Analysis Of Public Interest In Smartfren SIM Cards Using The K-Nearest Neighbors Method

Sri Tawanta Natalia Sembiring[1*], Volvo Sihombing[2], Deci Irmayani[3]

[1,2,3] Faculty of Science and Technology, Universitas Labuhanbatu, Sumatera Utara Indonesia.
[*]Corresponding Author:
Email: tawantasembiring1412@gmail.com

Abstract.

The use of Smartfren SIM cards is increasing along with the public's need for fast and stable internet services. However, a deep understanding of public interest in the SIM card is necessary to optimize marketing strategies and increase sales. Proper analysis can help companies identify potential target markets and develop effective marketing strategies. We chose the K-Nearest Neighbors method to analyze public interest in using Smartfren SIM cards. This study aims to develop and evaluate the K-Nearest Neighbors model in predicting public interest in using Smartfren SIM cards. This study uses a dataset containing information about Smartfren SIM card users. We divide the data into two sets: a training set for model building and a test set for evaluating model performance. We apply the K-Nearest Neighbors method to classify the data into two categories: interested and not interested. We evaluate the model performance using accuracy, precision, recall, and F1-score metrics. We present the evaluation results as a confusion matrix. The developed K-Nearest Neighbors model showed excellent performance with an accuracy of 94.29%, a precision of 94.20%, a recall of 100%, and an F1-score of 97.01%. These results indicate that the K-Nearest Neighbors model is effective in predicting people's interest in Smartfren SIM cards. The high recall value indicates that the model is able to identify all interested individuals without missing any, while the high precision value indicates that the model rarely makes false positive prediction errors. This study concludes that the K-Nearest Neighbors method is very effective for use in analyzing people's interest in using Smartfren SIM cards. We can rely on the developed model's strong performance for real-world applications in marketing strategies.

Keywords: *Confusion Matrix, K-Nearest Neighbors, Public Interest, SIM Cards, and Smartfren.*

I. INTRODUCTION

Mobile devices use a SIM (Subscriber Identity Module) card, a small smart card, to store user identity information, including phone numbers, contacts, and network data [1]. This card functions as a link between the mobile device and the operator's network, allowing users to make calls, send text messages, and use data services [2]. In Labuhanbatu Regency, the use of SIM cards from various telecommunications operators such as Telkomsel, Smartfren, XL, Exis, and IM3 is increasingly widespread. Telkomsel, one of Indonesia's largest operators, is known for its wide and stable signal coverage, making it the main choice for many people in the area. Telkomsel offers various service packages, ranging from calls and SMS to high-speed internet data. Smartfren, known for its unlimited data service, is also popular, especially among users who need unlimited internet access at affordable prices. Other operators, such as XL, Exis, and IM3, are also highly competitive, offering attractive packages with various advantages such as large quotas, economical prices, and increasingly wide networks. The variety of operator choices provides flexibility for the Labuhanbatu community to choose the SIM card that best suits their needs and preferences. Users often choose XL for its competitively priced data packages and good internet speeds. Exis, known for its affordable rates, is in high demand by students. IM3's promos and quota bonuses are also popular with users who want to maximize their spending. The existence of these different operators not only increases healthy competition in the telecommunications market, but also ensures that the people of Labuhanbatu Regency get the best service according to their needs, both for personal and professional communications.

Data mining is the process of exploring and analyzing large amounts of data to discover useful patterns and insights [3]. Data mining involves several steps, including data collection, preprocessing or data cleaning, applying analytical algorithms, and interpreting the results [4]. Techniques used in data mining include classification, clustering, associative, regression, and anomaly detection [5]. Data mining primarily aims to convert unstructured data into valuable insights that enhance decision-making [6]. Various fields such as marketing, healthcare, finance, and social sciences find applications of data mining. The data mining process often uses various statistical and machine learning methods to identify patterns in data that may not be apparent through simple analysis [7]. For instance, data mining in marketing identifies various customer segments based on their purchasing behavior, enabling the development of more effective marketing strategies. In healthcare, data mining can help identify risk factors for certain diseases based on patient data. Thus, data mining plays a vital role in uncovering hidden information in large and complex data, providing reliable insights for various practical applications [8].

One of the machine learning algorithms for classification and regression is the K-Nearest Neighbors (KNN) method [9], [10]. KNN operates by pinpointing the 'k' closest data points within the feature space of a newly classified data point. We measure the distance between data points using metrics like Euclidean, Manhattan, or Minkowski. After determining the 'k' nearest neighbors, the algorithm will classify the new data point based on the majority of class labels from those neighbors [11], [12]. The advantages of the KNN method are simplicity and high accuracy in cases where the data has a non-linear relationship. However, KNN has disadvantages, especially in terms of computational efficiency when dealing with large datasets, and is sensitive to irrelevant features or features that have different scales. Various practical applications like facial recognition, product recommendations, and anomaly detection often utilize KNN. In the context of classification, for example, KNN can help determine whether an email is spam or not based on features such as the frequency of certain keywords. Regression can also use this method to predict continuous values like house prices, taking into account characteristics like size, location, and number of rooms. Although simple, KNN necessitates careful selection of the 'k' value, as well as good data preprocessing to ensure optimal performance.

The author conducted this research using data mining techniques and the KNN method. The author chose the KNN method due to its strong consistency in calculating the proximity between new and old cases through weight matching [13]. The KNN method is able to provide accurate and accountable results [14]. KNN has the ability to classify data based on similar characteristics [15]. In this analysis, the author will collect data on public preferences for Smartfren SIM cards, including factors such as satisfaction with price, network quality, and user experience. We will analyze the collected data using KNN to classify and identify patterns of public interest in Smartfren SIM cards, enabling us to understand more detailed levels of acceptance and preferences. Using the KNN method, the author will be able to group data based on similarities in user preferences. This analysis will help to reveal groups of people who have a high interest in Smartfren SIM cards and the reasons behind their choices. We expect the results of this study to offer a deeper understanding of public perceptions of Smartfren SIM cards and the factors influencing their decisions. This information is not only useful for companies in improving services and marketing strategies but also provides potential users with an idea of what to expect from a Smartfren SIM card.

II. METHODS

We will conduct research on public interest in SIM cards using a data mining approach and the K-Nearest Neighbors (KNN) method. The KNN method will assist in classifying data based on different communities' preferences and characteristics. The data collected will include various variables such as price, network quality, and user experience. We will analyze this data using KNN to identify significant interest patterns among different community groups. The purpose of this analysis is to better understand what factors influence people's decisions to choose Smartfren SIM cards.

The data mining approach with the KNN method allows researchers to reveal hidden relationships in complex data. Through this analysis, researchers can identify segments of society that have high interest in Smartfren SIM cards and the reasons underlying their preferences. Telecommunications operators expect the results of this study to provide valuable insights for designing more effective marketing strategies and improving their services. In addition, this study will also provide prospective users with an overview of what to expect from Smartfren SIM cards, helping them make more informed decisions.

III. RESULT AND DISCUSSION

The study uses two data sets, one for training and the other for testing. Once we obtain the data, we will initially select it based on its suitability for this study. We have chosen the data to ensure the smooth execution of the study. The next stage is to design the KNN model. Figure 1 displays the design of the KNN classification model.

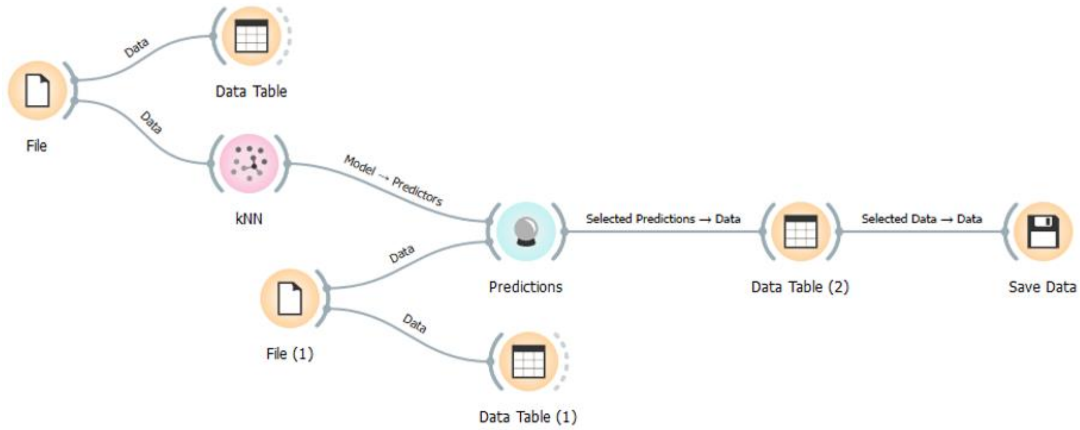

Fig 1. KNN Model Design

Figure 1 illustrates the design flow of the KNN model, which analyzes public interest in using Smartfren SIM cards. The File component shows the initial data source taken from an external file. This data comprises historical information from Smartfren SIM card users, providing pertinent information for analysis. We then enter the data from the file into a data table. Before the data undergoes further processing, this component serves as temporary storage. The kNN component displays how the KNN algorithm is applied to the data. KNN is a supervised learning algorithm used to classify data based on similarity. This algorithm searches for the closest data points (neighbors) of the sample to determine the class or category of the sample. The predictions component displays the KNN model's predicted results. This prediction is based on the similarity between new data and existing historical data. The KNN process enters historical data into the Data Table component (1) for comparison, which the model then stores and retrieves.

The File component (1) displays additional files that may contain additional or complementary data for analysis. The predictions component enters the prediction result into Data Table (2) for further processing or analysis. Once predicted and processed, the Save Data component stores the final data into a file for further analysis or documentation. The process starts by extracting data from an external file containing Smartfren SIM Card user information. We then enter this data into the data table. The KNN model uses the data in the Data Table as its input. The KNN algorithm processes this data by comparing each sample with historical data in the Data Table (1). The KNN model makes predictions about each sample's category or class based on the data's similarity or proximity. The Predictions component stores the prediction results. We enter the prediction results from the KNN model into the Data Table (2). The Save Data component then saves this data into a file for further analysis or documentation. The accuracy of the KNN model is highly dependent on the number of neighbors selected and the features used to compare the data. Therefore, choosing the right parameters is crucial to get accurate results.

Predicted

Actual	Interested	Not Interested	∑
Interested	65	0	65
Not Interested	4	1	5
∑	69	1	70

Fig 2. Confusion Matrix

Figure 2 displays the confusion matrix from the analysis of public interest in Smartfren SIM cards using the KNN method. A confusion matrix is a tool for evaluating the performance of a classification model. It shows the number of correct and incorrect predictions made by the model by comparing the predictions with the actual values. The confusion matrix consists of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The KNN model correctly predicts TP, or the amount of data that actually shows interest in using Smartfren SIM Cards, as "interested." In this context, the model predicts 65 out of 70 samples as truly interested. The model uses TP as a key indicator to accurately identify the majority of individuals who show interest in the product under study. The model accurately predicts TN, or the amount of data that genuinely lacks interest in using Smartfren SIM Cards, as "Not Interested." In this case, the model correctly predicts only one sample as truly not interested. TN shows the model's ability to identify individuals who are not interested, although in this case the number is very small. The model predicts the number of data points as "interested" even though they actually show no interest in using Smartfren SIM Cards.

The model predicts interest in Smartfren SIM Cards for 4 samples, despite their actual lack of interest. FP is a Type I error where the model predicts too "optimistically". When FP prioritizes resources for individuals who are not genuinely interested, it can pose a problem. FN represents the quantity of data points that genuinely express interest in using Smartfren SIM Cards, yet the model predicts them as "Not Interested". This result indicates that no samples, predicted as not interested (FN = 0), actually show interest in the Smartfren SIM Card. FN is also known as Type II error, and in marketing, this can mean missing the opportunity to target individuals who are actually interested in the product.The high TP value indicates that the KNN model is very effective in identifying individuals who are interested in Smartfren SIM Cards. This is critical for marketing strategies that want to target this segment correctly. The low TN value and the presence of FP indicate that the model is less efficient in identifying individuals who are not interested. Although there are only four cases of FP, this may mean that the marketing campaign targeted some individuals who were not interested, causing cost inefficiencies. The absence of FN indicates that the model did not exclude individuals who were truly interested. Ensuring the correct identification of all individuals who may respond positively to the marketing campaign is crucial. Overall, this confusion matrix shows that the KNN model in this study did a very good job of identifying interest in Smartfren SIM cards, but there is still room for improvement, particularly in identifying those who are not interested.

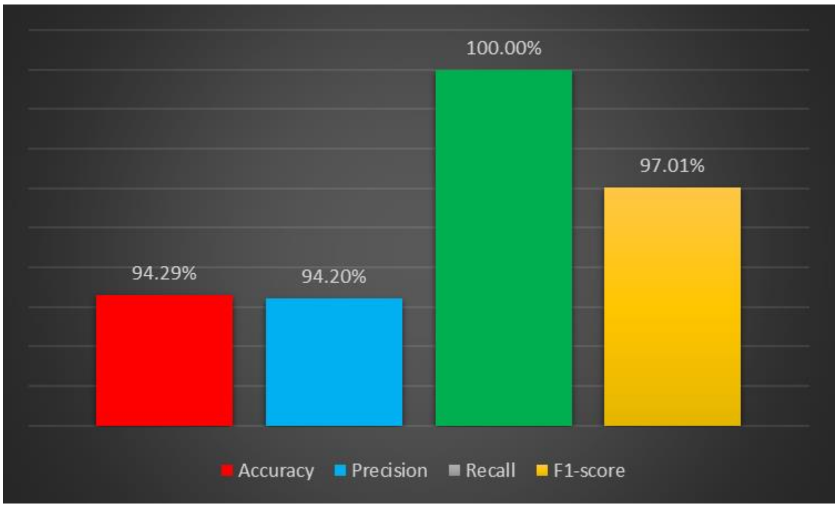

Fig 3. Performance Evaluation

The figure 3 shows the performance evaluation of the K-Nearest Neighbors (KNN) model used in this study. This evaluation includes key metrics such as accuracy, precision, recall, and F1-score displayed in the form of bar charts. The following provides a complete and detailed explanation of each metric, as well as an in-depth analysis of the results of this performance evaluation. Accuracy is defined as the total proportion of correct predictions (both positive and negative) from the total number of cases. An accuracy of 94.29% indicates that the model is very good at predicting interest in Smartfren SIM Cards as a whole. High accuracy means that most of the model's predictions are correct, both positive and negative. Precision is the proportion of correct positive predictions from all positive predictions generated by the model. A precision of 94.20% indicates that of all predictions stating that someone is interested in Smartfren SIM Cards, 94.20% of them are actually interested. In the context of marketing, this is crucial to prevent the waste of resources on the incorrect target. A high precision value indicates that the model is able to reduce Type I errors (false positives), which is a case where the model predicts "interested" when it is actually "not interested.

" Recall measures the proportion of correct positive predictions out of all cases that are actually positive. Perfect recall (100%) indicates that the model successfully identified all individuals who are truly interested. In the marketing context, this is crucial as it guarantees the identification of all potential interested customers. High recall indicates that the model is able to reduce Type II errors (false negatives), which is a case where the model fails to identify individuals who are truly interested. The F1-Score is the harmonic average of precision and recall. An F1-Score of 97.01% indicates an excellent balance between precision and recall. This means that the model is not only effective at identifying interested individuals (high recall) but also efficient at reducing false positive prediction errors (high precision). F1-Score is very useful in situations where there is a class imbalance, such as a larger number of "interested" data than "not interested.".The KNN model's performance evaluation results show excellent performance in terms of accuracy, precision, recall, and F1-score. High accuracy indicates that the KNN model is overall effective at predicting interest in Smartfren SIM Cards. This implies that we can rely on the model to accurately predict most of the tested data. In a marketing context, high accuracy means that campaigns based on model predictions will be on target and efficient in reaching audiences who are truly interested.

A high precision value indicates that the KNN model is effective in minimizing false positive prediction errors. In practice, this means that the model rarely makes mistakes in identifying individuals who are interested in Smartfren SIM Cards. The impact on marketing is significant as it allows the campaign to concentrate more on truly potential targets, thereby reducing advertising costs and minimizing wasted efforts. Perfect recall indicates that the KNN model does not miss individuals who are interested in Smartfren SIM Cards (False Negatives = 0). In a marketing context, this is crucial as it guarantees the inclusion of all interested individuals in the marketing campaign. This increases the chances of a marketing campaign being successful by reaching the entire potential audience. A high F1-Score indicates a fantastic balance between precision and recall. This means that the model is not only accurate in identifying interested individuals but also efficient in predicting relevant targets. In real applications, this balance ensures that marketing campaigns are not only targeted but also comprehensive and inclusive. With high accuracy and F1-score, the KNN model shows a strong and reliable performance in predicting people's interests. The absence of false negatives indicates that the model is very efficient in identifying all interested individuals, which is very important for marketing strategies. Despite the high precision, four cases of false positives suggest that the model still requires further improvement to minimize false positive prediction errors. The model can be influenced by the high number of positive data (interested) compared to negative data (not interested). We can further explore the use of data balancing techniques or other algorithms.

IV. CONCLUSION

This study aims to analyze public interest in using Smartfren SIM Cards using the K-Nearest Neighbors (KNN) method. Based on the performance evaluation results, the KNN model used in this study performed very well with high evaluation metrics. An accuracy of 94.29% indicates that the model is able to correctly classify public interest on most of the test data. Precision of 94.20% indicates the model's ability to identify individuals who are truly interested in Smartfren SIM Cards, minimizing false positive predictions. A recall of 100% indicates the model's ability to identify all interested individuals without missing any. F1-Score of 97.01% shows a good balance between precision and recall, ensuring that this model is not only accurate but also efficient in its predictions. This study successfully shows that the KNN method is effective in analyzing public interest in using Smartfren SIM cards.

We can rely on this model for more effective and efficient marketing strategies due to its high accuracy and recall. However, there is potential for further improvement through data balancing, hyperparameter optimization, and exploration of other algorithms. Further research following these recommendations can improve the model performance and ensure its success in a wider range of conditions and datasets. We propose several recommendations to enhance the model performance and validate the research results. Perform hyperparameter tuning to find the best configuration that can improve the model's performance. Conduct a thorough examination of the features utilized to guarantee the inclusion of all pertinent features.

REFERENCES

[1] N. Anwar, I. Riadi, and A. Luthfi, “Analisis SIM Card Cloning Terhadap Algoritma Random Number Generator,” ***J. Buana Inform***., vol. 7, no. 2, pp. 143–150, 2016.

[2] E. M. Sari Rochman, I. Imamah, and A. Rachmad, “Desain Model Peramalan Jumlah Kerusakan SIM CARD,” *Multitek Indones.*, vol. 12, no. 1, p. 20, 2018, doi: 10.24269/mtkind.v12i1.640.

[3] L. N. Rani, “Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit,” *INOVTEK Polbeng - Seri Inform.*, vol. 1, no. 2, p. 126, 2016, doi: 10.35314/isi.v1i2.131.

[4] Suwati, R. Yesputra, and A. Sapta, “Prediksi Kelancaran Pembayaran Angsuran Pada Koperasi Dengan Metode Naive Bayes Classifier,” ***Indones. J. Comput. Sci.***, vol. 11, no. 2, pp. 635–644, 2022, doi: 10.33022/ijcs.v11i2.3080.

[5] Muttaqin *et al.*, *Pengenalan Data Mining*, 1st ed., no. July. Yayasan Kita Menulis, 2023.

[6] A. Yudhistira, A. A. Aldino, and D. Darwis, “Analisis Klasterisasi Penilaian Kinerja Pegawai Menggunakan Metode Fuzzy C-Means (Studi Kasus : Pengadilan Tinggi Agama bandar lampung),**” *J. Ilm. Edutic Pendidik. dan Inform***., vol. 9, no. 1, pp. 77–82, 2022, doi: 10.21107/edutic.v9i1.17134.

[7] D. Firdaus, “Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer,” ***J. Format,*** vol. 6, no. 2, pp. 91–97, 2017.

[8] P. Primadona and R. Fauzi, “Penerapan Data Mining Pada Penjualan Produk Elektronik,” *Comput. Sci. Ind. Eng.*, vol. 9, no. 4, 2023, doi: 10.33884/comasiejournal.v9i4.7712.

[9] B. Hssina, A. Grota, and M. Erritali, “Recommendation system using the k-nearest neighbors and singular value decomposition algorithms,” ***Int. J. Electr. Comput. Eng.,*** vol. 11, no. 6, pp. 5541–5548, 2021, doi: 10.11591/ijece.v11i6.pp5541-5548.

[10] A. S. Abdelmagid and A. I. M. Qahmash, “Utilizing the Educational Data Mining Techniques Orange Technology ‘for Detecting Patterns and Predicting Academic Performance of University Students,’” *Inf. Sci. Lett.*, vol. 12, no. 3, pp. 1415–1431, 2023, doi: 10.18576/isl/120330.

[11] E. Y. Boateng, J. Otoo, and D. A. Abaye, “Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review,” ***J. Data Anal. Inf. Process.,*** vol. 08, no. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[12] R. Puspita and A. Widodo, “Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS,” ***J. Inform. Univ. Pamulang,*** vol. 5, no. 4, p. 646, 2021, doi: 10.32493/informatika.v5i4.7622.

[13] S. Sumarlin, “Implementasi Algoritma K-Nearest Neighbor Sebagai Pendukung Keputusan Klasifikasi Penerima Beasiswa PPA dan BBM,” ***J. Sist. Inf. Bisnis,*** vol. 5, no. 1, pp. 52–62, 2015, doi: 10.21456/vol5iss1pp52-62.

[14] S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, “Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa,” ***IJCIT (Indonesian J. Comput. Inf. Technol.,*** vol. 6, no. 2, pp. 118–127, 2021.

[15] H. A. Pratama, G. J. Yanris, and M. N. S. Hasibuan, “Implementation of Data Mining for Data Classification of Visitor Satisfaction Levels,” ***Sink. J. dan Penelit. Tek. Inform***., vol. 8, no. 3, pp. 1832–1851, 2023, doi: 10.33395/sinkron.v8i3.12674.