class distributions within the data.

Subsequently, the model's effectiveness is evaluated using the Stratified 5-Fold Cross-Validation method to guarantee an even data distribution across each fold. The core process is executed by applying various KNN Distance Metrics to determine the nearest neighbor classification. This sequence concludes with Performance Evaluation to measure the model's success rate, ultimately reaching the Performance Evaluation Target established as the final research standard.

In this study, Python serves as the core programming language for data processing and algorithm implementation. The development process is supported by the Anaconda IDE, which streamlines the integration of essential data science libraries and ensures the stability of the research environment

A. Dataset preparation

The dataset for this study was extracted from the Academic Information System (Sisak) at Politeknik Negeri Sriwijaya (Polsri), South Sumatra, Indonesia, covering the 2024–2025 academic year. The dataset encompasses 750 first-year vocational students from the Informatics Management department. The analyzed data represent the students' baseline academic and demographic conditions at the onset of their vocational education. The academic features consist of 11 attributes, including Semester Grade Point Average (SGPA), behavior, attendance, and core course grades. These attributes are specifically selected to capture the cognitive, affective, and psychomotor dimensions essential to vocational academic performance.

Additionally, demographic attributes are included, such as age, gender, previous school type, parental educational background, and parental income level. Integrating academic and demographic features provides a multidimensional representation of vocational student profiles. The target variable is the student performance category, classified into three labels: With Distinction, Very Satisfactory, and Satisfactory. This categorization is derived from the Cumulative Grade Point Average (CGPA) thresholds established by the institution. This data structure facilitates the application of classification models to analyze and predict student academic outcomes effectively.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.520
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

B. Data Preprocessing

The preprocessing phase was conducted prior to model implementation to ensure data quality and consistency. Systematic procedures were applied to prepare the dataset in accordance with the requirements of distance-based classification algorithms, specifically focusing on data cleaning, transformation, and scaling.

Data cleaning involved the removal of duplicate records and the handling of missing values using attribute-specific imputation strategies. This step is critical to mitigate bias and reduce noise that could degrade the classification model's performance. Subsequently, data transformation was performed on categorical academic and demographic attributes. These variables were converted into numerical representations using the One-Hot Encoding method. This approach was selected to represent discrete categorical data as binary vectors, ensuring that categorical information is preserved without imposing an artificial ordinal relationship. Consequently, categorical attributes are treated with equal weight to numerical attributes within the feature space, enabling the KNN algorithm to accurately distinguish patterns across student performance classes.

In the final stage, numerical attributes with varying ranges were scaled using the standardization method (Z-score normalization). This process transformed each numerical feature to have a mean of mu = 0 and a standard deviation of sigma = 1. Standardization is essential to eliminate the dominance of high-magnitude features that could distort distance calculations. This is particularly vital for distance metrics—including Euclidean, Manhattan, Chebyshev, and Minkowski—which are highly sensitive to feature scales. By implementing standardization, each feature contributes uniformly to the distance-weighted classification, thereby enhancing the stability and predictive accuracy of the KNN model.

C. Model Validation Using K-Fold Cross Validation

To ensure the stability and reliability of the classification model, this study employs the 5-Fold Cross Validation method. In this approach, the dataset is randomly partitioned into five equally sized subsets or "folds." The experimental process is executed five times; in each iteration, one fold is designated as the testing set, while the remaining four folds serve as the training set. Through this technique, every student record in the dataset is utilized as testing data at least once. The final performance metrics—including accuracy, precision, and recall—are calculated based on the average results across all five iterations. This approach effectively mitigates the risk of bias resulting from a single data split and ensures that the KNN model possesses robust generalization capabilities when predicting student academic performance on unseen data.

D. Handling Class Imbalance

In the academic performance dataset, a significant disparity in the distribution of student data across performance categories was identified, where certain classes, such as Very Satisfactory exhibit a much higher frequency compared to minority classes like the Satisfactory category. This class imbalance potentially causes the KNN algorithm to become biased in its predictions by favoring the majority class, thereby reducing the classification accuracy for smaller classes.

To address this challenge, this study implements the Synthetic Minority Over-sampling Technique (SMOTE) during the preprocessing stage. This technique works by creating new synthetic samples for the minority class based on feature proximity to its nearest neighbors, rather than merely replicating existing data. The implementation of SMOTE enables the model to learn more varied patterns within minority classes, thereby strengthening generalization capabilities and enhancing the stability of classification performance across all categories of vocational student academic achievement. Furthermore, the application of a distance-weighted voting scheme within the KNN algorithm serves as an additional mechanism to ensure that spatial proximity remains the primary priority in class determination, regardless of the differences in sample sizes between categories.

All preprocessing steps, including SMOTE and normalization, were applied only to the training data within each fold to prevent data leakage.

E. Distance Metrics in The KNN Algorithm

The primary classification algorithm employed in this study is KNN, with an emphasis on evaluating its parametric sensitivity. Four distinct distance metrics were systematically examined to determine their impact on classification performance.

Euclidean Distance

The Euclidean metric is the most commonly used distance measure in KNN classification. The concept of this metric is based on the straight-line distance between two points in a multidimensional space. Due to its high sensitivity to variations in feature scale, this metric necessitates a normalization process like the Z-score prior to its application in calculations. In the context of vocational student performance classification, the Euclidean metric is effective when the attributes have a relatively homogeneous distribution and the relationship between attributes is continuous

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_{i)}{}^2} \quad (1)$$

- *d(x, y)*: Represents the Euclidean distance between two data points, x and y.
- *n*: Denotes the total number of features or attributes within the dataset.
- x_i, y_i: Represent the coordinate values of the first and second data points, respectively, for the i^{th} feature.
- *i*: Refers to the index of the feature, ranging from 1 to *n*.

Manhattan Distance

The Manhattan metric calculates distance based on the sum of absolute differences between attributes. Compared to the Euclidean metric, this metric is more robust to outliers because it does not use squared differences in its calculation. The Manhattan metric is appropriate when data has high variability or when small changes in one attribute do not dominate the total distance.

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (2)$$

- *d(x, y)*: Represents the Manhattan distance between two data points, x and y.
- *n*: Denotes the total number of features or attributes analyzed in the study.
- x_i: The value of the i^{th} feature for the first data point.
- y_i: The value of the i^{th} feature for the second data point.
- $|x_i - y_i|$: The absolute difference between the values of the i^{th} feature for both points.
- *i*: The index representing each specific feature, ranging from 1 to *n*.

Minkowski Distance

The Minkowski distance is a generalized form of both Euclidean and Manhattan distances. With p=3, this metric assigns a larger penalty to extreme differences in attribute values compared to the Euclidean distance (p=2), and it is significantly more sensitive than the Manhattan distance (p=1). The use of Minkowski distance provides flexibility in adjusting the sensitivity of distance calculations to variations in attribute values. This makes it relevant for evaluating the impact of different distance metrics on the performance of the KNN algorithm

$$d(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{\frac{1}{p}}, p = 3 \quad (3)$$

- *d(x, y)*: Represents the Minkowski distance between two data points, *x* and *y*.
- n: Denotes the total number of features or attributes in the dataset.
- x_i, y_i: Represent the coordinate values of the first and second data points for the i^{th} feature.
- *p*: A parameter that defines the order of the Minkowski distance.
- i The index of the feature, ranging from 1 to n.

Chebyshev Distance

Chebyshev Distance measures distance based on the maximum difference in a single attribute. This metric emphasizes attributes with the largest differences and ignores the contributions of other attributes that are smaller. In classifying the academic performance of vocational students, Chebyshev Distance is useful for identifying cases where a single dominant factor (e.g., a specific GPA or IPS) is the main determinant of the proximity between data points.

$$d(x,y) = \max(|x_i - y_i|) \quad (4)$$

- *d(x, y)*: Represents the Chebyshev distance between two data points, *x* and *y*.
- n: Denotes the total number of features or attributes in the dataset.
- x_i, y_i: Represent the coordinate values of the first and second data points for the i^{th} feature.

F. Optimization of k-Values

The range of k-values evaluated consists of integers from 1 to 27 (k ∈ {1,2,3,4,...,27}to avoid tie conditions during the voting process. The selection of this range is based on the rule of thumb $k \approx \sqrt{n}$. With a total of n=750 training samples, $\sqrt{750}$. The

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i3.520
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

weighting scheme applied is distance-weighted voting, where the weight is defined as $w=1/d^2$, with d representing the distance between the test instance and its nearest neighbors.

G. Performance Evaluation

The performance of the KNN algorithm was evaluated by measuring accuracy, precision, recall, and F1-score for each combination of k-values and distance metrics tested, namely Euclidean, Manhattan, Minkowski (p=3), and Chebyshev. The k-value was varied in the range k=1 to k=27 to analyze the effect of the number of nearest neighbors on the accuracy of classifying the academic performance of vocational students. The evaluation results are presented in the form of tables and graphs comparing accuracy against k-values to show the model's performance trends for each distance metric. Accuracy is calculated as the proportion of correct predictions to the total test data.

Additionally, precision, recall, and F1-score metrics are used to provide a more in-depth picture of the quality of predictions for each student performance class. The F1-score serves as a complementary metric that combines precision and recall through harmonic mean, making it more representative when class distributions are imbalanced. The use of this combination of metrics allows for a more comprehensive analysis of the stability and effectiveness of the KNN model across various parameter configurations. Thus, not only is the overall accuracy level considered, but also the model's ability to distinguish between classes.

H. Target Performance Evaluation

The evaluation of target class performance was conducted to assess the effectiveness of the KNN model in classifying each student performance category using class-wise metrics such as precision, recall, and F1-score. The results show that classes with a larger number of samples achieve more stable and higher performance, indicating the model's ability to capture representative patterns when sufficient training data are available. Conversely, classes with fewer samples tend to exhibit lower performance due to limited neighborhood representation, which may bias predictions toward majority classes. This class-wise evaluation provides important insights into the strengths and limitations of the model and highlights the need for balanced data distribution and appropriate parameter tuning to enhance classification reliability across all target classes.

RESULTS AND DISCUSSION

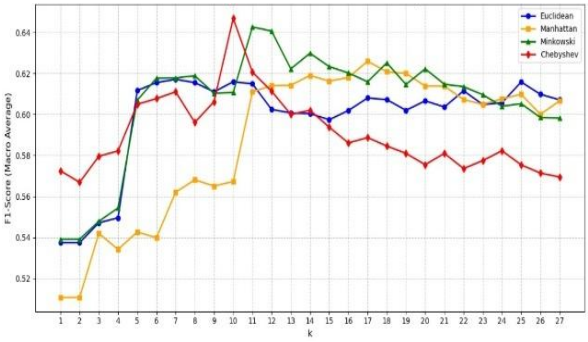

Figure 2. k-Value Optimization Based on F1-Score

Figure 2 presents the relationship between the value of k and the F1-score for each distance metric evaluated in the KNN modeling. The F1-score was selected as the primary metric because it effectively represents the balance between precision and recall, making it more informative than accuracy alone. The general pattern observed for the Euclidean, Chebyshev, Manhattan, and Minkowski metrics shows a rapid performance reaching a peak between k=10 and k=17 depending on the metric. Beyond this point, F1-scores tend to decrease gradually as k increases toward k=27. This pattern indicates that selecting excessively large values of k may reduce the model's ability to distinguish class boundaries in student performance classification.

Based on the evaluation of the F1-Score (Macro Average), the Chebyshev metric demonstrates the most superior performance in balancing precision and recall, reaching its peak at k=10 with a score of 0.6468. The graphical trend for this metric illustrates a sharp increase in efficiency starting from k=1 until reaching its optimality at k=10. This suggests that the application of maximum difference distance is highly effective in mapping vocational student achievement categories, particularly for classes characterized by imbalanced data distributions. However, the Chebyshev metric also exhibits high sensitivity to the increase in the number of neighbors, as its performance tends to degrade more rapidly after passing the optimal point compared to other distance metrics.

On the other hand, the Minkowski (p=3) and Euclidean metrics display a more consistent stability profile across medium to large k-values. Minkowski achieves its best performance at k=11

with a value of 0.6426, proving that the generalization of Minkowski distance can deliver highly competitive results in capturing student feature similarity patterns. Meanwhile, the Manhattan metric shows unique trend characteristics, where the model requires a broader neighbor coverage—specifically at k=17 (0.6259)—to achieve an optimal predictive balance. Overall, the success of the Chebyshev metric in achieving the highest F1-Score at k=10 reinforces the validity of selecting this model for your research, as it provides a fairer classification reliability across all academic achievement categories without being biased by the dominance of the majority class.

Table 1. Performance comparison of knn using different distance metrics

Metrics	k	Accuracy	Precision	Recall	F1-Score
Euclidean	11	0.8504	0.6044	0.6629	0.6149
Manhattan	17	0.8464	0.6173	0.6491	0.6259
Minkowski	11	0.8437	0.6283	0.7334	0.6426
Chebyshev	10	0.8464	0.6453	0.7120	0.6468

Table 1 show the performance evaluation results demonstrate that, in general, all distance metrics exhibit highly competitive capabilities, with accuracy rates exceeding 84%. This indicates that the KNN model is robust in classifying vocational student data. However, a more in-depth analysis reveals significant variations in balancing metrics, such as Recall and F1-Score. While the Euclidean metric recorded the highest global accuracy of 0.8504 at k=11, in the context of imbalanced data, accuracy alone is insufficient to fully represent the overall quality of the model.

Superiority in terms of sensitivity or coverage Recall was most evident in the Minkowski metric, which achieved a value of 0.7334, demonstrating its effectiveness in identifying students within specific categories and minimizing the risk of false negatives. Conversely, the Chebyshev metric proved to be the most reliable regarding predictive Precision with a score of 0.6453, indicating that this model possesses the highest confidence level when predicting specific achievement categories. Ultimately, the Chebyshev metric at k=10 is considered to provide the most balanced performance, as it attained the highest F1-Score (0.6468). This score represents the optimal harmonic mean between precision and prediction coverage for the classification of vocational student academic performance.

The classification performance using four distinct distance metrics (*Euclidean, Manhattan, Minkowski,* and *Chebyshev*) reveals varying degrees of effectiveness in handling extreme data imbalance between the *With Distinction* class (492 samples) and the *Satisfactory* class (6 samples).

Table 2. Classification results per class using euclidean metric

Class	Precision	Recall	F1 - Score	Support
With Distinction	0.95	0.80	0.91	492
Very Satisfactory	0.77	0.81	0.79	251
Satisfactory	0.07	0.33	0.12	6
Average	0.60	0.67	0.61	749

The *Euclidean* metric demonstrates robust performance for the majority class (*With Distinction*), achieving a high precision of 0.95. However, it struggles significantly with the minority *Satisfactory* class, yielding a recall of only 0.33 and a low F1-score of 0.12. This suggests that the straight-line distance calculation in the Euclidean space is highly sensitive to the dominant density of majority samples, resulting in a Macro Average F1-score of 0.61.

Table 3. Classification results per class using manhattan metric

Class	Precision	Recall	F1 - Score	Support
With Distinction	0.89	0.91	0.90	492
Very Satisfactory	0.79	0.74	0.77	251
Satisfactory	0.14	0.33	0.20	6
Average	0.61	0.66	0.62	749

The *Manhattan* metric provides a better balance for the majority class with an F1-score of 0.90. Notably, it improves the F1-score for the *Satisfactory* category to 0.20 compared to the Euclidean metric. While the recall for the minority class remains at 0.33, the increased precision (0.14) indicates a reduction in false-positive misclassifications for this specific category.

Table 4. Classification results per class using minkowski metric

Class	Precision	Recall	F1 - Score	Support
With Distinction	0.96	0.84	0.90	492
Very Satisfactory	0.73	0.86	0.79	251
Satisfactory	0.12	0.50	0.19	6
Average	0.61	0.73	0.63	749

The *Minkowski* metric (set at k=11) achieves the highest overall average performance among all tested metrics, with an Average Recall of 0.73 and a Macro F1-score of 0.63. Its primary strength lies in its ability to detect the minority *Satisfactory* class more effectively, reaching a recall value of 0.50. Furthermore, it shows excellent predictive coverage for the *Very Satisfactory* category with a recall of 0.86.

Table 5. Classification results per class using chebyshev metric

Class	Precision	Recall	F1 - Score	Support
With Distinction	0.90	0.91	0.91	492
Very Satisfactory	0.82	0.72	0.77	251
Satisfactory	0.12	0.50	0.19	6
Average	0.61	0.71	0.62	749

The *Chebyshev* metric shows exceptional stability for the *With Distinction* category (0.91 F1-score) and provides the highest precision for the *Very Satisfactory* category at 0.82. Similar to Minkowski, Chebyshev successfully identifies 50% of the actual minority data (*Satisfactory* recall of 0.50). However, due to its stricter selectivity, its overall Average F1-score (0.62) remains slightly below that of the Minkowski metric.

In summary, while all metrics reliably predict the majority class with over 90% accuracy, the Minkowski and Chebyshev metrics prove superior in addressing data imbalance. Both metrics afford proportional weight to the *Satisfactory* minority class without compromising the integrity of the dominant classes. The choice between them depends on institutional priorities: Minkowski offers broader inclusive coverage, while Chebyshev provides sharper predictive precision.

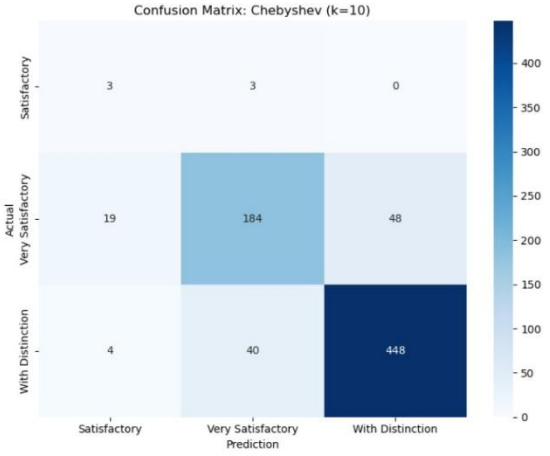

Figure 3. Confusion matrix chebyshev

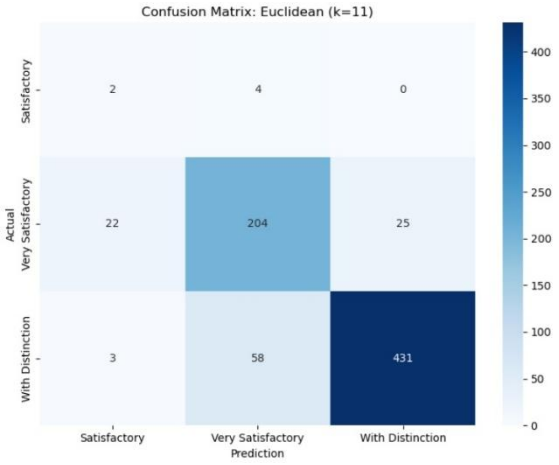

Figure 4. Confusion matrix euclidean

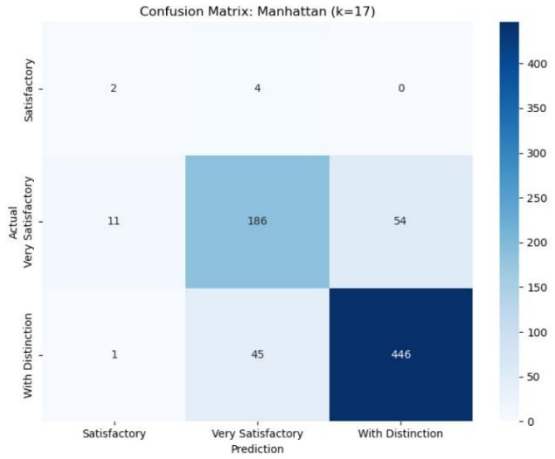

Figure 5. Confusion matrix manhattan

Figure 6. Confusion matrix minkowski

Figures 3 through 6 illustrate the Confusion Matrix results for each distance metric category evaluated in this study, namely Euclidean, Manhattan, Minkowski, and Chebyshev.

These matrices provide a detailed visualization of the classification performance for the three student academic performance categories: With Distinction, Very Satisfactory, and Satisfactory. Each matrix compares the actual labels from the dataset of 749 records against the predicted labels generated by the model.

- The diagonal elements represent the number of correctly classified instances for each performance class, reflecting the model's accuracy for that specific metric.
- The off-diagonal elements indicate the misclassifications, showing where the model confuses one performance category with another.

By analyzing these figures, we can evaluate how different distance metrics influence the model's ability to distinguish between academic achievement levels, particularly in handling the nuances of vocational student data

The evaluation through the Confusion Matrix provides deeper visibility compared to global accuracy, particularly in observing how the model handles data imbalance between the majority class (*With Distinction*) and the minority class (*Satisfactory*). The Chebyshev matrix (k=10) demonstrates a sharp level of selectivity, successfully identifying 3 out of 6 *Satisfactory* samples correctly, although the remaining samples were distributed into the *Very Satisfactory* class. While Chebyshev is highly robust in the *With Distinction* class with 448 correct predictions and excels at minimizing errors in the middle class, this metric tends to distribute minority class errors evenly to the classes above it. Conversely, the Euclidean matrix (k=11) exhibits a strong bias toward the majority class, showing the lowest performance in the minority class by capturing only 2 out of 6 *Satisfactory* samples. The characteristics of the Euclidean metric, which is susceptible to the dominance of majority data—as evidenced by 58 *With Distinction* samples being misclassified as *Very Satisfactory*—render it less effective for an early warning system requiring student intervention.

Meanwhile, the Manhattan matrix (k=17) shows a balanced but less sensitive performance in the lower class, where the detection of *Satisfactory* samples only reached 2 accurate samples, similar to the Euclidean metric. Despite displaying a competitive figure of 446 correct predictions in the *With Distinction* class, the Manhattan metric is considered to have failed in maximizing the potential of SMOTE to strengthen decision boundaries for the extreme minority class. On the other hand, the Minkowski matrix (k=11) emerges as the most inclusive and optimal model, as it successfully achieved the highest recall by correctly predicting 3 out of 6 *Satisfactory* samples. The superiority of the Minkowski metric is clearly visible in the middle class, with the highest number of correct predictions on the main diagonal at 216, while providing the best balance, which visually indicates the fairest model in providing predictive attention across all achievement categories.

In strategic summary, the presence of 2 or 3 correctly classified samples on the minority class

diagonal across all matrices proves that data augmentation through SMOTE successfully helped the model recognize rare class patterns that originally consisted of only 6 samples. Minkowski and Chebyshev are recommended as the best choices, where Minkowski excels at maximizing detection across all classes, while Chebyshev is better at maintaining precision in the high-achievement classes. However, further areas for improvement are required through advanced feature engineering to address the ambiguity between high-achieving and middle-achieving graduates, given the continued concentration of figures outside the main diagonal, such as the 75 *With Distinction* samples predicted as *Very Satisfactory* using the Minkowski metric.

CONCLUSIONS AND SUGGESTIONS

Conclusion

This study demonstrates that the performance of the KNN algorithm in classifying academic performance of vocational students is strongly influenced by the choice of distance metrics and the k parameter. The experimental results indicate that all evaluated metrics achieve satisfactory classification performance with accuracy exceeding 84%, yet exhibit significant differences in maintaining balanced performance on imbalanced data. The Chebyshev metric with k = 10 provides the best balance between precision and recall, while the Minkowski metric (p = 3) excels in detecting minority classes. In contrast, the Euclidean metric achieves the highest accuracy but tends to be biased toward the majority class.

These findings confirm that relying solely on accuracy is insufficient for evaluating model performance on imbalanced datasets; therefore, metrics such as F1-score and recall must be considered comprehensively. In addition, the application of SMOTE is shown to improve the model's ability to recognize minority classes, although it does not fully overcome the limitations caused by extreme data imbalance.

Overall, this study contributes a deeper understanding of parameter sensitivity in KNN and provides practical recommendations for selecting appropriate distance metrics and k values. The results are expected to support the development of more accurate, balanced, and adaptive academic prediction systems in vocational education environments.

Suggestion

Based on the results of this study, several recommendations can be proposed for future research and practical implementation. First, future studies are encouraged to explore other algorithms such as Random Forest, Support Vector Machine (SVM), or Gradient Boosting as benchmarks to assess the competitiveness of KNN in handling imbalanced vocational education data. Second, more advanced class imbalance handling techniques should be investigated, such as combining SMOTE with undersampling or utilizing ensemble-based methods to improve performance on minority classes.

Third, future research may consider applying feature selection or dimensionality reduction techniques, such as Principal Component Analysis (PCA), to reduce data complexity resulting from one-hot encoding and to enhance the effectiveness of distance calculations in KNN. Fourth, parameter sensitivity analysis can be extended by evaluating different values of p in the Minkowski metric or by implementing weighted KNN with alternative weighting schemes.

Additionally, it is recommended to use larger datasets with more balanced class distributions or include multi-cohort data to improve model generalization. From an implementation perspective, the findings of this study can serve as a foundation for developing early warning systems in vocational education institutions to more accurately and timely identify students at risk of academic performance decline.

REFERENCES

Abou Naaj, M., Mehdi, R., Mohamed, E. A., & Nachouki, M. (2023). Analysis of the Factors Affecting Student Performance Using a Neuro-Fuzzy Approach. *Education Sciences*, *13*(3). https://doi.org/10.3390/educsci13030313

Ali, M., & Koehler, T. (2020). *Evaluation of Indonesian Technical and Vocational Education in Addressing the Gap in Job Skills Required by Industry*.

Anadi, I., Havrda, D. E., Owens-Mosby, D. A., & Shelton, C. M. (2023). Evaluation of Academic and Nonacademic Factors of First-Generation Students Transitioning to a Pharmacy Program. *American Journal of Pharmaceutical Education*, *87*(12), 100598. https://doi.org/10.1016/j.ajpe.2023.100598

Astu, M., Pawitra, S., Hung, H., & Jati, H. (2024). *A*

Machine Learning Approach to Predicting On-Time Graduation in Indonesian Higher Education. *9*(2), 294–308.

Feng, G., Fan, M., & Chen, Y. (2022). Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining. *IEEE Access*, *10*, 19558–19571. https://doi.org/10.1109/ACCESS.2022.3151652

Johora, F. T., Hasan, M. N., Rajbongshi, A., Ashrafuzzaman, M., & Akter, F. (2025). An explainable AI-based approach for predicting undergraduate students academic performance. *Array*, *26*, 100384. https://doi.org/10.1016/j.array.2025.100384

Khamdun, K., Suparmi, S., Maridi, M., & Rusilowati, A. (2021). Development of vocational science learning devices to improve project based soft skills. *Linguistics and Culture Review*, *5*(S1), 201–213. https://doi.org/10.21744/lingcure.v5ns1.1348

Lakhdar, Y., EL-Bendadi, K., & Bakkas, B. (2024). A New Hybrid Model to Predict the Performance of Trainee Teachers Based on Clustering and Classification. *Journal of Computer Science*, *20*(9), 1020–1029. https://doi.org/10.3844/jcssp.2024.1020.1029

Manurung, J., Saragih, H., Prabukusumo, M. A., & Ahmad, E. (2025). *Optimizing the performance of the K-Nearest Neighbors algorithm using grid search and feature scaling to improve data classification accuracy*. *14*(2), 260–268.

Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering Analysis for Classifying Student Academic Performance in Higher Education. *Applied Sciences (Switzerland)*, *12*(19). https://doi.org/10.3390/app12199467

Pritasari, O. K., Suhartini, R., & Hasbi, A. (2026). *Technological Integration and Soft Skill Synergy in Vocational Education : A Data-Driven Model for Enhancing Hairdressing Work Competence Integración Tecnológica y Sinergia de Habilidades Blandas en la Educación Vocacional : un Modelo Basado en Datos para Mejorar la Competencia Laboral en Peluquería*. https://doi.org/10.56294/saludcyt2026263

8

Setiawan, A. (2022). Perbandingan Penggunaan Jarak Manhattan, Jarak Euclid, dan Jarak Minkowski dalam Klasifikasi Menggunakan Metode KNN pada Data Iris. *Jurnal Sains Dan Edukasi Sains*, *5*(1), 28–37. https://doi.org/10.24246/juses.v5i1p28-37

Shoaib, M., Sayed, N., Singh, J., Shafi, J., Khan, S., & Ali, F. (2024). AI student success predictor: Enhancing personalized learning in campus management systems. *Computers in Human Behavior*, *158*(February), 108301. https://doi.org/10.1016/j.chb.2024.108301

Staneviciene, E., Gudoniene, D., Punys, V., & Kukstys, A. (2024). A Case Study on the Data Mining-Based Prediction of Students' Performance for Effective and Sustainable E-Learning. *Sustainability (Switzerland)*, *16*(23). https://doi.org/10.3390/su162310442

Wati, E. F., Perangin-angin, E. S., & Sari, A. P. (2023). *Prediction of Student Graduation using the K-Nearest Neighbors Method*. *7*(158), 211–216.

Xiao, W., Ji, P., & Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. *Engineering Reports*, *4*(5), 1–23. https://doi.org/10.1002/eng2.12482

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00192-z

Yamin, M. (2026). *Vocational Education Model in Indonesian Vocational High Schools Based on Teaching Factory*. *1*(1), 10–16.

Yusof, R., Hashim, N., Abdul Rahman, N., Mohd Yunus, S. Y., & Aziz Fadzillah, N. A. (2022). Academic Performance Prediction Model Using Classification Algorithms: Exploring the Potential Factors. *International Journal of Academic Research in Progressive Education and Development*, *11*(3), 706–724. https://doi.org/10.6007/ijarped/v11-i3/14753