Implementation of the K-Nearest Neighbor Algorithm for Classifying Immigration Residence Permit Applicants at the Class I Special Immigration Office TPI Soekarno-Hatta

Nur Azizah[*1] , Henderi[2] , Berisno Hendro Pardamean Manik Raja[3]
[1,2,3] Faculty of Science and Technology, University of Raharja, Indonesia
E-mail: [*1]**nur.azizah@raharja.info**, [2]henderi@raharja.info, [3]berisno@raharja.info

Abstract

*This study aims to apply the K-Nearest Neighbor (KNN) algorithm in classifying immigration residence permit applicants at the Class I Special Immigration Office TPI Soekarno-Hatta, focusing on the algorithm's effectiveness and accuracy in categorizing residence permit applicants based on the types of residence permits: Visit Stay Permit (ITK), Limited Stay Permit (ITAS), and Permanent Stay Permit (ITAP). This study employs a quantitative, experiment-based approach utilizing a dataset of 17,212 residence permit applicant records consisting of 11 key attributes, such as nationality, visa type, residence permit type, gender, and age group.The research process began with data preprocessing stages, including data cleaning, normalization, and dataset splitting into training and testing sets with 80:20 and 70:30 partitioning scenarios. The KNN algorithm was implemented using a parameter of k=5k = 5k=5, chosen based on experimentation to achieve optimal performance. The model's performance evaluation was conducted using accuracy, precision, and recall metrics derived from a confusion matrix. The findings reveal that the KNN algorithm successfully classifies data with the highest accuracy of **96.95%** in the 80:20 dataset partition scenario and **96.84%** in the 70:30 scenario. The Visit Stay Permit (ITK) class demonstrated the best performance with a precision of **97.46%** and a recall of **99.97%**, whereas the Permanent Stay Permit (ITAP) class showed the lowest performance with a recall of **59.79%**, indicating challenges in recognizing patterns for this class. This study also identifies the advantages of the KNN algorithm, including its simplicity of implementation, flexibility in handling multiclass data, and effectiveness for low-dimensional datasets. However, the algorithm has limitations, such as sensitivity to imbalanced data distributions and high computational time for large datasets.*

Keywords *— K-Nearest Neighbor, Classification, Residence Permit, Immigration, RapidMiner, Accuracy, Precision, Recall*

1. INTRODUCTION

The management of residence permit applicants' data is one of the key aspects in immigration administration. In the current digital era, the number of residence permit applicants continues to increase, requiring an efficient system to manage and classify the data. According to data from the Directorate General of Immigration, in 2022, the number of residence permit applicants in Indonesia reached more than 500,000 people, marking a 15% increase compared to the previous year (Directorate General of Immigration, 2023). Proper data management not only facilitates administrative processes but also supports accurate decision-making.

The Class I Special Immigration Office TPI Soekarno-Hatta, as one of the main entry points to Indonesia, faces various challenges in classifying residence permit applicants. The current manual process often leads to delays and errors in data processing. This can result in inaccurate decisions regarding the issuance of residence permits, which in turn may affect public security and order. Therefore, a more efficient solution is needed to address these issues.

The role of information technology in enhancing the efficiency of the classification process is highly significant**.**With the application of algorithms and data mining methods, data processing can be automated and accelerated. One algorithm that can be applied is the K-Nearest Neighbor (KNN), which is known for its effectiveness in data classification. KNN has the capability to group data based on the proximity of new data to existing data, thereby aiding in better decision-making processes.

Based on the background description above, the problem formulation taken is:

1. How can the K-Nearest Neighbor (K-NN) algorithm be applied for the classification of immigration residence permit applicants?
2. How accurate is the K-NN algorithm in classifying residence permit applicant data?
3. What are the advantages and disadvantages of using the K-NN algorithm compared to other methods in classifying residence permit applicants?

2. RESEARCH METHOD

The research employs a quantitative approach with an experimental method to classify immigration stay permit applicants at the Soekarno-Hatta TPI Special Class I Immigration Office using the K-Nearest Neighbor (K-NN) algorithm. The study uses 17,212 primary data points with 12 attributes from 2024, collected from the Soekarno-Hatta International Airport Special Class 1 Immigration Office. The research methodology involves data collection, data pre-processing, K-NN algorithm implementation, model evaluation, and results analysis. Data pre-processing ensures data cleanliness and consistency for K-NN analysis. The K-NN algorithm is implemented by determining the 'k' value using RapidMiner (set to k=5) and calculating distances using Euclidean Distance as shown in Figure 1.

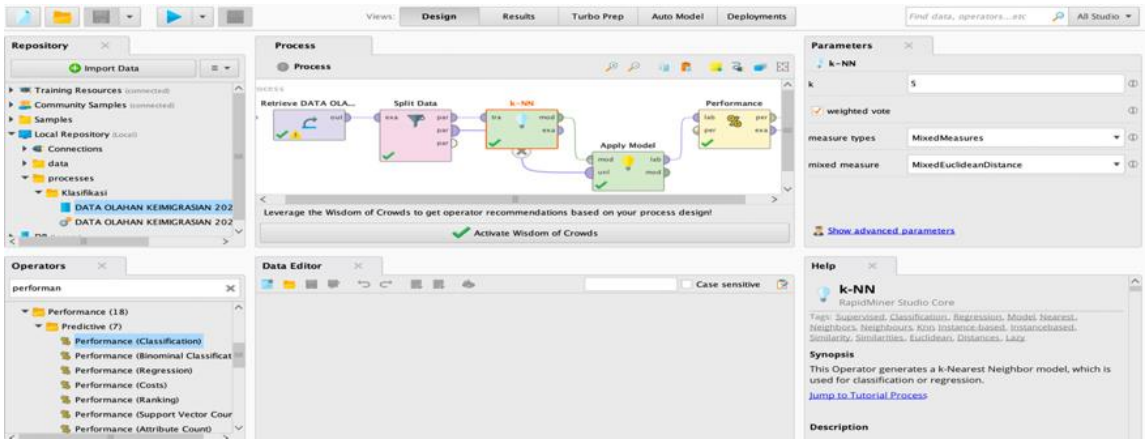

Figure 1. The process of determining the value of *k* in RapidMiner

The dataset is split into training and testing sets 80% training/20% testing to objectively evaluate the model's predictive performance as shown in Figure 2.

Split Type	Training Data (Count)	Testing Data (Count)
70% Training / 30% Testing	12048	5164
80% Training / 20% Testing	13769	3443

Figure 2. Comparison of training data and testing data split

Model evaluation utilizes accuracy, precision, and recall metrics, often employing a confusion matrix to assess classification performance. The system testing incorporates a Confusion Matrix to evaluate functionality, accuracy, and performance. Data analysis techniques include comprehensive data collection (full name, nationality, continent, visa type, stay permit type, gender, age group), preprocessing (cleaning, normalization, feature selection), dataset splitting, and k-fold cross-validation for robust model validation. Performance metrics such as accuracy, precision, recall, and F1-score are used for model evaluation, alongside system response time measurements.

3. RESEARCH RESULTS AND DISCUSSION

The study processed a dataset of 17,212 primary data points with 11 attributes, including nationality, visa type, stay permit type, and gender, using pre-processing techniques. A total of 5,164 data points were used for testing to evaluate the K-Nearest Neighbor (K-NN) method's effectiveness

K-Parameter Selection and Accuracy: The optimal 'k' value for the K-NN algorithm was determined using RapidMiner, with Euclidean Distance as the distance metric. For the 80% training (13,769 datasets) and 20% testing (3,443 datasets) scenario, a 'k' value of 5 yielded an accuracy of 96.95%. For the 70% training (12,048 datasets) and 30% testing (5,164 datasets) scenario, a 'k' value of 5 resulted in an accuracy of 96.84%. Model Evaluation (80:20 Split Scenario): The model's performance was evaluated using accuracy, precision, and recall, derived from a confusion matrix as show in figure 3.

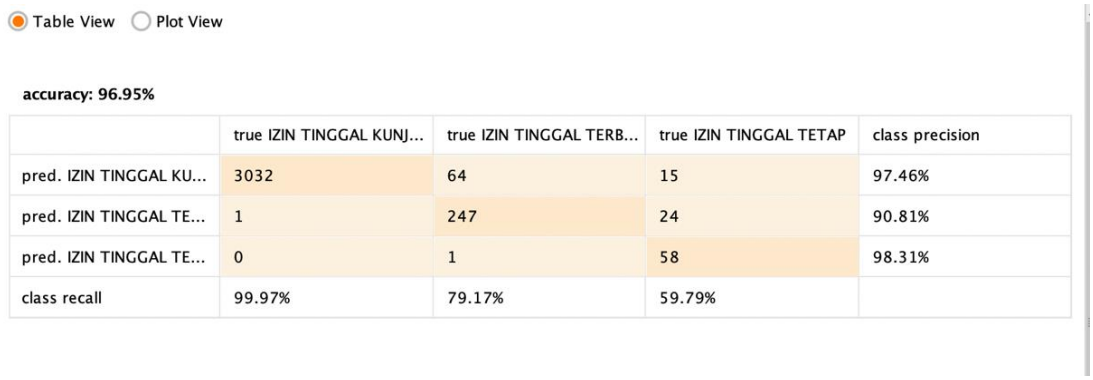
Table View Plot View

accuracy: 96.95%

	true IZIN TINGGAL KUNJ...	true IZIN TINGGAL TERB...	true IZIN TINGGAL TETAP	class precision
pred. IZIN TINGGAL KU...	3032	64	15	97.46%
pred. IZIN TINGGAL TE...	1	247	24	90.81%
pred. IZIN TINGGAL TE...	0	1	58	98.31%
class recall	99.97%	79.17%	59.79%	

Figure 3. Confuxion Matrix Result

Accuracy: The overall accuracy was 96.95%, indicating that the K-NN model correctly predicted almost 97% of the test data.

Precision:
Izin Tinggal Kunjungan (ITK - Visitor Stay Permit): 97.46%.
Izin Tinggal Terbatas (ITAS - Limited Stay Permit): 90.81%.
Izin Tinggal Tetap (ITAP - Permanent Stay Permit): 98.31%.
Recall (Sensitivity):
ITK: 99.97%, showing excellent performance for this class.
ITAS: 79.17%, indicating the model recognized nearly 80% of actual ITAS data.
ITAP: 59.79%, suggesting the model was less sensitive to this class and only recognized about 60% of actual ITAP data.
Error Analysis:
Some ITAS data were misclassified as ITK (64 data points), indicating feature overlap.
A small portion of ITAP data (58 data points) was not well predicted, suggesting this class is harder to distinguish.

3.1. Analysis of Results:

Model Performance: The combination of k=5 and an 80:20 dataset split yielded excellent accuracy, demonstrating that the chosen number of neighbors adequately recognized data patterns. Data Distribution Impact: The ITK class dominated recall (99.97%), while ITAP had the lowest recall (59.79%). This imbalance might stem from unequal data distribution among classes or less distinct features for the ITAP class. Recommendations for Improvement: To enhance overall model performance, especially for the ITAP class, normalization of features, data balancing techniques (oversampling/undersampling), and optimization of feature selection are recommended. Effectiveness and Advantages of KNN: The KNN algorithm proved highly effective for stay permit classification, achieving 96.95% accuracy in the 80:20 split scenario. Its effectiveness is attributed to high accuracy, strong precision for the dominant ITK class, good generalization ability despite imbalanced classes, and its simplicity and reliability. KNN's strengths include its non-parametric nature, ability to handle multi-class data, simple implementation, and adaptability to various datasets.

3.2. Classification Patterns and Challenges:

ITK Class: Shows clear patterns and highly distinguishing attributes, dominating model performance with 99.97% recall.
ITAS Class: Less distinct patterns than ITK, with 79.17% recall, and some misclassifications due to feature overlap with ITK.
ITAP Class: Least clear patterns, resulting in low recall (59.79%) as its attributes are similar to ITAS and ITK.

Key features influencing classification patterns are nationality, visa type, and stay permit type.

3.3. Constraints and Solutions:

Constraints: Data imbalance (ITK dominance, low ITAP count) leading to low ITAP sensitivity , pattern similarity between classes (ITAS and ITK overlap) causing prediction errors , and high computational time for large datasets due to distance calculations.

Solutions: Implement data balancing techniques (oversampling/undersampling) , normalize data and select relevant features to reduce class overlap , experiment with different 'k' values , and consider alternative algorithms like KD-tree for faster neighbor search on large datasets

4. CONCLUSION

This study successfully applied the K-Nearest Neighbor (KNN) algorithm to classify immigration stay permit applicants, achieving high accuracy (96.95% for 80:20 split, 96.84% for 70:30 split). While effective for Visitor and Limited Stay Permits, the algorithm showed lower sensitivity for Permanent Stay Permits (59.79% recall). KNN's strengths include its non-parametric nature, simplicity, and multi-class flexibility. However, limitations include high computational time for large datasets, sensitivity to class imbalance (low recall for ITAP), and reliance on feature scaling. Future work should focus on dataset refinement, algorithm optimization (e.g., trying different 'k' values or using KD-tree), evaluating alternative models, and developing an integrated system for real-time decision-making

5. SUGGESTED

To enhance the effectiveness of the KNN algorithm in classifying residence permit applicants, it is recommended to improve dataset balance through oversampling of minority classes and enrich the feature set with more relevant attributes. Optimization should involve experimenting with different *k* values and applying index-based methods like KD-tree to reduce computational time. Additionally, comparing KNN with other models such as Decision Tree, Random Forest, and SVM can help evaluate its relative performance. For practical implementation, a real-time classification system based on KNN should be developed to support fast and accurate decision-making at Immigration Offices. While RapidMiner is useful for initial experiments, transitioning to more flexible tools like Python or R is advised for advanced analysis and system integration.

6. REFERENCES

[1] Hamzah Naufal Zuhdi (2024)."Application of K-Nearest Neighbors Algorithm in Creditworthiness Evaluation: Case Study at Bank ABC". Indonesian Journal of Informatic Research and Software Engineering Vol. 4 No. 1

[2] Febrianda Putra (2024) "Application of K-Nearest Neighbor Algorithm Using Wrapper as Preprocessing for Determining Human Body Weight Information". JOURNAL OF MALCOM Vol. 4 No. 1

[3] Mey Susi Munthe, (2024) "Classification of Student Data Eligible for Assistance Using the K-Nearest Neighbors Method" in JUREKSI Journal Vol. 2 No. 1.
[4] Qurotul A'yuniyah (2024).”Application of K-Nearest Neighbor Algorithm for Classification of Student Majors at Sma Negeri 15 Pekanbaru". Indonesian Journal of Informatic Research and Software Engineering Vol. 3 No. 1.
[5] Rudianta Sitepu (2022).”Implementation of the K-Nearest Neighbor Algorithm for Credit Application Classification”. Journal of Information Systems, Informatics Engineering and Educational Technology Vol. 1 No. 2
[6] Putri AA, 2021 Application of data mining to predict fruit and vegetable sales using the K-Nearest Neighbor method (Case study: PT Sentral Brastagi Utama). Informatics and Information Engineering Resolution, 1(6), PP. 354-361.
[7] Implementation of K – MEANS Clustering on Rapidminer for accident prone analysis National Seminar on Applied Quantitative Research, PP.58 – 62
[8] Rizki Falutfi, H and Yogi, MA 2018. School Academic Mining Process using Rapid Miner. Matics , 10 (2) PP. 47 – 51
[9] Rerung, R,R, 2018. Application of data mining utilizing the Association rules method for product promotion j. Technol. Rekayasa 3(1), p.89.
[10] Saefudin, s., and Funanando, P, 2020. Application of book recommendation data mining using the Apriori algorithm, j journal of information systems. , 3 (1) , P. 89 .