PREDICTION OF HOTEL BOOKING CANCELLATION USING K-NEAREST NEIGHBORS (K-NN) ALGORITHM AND SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE)

Adli Abdillah Nababan [1], Miftahul Jannah [2] , Arif Hamied Nababan[3]
[1,2]Bisnis Digital, STMIK Pelita Nusantara, Medan, Indonesia
[3]Teknik Informatika, Universitas Prima Indonesia, Medan, Indonesia
E-mail : *adliabdillahnababan@gmail.com, miftahuljannah0077@gmail.com,
arifhamiednababan@unprimdn.ac.id

Abstract

Article Info
Receive, 01/08/22
Revised, 24/08/22
Accepted, 25/08/22

Cancellation of bookings puts considerable pressure on management decisions, in this case from the hospitality industry. Cancellation of bookings limits the correct prediction and is, therefore, a very important tool for revenue management performance. However, in recent times, thanks to the availability of considerable computing power through machine learning approaches, it has become possible to create more accurate models for predicting booking cancellations compared to using more traditional methods. Previous research has used several machine learning approaches, such as Decision Tree, Support Vector Machine, Deep Neural Network, Logistic Regression, and Random Forest to predict hotel cancellations. However, they have not addressed the class imbalance problem that exists in predicting hotel cancellations. In this study, we have provided a solution by introducing an oversampling technique to solve the class imbalance problem, together with the k-nearest neighbors algorithm to predict hotel booking cancellations better. The results of this study show that an increase in the performance of the method's accuracy increased by 3.88%, precision increased by 9.00%, recall increased by 10.00%, and F1-Score increased by 10.00% in the hotel booking dataset. It can be concluded that the SMOTE method with KNN has a better performance than only using the KNN method in predicting the cancellation of hotel reservations.

Keywords: prediction, hotel booking cancellation, k-nn, smote

1. Introduction

Along with the development of increasingly rapid technological advances, competition in terms of the marketing of goods and services is important in maximizing profits. With the development of the use of technology, the data collected in the database also of course accumulate from time to time.

There are many companies engaged in the hospitality sector for hotel reservations in Indonesia. To continue to survive in the competition, of course, each company does various ways to maintain customer loyalty by preparing innovations and evaluations to continue to provide the best service. People can make hotel reservations quickly and easily. However, this convenience can be detrimental to the company caused to various factors, one of which is data errors so that hotel reservations are not on target and lead to cancellation of hotel reservations. One technique that can overcome the data problem is using data mining techniques. Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases [1]. If the marketing strategy is carried out incorrectly, the company will spend a lot of time, effort, and money in vain. This will harm the company's profits.

Based on the explanation above, this study will discuss the classification of data in predicting the cancellation of hotel reservations in the future. The algorithm proposed in this study is the K-Nearest Neighbors algorithm. The K-Nearest Neighbors algorithm is a method used to classify objects based on training data that has the closest distance to the object [2]. The number of samples of hotel booking

cancellation data obtained from the keaggle repository datasets experienced an imbalance class where the number of successful bookings was more than bookings with cancellation status.
Research conducted by [3] on the topic of predicting hotel booking cancellations to increase company revenue. Based on the results of research with the model used to classify hotel reservations that are likely to be canceled using Boosted Decision Tree (BTS) and Decision Forest. The results of this study produce values that are considered very good.

Research conducted by [4] on the topic of predicting hotel booking cancellations using the Deep Neural Network Algorithm and Logistic Regression. In the experimental scenario that has been carried out, the results of the neural network after adding the learning rate show that the smaller the learning rate, the greater the accuracy, but we do not know what the optimal value is for the learning rate. By using the Logistic Regression algorithm by eliminating several attributes, the most influential accuracy levels are the country attribute and total of special requests, where previously the accuracy was 79.66%, an increase of 80.29% when the country attribute was omitted because there were 177 variations in that attribute.
Based on the research that has been done previously, it will be used as a reference or guide in this research. The difference between this study and previous research is in terms of the attributes used and in terms of the data used. The algorithm proposed in this study is the K-Nearest Neighbors algorithm. The K-Nearest Neighbors algorithm is a method used to classify objects based on training data that has the closest distance to the object [5].

Most of the classification algorithms tend to implicitly assume that the processed data has a balanced distribution, therefore the standard classifier is more inclined toward data with a dominant number of classes [6]. Synthetic Minority Oversampling Technique (SMOTE) is used in class balancing to apply the K-NN algorithm so that the classification does not lead to the majority of data. It is hoped that the combination of the SMOTE method can provide data that represents the original data and can increase the accuracy of the K-NN method to predict the cancellation of hotel reservations in the future. From this data, the company can see the pattern of orders that are indicated to be canceled. This research can also be used as an evaluation tool for hotel companies in minimizing losses that will occur by making policy/decision making based on new information obtained from the results of the data mining process [7].

2. Method

2.1 Research Stages

Stages are a process, namely a series of steps that are carried out in a planned and systematic way to get problem-solving or get answers to certain questions. In conducting this research, there are steps taken, namely identifying the problem, making formulations and problem boundaries so that the problem under study is clear and does not deviate from the problem, and determining the purpose and usefulness of the research so that this research runs well. The stages in this research can be described in the following figure:

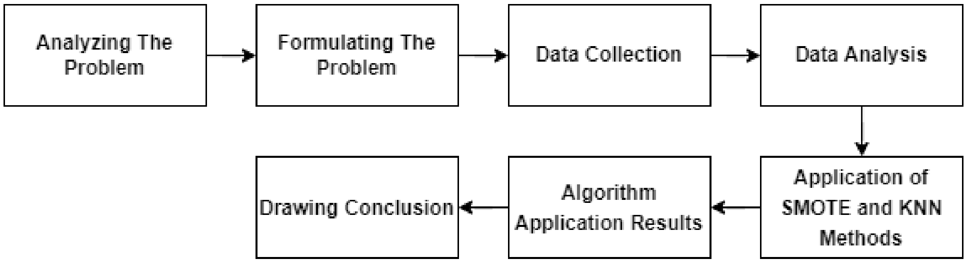

Figure 1. Research Stages

Based on the research stages in Figure 1, can be described as follows:

2.2 Analyzing the Problem

Analysis of the problem is the problem that will be discussed in this study and the data that will be used in the analysis and testing that will be carried out is the data used based on the restriction of hotel reservations.

2.3 Formulating the Problem

At this stage, the formulation of the problem obtained from this research is: how to predict the KKN and SMOTE algorithms for hotel booking cancellations.

2.4 Data Collection

a. **Literature Study:** Literature study by collecting, searching, and studying various reading materials and necessary data sourced from books on the KNN and SMOTE Algorithm.

b. **Observation:** This observation is done by observing the data that will be used in testing the KNN and SMOTE Algorithms.

2.5 Data Analysis

Data analysis aims to define the requirements needed in research and analysis of algorithm performance results. This is related to the prediction of hotel booking cancellations with the KNN and SMOTE algorithms.

2.6 Application of SMOTE and KNN Methods

This stage is to determine SMOTE and KNN in predicting the cancellation of hotel reservations. The SMOTE method works by looking for K-Nearest Neighbors (that is, K-nearest neighbors of data) for each data in the minority class, after that synthetic data is made as much as the desired percentage of duplication between the minor data and randomly selected K-Nearest Neighbors [8].

The SMOTE method is one of the most popular methods applied to handle data with class unbalanced conditions. An imbalanced class can be interpreted as the presence of data that has a greater tendency (majority class) than other data [9]. In hotel booking data, there are more successful bookings as the majority class, while bookings with cancel status are fewer or called the minority class. For this reason, SMOTE is used to balance class to produce a more accurate classification.

Synthetic data processing on data with different numerical feature values from data with categorical feature values, according to [10] Euclidean equation is used to measure the similarity of numerical data, while categorical data is measured using the Value Difference Metric (VDM) formula, namely:

$$d(V_1,V_2)=\sum_{i=1}^{N}\left|\frac{C_{1i}}{C_1}-\frac{C_{2i}}{C_2}\right| \text{...........}(1)$$

The steps in the classification of the K-Nearest Neighbor (KNN) according to [11],[12] are:

1. Determine the parameter k (number of closest neighbors).
2. Calculating proximity based on the Euclidean distance model to the given training data, with the equation:

$$\mathrm{D}(x,y) = ||x-y||_2\sqrt{\sum_{j=1}^{N}|x-y|^2} \text{ ....}(2)$$

3. Sorting the distance results obtained in ascending order (sequentially from high to low value).
4. Count the number of each class based on the k nearest neighbors.
5. The majority class is used as the class for the test data.

2.7 Algorithm Application Results

This stage measures the performance of the application of the KNN and SMOTE algorithms in predicting the cancellation of hotel reservations in terms of accuracy, precision, recall, and f1-measure.

2.8 Drawing Conclusion

At this stage, the researcher concludes and provides suggestions from the analysis stage of the application of the KNN and SMOTE algorithms in predicting the cancellation of hotel reservations.

3. Results and Discussion

In this study, the authors started with the SMOTE technique, the data which was originally a hotel booking cancellation data where there were more bookings with check-out status as the majority class and canceled as the minority class because the data was less. The state of this data is not balanced or is called an imbalance class. To balance the data, the SMOTE technique is used. With this method, we can make the dataset balanced without being too overfit, by making synthetic samples rather than duplicating samples. From the SMOTE technique process, new balanced data will be obtained. Furthermore, the data in a balanced state is processed by the K-NN calculation. The calculation process begins by determining the K value to be processed based on the specified K parameter value, then performing the Euclidean Data process to calculate the distance of each neighbor, then sorting the results by distance, starting from the smallest to the largest, then carrying out the process of determining or voting.

In general, this study will use resampling data with the SMOTE approach to handle hotel booking cancellation data, as shown in Figure 2. The stages of the K-NN and SMOTE algorithms can be seen in the image below:

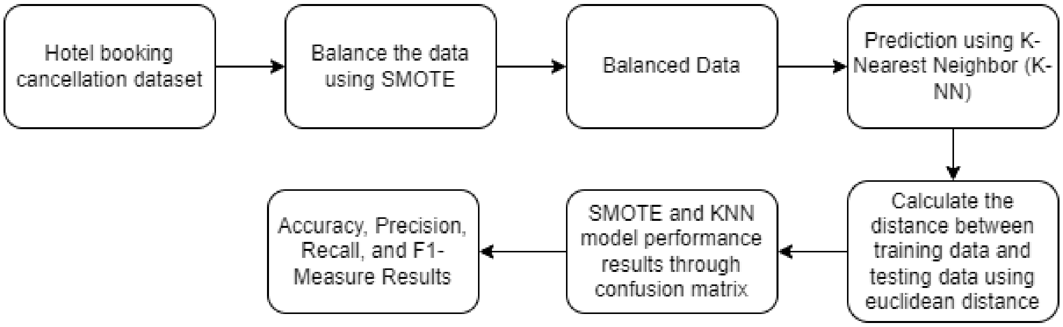
Hotel booking cancellation dataset
Balance the data using SMOTE
Balanced Data
Prediction using K-Nearest Neighbor (K-NN)
Calculate the distance between training data and testing data using euclidean distance
SMOTE and KNN model performance results through confusion matrix
Accuracy, Precision, Recall, and F1-Measure Results

Figure 2. SMOTE and KNN stages

3.1 Data Requirements

The first stage is data collection, where the hotel reservation cancellation dataset is used which consists of 2 classes with 16 features and 39770 instances. The details of the data used can be seen in table 1.

Table 1. Hotel Booking Dataset Features

No.	*Fitur*
1	*LeadTime*
2	*ArrivalDateYear*
3	*ArrivalDateWeekNumber*
4	*ArrivalDateDayOfMonth*
5	*StaysInWeekendNights*
6	*StaysInWeekNights*
7	*Adults*
8	*Children*
9	*Babies*

10	*PreviousCancellations*
11	*PreviousBookingsNotCanceled*
12	*BookingChanges*
13	*DaysInWaitingList*
14	ADR
15	*RequiredCarParkingSpaces*
16	*TotalOfSpecialRequests*

As shown in Table 1. The data will be used to determine how many minorities and majority classes there are. To see whether the suggested technique can produce higher accuracy results, this study will compare it with the K-NN method which uses a hotel reservation dataset. Table 2 shows the data set used by class distribution:

Table 2. Data Class Distribution

Data	***Features***	***Classes***	***Class Distribution***	
			P	**N**
Hotel Booking	16	2	28933	10826

3.2 Oversampling Process

The dataset will be oversampled using the Synthetic Minority Oversampling Technique (SMOTE) in the second step to balance the unbalanced quantity of data between the positive and negative classes. The data will be oversampled first using SMOTE, Table 3 contains details for the new dataset.

Table 3. Detailed Data After SMOTE

Data	***Features***	***Classes***	***Class Distribution***	
			P	**N**
Hotel Booking	16	2	28938	28938

3.3 Testing

The next step is to use K-Nearest Neighbor to perform calculations using a confusion matrix between training and test data on each dataset (K-NN). We opted for an 80% and 20% split strategy to maintain feature continuity. Specifically, for each sample of data, we use the first 80% for training and the remaining 20% for testing.

3.4 Evaluation

Finally, compare the new data set using 80/20 data separation in conducting the data classification process to see whether SMOTE and K-NN and K-NN without SMOTE have better performance in terms of Accuracy, Precision, Recall, and F-1 Score.

3.5 Result

Based on Figure 3 the average accuracy of the data, the accuracy value for K-NN in the Hotel Booking dataset is 79.35%, while SMOTE + KNN is 83.23%.

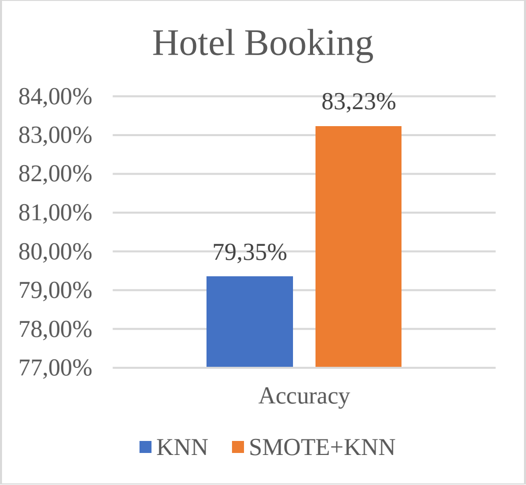

Figure 3. Average Accuracy from the data

Table 4 shows that SMOTE+KNN outperformed K-NN in terms of precision, recall, and F1-Score when applied to the Hotel Booking dataset, although the scores were not superior in terms of accuracy. Precision up 18.00%, recall up 28.00% and F1-Score up 27.00% in the Hotel Bookers dataset.

Table 4. Performance of K-NN and SMOTE+KNN

Performance	***Hotel Booking***		
	K-NN	***SMOTE+ K-NN***	***Avg Increase***
Accuracy	79,35%	83,23%	3,88%
Precision	74,00%	83,00%	9.00%
Recall	73,00%	83,00%	10.00%
F1-Score	73,00%	83,00%	10.00%

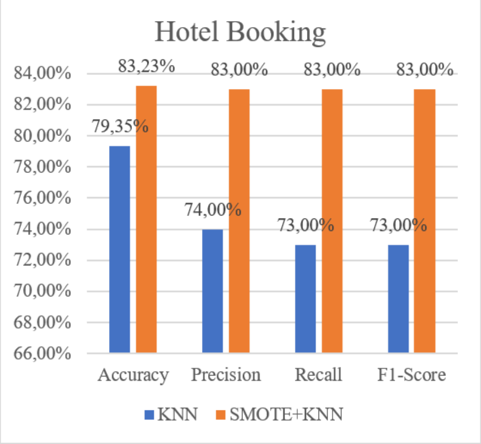

Figure 4. Performance Comparison of K-NN and SMOTE+KNN

4. Conclusion

Hotel booking data in the target class “Cancelled” can be handled using SMOTE. The proposed method can empower hotel booking supervisors to calculate their losses arising from continued cancellation of hotel bookings and limiting issues related to overbooking (redistribution fees, money, or administrative payments.Based on the explanation in the previous section, it can be concluded that the use of SMOTE can improve the performance of the K-NN method for hotel booking data. The accuracy produced by SMOTE with KNN is 83.23% compared to KNN without using SMOTE which is 79.35%, the recommended technique is to improve classification performance, the proposed method can outperform K-NN using SMOTE in terms of precision, recall, and accuracy. F1-Score when applied to the Hotel Bookings dataset. Accuracy is up 3.88%, Precision is up 9.00%, Recall is up 10.00%, and F1-Score is up 10.00% on the hotel booking dataset. It can be concluded that the SMOTE method with KNN has a better performance than using only the KNN method.

REFERENCE

[1] P. Butka, P. Bednár, and J. Ivančáková, “Methodologies for Knowledge Discovery Processes in Context of AstroGeoInformatics,” *Knowl. Discov. Big Data from Astron. Earth Obs. Astrogeoinformatics*, pp. 1–20, 2020, doi: 10.1016/B978-0-12-819154-5.00010-2.

[2] R. N. Yusra and O. S. Sitompul, “InfoTekJar : Jurnal Nasional Informatika dan Kombinasi K-Nearest Neighbor (KNN) dan Relief-F Untuk Meningkatkan Akurasi Pada Klasifikasi Data,” vol. 1, pp. 0–5, 2021.

[3] N. Antonio, A. de Almeida, and L. Nunes, “Predicting hotel booking cancellations to decrease uncertainty and increase revenue,” *Tour. Manag. Stud.*, vol. 13, no. 2, pp. 25–39, 2017, doi: 10.18089/tms.2017.13203.

[4] Y. Azhar, G. A. Mahesa, and M. C. Mustaqim, “Prediction of hotel bookings cancellation using hyperparameter optimization on Random Forest algorithm,” *J. Teknol. dan Sist. Komput.*, vol. 9, no. 1, pp. 15–21, 2021, doi: 10.14710/jtsiskom.2020.13790.

[5] N. Z. Dina and R. S. Marjianto, “PREDIKSI PENENTUAN PENERIMA BEASISWA DENGAN METODE KNEAREST NEIGHBOURS (Studi Kasus: Program Studi Sistem Informasi Fakultas Vokasi Universitas Airlangga),” *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 2, no. 2, pp. 135–139, 2018, doi: 10.30743/infotekjar.v2i2.269.

[6] E. Sutoyo and M. A. Fadlurrahman, “Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network,” *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 3, p. 379, 2020, doi: 10.26418/jp.v6i3.42896.

[7] D. L. Panjaitan and P. M. Hasugian, “Implementation of K-Nearest Neighbor Algorithm for Classification of Class Placement At Junior High School , Padang Month Issn : 2302-9706,” *J. Infokum*, vol. 10, no. 1, pp. 43–49, 2021.

[8] R. Perangin-angin, E. J. G. Harianja, and I. K. Jaya, “Pendekatan Level Data untuk Menangani Ketidakseimbangan Data Menggunakan Algoritma K-Nearest Neighbor,” *J. TIMES*, vol. IX, no. 1, pp. 22–32, 2020.

[9] H. Hairani, K. E. Saputro, and S. Fadli, “K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes,” *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[10] K. U. Syaliman, “ENHANCE THE ACCURACY OF K-NEAREST NEIGHBOR (K-NN) FOR UNBALANCED CLASS DATA USING SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) AND GAIN RATIO (GR),” vol. 10, no. 1, pp. 188–195, 2021.

[11] A. A. Nababan, M. Khairi, and B. S. Harahap, “Implementation of K-Nearest Neighbors (KNN) Algorithm in Classification of Data Water Quality,” vol. 6, no. 36, pp. 30–35, 2022.

[12] Y. Yuliska and K. U. Syaliman, “Peningkatan Akurasi K-Nearest Neighbor Pada Data Index Standar Pencemaran Udara Kota Pekanbaru,” *IT J. Res. Dev.*, vol. 5, no. 1, pp. 11–18, 2020, doi: 10.25299/itjrd.2020.vol5(1).4680.