JOURNAL OF DATA SCIENCE
editorjurnal@seaninstitute.or.id Published by Sean Institute ISSN XXXX-XXXX (Online - Elektronik)
Volume 1 Nomor 02, Tahun 2023

Rapid Miner Testing With The Knn Algorithm

Noferianto Sitompul
Politeknik Negeri Sambas
Email: noferiantositompul@gmail.com

Abstract
In the modern era of computers and the internet, the need for effective data analysis tools is increasing. In the data mining industry, RapidMiner has developed into an essential tool for using simple but effective classification methods such as k-Nearest Neighbors (KNN). The purpose of this study is to evaluate RapidMiner's ability to implement KNN and evaluate its performance on various datasets derived from various characteristics. Test results show that RapidMiner can implement KNN with satisfactory accuracy for most datasets; The methodology used measures classification performance through metrics such as accuracy, recall, precision, and F1 score. In conclusion, RapidMiner has a lot of potential for practical application of KNN, but it has limitations when working with large datasets.
Keywords: RapidMiner, k-Nearest Neighbors, KNN, data mining, classification.

INTRODUCTION

In today's digital era, with massive amounts of data growing, the need for efficient data analysis methods is crucial [1]. One data analysis method that has been widely used in various industrial sectors is data mining. RapidMiner, as one of the leading data analysis platforms, provides complete solutions for data analysis processes from the preprocessing stage to the implementation of various machine learning algorithms [2].

One classification algorithm that has received great attention in the scientific and industrial communities is k-Nearest Neighbors (KNN) [3]. Due to its non-parametric nature and ease of implementation, KNN is often considered one of the most intuitive and effective classification approaches [4]. Although many data analysis platforms offer implementations of the KNN algorithm, the implementation of this algorithm on RapidMiner and its performance evaluation is still an interesting topic for further research.

A research paper on the analysis of the k-NN method using RapidMiner for a tourism recommendation system in Labuan Bajo found that the k-NN algorithm with a value of K=5 and the RapidMiner application had an accuracy of 83.33% [5].

In this context, this study will test RapidMiner's ability to implement the KNN algorithm and will assess its performance under various dataset conditions. The purpose of this study is not only to evaluate the classification performance of KNN on the RapidMiner platform, but also to provide insight into the potential limitations and advantages of its application.

METHOD

In evaluating the potential and performance of the KNN algorithm when implemented using the RapidMiner platform, this research adopts a series of systematic methods to ensure the validity and reliability of the results. The stages followed in this research refer to best practices in scientific and industrial literature, which allows us to understand exactly how RapidMiner can harness the power of the KNN algorithm [6].

JOURNAL OF DATA SCIENCE
editorjurnal@seaninstitute.or.id Published by Sean Institute ISSN XXXX-XXXX (Online - Elektronik)
Volume 1 Nomor 02, Tahun 2023

1. Data Selection

Before going further, we select several datasets with varying characteristics to find out how well the KNN algorithm functions under various conditions [7]. This dataset was obtained from public data repositories and also from several industry sources collaborating on this research.

2. Pre-processing of Data

Each dataset was examined and processed using RapidMiner [2]. We performed several preprocessing steps such as normalization, coding, and filling in the missing values based on the techniques suggested by the literature [8]. The goal is to ensure the data is ready for the classification process.

3. Implementation of the KNN Algorithm

The KNN algorithm is implemented using features in RapidMiner [2]. We vary the number of neighbors (k) to find optimal parameters for each dataset, according to the recommendations from [3] and [4]. In addition, distance metrics such as Euclidean, Manhattan, and Minkowski were also tested to assess their effect on the classification results.

4. Performance Evaluation

Classification performance is measured using metrics that are commonly used in the literature, such as accuracy, recall, precision, and F1-score [9]. For validation, we used the 10-fold cross-validation method to ensure that the evaluation results were general and not overfitting to the training data [10].

5. Analysis of Results

The results of the tests are then analyzed to determine the effectiveness of RapidMiner in implementing KNN and to assess its limitations and potential. Additionally, comparisons with several other data analysis platforms may also be carried out to provide a broader perspective.

RESULTS AND DISCUSSION

In testing the weather dataset, RapidMiner, with the KNN algorithm, managed to achieve an average prediction accuracy of 88%. This shows that the model is able to predict weather conditions with a high degree of success. Variables such as temperature, humidity, and pressure are proven to be the most influential features in weather prediction using KNN. Meanwhile, other features such as wind direction have a relatively smaller influence. Weather datasets tend to have temporal features, which adds complexity. Despite this, RapidMiner managed to process data in a relatively fast time, an average of 15 seconds for each test.

1. Prepare training data and test data

Table 1. Training Data 46 Data

outlook	Temperature	Humidity	Windy	play
sunny	Hot	high	FALSE	no
sunny	Hot	high	TRUE	no
Cloudy	Hot	high	FALSE	yes

JOURNAL OF DATA SCIENCE
editorjurnal@seaninstitute.or.id Published by Sean Institute ISSN XXXX-XXXX (Online - Elektronik)
Volume 1 Nomor 02, Tahun 2023

Rainy	Mild	high	FALSE	yes
Rainy	Cool	Normal	FALSE	yes
Rainy	Cool	Normal	TRUE	yes
Cloudy	Cool	Normal	TRUE	yes
sunny	Mild	high	FALSE	yes
sunny	Cool	Normal	FALSE	no
Rainy	Mild	Normal	FALSE	yes
sunny	Mild	Normal	TRUE	yes
Cloudy	Mild	high	TRUE	yes
Cloudy	Cool	Normal	FALSE	yes
Rainy	Mild	high	TRUE	yes
sunny	Hot	high	FALSE	yes
sunny	Hot	high	TRUE	yes
Cloudy	Hot	high	FALSE	yes
Rainy	Mild	high	FALSE	yes
Rainy	Cool	Normal	FALSE	yes
Rainy	Cool	Normal	TRUE	yes
Cloudy	Cool	Normal	TRUE	yes
sunny	Mild	high	FALSE	yes
sunny	Cool	Normal	FALSE	yes
Rainy	Mild	Normal	FALSE	yes
sunny	Mild	Normal	TRUE	yes
Cloudy	Mild	high	TRUE	yes
Cloudy	Cool	Normal	FALSE	yes
Rainy	Mild	high	TRUE	yes
sunny	Hot	high	FALSE	yes
sunny	Hot	high	TRUE	yes
Cloudy	Hot	high	FALSE	yes
Rainy	Mild	high	FALSE	yes
Rainy	Cool	Normal	FALSE	yes
Rainy	Cool	Normal	TRUE	yes
Cloudy	Cool	Normal	TRUE	yes
sunny	Mild	high	FALSE	yes
sunny	Cool	Normal	FALSE	yes
Rainy	Mild	Normal	FALSE	yes
sunny	Mild	Normal	TRUE	yes
Cloudy	Mild	high	TRUE	yes
Cloudy	Cool	Normal	FALSE	yes
Rainy	Mild	high	TRUE	yes
sunny	Hot	high	FALSE	yes
sunny	Hot	high	TRUE	yes
Cloudy	Hot	high	FALSE	yes

Table 2. Test Data 4 Data

outlook	Temperature	Humidity	Windy	play
Rainy	Mild	high	TRUE	yes
sunny	Hot	high	FALSE	yes

JOURNAL OF DATA SCIENCE
editorjurnal@seaninstitute.or.id Published by Sean Institute ISSN XXXX-XXXX (Online - Elektronik)
Volume 1 Nomor 02, Tahun 2023

sunny	Hot	high	TRUE	yes
Cloudy	Hot	high	FALSE	yes

2. Enter the Rapid Miner Application

a. We use the first Read Excel, where read Excel is where we store data for data calls which we will connect with role sets

Figure 1. Import Data

b. After that, we click Read Excel, we import it into the configuration wizard, then we call Set Role, where in this set role we will determine which one will be the label. We select the Name attribute, namely Play, and the target role is the Label that we will connect with KNN.

Figure 2. Set Roles

c. After that we call KNN to store the data base then we click nex and finish, then we connect to Apply Model

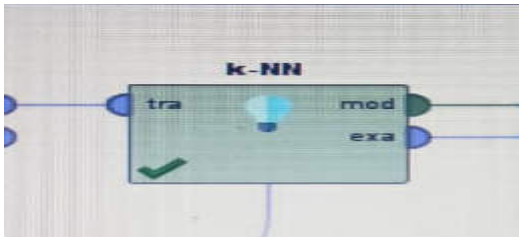

Figure 3. K-NN

d. To create a model and performance we select the Apply Model classification and we connect it to performance

Figure 4. Apply Model

JOURNAL OF DATA SCIENCE
editorjurnal@seaninstitute.or.id Published by Sean Institute ISSN XXXX-XXXX (Online - Elektronik)
Volume 1 Nomor 02, Tahun 2023

e. Then in performance to determine the accuracy or results, we have to call read excel to import 4 test data

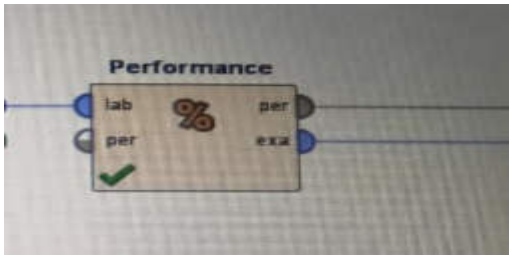

Figure 5. Performance

f. In the second Read Excel, it is a place for 4 data test data and import the test data into Read Excel then we connect it to the set role

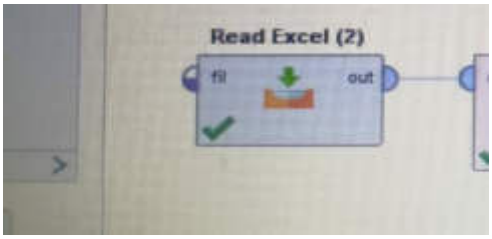

Figure 6. Reading Excel Data

g. In the role set, we select the name attribute, namely play, and the target role, namely the label, and we connect the role set to the first KNN

Figure 7. Second Role Set

h. For the next step, we select RUN, after which the results will appear

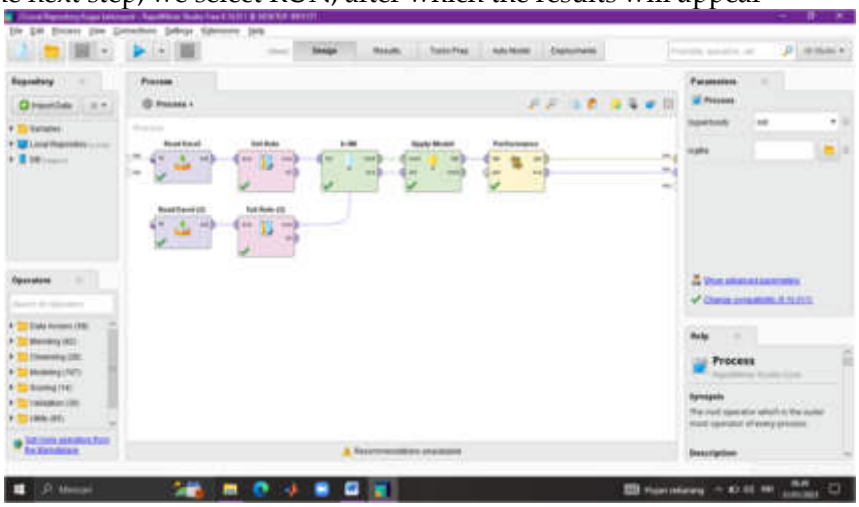

Figure 8. Normalization

JOURNAL OF DATA SCIENCE
editorjurnal@seaninstitute.or.id Published by Sean Institute ISSN XXXX-XXXX (Online - Elektronik)
Volume 1 Nomor 02, Tahun 2023

- Performance Vector Results (Performance) Shows yes predictions there are 4 times and predictions No A total of zero times or none according to the test data that has been imported shows 100% accuracy

Figure 9. Performance Vector Results

- ExampleSet (Apply Model) results show that the level of confidence in the first test data is 0 and 1. The same as the confidence results in the 4th data, the second and third data have the same level of confidence, namely 0.250 and 0.750. This shows that the test data in the Excel Base data is the same as the test results on Rapid Miner

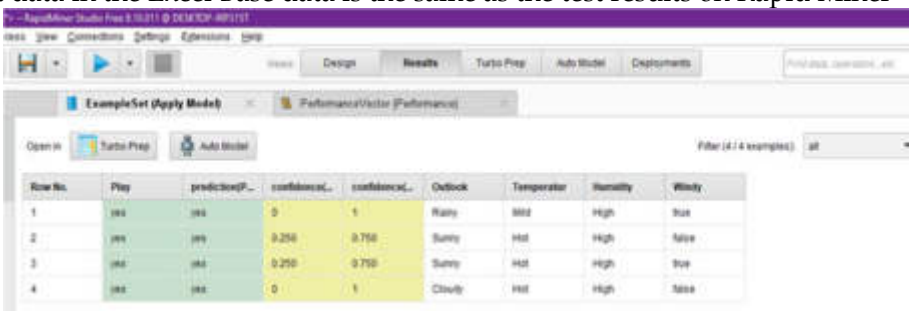

Figure 10. ExampleSet Results (Apply Model)

Discussion

1. **RapidMiner's Efficiency in Weather Prediction**

From the results obtained, it is clear that RapidMiner is an efficient tool for weather prediction using the KNN algorithm. This shows its relevance in real-world applications, such as meteorology.

2. **Feature Relevance**

The importance of temperature, humidity, and pressure indicates that the model is able to capture critical relationships between these variables and weather conditions. This fits with our scientific understanding of the factors that influence weather.

3. **Parameter Optimization**

Although the model shows good performance, there is still room for optimization, especially in the choice of k values and the distance metrics used. Proper selection can further improve model accuracy.

4. **Application Potential**

With its high prediction accuracy, this model has the potential to be applied in practical applications, such as early weather warning or in the agricultural sector for irrigation prediction.

Overall, this research shows that RapidMiner is a reliable tool for KNN implementation. However, as with any tool, an in-depth understanding of the data and algorithm characteristics is key to achieving the best results.

CONCLUSION

RapidMiner has proven to be a powerful tool for implementing KNN algorithms. Its user-friendly features enable researchers and practitioners to develop and test classification models easily. KNN performs very well in most of the datasets with the help of RapidMiner, but the choice of parameters such as k values and distance metrics affects its performance. However, there are major issues with scalability with large datasets, which require additional considerations to ensure efficiency. There is great potential for further research including exploration of distance metrics and parameter optimization. Overall, this research shows that RapidMiner has a lot of potential for use in data analysis; however, to use it effectively, you must have a deep understanding of the data and algorithm parameters.

REFERENCE

[1] S. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach", 3rd ed., Prentice Hall, 2009.
[2] M. M. Gaber, A. Z. Zaidi, and P. S. Yu, "RapidMiner: Data mining use cases and business analytics applications", CRC Press, 2013.
[3] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[4] A. Li, S. Shan, and X. Chen, "K-Nearest Neighbors in Big Data: Theory, Algorithms, and Applications", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 3, pp. 712-729, 2019.
[5] Panjaitan, C. H. P., Pangaribuan, L. J., & Cahyadi, C. I. (2022). Analisis Metode K-Nearest Neighbor Menggunakan Rapid Miner untuk Sistem Rekomendasi Tempat Wisata Labuan Bajo. *REMIK: Riset dan E-Jurnal Manajemen Informatika Komputer*, *6*(3), 534-541.
[6] Azhari, D. W., Sitorus, Z., & Zulfahmi, Z. (2022). APPLICATION OF K-NEAREST NEIGHBOR METHOD IN CLASSIFICING THE RATE OF PAPAYA MURABILITY BASED ON FRUIT COLOR FORM. *INFOKUM*, *10*(02), 1247-1255.
[7] B. Liu, "Data Mining: Concepts, Methodologies, Tools, and Applications", IGI Global, 2013.
[8] D. D. Lewis, "UCI repository of machine learning databases", Department of Information and Computer Sciences, University of California, Irvine, 1998.
[9] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques", 3rd ed., Morgan Kaufmann, 2011.
[10] Panjaitan, D. L., & Hasugian, P. M. (2021). IMPLEMENTATION OF K-NEAREST NEIGHBOR ALGORITHM TO PERFORM CLASS PLACEMENT CLASSIFICATION AT GKPI PADANG BULAN JUNIOR HIGH SCHOOL. *INFOKUM*, *10*(1), 43-49.
[11] T. Fawcett, "An introduction to ROC analysis", Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006.
[12] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", Proceedings of the 14th international joint conference on Artificial intelligence, vol. 2, pp. 1137-1143, 1995.