OPTIMIZATION OF TIME DEPOSIT CLASSIFICATION USING KNN ALGORITHM WITH INFORMATION GAIN FEATURE SELECTION

*[1]**Anggun Pambudi**,[2] **Fikri Budiman**,[3] **M. Arief Soeleman**
[4]**Purwanto**,[5] **Aris Marjuni**,[6] **A. Zainul Fanani**

[1]*Faculty of Computer Science,Institution, Dian Nuswantoro University,Semarang, Indonesia*
[2]*Faculty of Computer Science,Institution, Dian Nuswantoro University,Semarang, Indonesia*
[3]*Faculty of Computer Science,Institution, Dian Nuswantoro University,Semarang, Indonesia*
[4]*Faculty of Computer Science,Institution, Dian Nuswantoro University, Semarang, Indonesia*
[5] *Faculty of Computer Science,Institution, Dian Nuswantoro University,Semarang, Indonesia*
[6] *Faculty of Computer Science,Institution, Dian Nuswantoro University,Semarang, Indonesia*

Author's email:
[1]*anggun.pambudi@gmail.com* ;[2] *fikri.budiman@dsn.dinus.ac.id* ;[3] *ARIEF22208@gmail.com* ,
[4] *purwanto@dsn.dinus.ac.id* ;[5] *aris.marjuni@dsn.dinus.ac.id* ; [6] *a.zainul.fanani@dsn.dinus.ac.id*
**Corresponding author:* [1]*anggun.pambudi@gmail.com*

Abstract. *A broad marketing strategy in the banking and financial services industry to serve business, investment, and banking companies has the opportunity to carry out promotions and marketing strategies efficiently with the help of technology. Customer identity that provides a promotional response to a product is an important part of marketing. Data mining can provide solutions to these problems. With the method used in completing direct marketing, there are still data features that have little effect on performance in direct marketing so it provides a low level of accuracy to increase the level of performance is done by reducing several features that do not affect direct marketing by using add information gain on the KNN algorithm to get good performance. The first step is to model with the KNN algorithm in a limited way using 5-fold cross-validation to produce the best K value, namely K = 9 with an accuracy of 89.72%, 98% recall, and 91% precision. The addition of information gain to the KNN algorithm at K=9 using 5-fold-cross-validation produces an accuracy value of 90.49%, a recall of 97 %, and 92% precision so that the addition of information gain to the KNN algorithm can increase the accuracy value and provide an increase in the proportion of precision.*

Keywords: *Data mining, KNN, Feature Selection, Information Gain, Bank Telemarketing*

1. INTRODUCTION

Business, investment, and banking businesses have the chance to efficiently implement promotional and marketing plans in the financial services and banking industries by using technology. These issues can be solved via data mining. The relational database can be used to store data. Data mining may be used to boost targets when choosing potential clients and is highly successful for marketing. The act of gathering, purifying, processing, analyzing, and concluding data is known as data mining (Aggarwal 2015; Koumétio and Toulni 2021; Moro, Cortez, and Rita 2014)..

Telemarketing has already been the subject of studies. In this experiment, Koumetio & Hamza Touln enhance the KNN model for direct marketing in smart cities (Koumétio & Toulni 2021). In their study, Kim K.H., Lee C.S., Jo. S.M., and Cho. S.B. used direct marketing datasets to predict the success of banking telemarketing (Kim, Lee, and Jo 2015). Cortez, P., Rita, and Moro, Research on applying the NN algorithm to predict telemarketing success based on data (Moro, Cortez, and Rita 2014). Zeinulla, Bekbayeva, and Yazici's research on the same dataset as other studies compared classification models that use the KNN, LR, ANN, NB, and SVM algorithms to predict bank telemarketing (Zeinulla, Bekbayeva, and Yazici, 2018). An approach to data modeling for classification problems in banking telemarketing predictions using ANN, SVM, NB, DT, and LR algorithms was developed by Tekouabou, S.C.K., Cherif, W., and Silken, H. using direct marketing dataset from Portuguese banking institutions. The research focused on optimizing predictions on telemarketing target telephones with classification techniques using NB, DT, ANN, and SVM algorithms by combining with normalization.In 2019, Koumetio, Cherif, and Hassan.

Data aspects still have little effect on direct marketing achievement, which has resulted in a low degree of accuracy despite the wide range of approaches employed in past studies to complete direct marketing. Quality prospect data, an understanding of consumer behavior, and the application of machine learning to identify prospects with a better likelihood of becoming clients are the key criteria in direct marketing.A 2019 study by Teouabou, Cherif, and Silkan. The features of the input data, such as bank customer information, which includes numerous aspects, are one of the primary elements that influence prediction performance. When the details of the data are not as important, the degree of prediction performance declines.A 2019 study by Teouabou, Cherif, and Silkan.

The goal of this investigation is to use information gain to reduce the number of unnecessary characteristics, which will improve the performance of the KNN algorithm in direct marketing. Due to the availability of several pointless characteristics, the KNN method had previously performed less accurately than other algorithms like Random Forest, NN, and ANN. By utilizing information gain, the research intends to improve the efficiency of the KNN algorithm by lowering the number of pointless features. This will be performed by restricted modeling using the KNN method, which will be assessed and enhanced to get the ideal K value. The KNN algorithm will then be enhanced using information gain, and its performance will be evaluated using k-fold cross-validation. The increased KNN algorithm's performance will be assessed in the study together with that of other direct marketing algorithms including Random Forest, NN, and ANN. The research's overall goal is to address the problem of the KNN algorithm's reduced accuracy caused by unimportant features and to offer a more effective and efficient way for direct marketing that uses the KNN algorithm with information gain.

In order to improve the precision of findings, Information Gain is applied to eliminate features that have no impact on direct marketing. It is helpful to know that the approach is superior to the current ways by comparing the accuracy results between KNN and Information Gain.(Tekouabou, Cherif, and Silkan 2019)

2. LITERATURE REVIEW

2.1 Related Research

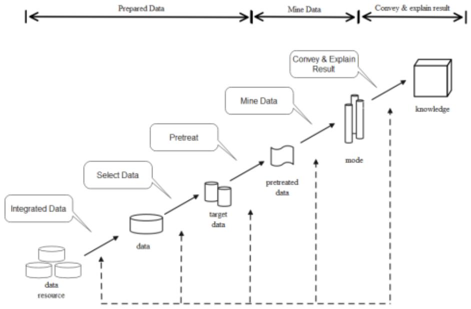
The effectiveness of the algorithm is assessed by ROC (Receiving Operator Characteristic) and curve analysis CAP (Cumulative Analysis Profile), the accuracy of the algorithm, and the scalability of the algorithm. Several related studies have been conducted in the past, such as those conducted by Elzan Zeinulla, Karina Bekbayeva, and Adnan Yazici, 2018, which discuss evaluating several types of classification models for predicting bank telemarketing campaigns where customers are likely to subscribe to deposits. According to Zeinulla, Bebayeva, and Yazici (2018), the Random Forest algorithm generated the greatest accuracy in their study (90.88%), with a percentage of 50% positive observation (CAP Curve) of 95.83%. Research conducted by Moro. S, Cortez. P, and Rita P, 2014 on data-based approach research to predict success in telemarketing with the best performance at AUC 0.8 and ALIFT 0.67 for the NN algorithm (Moro, Cortez, and Rita 2014).

The upgraded KNN model is used to optimize predictions on telemarketing telephone targets for banking time deposit products in smart cities, according to research by Stephane Cedric Tekouabou Koumetio and Hamza Toulni published in 2021. By choosing important qualities and normalizing them, simple algorithms with increased data processing capabilities may be used to improve performance and accelerate the research process. For f-measures, the performance outcomes of the suggested technique have improved by an average of more than 96.91% while processing time has decreased (Koumétio and Toulni, 2021). In his study using the direct marketing dataset for memories of success in banking telemarketing using a deep convolutional neural network, Kim K.H., Lee C.S., Jo. S.M., and Cho. S.B. produced an accuracy value of 76.70% compared to other algorithms (Kim, Lee, and Jo 2015). To choose the algorithm with the best performance, Saputra. E. P. conducted research in 2017 that forecasts the success of bank telemarketing. According to performance data of the suggested strategy, the best accuracy is 91.80% (Saputra 2017). The proposed approach by combining classification algorithms with data normalization can perform well in f-measures and reach 60.12% of customers (Ismail Fawaz et al. 2019), and research in 2019, with the same dataset discussing approaches to data mode, shows that the proposed methodology by combining classification algorithms with data normalization can perform well in f-measures and reach 60.12% of customers. (Tekouabou, Cherif, and Silkan 2019) Highest accuracy and F-measure of 100% with DT without normalization.

Enhancing the KNN algorithm's capacity for direct marketing feature optimization would result in substantial and more accurate performance when categorizing target clients for bank telemarketing. Therefore, researchers will enhance the KNN algorithm's performance by incorporating information gain into the algorithm, which aims to obtain the best accuracy and can be used as a recommendation to banking staff so that they can quickly identify suitable potential customers for time deposit products, can improve performance businesses, and can lower marketing operational costs.

2.2 Data Mining

Finding useful information from a set of data is known as data mining. According to P. Sundari and K. Thangadurai (2010), data mining is the process of obtaining previously undiscovered facts or information from vast volumes of usable data. According to Zhang et al. (2015), data mining is a broad term for a difficult and disjointed research process that uses an algorithmic model to search for undiscovered important information. Below is a description of data mining's fundamental procedure:



Gambar 2.1 Alur Proses Dasar Pada Data Mining

2.3 Normalization

In data mining, normalization is the first step. Data is obtained by using different numbers, such as doubles and integers. Data can be standardized with desired values throughout the normalization process (Z. Han et al. 2017). The following method might be used to calculate normalized values for all characteristics with a range between 0 and 1:

$$z = \frac{Max\ v_i - v_i}{\mathrm{Max}\, v_i - \mathrm{Min}\, v_i}$$

The value of v_i is the actual value of attribute *i*, and the maximum and minimum are taken for all examples of the training set.

2.4 KNN Algorithm

When estimating and predicting a model or approach from supervised learning (Punjabi and Prajapati 2018; Puspadini 2020; Syadzali, Suryono, and Endro Suseno 2020) or a non-parametric method for classification that surrounds it, the KNN algorithm is utilized. and the separation between the two points for the ideal k value, where k is the number of training samples that are closest to one another (Zeinulla, Bebayeva, and Yazici, 2018).

KNN uses the calculation method with the shortest distance using Euclidean with the equation formula and the following steps:

a. Determining K Parameter Values

b. Calculating Euclidean Distance

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (Xi - Yi)^2}$$

Note: Matrix D(X,Y) is the distance between the two vectors X and Y. X is sample data and Y is test data.

c. Sort the distance results and determine the nearest neighbors

d. Using the simple majority of the nearest neighbor class as the prediction value of the temporary new data to find the predicted value of k-NN is calculated by the equation:

$$Y = \frac{1}{K} \sum_{\boldsymbol{i=1}}^{\boldsymbol{K}} \boldsymbol{yi}$$

2.5 Information Gain

The first step in the Information Gain process is to find the Entropy value shown in the equation:

$$Entropy(S) = \sum_{i}^{c} - P_{i\ n} Log\ _2 P_i$$

Note :

c	: accumulated grades from the classification class
Pi	: accumulation of samples from the class i.

Then after the value $Entropy$ is obtained, the Information Gain calculation process can be carried out with the following formula:

$$Gain(S,A) = Entropy(S) - \sum_{values(A)} \frac{|S_p|}{|S|_n} Entropy(S_p)$$

Note :

Gain(S, A)	: Value Gain from attribute
A	: Attribute
V	: the total value of attribute *A*
(A)	: possible values of set A
Sv	: Number of sample values from v
S	: The total number of data samples
Entropy (Sv)	: $Entropy$ value sample v

2.6 Model Validation and Evaluation

In evaluating the performance accuracy of the classification model using the confusion matrix. Testing on the classification model is expected to run by providing performance with a good level of accuracy and producing the smallest error value. Confusion Matrix in table 2.1.

Table 2.1 Confusion Matrix for Two Classes

	Kelas Prediksi		
True Class	**Positif**	**Negatif**	**Total**
Positif	TP=*True Postive*	FN=*False Negative*	P
Negatif	FP=*False Postive*	TN=*True Negative*	N
Total	p’	n’	N

There are 4 additional terms which are "building blocks" to calculate the number of calculations in the evaluation (J. Han, Kamber, and Pei 2012).

a. True Positive (TP) : Refers to a correctly labeled positive tuple. TP is the number of True Positives.
b. True Negative (TN) : Negative tuples that are correctly labeled during classification. TN is the number of True Negative.
c. False Positive (FP) : Negative tuples that are incorrectly labeled as positive. FP is the number of False Negatives.
d. False Negative (FN) : Positive tuples that are incorrectly labeled as negative. FN is the number of False Negatives.

The analysis of classes using a confusion matrix to find tuples from class differences. When the classification makes accurate predictions, TP and TN convey information, whereas FP and FN do the same when the classification makes inaccurate predictions (ALPAYDIN 2014; Wang et al. 2014).

The Confusion Matrix, measurement metrics are made to obtain accuracy values with the formula (P. Sundari and K. Thangadurai 2010; Wang et al. 2014; Zhang et al. 2015):

$$akurasi = \frac{TP+TN}{TP+TN+FP+FN}$$

There are also effective measurement methods in the classification process :

a. *Precision*

Precision is the degree of accuracy at what percentage of tuples that have been labeled positive classifications which are actually positive.

$$precision = \frac{TP}{TP+FP}$$

b. *Recall*

Recall is the percent completeness of positive tuples that have been classified with positive labels.

$$recall = \frac{TP}{TP+FN}$$

3. METHODOLOGY

Secondary data, or actual data that is already present in the UCI repository, served in the present study. The collection of information was collected from Portuguese financial institutions between 2008 and 2013. This dataset focuses on Portuguese direct marketing to financial institutions. Calls are the basis of this marketing. To determine if the consumer would accept the product and subscribe or not, more than one interaction with the client must take place. 4,521 records with 16 regular qualities and 1 label are present.

Tabel 3.1 Detail Dataset Bank Telemarketing

	Nama Atribut	Keterangan	Tipe
1	Age	Usia klien	Numeric
2	Job	Jenis pekerjaan klien	Polynominal

3	Marital	Status pernikahan	Polynominal
4	Education	Pendidikan terakhir klien	Polynominal
5	Default	Apakah klien telah memiliki kredit	Polynominal
6	Balance	Saldo rata-rata tahunan (dalam euro)	Numeric
7	Housing	Apakah memiliki kredit perumahan?	Polynominal
8	Loan	Apakah memiliki pinjaman pribadi?	Polynominal
9	Contact	Jenis komunikasi saat dihubungi	Polynominal
10	Day	Terakhir dihubungi pada setiap bulan	Numeric
11	Month	Bulan terakhir dihubungi pada 1 tahun	Polynominal
12	Duration	Durasi terakhir saat dihubungi (dalam detik)	Numeric
13	Campaign	Jumlah kontak yang dilakukan selama kampanye dan untuk klien	Numeric
14	Pdays	Banyaknya hari setelah klien terakhir dihubungi.	Numeric
15	Previous	Banyaknya kontak yang dilakukan sebelum promosi.	Numeric
16	Poutcome	Hasil dari promosi sebelumnya	Polynominal
17	Output	Apakah klien telah berlangganann deposito berjangkan	Polynominal

This study will employ some methodologies, each of which will be related to others. The graphic displays the suggested method's flow chart in figure 3.1.

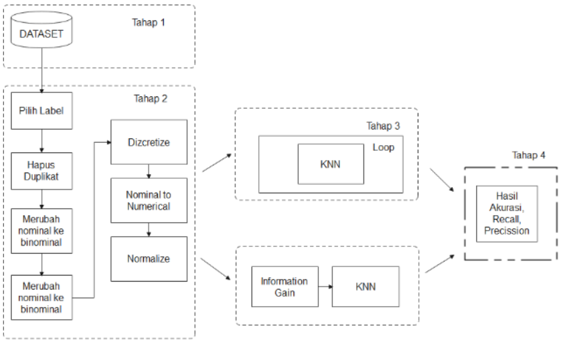


Figure 3.1 Proposed Method

In Figure 3.1 about the proposed method, can be explained among others:

1. stage 1 prepared using a Portuguese Bank Marketing dataset that had been downloaded from the UCI repository. The dataset consists of 4,521 records with a total of 16 regular attributes and 1 label that separates customers who have subscribed to time deposits from those who have not.
2. Before modeling is done, it is necessary to select labels, check empty data, duplicate data, and change data from polynomial to numeric by using replace on object-type attributes such as job, marital status, education, default, housing, loan, contact, month, output, and y. Stage 2 involves preprocessing the data, but the data still contains anomalies. Normalize the data next.
3. Stage 3 runs the initial test using the constrained KNN algorithm in the manner described below:
   a. Determine the value of K.
   b. Calculating the shortest distance between the testing data and the training data.
   c. Identifies the K sequence's shortest path by sorting the data..
   d. Insert the proper class value inside.

e. Identify the class with the greatest number of classes from its nearest neighbor, and designate it as the evaluated class.

The stages that follow in the second experiment using Information Gain:

a. Determine the root entropy.
b. Determine the formula for computing the root gain value.
c. Select a characteristic based on the value with the greatest potential advantages.
d. Using the same approaches as in the previous experiment, model the relevant characteristics based on the greatest gain value.

4. Stage 4 analyzes the confusion matrices to compare the accuracy, recall, and precision performance assessment outcomes of the KNN model with the Information Gain in the KNN method.

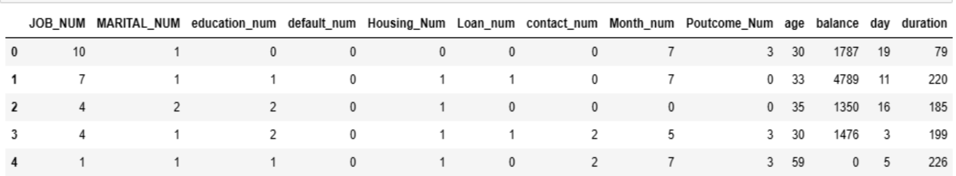
4. RESULTS AND DISCUSSION

1) Data Collecting Data

In this study, the possibility that retail bank clients will join the program will be calculated utilizing information gained through KNN, which may be utilized for direct marketing categorization. The dataset utilized has 4,521 entries and is the Portuguese retail bank set of data from the UCI Repository. It has 16 regular characteristics and 1 label attribute.

2) Pre Processing Data

Data is preprocessed when the dataset is acquired by doing numerous data checks, converting the object data type to float64, and normalizing the data. Python version 3.9 is used for preprocessing. After data preparation, 4,521 records with 16 regular characteristics and 1 label attribute were the end result.

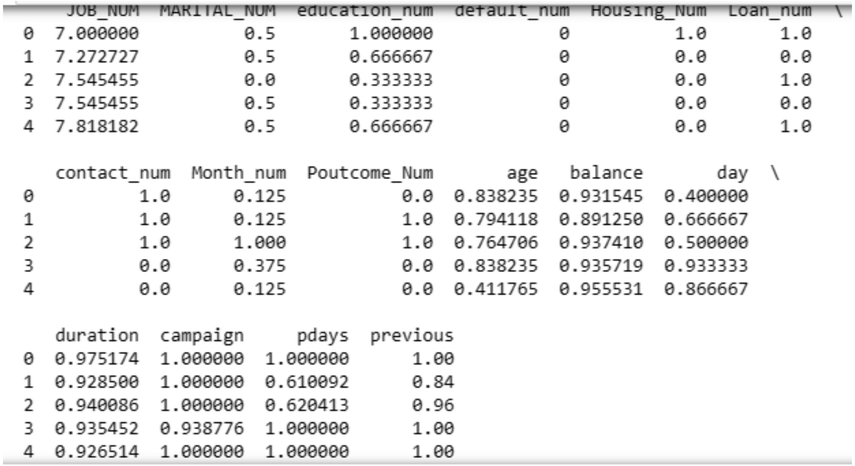


	JOB_NUM	MARITAL_NUM	education_num	default_num	Housing_Num	Loan_num	contact_num	Month_num	Poutcome_Num	age	balance	day	duration
0	10	1	0	0	0	0	0	7	3	30	1787	19	79
1	7	1	1	0	1	1	0	7	0	33	4789	11	220
2	4	2	2	0	1	0	0	0	0	35	1350	16	185
3	4	1	2	0	1	1	2	5	3	30	1476	3	199
4	1	1	1	0	1	0	2	7	3	59	0	5	226

Figure 4.1 The results of changing the data type to become numeric in the dataset

The next step after changing the data type is to use the algorithm for normalizing the data for each attribute.

Several characteristics with redundant data are subjected to the data normalization procedure. Additionally, the dataset is only loosely modeled using the KNN method, assessed, and the next stage is to add information gain to the KNN algorithm modeling. Figure 4.2 displays the normalizing findings.



```
     JOB_NUM  MARITAL_NUM  education_num  default_num  Housing_Num  Loan_num  \
0  7.000000          0.5       1.000000            0          1.0       1.0
1  7.272727          0.5       0.666667            0          0.0       0.0
2  7.545455          0.0       0.333333            0          0.0       1.0
3  7.545455          0.5       0.333333            0          0.0       0.0
4  7.818182          0.5       0.666667            0          0.0       1.0

   contact_num  Month_num  Poutcome_Num       age   balance       day  \
0          1.0      0.125           0.0  0.838235  0.931545  0.400000
1          1.0      0.125           1.0  0.794118  0.891250  0.666667
2          1.0      1.000           1.0  0.764706  0.937410  0.500000
3          0.0      0.375           0.0  0.838235  0.935719  0.933333
4          0.0      0.125           0.0  0.411765  0.955531  0.866667

   duration  campaign     pdays  previous
0  0.975174  1.000000  1.000000      1.00
1  0.928500  1.000000  0.610092      0.84
2  0.940086  1.000000  0.620413      0.96
3  0.935452  0.938776  1.000000      1.00
4  0.926514  1.000000  1.000000      1.00
```

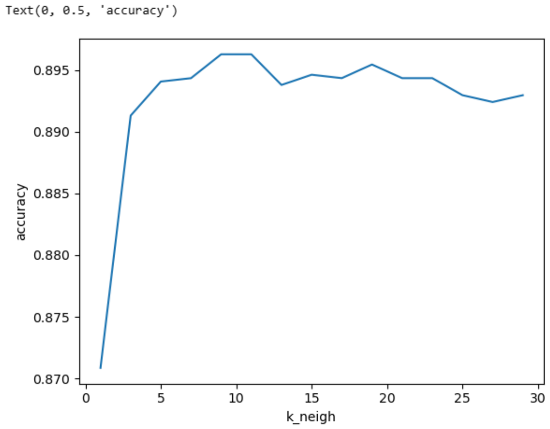
Figure 4.2 Hasil Normalisasi Data

3) Using the KNN Algorithm for Modeling

The first experiment employs modeling with the KNN method to exclude the range with even values and identify the best K value using 5-fold cross-validation with the range K = 1 to 31. To model the data and get the optimal K value, the following steps are taken using Python 3.9:

a. It contains 3,616 data for training and 905 data to be evaluated in the normalized dataset, which has been separated into 80% of the training data and 20% of the data for testing. With a range of K = 1 to 31, the training data is generated using Euclidean for the shortest distance in the KNN method by removing even values to prevent the same results. Each input unit first performs (Xi, i = 1..n) on the training data and (Yi, i = 1..n) on the testing data. The total of the reductions from (Xi, i = 1..n) and (Yi, i = 1..n) is then squared, and lastly the square root. The shortest distance to the sequence K is then determined after sorting the distance. The largest number of classes from the closest neighbors are then sought for in the distance calculation results from the distance calculation that has been sorted, and it is decided that class is the class being evaluated. According to the findings of the search equation, the class that is being evaluated is the one with the greatest number of classes (determined by a majority vote) from the closest neighbor.

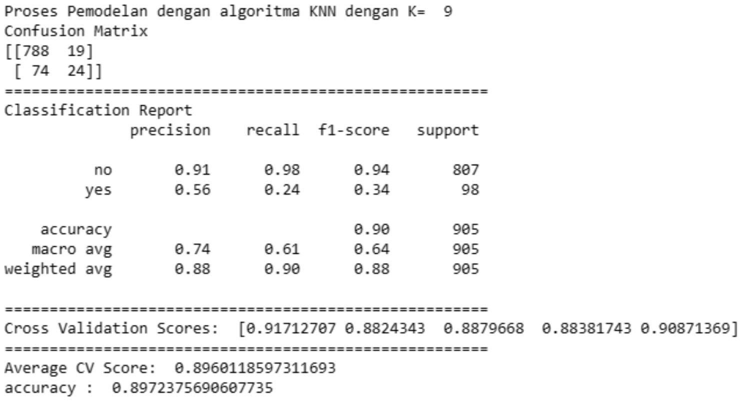
b. The KNN algorithm's highest K value, with accuracy values of 89.72%, 98% recall, and 91% precision, had been obtained for the best K search computation. Figure 4.3 shows the curve of the optimal K value, and Figure 4.4 displays the accuracy, recall, and precision results.



```
Text(0, 0.5, 'accuracy')
```

Figure 4.3 Graph of Accuracy Value Successful using KNN Algorithm

c. Implement 5-fold cross-validation as evaluating



```
Proses Pemodelan dengan algoritma KNN dengan K=  9
Confusion Matrix
[[788  19]
 [ 74  24]]
====================================================
Classification Report
              precision    recall  f1-score   support

          no       0.91      0.98      0.94       807
         yes       0.56      0.24      0.34        98

    accuracy                           0.90       905
   macro avg       0.74      0.61      0.64       905
weighted avg       0.88      0.90      0.88       905

====================================================
Cross Validation Scores:  [0.91712707 0.8824343  0.8879668  0.88381743 0.90871369]
====================================================
Average CV Score:  0.8960118597311693
accuracy :  0.8972375690607735
```

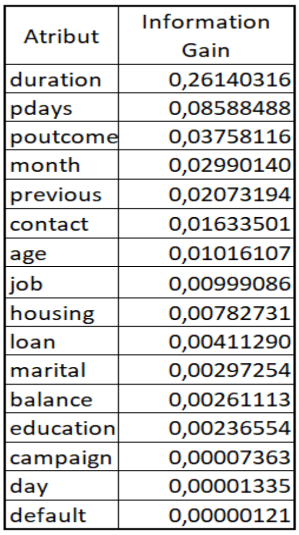
Figure 4.4 Outcomes for the KNN Algorithm Experiment for Accuracy Value

4) The Modeling Using KNN Algorithms with Information Gain
Modeling by adding information gain to the KNN algorithm with a value of K = 9 and using 5-fold cross-validation. In this modeling, the data is divided into 80% training data and 20% testing data. Steps for modeling the KNN algorithm and adding information gain, with the following details:

a. The normalized dataset is divided into 80% training data and 20% testing data, so the training data is 3,616 data and the testing data is 905 data.

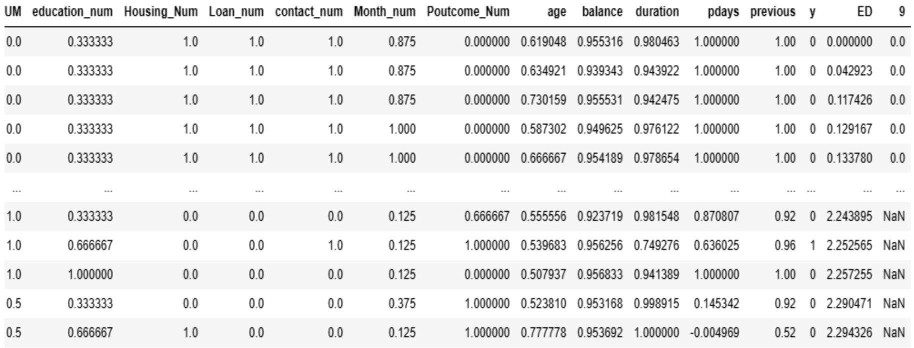
b. To uncover significant characteristics, add information gain to the training data. use a dataset with 16 properties. The qualities received with each value in the computation of information gain. The traits that have a high information gain value are the ones that are picked to have an impact on direct marketing. As shown in table 4.1, information gain outcomes may be compiled and classified from large to small. The campaign, day, and default attributes are three qualities of little value, and as a result, they have no impact on direct marketing. The remaining 13 qualities, such as length, pdays, poutcome, month, prior, contact, age, job, housing, loan, marriage, balance, and education, have substantial information gain values over 0.001.

Table 4.1 Results of Information Gathering and Order



Atribut	Information Gain
duration	0,26140316
pdays	0,08588488
poutcome	0,03758116
month	0,02990140
previous	0,02073194
contact	0,01633501
age	0,01016107
job	0,00999086
housing	0,00782731
loan	0,00411290
marital	0,00297254
balance	0,00261113
education	0,00236554
campaign	0,00007363
day	0,00001335
default	0,00000121

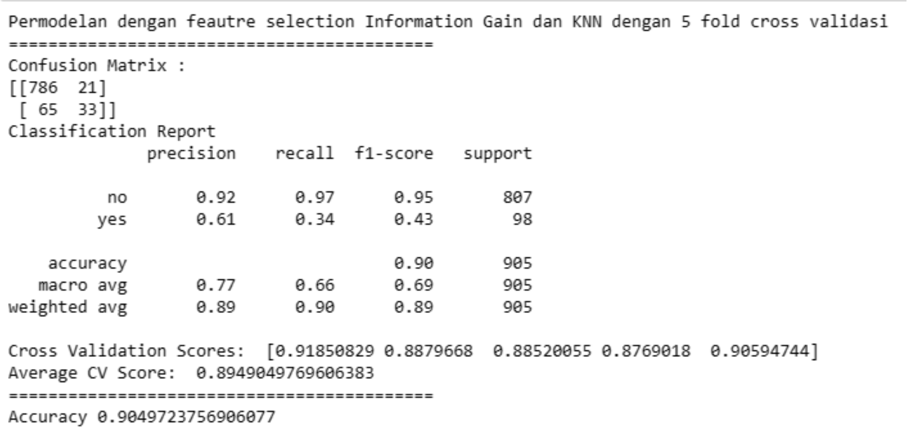
makes use of 13 characteristics that affect information gain and the KNN method with K = 9 to model the data. The shortest distance is then determined using the training data using Euclidean. Each input unit first performs (Xi, i = 1..n) on the training data and (Yi, i = 1..n) on the testing data. The total of the reductions from (Xi, i = 1..n) and (Yi, i = 1..n) is then squared, and lastly the square root. The distance that is closest to the order of K is then determined after sorting the distance. After sorting it, a search for K = 9 classes from the closest neighbors is conducted. The search equation's outcomes for the class are shown in Figure 4.5.



UM	education_num	Housing_Num	Loan_num	contact_num	Month_num	Poutcome_Num	age	balance	duration	pdays	previous	y	ED	9
0.0	0.333333	1.0	1.0	1.0	0.875	0.000000	0.619048	0.955316	0.980463	1.000000	1.00	0	0.000000	0.0
0.0	0.333333	1.0	1.0	1.0	0.875	0.000000	0.634921	0.939343	0.943922	1.000000	1.00	0	0.042923	0.0
0.0	0.333333	1.0	1.0	1.0	0.875	0.000000	0.730159	0.955531	0.942475	1.000000	1.00	0	0.117426	0.0
0.0	0.333333	1.0	1.0	1.0	1.000	0.000000	0.587302	0.949625	0.976122	1.000000	1.00	0	0.129167	0.0
0.0	0.333333	1.0	1.0	1.0	1.000	0.000000	0.666667	0.954189	0.978654	1.000000	1.00	0	0.133780	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1.0	0.333333	0.0	0.0	0.0	0.125	0.666667	0.555556	0.923719	0.981548	0.870807	0.92	0	2.243895	NaN
1.0	0.666667	0.0	0.0	1.0	0.125	1.000000	0.539683	0.956256	0.749276	0.636025	0.96	1	2.252565	NaN
1.0	1.000000	0.0	0.0	0.0	0.125	0.000000	0.507937	0.956833	0.941389	1.000000	1.00	0	2.257255	NaN
0.5	0.333333	0.0	0.0	0.0	0.375	1.000000	0.523810	0.953168	0.998915	0.145342	0.92	0	2.290471	NaN
0.5	0.666667	1.0	0.0	0.0	0.125	1.000000	0.777778	0.953692	1.000000	-0.004969	0.52	0	2.294326	NaN

Figure 4.5 The results to figure out how many K = 9 nearest

The accuracy, recall, and precision were all 92% after adding the information gain feature to the KNN algorithm using 5-fold cross-validation. Figure 4.6 shows the performance outcomes of this simulation.



```
Permodelan dengan feautre selection Information Gain dan KNN dengan 5 fold cross validasi
==========================================
Confusion Matrix :
[[786  21]
 [ 65  33]]
Classification Report
              precision    recall  f1-score   support

          no       0.92      0.97      0.95       807
         yes       0.61      0.34      0.43        98

    accuracy                           0.90       905
   macro avg       0.77      0.66      0.69       905
weighted avg       0.89      0.90      0.89       905

Cross Validation Scores:  [0.91850829 0.8879668  0.88520055 0.8769018  0.90594744]
Average CV Score:  0.8949049769606383
==========================================
Accuracy 0.9049723756906077
```

Figure 4.6 Information Gain Experiments on the KNN Algorithm Results

5) Research Results

The data were preprocess in the first experiment. The detail of the nominal to numeric data transition. The data is normalized when the conversion is complete as shown in Figure 4.3. The KNN method was employed in a restricted manner in an experiment that used 16 regular characteristics and 1 label attribute. The experiment employs the odd number k = 1 - 31 as the K value. The graph in Figure 4.3 displays the results of the experiments that gave the highest accuracy value at K = 9.

The greatest accuracy in this experiment was obtained at k = 9 with a precision (No) of 91% and a recall (No) of 98% using 5-fold-cross-validation. Figure 4.4 displays the outcomes of these tests.

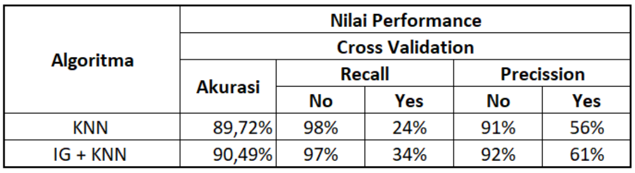
The researcher then modified the KNN algorithm experiment by include information gain in the second trial with the same K value. Calculating the root entropy value and the information gain weighting value are both necessary steps in determining the initial information gain weighting value. The information gain weighting findings were then sorted in ascending order, and 13 qualities were found to have a high information gain weight value and so have an impact on direct marketing. The following factors can have a significant impact: length, pdays, output, month, prior, contact, age, employment, housing, loan, marital status, balance, and education. While three attributes, such as day, campaign, and default, which have a negligible weight value, have no impact. Table 4.1 displays the findings for ascending and weighting.

CONCLUSION

In order to get the highest accuracy outcomes in banking direct marketing, this research contrasts the KNN algorithm with knowledge gain to optimize KNN. The 4,521 entries in the dataset have 16 regular characteristics and 1 label. Some of the dataset's data are still of the nominal and categorical data types, therefore pre-processing of the data must be done before modeling is used. After preprocessing, the dataset is normalized, followed by a restricted application of the KNN method to model the data, and finally, an assessment. The best K value in this modeling was K = 9, with accuracy of 89.72%, recall of 98%, and precision of 91%.

Modeling by Information Gain: Given a dataset with 16 attributes, information gain features are chosen into 13 attributes that are relevant to direct marketing, such as duration, pdays, poutcome, month, previous, contact, age, job, housing, loan, marital status, and balance. The accuracy value of the feature selection was 90.49%, the recall rate was 97%, and the precision rate was 92%. The feature selection results were modeled using the KNN method with a value of K = 9. Table 5.1 can be used to compare performance results. According to the research, the accuracy percentage outcomes of the KNN algorithm may be increased in comparison to the KNN algorithm modeling.

Tablel 0.1 Performance from Classification Model



Algoritma	Nilai Performance				
	Cross Validation				
	Akurasi	Recall		Precission	
		No	Yes	No	Yes
KNN	89,72%	98%	24%	91%	56%
IG + KNN	90,49%	97%	34%	92%	61%

Following the completion of this research, it is possible to carry out a number of recommendations for additional research in its development, including the proposed strategy that can be applied to banking direct marketing in order to provide convenience in product marketing, and additional research by comparing using several classification algorithms with other feature selection.

REFERENCES

Aggarwal, Charu. C. 2015. 14 Cancer Letters *Data Mining The Textbook*. https://linkinghub.elsevier.com/retrieve/pii/030438358190152X.

ALPAYDIN, ETHEM. 2014. *INTRODUCTION TO MACHINE LEARNING*. third edit.

Gorunescu, Florin. 1369. *Data Mining - Intelligent System Reference Library Vol 12*.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. Data Mining: Concepts and Techniques *Data Mining: Concepts and Techniques*.

Han, Zongtao et al. 2017. “A FAST ITERATIVE FEATURES SELECTION FOR THE K-NEAREST NEIGHBOR.” : 5810–13.

Ismail Fawaz, Hassan et al. 2019. “Deep Learning for Time Series Classification: A Review.” *Data Mining and Knowledge Discovery* 33(4): 917–63.

Kim, Kee-hoon, Chang-seok Lee, and Sang-muk Jo. 2015. “Predicting the Success of Bank Telemarketing Using Deep Convolutional Neural Network.” : 314–17.

Koumétio, Cédric Stéphane Tékouabo., and Hamza . Toulni. 2021. 971 *Improving KNN Model for Direct Marketing Prediction in Smart Cities*. https://link.springer.com/10.1007/978-3-030-72065-0.

Koumetio, Cedric Stephane Tekouabou, Walid Cherif, and Silkan Hassan. 2019. “Optimizing the Prediction of Telemarketing Target Calls by a Classification Technique.” *Proceedings - 2018 International Conference on Wireless Networks and Mobile Communications, WINCOM 2018*.

Moro, Sérgio, Paulo Cortez, and Paulo Rita. 2014. “A Data-Driven Approach to Predict the Success of Bank Telemarketing.” *Decision Support Systems*. http://dx.doi.org/10.1016/j.dss.2014.03.001.

P. Sundari, and K. Thangadurai. 2010. “An Empirical Study on Data Mining Applications.” *Global Journal of Computer Science and Technology* 10(5): 23–27.

Punjabi, Mamta, and Gend Lal Prajapati. 2018. “Lazy Learner and PCA: An Evolutionary Approach.” *Proceedings of Computing Conference 2017* 2018-Janua(July): 312–16.

Puspadini, Ratih. 2020. “Feature Selection on K-Nearest Neighbor Algorithm Using Similarity Measure.” : 226–31.

Saputra, E P. 2017. “PREDIKSI KEBERHASILAN TELEMARKETING BANK UNTUK.” 2(2): 66–72.

Syadzali, Chashif, Suryono Suryono, and Jatmiko Endro Suseno. 2020. “Business Intelligence Using the K-Nearest Neighbor Algorithm to Analyze Customer Behavior in Online Crowdfunding Systems.” *E3S Web of Conferences* 202: 1–7.

Tekouabou, Stéphane Cédric Koumetio, Walid Cherif, and Hassan Silkan. 2019. “A Data Modeling Approach for Classification Problems: Application to Bank Telemarketing Prediction.” *ACM International Conference Proceeding Series* Part F1481: 1–7.

Wang, Aiguo et al. 2014. “Incremental Wrapper Based Gene Selection with Markov Blanket.” *Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014*: 74–79.

Zeinulla, Elzhan, Karina Bebayeva, and Adnan Yazici. “Comparative Study of The Clasification Models for Prediction of Bank Telemarketing.” : 1–5.

Zhang, Hong et al. 2015. “The Application of Data Mining in Finance Industry Based on Big Data Background.” *Proceedings - 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security and 2015 IEEE 12th International Conference on Embedded Software and Systems, H*: 1536–39.