Jurnal Publikasi Ilmu Komputer dan Multimedia

E-ISSN: 2808-8999
P-ISSN: 2808-9375

Application of the K-Nearest Neighbor Algorithm in the Data Mining Process to Predict Drug Sales at Pratama Haji Medan-Pancing Clinic

Indra Nasution[1]*, Rezkinah Rambe[2] , Khairil Putra[3], Muhammad Dafa[4], and Muhammad Syahputra Novelan[5]

[1] Universitas Pembangunan Panca Budi; Email: sutanindrabungsu@gmail.com
[2] Universitas Pembangunan Panca Budi; Email: renarrambe@gmail.com
[3] Universitas Pembangunan Panca Budi; Email: khairilputra0805@yahoo.co.id
[4] Universitas Pembangunan Panca Budi; Email: dafaalfikri75@gmail.com
[5] Universitas Pembangunan Panca Budi; Email: putranovelan@dosen.pancabudi.ac.id
***Author Correspondence**

Abstract: Pratama Haji Medan-Pancing Clinic is a healthcare facility that routinely sells medications to patients. However, the current manual drug inventory management process poses risks such as delayed procurement and overstocking. To address this issue, this study aims to implement a data mining approach using the K-Nearest Neighbor (KNN) algorithm to predict drug sales at Klinik Pratama Haji Medan-Pancing. A quantitative research method was employed, utilizing historical drug sales data from the past two to three years. The data underwent a thorough process of assessment, cleaning, and transformation before being processed using the K-Neighbor Classifier from the scikit-learn library. The results demonstrated that the KNN method achieved a prediction accuracy rate of 88.9%, indicating its effectiveness in forecasting drug sales. By implementing this predictive system, Klinik Pratama Haji Medan-Pancing can improve the efficiency of inventory management, reduce the risk of stock shortages or surpluses, and support faster, data-driven decision-making. In conclusion, the KNN algorithm proves to be a feasible predictive solution for drug sales systems in clinics and holds potential for further development in intelligent and integrated pharmacy management.

Keywords: K-Nearest Neighbor; data mining; sales prediction; drug inventory.

Abstract: *Klinik Pratama Haji Medan-Pancing merupakan salah satu fasilitas kesehatan yang rutin melakukan penjualan obat-obatan kepada pasien. Namun, pengelolaan stok obat yang masih dilakukan secara manual menimbulkan risiko keterlambatan pengadaan dan kelebihan stok, sehingga diperlukan sistem yang mampu memprediksi kebutuhan penjualan obat secara akurat. Penelitian ini bertujuan untuk menerapkan metode data mining dengan algoritma K-Nearest Neighbor (KNN) dalam memprediksi penjualan obat di Klinik Pratama Haji Medan-Pancing. Penelitian ini menggunakan pendekatan kuantitatif dengan pengumpulan data penjualan obat selama dua hingga tiga tahun terakhir yang kemudian diolah menggunakan K-Neighbor Classifier dari pustaka scikit-learn. Data yang digunakan melewati proses pemeriksaan, pembersihan, dan transformasi sebelum dilakukan klasifikasi dan prediksi. Hasil penelitian menunjukkan bahwa metode KNN mampu memberikan tingkat akurasi prediksi sebesar 88,9%, yang menunjukkan efektivitas metode ini dalam membantu klinik memperkirakan kebutuhan obat secara lebih tepat. Dengan penerapan sistem prediksi berbasis KNN, Klinik Pratama Haji Medan-Pancing dapat meningkatkan efisiensi dalam pengelolaan stok obat, mengurangi risiko kekurangan maupun kelebihan stok, serta mendukung pengambilan keputusan yang lebih cepat dan berbasis data. Kesimpulannya, algoritma KNN terbukti layak digunakan sebagai solusi prediktif dalam sistem penjualan obat di klinik, dan dapat dikembangkan lebih lanjut untuk pengelolaan farmasi yang lebih cerdas dan terintegrasi.*

Kata Kunci: *K-Nearest Neighbor; data mining; prediksi penjualan; pengelolaan obat*

Received: 01 Februari 2025
Revised: 10 Maret 2025
Accepted: 30 April 2025
Published: Mei 2025

DOI: https://doi.org/10.55606/jupikom.v4i2.4081

1. Introduction

The Pratama Haji Medan-Pancing Clinic located on Jl. Williem Iskandar No.113 E, Sidorejo, Medan Tembung District, is one of the first-level health service facilities that has an important role in helping the community in handling disease complaints. As a healthcare provider, this clinic needs to ensure the availability of adequate stock of medicines and in accordance with patient demands. Drug sales activities in this clinic occur every day and require a good recording system so that the data produced can be easily analyzed and used for decision-making related to stock management [1]. However, in practice, the Pratama Haji Medan-Pancing Clinic still uses a manual recording method using a spreadsheet application. The use of this manual recording system has several drawbacks, such as being less efficient, prone to input errors, and making it difficult to search and analyze sales data historically.

Several methods and approaches have been used in previous research to overcome the problem of sales prediction, one of which is the classification algorithm in data mining techniques. One of the methods that is often used is the K-Nearest Neighbor (KNN) algorithm, which is proven to be able to provide fairly accurate prediction results. This method classifies data based on the proximity or similarity of new data to previous data that is already known to the class [2] [3] . Research by [4] Shows 100% accuracy in data testing. Meanwhile, research by [5] successfully implemented KNN to predict the sales of the best-selling motorcycles with an accuracy rate of 96.15%, and Mulyati et al. (2020) successfully used this method to predict national exam passing. Based on these results, KNN has advantages in ease of implementation and effectiveness in classification of similarity-based data.

However, KNN also has disadvantages, such as reliance on the quality and volume of historical data, as well as degraded performance when handling data with high dimensions or very large amounts of data [6] [7]. However, in the context of drug sales data at the Medan-Pancing Pratama Haji Clinic which tends to be structured and not too large, the KNN is still relevant and can be used optimally [8]. The main problem faced by the Pratama Haji Medan-Pancing Clinic is the absence of a prediction system to support decisions in drug stock management. The inability to manage stocks appropriately has the potential to cause a shortage of drugs needed by patients or a buildup of drugs that are rarely needed, even to the point of causing losses due to expiration.

As a solution to this problem, this study proposes an approach to predict drug sales using the K-Nearest Neighbor (KNN) algorithm. By implementing this method, it is hoped that the Pratama Haji Medan-Pancing Clinic can predict the type and amount of drugs that are most needed based on historical sales patterns. The results of this prediction will help in making better decisions regarding the ordering and provision of drug stock. Therefore, this study focuses on the application of the KNN method in building a decision support system that is able to improve operational efficiency and service to patients at the Pratama Haji Medan-Pancing Clinic.

2. Literature Review

2.1. Data Mining

Data Mining is a form of data mining that is used to extract knowledge from large amounts of data [9]. Data mining is necessary in making predictions for relationships that have meanings, patterns, and tendencies by examining a large set of data stored in storage using statistical or mathematical pattern recognition techniques [10]. In Data Mining, the process of finding patterns or useful information from data that has been selected or processed is called Knowledge Data Discovery (KDD) [11].

2.2. K-Nearest Neighbor Algorithm

K-Nearest Neighbor is a method commonly used in Data Mining projects [12]. This method uses the Supervised Learning algorithm. Supervised Learning involves using data that has been marked from previous outcomes [13]. The goal of Supervised Learning is to train computer models that can learn patterns in data and make accurate predictions of unknown data [14]. The goal of this algorithm is to classify new objects based on attributes and training data. This algorithm works based on the shortest distance from the data request to the training data to determine its KNN [15]. One way to calculate the proximity or distance of each piece of data or neighbors in the data is to use the Eucledian Distance method [16].

Euclidean Distance is a method often used to calculate distances between singles. This distance is used to test the interpretation of the approximate distance between two objects [17]. The formula for calculating Eucledian Distance is as follows:

$$\boldsymbol{D(x,y)} = \sqrt{\sum_{i=1}^{n} \mathbf{Xi} - \mathbf{Yi}^2} \quad (1)$$

Where:

$\boldsymbol{D(x,y)}$: Eucledian distance between two points x and y

n : Eucledian space dimensions

Xi,Yi : Coordinates of x and y points in the i-i dimension

In classifying using the KNN algorithm, we must determine the value of the k parameter, the value of k in the KNN is the number of closest neighbors, if k is worth 1, then the class of one nearest training data will become the class for the new test data, if k is worth 3, the three closest training data will be taken to become the class for the new test data [18].

2.3. Confusion Matrix

The Confusion Matrix is a table that declares the classification of the correct number of tests and the number of incorrect tests [19]. From the definition of the confusion matrix, several points in the confusion matrix are used to calculate precision, recall, and f1 score. Precision is a comparison between True Positive (TP) and the amount of data that is predicted to be positive, mathematically it can be seen below [20]:

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (2)$$

For the recall itself, it is a comparison between true positive (TP) and the amount of data that is actually positive. It can be mathematically stated as follows:

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (3)$$

While the F1 Score is the middle value of precision and recall. The best value of the F1 Score is 1 and the worst value is 0, mathematically it can be written as follows:

$$\frac{\mathbf{1}}{\boldsymbol{F1}} = \frac{\mathbf{1}}{\mathbf{2}}\left(\frac{\mathbf{1}}{\boldsymbol{Precision}} + \frac{\mathbf{1}}{\boldsymbol{Recall}}\right) \quad (4)$$

A good F1 Score value indicates that our classification model has good precision and recall.

2.4. Pre-Processing of Data

Data pre-processing is a stage to process raw data by eliminating some annoying problems during data processing [21]. This is due to data that is inconsistent in format. Through this process, the modeling of the KNN algorithm will run more effectively and efficiently [22]. The stages in pre-processing data are:

1. First of all, in the early stages of data preprocessing, the essential step is to perform data cleaning. This process involves re-selecting raw data to eliminate incomplete, irrelevant, or inaccurate entries. By doing this, we can avoid misunderstandings when analyzing such data.
2. The next step is data integration, which is necessary because data preprocessing involves combining data from various sources into a single dataset. It is important to ensure that data from different sources has a uniform format.
3. After that, we move on to the data transformation stage. As explained earlier, data coming from different sources may have a variety of formats. Therefore, it is necessary to make

format adjustments so that all data collected have a uniform structure, facilitating the data analysis process.

4. The final stage in data preprocessing is to reduce the amount of data, known as data reduction. The main goal is to reduce the sample of data without changing the results of the analysis. There are three techniques that can be applied at this stage, namely dimensionality reduction, numerosity reduction, and data compression.

2.5. Z-Score Normalization

Data normalization is a part of data preprocessing where the values in the dataset are readjusted to ease the processing process. This process is important because datasets often have different ranges of values for each of their attributes. Significant differences in the range of values between attributes can impair the optimal performance of attributes in a dataset. Therefore, normalization is carried out to equalize the scale of attribute values so that the data analysis process becomes more efficient [23].

Z-score normalization is a normalization technique in which data values are adjusted based on the mean value and standard deviation of the data [24]. In Z-Score, the data undergoes a transformation or change to create a new range of values, based on the range of values that pre-existed in the dataset. The formula used in Z-score is as follows:

$$\boldsymbol{Z = \frac{X-\mu}{\sigma}} \tag{5}$$

Where

Z : Z-Score

X : Data value

μ : Average data

σ : Standard deviation

The formula for finding Standard Deviation is as follows:

$$\boldsymbol{\sigma = \sum \frac{X_1-\mu}{n}} \tag{6}$$

Where

Xi : Data value

μ : Average data

σ : Standard deviation

n : Amount of data

3. Method

3.1. Research Object

This study implements Data Mining techniques using the K-Nearest Neighbor (KNN) method to predict the sales of drugs at the Pratama Haji Medan-Pancing Clinic. The purpose of this study is to build a prediction system that can help clinics in managing drug stocks more efficiently. The research method used in this study is a quantitative method, because the author relies on numerical data in the form of sales history to carry out the prediction process. The data will later be analyzed and used as a basis for building a predictive system based on the KNN algorithm.

3.2. Literature Studies

The literature study stage is carried out to understand more deeply the needs needed to solve the problem being studied, as well as to examine the method to be used, namely K-Nearest Neighbor. In this process, the author studies various references such as scientific journals, previous research articles, and relevant books. The goal is to obtain a strong and in-depth theoretical foundation as a foundation in the implementation of research and implementation of drug sales prediction systems.

3.3. Data Gathering

The data collection process aims to obtain information and facts relevant to the research topic. The data collected is in the form of historical data on drug sales at the Pratama Haji Medan-Pancing Clinic for the past two to three years. To obtain accurate and comprehensive data, the author conducted direct observations at the clinic location and conducted interviews with related parties. This interview aims to gain a deeper understanding of the management process and drug sales system that applies at the clinic.

3.4. Assessing Data

The data examination stage is carried out to identify potential problems contained in the data, as well as ensure that the data has good quality before being used in the analysis process. According to [25] Some common problems in data include missing values, which are missing values and are usually marked with NaN, which can be identified using the isnull() or isna() function combined with sum(). In addition, there are also invalid values which refer to data that do not make sense, duplicate values which are duplicate data that can be recognized by the duplicated() method, inaccurate values due to recording errors, and inconsistent values, which are inconsistent values in units or formats.

3.5. Data Cleansing and Transformation

After the check, the next stage is data cleaning. Issues found in the data, such as missing, duplicate, or inconsistent values, will be addressed using cleanup techniques. After the data is cleaned, the next process is data transformation, which is the process of converting raw data into a more structured form and ready to be used in the analysis process. This transformation includes creating data aggregations or summaries to match the needs of the K-Nearest Neighbor algorithm in the data mining process. This stage is important so that data can produce accurate and useful predictions in the context of managing drug stocks at the Pratama Haji Medan-Pancing Clinic.

3.6. Application of the K-Nearest Neighbor Method.

K-Nearest Neighbor classifies objects based on learning data that is closest to the object. The selection of attributes consists of neighbors n (commonly called k). The K parameter is determined based on the K value with the best performance during data training [26]. After the data has gone through the process of examination, cleaning, and transformation, the following is an example of the application of the k-nearest neighbor method in related research. Based on the number of types of drugs sold from 2024, which totals 45 different types of drugs, the classification class is determined based on the range of drug sales.

4. Results and Discussion

4.1. Research Data Processing

The first step in the application of the K-Nearest Neighbor Algorithm is to process the data obtained from the research location, the data is a sales list that is inputted using the Microsoft Excel application as in the table below:

Table 1. Sales Data of Pratama Haji Medan-Pancing Clinic in 2023

Date	Drug Name	Price	Sum	Total
01/01/2024	Novell Fluconazole 150mg Capsules	25000	1	25000
01/01/2024	Amlodipine Dexa 5mg Tablets	15000	2	30000
01/01/2024	Esomeprazole Etercon 40mg Tablet	16500	4	66000
01/01/2024	Metronidazole Novell 500mg Tablet	3000	5	15000
01/01/2024	Dexamethasone 0.5mg Tablets	5500	11	60500
01/01/2024	Metoclopramide 10mg Tablet	4000	12	48000
01/01/2024	Diazepam 5mg Tablets	8000	13	104000
01/01/2024	Methylprednisolone 4mg Tablet	6000	15	90000
	…….			
31/12/2024	Ibuprofen 200mg Tablets	5000	22	110000

31/12/2024	Sodium Valproate 500mg Tablet	6000	23	138000
31/12/2024	Metformin 500mg Tablet	4000	23	92000
31/12/2024	Ambroxol Syrup 60ml	8000	24	192000
31/12/2024	Dexymox Forte 500Mg	1100	25	27500
31/12/2024	Ofloxacin 200mg Novell Tablets	1000	34	34000

4.2. Application of K-Nearest Neighbor

K-Nearest Neighbor classifies objects based on learning data that is closest to the object. The distance itself is calculated based on the eucledian distance formula, the following is the data that will be used to classify:

Table 2. Data that has been processed

Drug Name	Price (Rp)	Sum Sold	Classification
Acyclovir KF 400mg Tablet	-0,4456	1,4484	Very Popular
Alerhis Loratadine 10mg Capsules	0,2280	-1,4438	Less Popular
Ambroxol Syrup 60ml	-0,1492	0,0106	Bestselling
Amlodipine Dexa 5mg Tablets	0,2280	-1,4273	Less Popular
Amoxicillin IF 500mg	-0,5264	1,1944	Very Popular
Atorvastatin Pratapa Nirmala 20 mg Tablets	-0,2839	0,1177	Bestselling
Betahistine Novell 6mg Tablets	-0,4994	0,9032	Bestselling
Bisoprolol Fumarate Dexa	-0,3917	0,4006	Bestselling
Cardio Aspirin 100mg Tablets	-0,3917	0,0724	Bestselling
Cefadroxil Berno 500mg Capsules	-0,4725	1,7808	Very Popular
Cefadroxil if D.Syr 60ml 125mg/5ml	0,4705	-1,4094	Less Popular
Cefixime Dexa 100mg Capsules	-0,3917	-0,0993	Bestselling
Cendesartan Dexa 16mg Tablets	-0,1761	0,5338	Bestselling
Cetirizine Dexa 5mg Syrup 60ml	0,2280	-1,4767	Less Popular
Dexamethasone 0.25mg Tablets	-0,4186	0,3196	Bestselling
Dexamethasone 0.5mg Tablets	-0,2839	0,4377	Bestselling
Dexymox Forte 500Mg	-0,5210	1,4965	Very Popular
Diazepam 5mg Tablets	-0,1492	0,0833	Bestselling
Diphenhydramine 25mg Tablet	-0,3917	-0,0018	Bestselling
Esomeprazole Etercon 40mg Tablet	0,3088	-1,5221	Less Popular
Fenofibrate Medikon 200mg Tablet	-0,2300	0,2715	Bestselling
Novell Fluconazole 150mg Capsules	0,7668	-1,4204	Less Popular
Furosemide FM 40Mg	-0,5533	0,8249	Bestselling
Hydrochlorothiazide 25mg Tablet	-0,1223	0,1891	Bestselling
Ibuprofen 200mg Tablets	-0,3108	0,3923	Bestselling
Kifarox 500mg Tablets	0,4974	-1,3490	Less Popular
Lansoprazole 30mg Capsule	-0,0953	-0,1034	Bestselling
Laserin Syrup 110ml	1,0362	-1,4122	Less Popular
Laserin Syrup 60ml	0,3896	-1,4644	Less Popular
Levofloxacin Novell 500mg Tablets	-0,2031	0,1657	Bestselling
Lorazepam Novell 2mg Tablet	-0,0953	0,2056	Bestselling
Metformin 500mg Tablet	-0,3647	0,5956	Bestselling
Methylprednisolone 4mg Tablet	-0,2570	0,1053	Bestselling
Metoclopramide 10mg Tablet	-0,3647	0,0600	Bestselling
Metronidazole 250mg/5ml Syrup 60ml	0,4435	-1,4204	Less Popular
Metronidazole Novell 500mg Tablet	-0,4186	-0,1584	Bestselling
Ofloxacin 200mg Novell Tablets	-0,5264	1,7066	Very Popular
Omeprazole if 20mg capsules	-0,4994	1,5130	Very Popular
Paracetamol IF 15 ml	0,2280	-1,3696	Less Popular
Paracetamol IF 500mg Tablets	-0,3108	0,3099	Bestselling
Paracetamol Chemical Farma	-0,1492	0,0339	Bestselling
Propepsa 500mg/5ml Sucker 100 ml	6,1552	-1,4424	Less Popular
Propranolol Dexa 100 mg Tablets	-0,5533	1,7231	Very Popular
Simvastatin 20mg Tablet	-0,1761	0,4473	Bestselling
Sodium Valproate 500mg Tablet	-0,2570	0,1781	Bestselling

The next step is to divide the data into 80% training data and 20% test data. The sharing of training data and test data is facilitated by the train test split method from Scikit-Learn. However, before dividing the data, the classification column must first be changed to numerical data, so that it can be more easily processed by the algorithm. The coding for each classification class is as follows:

Table 3. Numeric Code Table

Numeric Code	Class
0	Less Popular
1	Bestselling
2	Very Popular

Here is a table of the data that has been trained and tested:

Table 4. Drill Data Table

Drug Name	Price (Rp)	Quantity Sold	Classification
Cefixime Dexa 100mg Capsules	-0,392	-0,099	Bestselling
Betahistine Novell 6mg Tablets	-0,499	0,903	Bestselling
Simvastatin 20mg Tablet	-0,176	0,447	Bestselling
Dexamethasone 0.25mg Tablets	-0,419	0,320	Bestselling
Levofloxacin Novell 500mg Tablets	-0,203	0,166	Bestselling
Cefadroxil if D.Syr 60ml 125mg/5ml	0,470	-1,409	Less Popular
Acyclovir KF 400mg Tablet	-0,446	1,448	Very Popular
Omeprazole if 20mg capsules	-0,499	1,513	Very Popular
Paracetamol IF 15 ml	0,228	-1,370	Less Popular
Diphenhydramine 25mg Tablet	-0,392	-0,002	Bestselling
Paracetamol IF 500mg Tablets	-0,311	0,310	Bestselling
Cendesartan Dexa 16mg Tablets	-0,176	0,534	Bestselling
Methylprednisolone 4mg Tablet	-0,257	0,105	Bestselling
Cefadroxil Berno 500mg Capsules	-0,472	1,781	Very Popular
Propranolol Dexa 100 mg Tablets	-0,553	1,723	Very Popular
Dexymox Forte 500Mg	-0,521	1,496	Very Popular
Furosemide FM 40Mg	-0,553	0,825	Bestselling
Metronidazole 250mg/5ml Syrup 60ml	0,444	-1,420	Less Popular
Lorazepam Novell 2mg Tablet	-0,095	0,206	Bestselling
Cetirizine Dexa 5mg Syrup 60ml	0,228	-1,477	Less Popular
Fenofibrate Medikon 200mg Tablet	-0,230	0,271	Bestselling
Metoclopramide 10mg Tablet	-0,365	0,060	Bestselling
Propepsa 500mg/5ml Sucker 100 ml	6,155	-1,442	Less Popular
Ibuprofen 200mg Tablets	-0,311	0,392	Bestselling
Amlodipine Dexa 5mg Tablets	0,228	-1,427	Less Popular
Alerhis Loratadine 10mg Capsules	0,228	-1,444	Less Popular
Atorvastatin Pratapa Nirmala 20 mg Tablets	-0,284	0,118	Bestselling
Metformin 500mg Tablet	-0,365	0,596	Bestselling
Kifarox 500mg Tablets	0,497	-1,349	Less Popular
Sodium Valproate 500mg Tablet	-0,257	0,178	Bestselling
Laserin Syrup 60ml	0,390	-1,464	Less Popular
Cardio Aspirin 100mg Tablets	-0,392	0,072	Bestselling
Hydrochlorothiazide 25mg Tablet	-0,122	0,189	Bestselling

Dexamethasone 0.5mg Tablets	-0,284	0,438	Bestselling
Lansoprazole 30mg Capsule	-0,095	-0,103	Bestselling
Amoxicillin IF 500mg	-0,526	1,194	Very Popular

Table 5. Test Data Table

Drug Name	Price (Rp)	Quantity Sold	Classification
Novell Fluconazole 150mg Capsules	0,767	-1,420	Less Popular
Metronidazole Novell 500mg Tablet	-0,419	-0,158	Bestselling
Esomeprazole Etercon 40mg Tablet	0,309	-1,522	Less Popular
Laserin Syrup 110ml	1,036	-1,412	Less Popular
Ofloxacin 200mg Novell Tablets	-0,526	1,707	Very Popular
Paracetamol Chemical Farma	-0,149	0,034	Bestselling
Bisoprolol Fumarate Dexa	-0,392	0,401	Bestselling
Diazepam 5mg Tablets	-0,149	0,083	Bestselling
Ambroxol Syrup 60ml	-0,149	0,011	Bestselling

Next, we will calculate the Euclidean distance in the data, The process carried out is the calculation of the classification of the data by calculating the distance of each test data with all the training data. The following is an example of a case of Euclidean distance calculation for the first data in the test data with the first data in the train data using the Eucledian distance formula.

$$D(x,y) = \sqrt{(-0{,}392\ -\ 0{,}767)^2 + (-0{,}099\ -\ (-1{,}420)^2}$$
$$D(x,y) = \sqrt{1{,}343281\ +\ 1{,}745041}$$
$$D(x,y) = \sqrt{3{,}088322}$$
$$D(x,y) = 1{,}757$$

This process continues by calculating the distance between each test data and all the data in the training dataset. Next, the results are sorted from closest to farthest. After that, a class label that corresponds to the original label on the training data is assigned to each test data in order of distance. Here are the results of the data sequencing based on Euclidean distances.

Table 6. Eucledian Distance Sequencing and Labeling

Drug Name	Eucledian Distance	Classification
Kifarox 500mg Tablets	0,2787	Less Popular
Cefadroxil if D.Syr 60ml 125mg/5ml	0,2966	Less Popular
Metronidazole 250mg/5ml Syrup 60ml	0,3233	Less Popular
Laserin Syrup 60ml	0,3797	Less Popular
Amlodipine Dexa 5mg Tablets	0,5389	Less Popular
Alerhis Loratadine 10mg Capsules	0,5393	Less Popular
Paracetamol IF 15 ml	0,5412	Less Popular
Cetirizine Dexa 5mg Syrup 60ml	0,5418	Less Popular
Lansoprazole 30mg Capsule	1,5741	Bestselling
Cefixime Dexa 100mg Capsules	1,7571	Bestselling
Diphenhydramine 25mg Tablet	1,8316	Bestselling
Methylprednisolone 4mg Tablet	1,8374	Bestselling
Hydrochlorothiazide 25mg Tablet	1,8388	Bestselling
Lorazepam Novell 2mg Tablet	1,8404	Bestselling
Levofloxacin Novell 500mg Tablets	1,8592	Bestselling
Atorvastatin Pratapa Nirmala 20 mg Tablets	1,8627	Bestselling
Metoclopramide 10mg Tablet	1,8634	Bestselling
Cardio Aspirin 100mg Tablets	1,8896	Bestselling
Sodium Valproate 500mg Tablet	1,8983	Bestselling
Fenofibrate Medikon 200mg Tablet	1,9637	Bestselling
Paracetamol IF 500mg Tablets	2,0385	Bestselling

Simvastatin 20mg Tablet	2,0922	Bestselling
Dexamethasone 0.25mg Tablets	2,1054	Bestselling
Ibuprofen 200mg Tablets	2,1089	Bestselling
Dexamethasone 0.5mg Tablets	2,1346	Bestselling
Cendesartan Dexa 16mg Tablets	2,1698	Bestselling
Metformin 500mg Tablet	2,3119	Bestselling
Furosemide FM 40Mg	2,6047	Bestselling
Betahistine Novell 6mg Tablets	2,6463	Bestselling
Amoxicillin IF 500mg	2,9171	Very Popular
Acyclovir KF 400mg Tablet	3,1145	Very Popular
Dexymox Forte 500Mg	3,1886	Very Popular
Omeprazole if 20mg capsules	3,1950	Very Popular
Propranolol Dexa 100 mg Tablets	3,4095	Very Popular
Cefadroxil Berno 500mg Capsules	3,4327	Very Popular
Propepsa 500mg/5ml Sucker 100 ml	5,3884	Less Popular

The final step is to perform a class classification for the test data based on the nearest distance value and the predetermined k-value (K value = 3). So that 3 closest neighbors will be taken from the data sequencing based on Eucledian distance. The prediction results of the classification class can be seen by looking at the comparison of existing classes, the most results will be taken as a classification from the test data. Here are 3 data with the nearest eucledian distance taken as a prediction result.

Table 7. Test One Data Prediction Results with k = 3

Drug Name	Eucledian Distance	Classification
Kifarox 500mg Tablets	0,2787	Less Popular
Cefadroxil if D.Syr 60ml 125mg/5ml	0,2966	Less Popular
Metronidazole 250mg/5ml Syrup 60ml	0,3233	Less Popular

The next step is to calculate the evaluation metrics and assess the ability of the KNN algorithm to predict drug sales using accuracy, precision, recall, and F1 score metrics. The first step is to calculate the number of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) along with the distribution of the calculation based on the results of the existing data.

4.3. Classification of Results

For the overall data, the total TP, TN, FP, and FN of all categories are as follows:

1. Bestselling
   TP : Predicted "Bestselling" and actual "Selling" = 4
   FP : Predict "Sell", actual not "Sell" = 0
   FN : Prediction is not "Sell", actual "Run" = 0
2. Less Popular
   TP : Predicted "Less Selling" and actual "Less Selling" = 3
   FP : Prediction of "Less Selling", actual not "Less Selling" = 1
   FN : Prediction is not "Less Selling", actual "Less Selling" = 0
3. Very Popular
   TP : Prediction of "Very Selling" and actual "Very Selling" = 1
   FP : Prediction "Very Selling", actual not "Very Selling" = 0
   FN : Prediction is not "Very Selling", actual "Very Selling" = 0

4.4. Calculation of Metrics

From the calculations in the previous section, the calculation of accuracy, precision, recall, and F1 score is as follows:

1. Precision Calculation

$$\boldsymbol{Precision = \frac{TP}{TP+FP}} \quad (2)$$

$$Precision = \frac{8}{8+1}$$
$$Precision = 0{,}889$$

2. Recall Calculation

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (3)$$
$$Recall = \frac{8}{8}$$
$$Recall = 1$$

3. F1 Score Calculation

$$\frac{\mathbf{1}}{\boldsymbol{F1}} = \frac{\mathbf{1}}{\mathbf{2}}(\frac{\mathbf{1}}{\boldsymbol{Precision}} + \frac{\mathbf{1}}{\boldsymbol{Recall}}) \quad (4)$$
$$\frac{1}{F1} = \frac{1}{2}(\frac{1}{0{,}0889} + \frac{1}{1}$$
$$F1 = 0{,}0941$$

4. Calculation of Accuracy

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$
(5)
$$Accuracy = \frac{8}{9}$$
$$Accuracy = 0{,}0889$$

From the above results, it can be concluded as follows:

1. Accuracy (0.889 or 88.9%): The model has a fairly high accuracy. This shows that most of the predictions made by the model match the actual labels of the data.
2. Recall (1.0 or 100%): A perfect recall indicates that the model successfully identified all true positive cases from the data. In this context, it means that all drugs that should be categorized as “Bestselling”, “Not Selling”, or “Very Selling” have been successfully identified by the model without missing anything.
3. F1 Score (0.941 or 94.1%): A high F1 Score indicates that the model has a good balance between precision and recall.

5. Conclusion

Based on the results of discussions and trials conducted in implementing data mining using the K-Nearest Neighbor (KNN) algorithm to predict drug sales at the Pratama Haji Medan-Pancing Clinic, it can be concluded that the data mining process with the help of the K-Neighbor Classifier from the scikit-learn library has proven to be effective in predicting and classifying drug sales, with a prediction accuracy rate of 88.9%. The application of this method provides real benefits for the Pratama Haji Medan-Pancing Clinic in accelerating the process of managing drug stocks, because the predictive results produced are able to replace the need for manual data checking. This not only increases operational efficiency, but also helps clinics in anticipating drug needs more accurately and in a planned manner.

Reference

[1] U. A. Putri, A. B. Prasetijo, and C. T. Purnami, “Sistem informasi manajemen logistik obat di pelayanan farmasi puskesmas: Literature review,” *Media Publ. Promosi Kesehat. Indones.*, vol. 6, no. 6, pp. 1016–1024, 2023.

[2] F. Farahdinna and M. N. Shofy, “IMPLEMENTASI K-NEAREST NEIGHBOR UNTUK KLASIFIKASI BUNGA IRIS,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 2, pp. 3510–3514, 2025.

[3] A. A. Baihaqi and M. Fakhriza, “K-Nearest Neighbors (KNN) to Determine BBRI Stock Price,” *Sist. J. Sist. Inf.*, vol. 14, no. 2, pp. 969–984, 2025.

[4] A. Andri, “Penerapan Algoritma K-Nearest Neighbor Untuk Prediksi Penjualan Obat Pada Apotek Kimia Farma Atmo Palembang,” in *Bina Darma Conference on Computer Science (BDCCS)*, 2020, pp. 199–208.

[5] R. Rismala, I. Ali, and A. R. Rinaldi, “Penerapan Metode K-Nearest Neighbor Untuk Prediksi Penjualan Sepeda

Motor Terlaris," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 585–590, 2023.

[6] R. F. Putra *et al.*, *Data Mining: Algoritma dan Penerapannya*. PT. Sonpedia Publishing Indonesia, 2023.

[7] S. Sudirwo *et al.*, *Artificial Intelligence: Teori, Konsep, dan Implementasi di Berbagai Bidang*. PT. Sonpedia Publishing Indonesia, 2025.

[8] A. Wibowo, "Cara Mudah Menganalisis Big Data," *Penerbit Yayasan Prima Agus Tek.*, pp. 1–159, 2024.

[9] V. P. Virza, G. T. Pranot, and F. E. Putra, "Klasifikasi Kebutuhan Sparepart Dengan Algoritma K-Nearest Neighbor Untuk Meningkatkan Penjualan Sparepart," *Bull. Inf. Technol.*, vol. 4, no. 3, pp. 287–293, 2023.

[10] M. S. Safira, N. Rahaningsih, and R. D. Dana, "Penerapan Data Mining Untuk Klasifikasi Penjualan Obat Menggunakan Metode K-Nearest Neighbor," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 380–385, 2024.

[11] P. H. Putra and M. S. Novelan, "PERANCANGAN APLIKASI PENENTUAN KUALITAS SAYURAN BERDASARKAN WARNA MENGGUNAKAN DATA MINING," in *Scenario (Seminar of Social Sciences Engineering and Humaniora)*, 2021, pp. 103–109.

[12] D. W. Azhari, Z. Sitorus, and Z. Zulfahmi, "APPLICATION OF K-NEAREST NEIGHBOR METHOD IN CLASSIFICING THE RATE OF PAPAYA MURABILITY BASED ON FRUIT COLOR FORM," *INFOKUM*, vol. 10, no. 02, pp. 1247–1255, 2022.

[13] F. Amalini and A. M. L. Harefa, "Application of Linear Regression Method in Predicting Veil Sales (Case Study: Fauzan Kerudung Shop)," *J. Data Sci. Technol. Artif. Intell.*, vol. 1, no. 1, pp. 13–17, 2024.

[14] J. C. Mestika, M. O. Selan, and M. I. Qadafi, "Menjelajahi Teknik-Teknik Supervised Learning untuk Pemodelan Prediktif Menggunakan Python," *vol*, vol. 99, pp. 216–219, 2022.

[15] W. Yustanti, "Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah," *J. Mat. Stat. dan Komputasi*, vol. 9, no. 1, pp. 57–68, 2012.

[16] A. A. Nababan, M. Jannah, and A. H. Nababan, "Prediction Of Hotel Booking Cancellation Using K-Nearest Neighbors (K-Nn) Algorithm And Synthetic Minority Over-Sampling Technique (Smote)," *INFOKUM*, vol. 10, no. 03, pp. 50–56, 2022.

[17] M. Jannah, A. A. Nababan, and Y. S. Ningsi, "Penerapan Metode K-Nearest Neighbor Dalam Identifikasi Jenis Ikan Salmon Yang Dapat Dikomsumsi Untuk Bahan Mpasi Bayi," *J. Teknol. Sist. Inf. dan Sist. Komput. TGD*, vol. 6, no. 2, pp. 636–644, 2023.

[18] S. Mulyati, S. M. Husein, and R. Ramdhan, "Rancang bangun aplikasi data mining prediksi kelulusan ujian nasional menggunakan Algoritma (Knn) K-Nearest Neighbor dengan metode Euclidean Distance pada SMPN 2 Pagedangan," *JIKA (Jurnal Inform.*, vol. 4, no. 1, pp. 65–73, 2020.

[19] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes classifier dan confusion matrix pada analisis sentimen berbasis teks pada Twitter," *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 5, no. 2, pp. 697–711, 2021.

[20] A. A. Putri, "Perbandingan Metode Naive Bayes, KNN dan SVM pada Analisis Sentimen Jasa Transportasi Online= Comparison of Naive Bayes, KNN dan SVM for Sentiment Analysis of Online Transportation Services." Universitas Hasanuddin, 2023.

[21] S. F. Amrilah, D. Krisbiantoro, and A. Prasetyo, "Penerapan Metode K-Nearest Neighbors dan Naïve Bayes pada Analisis Sentimen Pengguna Aplikasi Bstation melalui Platform Playstore," 2024.

[22] F. Sulianta, *Dasar & Konsep Data Science*. Feri Sulianta, 2024.

[23] R. G. Whendasmoro and J. Joseph, "Analisis Penerapan Normalisasi Data Dengan Menggunakan Z-Score Pada Kinerja Algoritma K-NN." 2022.

[24] S. E. Saqila, I. P. Ferina, and A. Iskandar, "Analisis Perbandingan Kinerja Clustering Data Mining Untuk Normalisasi Dataset," *J. Sist. Komput. dan Inform. Hal*, vol. 356, p. 365, 2023.

[25] M. R. A. Prasetya and A. M. Priyatno, “Penanganan Imputasi Missing Values pada Data Time Series dengan Menggunakan Metode Data Mining,” *J. Inf. Dan Teknol.*, pp. 52–62, 2023.

[26] A. Pratiwi, A. T. Sasongko, and D. K. Pramudito, “ANALISIS PREDIKSI GILINGAN PLASTIK TERLARIS MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR DI CV MENEMBUS BATAS,” *J. Inform. Teknol. dan Sains*, vol. 5, no. 3, pp. 437–445, 2023.