Journal of Information Systems and Informatics
Vol. 5, No. 2, June 2023 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v5i2.485 Published By DRPM-UBD

Forecasting Brown Sugar Production Using k-NN Minkowski Distance and Z-Score Normalization

Elindra Ambar Pambudi[1], Akhitya Ghany Fahrezi[2], Abid Yanuar Badharudin[3]

[1,2,3]Informatics Engineering Departement, Universitas Muhammadiyah Purwokerto, Banyumas, Indonesia
Email: [1]elindraambarpambudi@ump.ac.id, [2]akhityaghanyfahrezi.sanez@gmail.com, [3]zaidsoft.indonesia@gmail.com

Abstract

The demand for brown sugar products often falls below the level of production, resulting in unsold goods when market demand surpasses the production capacity. This paper addresses the challenge faced by many brown sugar businesses in estimating production yields. Another issue, apart from production uncertainty, is the presence of a dataset with a significant nominal range. The study focuses on a specific brown sugar producing company in Indonesia. To address the production estimation problem, this research proposes the use of k-NN supervised learning as a forecasting method. However, instead of relying solely on k-NN, the study suggests employing z-score normalization to handle the dataset's large nominal range. The production data used for analysis spans from March 2019 to February 2022, comprising 144 weekly records. The dataset is divided into training and testing data, employing an 8:2 split validation ratio. The proposed method consists of several steps, including data normalization using z-score, processing k-NN based on the Minkowski distance, and concluding with the de-normalization process. The results demonstrate the successful implementation of the proposed method in predicting production levels. The evaluation indicates an average error margin of 3.34%, which is below the 5% threshold. The evaluation of predictive data for k-NN with z-score normalization proves effective in forecasting brown sugar production uncertainty and addressing the challenge of a large nominal range.

Keywords: Forecasting, KNN, Z-Score Normalization, Minkowski.

1. INTRODUCTION

The brown sugar industry represents a traditional household trade that has been passed down through generations. The production process involves simple methods and equipment [1]. Sugar is a crucial staple in the Indonesian diet, consumed for its energy, flavor-enhancing properties, and as a raw material in the food and beverage industry [2]. These factors contribute to the extensive market advantage enjoyed by brown sugar. However, the consumption patterns of brown sugar exhibit significant fluctuations. Demand for brown sugar products often fails

to match production levels, with instances where market demand exceeds production capacity. These challenges in estimating sales and production levels are commonly faced by brown sugar businesses. Consequently, there is a need for a technique to predict weekly production levels.

Forecasting entails making informed predictions about the future state of a subject, phenomenon, or process [3]. Prediction and estimation play a vital role as they underpin the majority of business decisions, relying on forecasts of future outcomes [4]. Forecasting techniques employ mathematical calculations to anticipate future situations. The significance of forecasting stems from its crucial role in operations and production management [5]. Building upon previous research [3]–[5], we propose a forecasting approach to address the uncertainty associated with brown sugar production, as highlighted in the preceding paragraph.

Several studies have been conducted on the prediction of sugar production, including the work presented in [6]. Its focuses on predicting crop production yields using K-Means and a modified version of the k-NN algorithm. The research primarily revolves around determining categories through clustering and classifying input data into five specific categories using the modified k-NN approach. Another study by [7] employs data mining techniques to develop a model for estimating sugarcane yield, considering variables such as meteorological factors and crop management. Data mining techniques, including random forest, boosting, and support vector machines, are utilized in this study. Notably, papers [6] and [7] primarily concentrate on the prediction methods, with [6] focusing on predicting major crop yields and [7] discussing the development of a model to predict sugarcane yield based on meteorological and crop management variables. However, neither of these studies delves into detail regarding the preprocessing procedures. Thus, we identify a gap in utilizing preprocessing techniques, particularly data normalization, in the context of brown sugar production datasets.

In this research, we employ proposed method the K-Nearest Neighbor (k-NN) machine learning approach due to its proven accuracy, efficiency, and low error ratio [8]. K-Nearest Neighbor is known for its simplicity in implementation [9], fast training on data, and robustness even without feature selection [10]. Additionally, it performs well on small datasets [11], [12], making it widely used in forecasting techniques for various problems. While the algorithm is considered straightforward, researchers have continuously sought to improve its predictive performance. K-Nearest Neighbor is applicable to both regression and classification tasks and has been successfully implemented in diverse fields [13]–[15]. Furthermore, paper [16] demonstrates the suitability of k-NN for time series forecasting. Considering the reliability of k-NN in forecasting and the small dataset available for our research, we have chosen k-NN as the preferred algorithm for our forecasting purposes.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

The k-NN algorithm operates by determining the k-NN results based on the shortest distance. Various methods exist for approximating distance, including the Minkowski distance approach. While the Euclidean distance is commonly employed by most k-NN algorithms, this research introduces an update by utilizing the Minkowski distance approach. Mailagaha Kumbure and Luukka (2021) have suggested that the Euclidean distance is often suboptimal for practical problems, and better outcomes can be achieved by generalizing it. By employing the Minkowski distance approach, the proposed method can identify more suitable nearest neighbors for the target [17].

Another challenge encountered, apart from the uncertainty in sugar production, is the presence of attributes with a relatively large nominal range. Attributes with large value ranges can unjustly dominate the results of the classification process solely due to their numerical magnitude. Hence, normalization is necessary to balance the range of values [18]. In order to address the challenges associated with estimating production levels under uncertainty and dealing with a dataset characterized by a significant nominal range, this paper proposes a technique that combines the use of the Minkowski distance k-NN and z-score normalization to forecast weekly production levels.

2. METHODS

2.1 Collection Data

Collection of data sets of brown sugar production starts from March 2019 to February 2022. The obtained production data is weekly data with a total of 144. The data set that has been collected will be split into data train and data test. Splitting the data using common ratio of 8:2, where the percentage of 80% training data and 20% testing data.

2.2 Proposed Method

The K-Nearest Neighbor method is widely recognized as a popular classification algorithm used for predicting classes based on neighboring records or samples [19]. In the proposed method, the original image processing involves a series of steps, as illustrated in Figure 1.

Step 1: Normalization process to produce a range balanced value using z-score normalization.
Step 1: The next step is determining k value (the number closest neighbor)
Step 2: Calculate the Minkowski distance for each object against the given training data. Minkowski Distance or Minkowski metric is a metric in a normed vector space that can be thought of as a generalization of the Euclidean distance and the Manhattan distance [20]. Minkowski distance

with the symbol p (where p is an integer) between the two-point shown in equation 1.

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

Where d is distance between x and y , i = each data, n = amount of data x_i = data at the center of the i cluster, y_i = data on each data to i, p = powers.

Step 3: Sorting objects into groups that have the smallest distance, and check objects from nearest neighbors (range of K values).

Step 4: Calculating the average of object values in k range using the closest Nearest Neighbor category (K range), then the calculated query instance value can be predicted. Equation 3 below shows the average calculation of object values in the K range.

$$Y=\frac{1}{k}\sum_{i=1}^{k}Yi \quad (2)$$

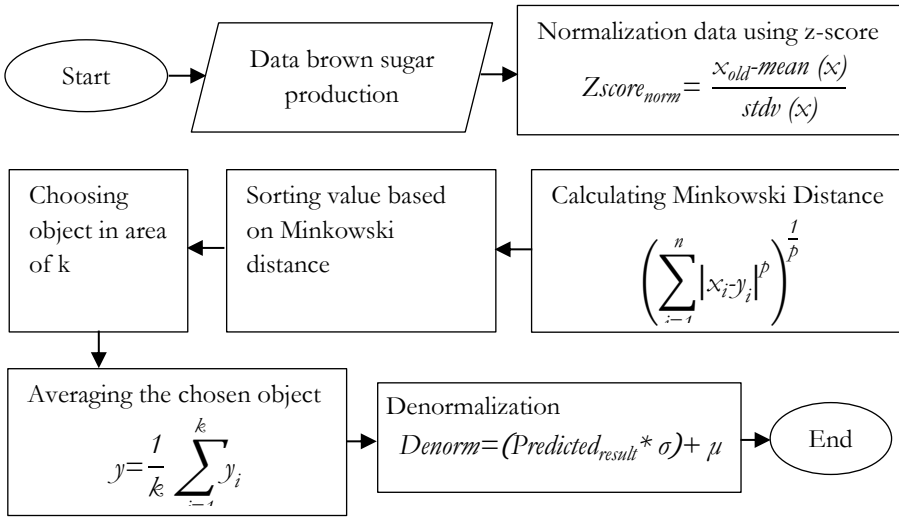

Figure 1. The Proposed Method

3.1. Evaluation

Evaluation of prediction of this paper uses margin error according paper [12]. *Margin Error* (ME) can be used for knowing the difference between prediction and real data.

$$ME=\frac{Value_{pred}-Value_{real}}{Value_{real}}x\ 100 \quad (3)$$

3. RESULTS AND DISCUSSION

The input data will undergo calculation using the k-NN method to forecast production results. The steps involved in the production forecasting calculation are as follows:

3.1. Normalization

Since the data in many attributes exhibit varying ranges, it is necessary to normalize them. To achieve this, the datasets are transformed using the z-score method, which involves calculating the minimum and maximum values for each attribute. The modified values resulting from this normalization process can be observed in Table 2. Notably, these processed values demonstrate a balanced range. For reference, Table 1 displays the original values prior to any preprocessing steps.

Table 1. Sample Data Normalization

x week of *n* month	Data Production (/kg)	Data Normalization
1st Week in 3rd month 2019	6510	-1,28712
2nd Week in 3rd month 2019	7720	-0,673768
3rd Week in 3rd month 2019	7140	-0,967771
4th Week in 3rd month 2019	8030	-0,516628
1st Week in 4th month 2019	8330	-0,364558
2nd Week in 4th month 2019	8330	-0,364558
3rd Week in 4th month 2019	6070	-1,51016
4th Week in 4th month 2019	6710 Kg	-1,18574
…	…	…
1st Week in 2nd month 2022	9880 Kg	0,421141
2nd Week in 2nd month 2022	8430 Kg	-0,313867
3rd Week in 2nd month 2022	9530 Kg	0,243725
4th Week in 2nd month 2022	9250 Kg	0,101792

3.2. K-NN

The fundamental principle underlying k-NN is to identify the shortest distance between the data under evaluation and its nearest neighbors. Following the data split and normalization steps, the algorithm proceeds to calculate the distance between the attribute values of the testing data and each training data attribute using the Minkowski distance Eq. 1.

Table 2. Minkowski Distance

No	Calculation Minkowski Distance	Result
109	ABS(((-0.775148 - -0.92215)^3)+((-1.59126 - -0.316402)^3)+((-1.21615 - -1.0945)^3)+((-1.57605 - -1.72559)^3)+((-2.13365 - -0.182073)^3)+((-1.73826 - -	0.210973

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

No	Calculation Minkowski Distance	Result
	0.24797)^3)+((-1.36822 - -0.620543)^3)+((0 - -0.531835)^3)+((0 - -0.577457)^3)+((0 - -0.785286)^3)+((0 - -1.8143)^3)+((0 - 0.208242)^3)+((0 - -0.882612)^3)+((0 - 0.177828)^3)+((0 - 0.0713782)^3)+((0 - 0.411003)^3)+((0 - -0.567319)^3)+((0 - -0.648423)^3)+((0 - -1.40878)^3)+((0 - -0.349351)^3)+((0 - -1.05394)^3)+((0 - -0.544508)^3)+((0 - -1.02353)^3)+((0 - 0.436348)^3)+((0 - -0.653492)^3)+((0 - 0.421141)^3)+((0 - -0.313867)^3)+((0 - 0.243725)^3)+((0 - 0.101792)^3))^(1/3)	
86	ABS(((0.89256 - -0.92215)^3)+((0.0206879 - -0.316402)^3)+((-1.29726 - -1.0945)^3)+((0.37045 - -1.72559)^3)+((0.198104 - -0.182073)^3)+((0.512383 - -0.24797)^3)+((-1.74333 - -0.620543)^3)+((0.580815 - -0.531835)^3)+((-1.53043 - -0.577457)^3)+((1.17389 - -0.785286)^3)+((0.689799 - -1.8143)^3)+((0.649247 - 0.208242)^3)+((-1.57605 - -0.882612)^3)+((-0.541973 - 0.177828)^3)+((-0.780217 - 0.0713782)^3)+((-0.430455 - 0.411003)^3)+((-1.42398 - -0.567319)^3)+((-0.607871 - -0.648423)^3)+((-0.209953 - -1.40878)^3)+((-0.425386 - -0.349351)^3)+((-3.16266 - -1.05394)^3)+((-0.460869 - -0.544508)^3)+((-0.785286 - -1.02353)^3)+((-0.775148 - 0.436348)^3)+((-1.59126 - -0.653492)^3)+((-1.21615 - 0.421141)^3)+((-1.57605 - -0.313867)^3)+((-2.13365 - 0.243725)^3)+((-1.73826 - 0.101792)^3))^(1/3)	0.906285
N (rank)	…	…

a) Sorting process is carried out ascending based on result of minkowski distance. The value of the distance between data test and data train is sorted from the lowest value. This paper used k=3.

Table 3. Sorting Proccess

No	Calculation Minkowski Distance	Rank
109	0.210973	1
86	0.906285	2
108	1.17139	3
N (rank)	…	…

b) Calculating mean object in range of k

To calculate the mean object, Eq. 2 is employed. Here, we provide the sample mean object value within the range of k = 3, with neighbor values derived from the testing data. Based on the results presented in Table 3, we can conclude that the data utilized corresponds to numbers 1 (neighbor = 0.210973), 2 (neighbor =

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

0.906285), and 3 (neighbor = 1.17139). As a result, the predicted value for the week of March 2022, month 3, can be computed as follows:

$$\frac{(0{,}436347719 + 0{,}071378161 + -1{,}023530512)}{3} = -0{,}171934878$$

Once the predicted value has been obtained, the subsequent step involves denormalizing the prediction value to its original, real value.

3.3. Evaluation

During the evaluation stage, a set of test data is classified using the established classification model. The classification process employs the Minkowski distance measurement. It is important to note that the test data utilized in this phase is distinct from the data used for training. A comparison between the evaluation results of the proposed method and the standard k-NN can be observed in Table 4.

Tabel 4. Evaluation prediction of data production using KNN

ID *term*	Real Data	KNN		KNN ZScore	
		Prediction	Margin Error	Prediction	Margin Error
1st Week in 3rd month 2022	8920	9280	4,1%	8710	2,35%
2nd Week in 3rd month 2022	8090	8635	6,7%	8290	2,47%
3rd Week in 3rd month 2022	8590	7471	13%	8233	4,15%
4th Week in 3rd month 2022	8890	8141	8,4%	8461	4,81%
1st Week in 4th month 2022	8100	7938	1,9%	7978	1,5%
2nd Week in 4th month 2022	8330	8926	7,1%	9043	8,56%
3rd Week in 4th month 2022	8350	8128	2,6%	8128	2,65%
4th Week in 4th month 2022	8500	7911	6,9%	8480	0,23%

Referring to the data presented in Table 4, the prediction results indicate that the margin of error for the aforementioned data falls below 5%. The utilization of z-score normalization as a preprocessing technique yields an average margin of error of 3.34%. Consequently, it can be inferred that the prediction model, combined with the z-score normalization, demonstrates a satisfactory level of accuracy.

The provided data in Table 4 presents the prediction results and margin of error for both the standard k-NN and the k-NN with z-score normalization methods.

For the standard k-NN predictions, the margin of error ranges from 1.9% to 13%. Meanwhile, the k-NN with z-score normalization shows a lower margin of error, varying from 0.23% to 8.56%. Comparing the two methods, it is evident that the k-NN with z-score normalization consistently yields lower margin errors across the predictions. This indicates that the application of z-score normalization effectively reduces the variability and improves the accuracy of the production forecasts.

The results demonstrate that the proposed k-NN with z-score normalization approach provides more reliable predictions compared to the standard k-NN. The average margin of error achieved using the z-score normalization technique is 3.34%, indicating a high level of precision in forecasting the production levels of brown sugar.

Overall, these findings support the effectiveness of employing the k-NN algorithm with z-score normalization for estimating production levels in the brown sugar industry. The reduced margin of error implies improved decision-making capabilities and better planning for brown sugar businesses, ultimately leading to more efficient operations and resource allocation.

4. CONCLUSION

The K-Nearest Neighbor algorithm can predict the amount of brown sugar production, so that it can help overcome and minimize excess product availability and can make decisions to increase or reduce brown sugar production. In some case the result of KNN has weaknesses, it caused the minimize the dataset, so we have to collect more dataset, and then analysis k, maybe we can use more k value The K-Nearest Neighbor algorithm with data preprocessing using z-score normalization is better than K-Nearest Neighbor without data preprocessing, with an average margin of error of 3.34%.

REFERENCES

[1] A. Rifa'i, I. M. Sudarma, and Widhianthini, "Strategi Pengembangan Usaha Industri Gula Merah Tebu di Kabupaten Tulungagung Provinsi Jawa Timur," *Jurnal Agribisnis dan Agrowisata* , vol. 8, no. 3, 2019.

[2] H. Gula Pasir Di Jakarta, A. Andri Wiliyana, and M. Yamin Darsyah, "Perbandingan Metode Arima Dan Moving Average Pada Kasus Comparison of The Use of Arima And Moving Average Methods in the Case of Granulated Sugar Price in Jakarta," in *Prosiding Seminar Nasional Mahasiswa Unimus*, 2018, pp. 361–367.

[3] T. K. Belyaeva, E. E. Egorov, T. K. Potapova, T. L. Shabanova, and M. Y. Shlyakhov, "Implementation of the Division Model of Pedagogical Labor in the Teacher Training System of a New Type," in *the 21st Century from the Positions of Modern Science: Intellectual, Digital and Innovative Aspects*, E. G. Popkova and B. S. Sergi, Eds., Cham: Springer International Publishing, 2020, pp. 430–438.

[4] P. Gupta *et al.*, "Implementation of demand forecasting – A comparative approach," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1742-6596/1714/1/012003.

[5] Y. Y. Fan, "Demand Prediction of Production Materials And Simulation Of Production Management," *International Journal of Simulation Modelling*, vol. 21, no. 4, pp. 720–731, Dec. 2022, doi: 10.2507/IJSIMM21-4-CO20.

[6] A. Suresh, P. G. Kumar, and M. Ramalatha, "Prediction of major crop yields of Tamilnadu using K-means and Modified KNN," in *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, 2018, pp. 88–93. doi: 10.1109/CESYS.2018.8723956.

[7] R. G. Hammer, P. C. Sentelhas, and J. C. Q. Mariano, "Sugarcane Yield Prediction Through Data Mining and Crop Simulation Models," *Sugar Tech*, vol. 22, no. 2, pp. 216–225, 2020, doi: 10.1007/s12355-019-00776-z.

[8] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient kNN classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, 2016, doi: https://doi.org/10.1016/j.neucom.2015.08.112.

[9] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans Neural Netw Learn Syst*, vol. 29, no. 5, pp. 1774–1785, May 2018, doi: 10.1109/TNNLS.2017.2673241.

[10] A. Pamuji, "Performance of the K-Nearest Neighbors Method on Analysis of Social Media Sentiment," *JUISI*, vol. 07, no. 01, 2021.

[11] M. Suyal and P. Goyal, "A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning," *International Journal of Engineering Trends and Technology*, vol. 70, no. 7. Seventh Sense Research Group, pp. 43–48, Jul. 01, 2022. doi: 10.14445/22315381/IJETT-V70I7P205.

[12] R. N. Sukmana, A. Abdurrahman, and Y. Wicaksono, "Implementasi K-Nearest Neighbor Untuk Menentukan Prediksi Penjualan:(Studi Kasus: Pt Maksiplus Utama Indonesia)," *Jurnal Teknologi Informasi dan Komunikasi*, vol. 9, no. 2, pp. 31–37, 2020.

[13] M. Nanja and P. Purwanto, "Metode K-Nearest Neighbor Berbasis Forward Selection Untuk Prediksi Harga Komoditi Lada," *Pseudocode*, vol. 2, no. 1, pp. 53–64, Aug. 2015, doi: 10.33369/pseudocode.2.1.53-64.

[14] D. Srianto and E. Mulyanto, "Perbandingan K-Nearest Neighbor Dan Naive Bayes Untuk Klasifikasi Tanah Layak Tanam Pohon Jati," *Techno. Com*, *15*(3), pp.241-245 2016.

[15] L. Yang, S. Liu, S. Tsoka, and L. G. Papageorgiou, "A regression tree approach using mathematical programming," *Expert Syst Appl*, vol. 78, pp. 347–357, 2017, doi: https://doi.org/10.1016/j.eswa.2017.02.013.

[16] S. Tajmouati, B. El Wahbi, A. Bedoui, A. Abarda, and M. Dakkoun, "Applying k-nearest neighbors to time series forecasting: two new approaches," Mar. 2021.

[17] M. Mailagaha Kumbure and P. Luukka, "A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance," *Granular Computing*, vol. 7, no. 3, pp. 657–671, 2022, doi: 10.1007/s41066-021-00288-w.

[18] S. Sinsomboonthong, "Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification," *Int J Math Math Sci*, vol. 2022, p. 3584406, 2022, doi: 10.1155/2022/3584406.

[19] B. Zaman, A. Rifai, and M. Hanif, "Komparasi Metode Klasifikasi Batik Menggunakan Neural Network Dan K-Nearest Neighbor Berbasis Ekstraksi Fitur Tekstur", *J. Inf. Syst. Informatics*, vol. 3, no. 4, pp. 582-595, Dec. 2021. doi: https://doi.org/10.51519/journalisi.v3i4.213

[20] V. B. S. Prasath *et al.*, "Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier -- A Review," Aug. 2017, doi: 10.1089/big.2018.0175.