Application of the KNN Algorithm for Predicting Data Card Sales at PT. XL Axiata Makassar

Medy Wisnu Prihatmono[1], Sitti Arni[2], Johar Nur Iin[3] Dikwan Moeis[4]
Information Systems, STMIK Profesional[1,2,3,4]
Jl.A.P.Pettarani No.27 Makassar, 0411-431139/Indonesia[1,2,3,4]
e-mail: medy_wisnu_prihatmono@stmikprofesional.ac.id[1], sitti_arni@stmikprofesional.ac.id[2], iinstmikpro@gmail.com3, dikwan_moeis@stmikprofesional.ac.id[4]

To cite this document:

Prihatmono, M. W. ., Arni, S., Iin, J. N. ., & Moeis, D. (2022). Application of the KNN Algorithm for Predicting Data Card Sales at PT. XL Axiata Makassar. Conference Series, 4(1), 59–64. https://doi.org/10.34306/conferenceseries.v4i1.692

Hash:ABCoDl3kNpTmGKwlNCHQV5ynHwFrnSgXK2EojKCz13CBVoVQniXaChTzdZPXzdyx

Abstract

XL Axiata Company, TBK. Is an internet access service provider from the many operators in Indonesia and has to provide the best service in communication in the ease of communication services. The mainstay of communication service products owned by XL Axiata, TBK for the wider community is XL Prioritas Unlimited which includes three types of packages, namely silver, gold and platinum. People using these three packages have different levels of demand. Order the number of products ordered every month by the company, should be made based on the prediction of the number of previous sales. This study was conducted to assess product sales in the form of K-Nearest Neighbor (KNN) classification modeling. The number of datasets used is 500 records with details of 90% training and 10% testing. The results obtained are predictions of 0.98%.

Keywords: Card, Prediction, K-Nearest Neighbor

1. Pendahuluan

In Indonesia, the dominance of cellular telecommunications operators competes with each other, by competing to provide the best service and internet package features with various facilities. [1]. Various names of cellular operator providers that provide cellular services for the Indonesian people include Telkomsel, Indosat Ooredoo, XL, 3 and Smartfreen. By providing conditions of fairly tight competition, providers in marketing their products use various marketing strategies to win healthy competition in getting consumers. Before buying intention is made by consumers, sometimes consumers usually collect information about product services from advertisements and also add other information obtained based on the experience of those who have used it.[2] One of the cellular providers is the XL cellular operator company, by issuing XL Prioritas Unlimited cards, namely silver, gold and platinum. However, there is a separate problem with stock availability, where for 2 districts, namely Maros and Gowa, they still receive product supply from Makassar city, so there is a problem with the availability of this service product in Makassar city itself. By making predictions using old data that is used as a reference so that product availability can be done. In making predictions, the K-Nearest Neighbor (KNN) Algorithm is used, the reason for using this algorithm is very simple and works by doing the shortest distance. The K-Nearest Neighbor (KNN) algorithm groups the calculation results with the training data that has the most relatives in the specified range value. The distance between the training data and the test data is calculated using the Euclidean equation.[3] In this study, predictions will be made using the criteria of type, quota, price and age, on the availability of XL Prioritas Unlimited products, namely: silver, gold and platinum for the Makassar area.

2. Research methods

2.1 Data Mining

Data Mining is a data mining activity that utilizes the processes of mathematical functions, statistics and machine learning to extract and identify useful information [4]. Data mining activity is data preparation which is divided into three stages, namely data is selected, cleaned and preprocessed following the guidelines and knowledge of domain experts. Data mining is categorized into five parts, namely estimation, prediction, classification, clustering and association. Using the right algorithm for data mining based on the five categories above can explore integrated data to facilitate the introduction of valuable information [5]. In mining data, if the data used is of large capacity, the data processed will be even greater [6]

2.2 Machine Learning

A simple understanding of machine learning is the process of learning knowledge from experience or without human supervision. For the supervised learning process, the output is predicted for input by learning from pairs of input and output labeled: where the program learns from examples of correct answers, while learning without supervision, the program will not experience a learning process using data labels [7]

2.3 K-Nearest Neighbor

The K-Nearest Neighbor Algorithm (k-NN or KNN) is an algorithm that is carried out for classification activities against targets based on learning data (neighbors) that are closest to the target. For targets that are far or near the neighbor, the calculation process starts by using the Euclidean distance [8]. The K-Nearest Neighbor Algorithm (k-NN or KNN) is divided into two parts, namely learning (training) and testing (testing). In the part for learning the process that occurs is storing feature vectors and classifications from learning data, for classification, features with the same model will be calculated for data to be tested (whose classification is unknown). If the distance with the new vector to all vectors of the learning data is calculated, and the number of k for the closest neighbors is taken. For the neighbor distance calculation process using the Euclidean algorithm as shown in the example equation 1.[3]

$$euc = \sqrt{((a_1 - b_1)^2 + \ ... + a\_n - b\_n)^2)} \quad (1)$$

Where a =a1,a2, .. , an, dan b = b1,b2, .., bn represents the n attribute values of the two records. For attributes with category values [5]. Later a point will be predicted by type based on the most classifications of the surrounding neighbors, for an example of the discussion, it can be seen in Figure 1.

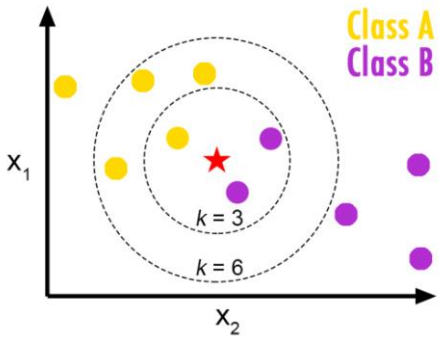

Figure 1. Illustration of the use of the value of k in the KNN. method

The best value of k for the KNN algorithm depends on the data. Basically, the highest k value will reduce the noise effect for the classification, but make the boundaries between each classification more blurred. The best value for k can be chosen by optimizing parameters, for example when using cross-validation.

2.4 Python

Bahasa pemrograman python adalah bahasa pemrograman yang berorientasi obyek, Kehadiran python dengan libarary-library standart yang dapat diperluas serta kemudahan untuk dapat dipelajari dengan singkat. Python boleh dikatakan sebagai bahasa pemrograman yang menggabungkan kapabilitas, kemampuan, dengan sintaksis kode yang sangat jelas, dan memiliki fitur fungsionalitas pustaka standar yang sangat besar serta memiliki komprehensif yang cakupannya cukup luas.[9]

2.5 Performance

The performance results of each classification process carried out can be done by testing with the values of accuracy, precision, recall and f-measure [10]. The definition of accuracy is the level of closeness between the predicted value and the actual value, while the accuracy formula is explained in equation 2 as follows:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2)$$

The definition of precision is a match for a request for information with an answer to that request. For the precision formula equation can be explained in equation 3 as follows:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

The definition of recall is a ratio to the relevant system that is already available, while the precision formula equation can be explained in equation 4 as follows:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

The definition of the F-measure is an equation formula which is usually called the F1-Score, the F-measure formula can be explained in equation 5 as follows:

$$F-Measure = 2\,\frac{Presisi\ x\ recall}{Presisi+Recall} \quad (5)$$

2.6 Design / Model

In this research, the author designs a model according to the flow process of the K-Nearest Neighbor algorithm, the first model process is to collect the dataset, then divide the dataset or usually called split. Implementation of the K-Nearest Neighbor method to calculate the performance of the K-Nearest Neighbor method from the data set. The flow of the model is shown in Figure 2

p-ISSN : 2774-9576
e-ISSN : 2747-2981

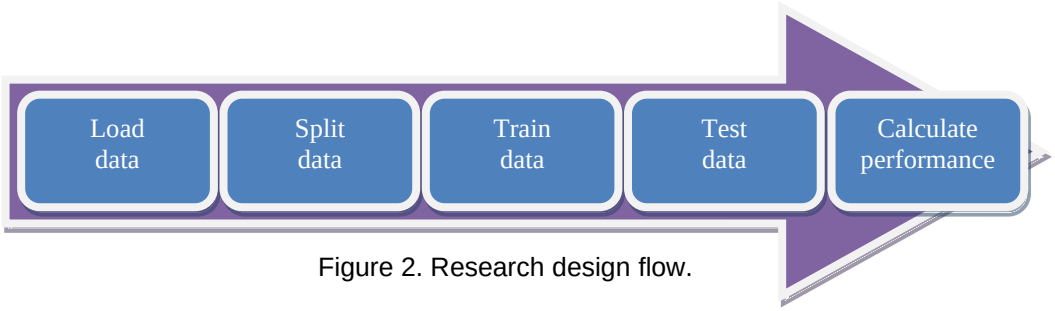

Figure 2. Research design flow.

2.7 Sample and Data

In this study, the data used was obtained from the XL Axiata Company, TBK with the office address of XL Center Pettarani Jl.A. Pettarani No.68 Makassar, where the data set was obtained in the sales marketing section for the XL Prioritas Unlimited data card which consists of three packages, namely Silver, Gold and Platinum. The dataset taken is 500 sales record sets with a period taken from July – December 2021

2.8 Analysis Techniques

In this study, the performance calculation process of the K-Nearest Neighbor Algorithm (k-NN or KNN) is carried out, where the performance to be tested and measured is accuracy, precision, recall and f-measure, using the process of equations 2, 3 and 4. *Scikit-learn library* is a machine learning tool used in this research. The first step is to collect data, in which the data is as much as 500 data card data. The second step is to split the data where 75% is used as training data and 25% for testing data, while the detailed description is shown in table 1.

Dataset	Amount	Class	Splits
Data 1	500	Silver Gold Platinium	90 % Training 10 % Testing

Table 1. Distribution of data card dataset

3. Results and Discussion

The next process is to apply the K-Nearest Neighbor (k-NN or KNN) algorithm using testing data and training data that have been prepared in advance. The next step is to calculate the performance of all testing data with various neighboring simulations on the K-Nearest Neighbor Algorithm (k-NN or KNN). Table 2 shows some of the main commands used, namely the scikit-learn library.

Tabel 2. Source code implementasi metode KNN

Des	**SourceCode**
Load	data =pd.read_csv("XLPrioritasUnlimited.csv")
Split	x = data.iloc[:,3:6].values y = data.iloc[:, 6].values from sklearn.model_selection import train_test_split x_train, x_test, y_train, y_test = train_test_split (x, y, test_size=0.10, random_state=0)
Train	from sklearn.neighbors import KNeighborsClassifier classifier = KNeighborsClassifier(n_neighbors =5, metric ='minkowski', p = 2) classifier.fit(x_train, y_train)
Test	y_pred = classifier.predict(x_test)
Result	from sklearn.metrics import classification_report, confusion_matrix print(confusion_matrix(y_test, y_pred)) print(classification_report(y_test, y_pred)) accuracy_score(y_test, y_pred)

3.1 Table Matrix

The results of the matrix carried out for silver, gold and platinum which consist of precision, recall, f1-score and support, these results provide information about the performance testing in this study simulated on various neighboring values of the KNN method, while the detailed description is shown in Fig. table 3.

	precision	recall	f1-score	support
GOLD	0.94	1.00	0.97	15
PLATINUM	1.00	1.00	1.00	17
SILVER	1.00	0.94	0.97	18
accuracy			0.98	50
macro avg	0.98	0.98	0.98	50
weighted avg	0.98	0.98	0.98	50

```
Out[98]: 0.98
```

Table 3. Results of the KNN method performance test

4. Conclusions and Advice

Based on the results of this study, the authors can draw several conclusions, namely by simulating KNn with a dataset of 500, resulting in a prediction accuracy of 0.98%. Based on the conclusions above, the authors suggest several things so that they can be taken into consideration for further research. Further research can try cross validation to find new performance values by doing various data simulations.

References

[1] U. Baetulloh, A. I. Gufroni, and R. -, “Penerapan Metode Association Rule Mining Pada Data Transaksi Penjualan Produk Kartu Perdana Kuota Internet Menggunakan Algoritma Apriori,” *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 10, no. 1, pp. 173–188, 2019, doi: 10.24176/simet.v10i1.2890.

[2] M. A. Tito, “PENGARUH INOVASI PRODUK DAN HARGA TERHADAP NIAT BELI KARTU XL 4G LTE. (Studi Pada Pengguna Smartphone 4G Di WTC Surabaya),” vol. 4, pp. 1–8, 2016.

[3] M. M. Baharuddin, H. Azis, and T. Hasanuddin, “Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca,” *Ilk. J. Ilm.*, vol. 11, no. 3, pp. 269–274, 2019, doi: 10.33096/ilkom.v11i3.489.269-274.

[4] A. A. Karim, H. Azis, and Y. Salim, “Kinerja Metode C4.5 dalam Penyaluran Bantuan Dana Bencana,” *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 84–87, 2018.

[5] H. Leidiyana, “Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor,” *J. Penelit. Ilmu Komputer, Syst. Embed. Log.*, vol. 1, no. 1, pp. 65–76, 2013.

[6] N. Fadhillah, Huzain Azis, and D. Lantara, “Validasi Pencarian Kata Kunci Menggunakan Algoritma Levenshtein Distance Berdasarkan Metode Approximate String Matching,” *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 129–133, 2018.

[7] R. putri Indahningrum, *Mastering Machine Learning with scikitlearn*, vol. 2507, no. 1. 2020.

[8] A. Fitria and H. Azis, “Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier,” *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 102–106, 2018.

[9] Fitri, K. R. R, A. Rahmansyah, and W. Darwin, “Penggunaan Bahasa Pemrograman

Python Sebagai Pusat Kendali Pada Robot 10-D," *5th Indones. Symp. Robot. Syst. Control*, pp. 23–26, 2017.

[10] C. A. Ul Hassan, M. S. Khan, and M. A. Shah, "Comparison of machine learning algorithms in data classification," *ICAC 2018 - 2018 24th IEEE Int. Conf. Autom. Comput. Improv. Product. through Autom. Comput.*, no. September, pp. 1–6, 2018, doi: 10.23919/IConAC.2018.8748995.