Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Classification of Customer Decisions in Opening Deposits Using the C4.5 Algorithm Implemented in Python

Chaerul Hidayat[1]*, Fabian Eka Prasetyo[2], Elkin Rilvani[3]

[1,2,3] *Informatics Engineering, Pelita Bangsa University*
chaerulhidayat70@gmail.com[1]*, fabian77yoroi@gmail.com[2], elkin.rilvani@pelitabangsa.ac.id[3]

Abstract

The banking industry needs to understand customer behavior to improve marketing strategies, particularly for deposit products. This study aims to develop a classification model to predict customer decisions in opening deposits using the C4.5 decision tree algorithm. The dataset consists of secondary banking data, including attributes such as age, occupation, marital status, education, account balance, and deposit decisions. The research adopts the Knowledge Discovery in Databases (KDD) process, from data cleaning to model implementation using Python and the chefboost library. Based on 11,162 customer records, the model achieved 56.69% accuracy, 55.63% precision, and 42.50% recall. Evaluation through a confusion matrix resulted in 2,248 True Positives (TP), 4,080 True Negatives (TN), 1,793 False Positives (FP), and 3,041 False Negatives (FN). The findings suggest that the C4.5 algorithm serves as a baseline approach for predicting customer behavior in deposit decisions. However, further improvements are needed to enhance its performance. This study contributes to the development of data-driven decision support systems in the banking sector.

***Keywords**: Data Mining; C4.5; classification; Python; deposits; banking.*

1. Introduction

In today's digital age and with the rapid development of information technology, the banking industry is required to be able to manage and utilize customer data in a more effective, efficient, and structured manner in order to enhance its competitiveness in an increasingly competitive market. Efforts to improve service quality through targeted strategies are key factors in maintaining and expanding market share. Deposits, as one of the savings products that offer both fund security and competitive returns over a specific period, remain the top choice for individuals seeking relatively low-risk investments. However, it is important to recognize that not all customers have the same interest or potential to open deposits, necessitating a deep segmentation and classification approach based on behavior, preferences, and demographic and financial characteristics of prospective customers. As a result, banks can accurately identify the most potential customer groups, enabling marketing strategies and services to be more focused, personalized, and maximally impactful in increasing deposit account openings and maximizing customer retention. The use of data mining technology and classification algorithms, such as C4.5, is a highly useful tool in this process, as it can uncover important patterns from large and complex customer data while optimizing data-driven business decision-making that is accurate and accountable [1].

Manual or conventional customer data processing is no longer adequate for handling the large and complex volumes of data commonly encountered by banking institutions today. These traditional methods often pose various challenges, including slow processes, high risk of input errors, difficulty in searching and validating data, and limitations in conducting in-depth analysis for strategic decision-making. Therefore, the use of data mining techniques with classification algorithms such as C4.5 is crucial as a modern solution for extracting hidden patterns in data, enabling faster, more accurate, and evidence-based decision-making. The C4.5 algorithm is known for its ability to build effective and accurate decision trees, and it can accommodate both numerical and categorical data in its process. Additionally, the resulting models are easy to understand and interpret by stakeholders, thereby supporting transparency and clarity in the business decision-making process. As a result, the implementation of the C4.5 algorithm provides significant added value in enhancing operational efficiency and service quality in the banking sector [2].

Research conducted in Indonesia over the past five years has consistently shown that the C4.5 algorithm is one of the main methods chosen for classifying customers with the potential to open deposits. For example, a study by Handayani and Nuryuliani (2023) demonstrated that the C4.5 algorithm can achieve a prediction accuracy rate of 91.95%, which represents a significant advantage compared to the Naive Bayes method, which had a lower accuracy rate in that study. Additionally, research published in the Intech Scientific Journal reinforces this finding by showing that the C4.5-based Decision Tree algorithm provides more accurate and reliable classification results compared to Naive Bayes and k-Nearest Neighbor in predicting the potential of customers to open term deposits. This reflects C4.5's ability to effectively manage numerical and categorical data and produce models that are easy to interpret and highly suitable for application in the context of complex banking customer data. Therefore, the C4.5 algorithm has not only been empirically proven to have high performance but also provides valuable insights for strategic decision-making in the field of time deposit product marketing [3].

By utilizing attributes such as occupation, status, education, income, credit history, and marketing contacts, the C4.5 classification model can quickly and accurately map the characteristics of potential customers. This allows banks to allocate marketing resources more efficiently and increase customer conversion effectiveness. This study aims to develop a classification model using the C4.5 algorithm that can be used to predict customer decisions in opening deposits, thereby contributing to improving marketing effectiveness and banking business growth in Indonesia.

2. Literature Review

2.1. Data Mining

Data mining is a multidisciplinary analytical process that involves a combination of various fields of study, such as statistics, machine learning, database systems, and artificial intelligence. The primary objective of this process is to uncover hidden patterns, trends, and significant correlations within large and complex datasets that cannot be directly identified using conventional methods. By applying this approach, organizations and institutions can gain deeper insights to support data-driven decision-making processes. This enables higher operational efficiency, improved accuracy in predictions, and the formulation of more adaptive strategies in response to market dynamics or user behavior [4].

2.2. Algorithm C4.5

The C4.5 algorithm is one of the decision tree-based classification methods developed by Ross Quinlan as an improvement on the ID3 algorithm. C4.5 uses gain ratio as the attribute selection criterion, which is more effective than information gain in addressing bias in attributes with many values. This algorithm can also handle both numerical and categorical attributes, manage missing values, and perform pruning to avoid overfitting. The model output consists of decision rules that are easy to interpret, making it highly useful in operational and business decision-making contexts [5].

2.3. Python

Python is a high-level programming language first developed by Guido van Rossum and officially released in 1991. With the advancement of information technology, Python has become one of the most popular and widely used programming languages across various fields, particularly in scientific computing and data science. Python is known for its versatility and flexibility, as it supports various programming paradigms such as procedural, object-oriented, and functional programming. Its popularity continues to grow due to its simple yet powerful syntax, as well as its extensive ecosystem of libraries. In the context of Machine Learning and Deep Learning, Python is the top choice thanks to the availability of various leading libraries such as TensorFlow, Keras, PyTorch, and scikit-learn, which significantly accelerate the development of predictive and data-driven analytical models [6].

2.4. Machine Learning

Machine learning is a branch of computer science that focuses on developing algorithms that enable computers to "learn" automatically from historical data without explicit programming for each case. This approach is often referred to as learning from data, where the system uses past data to form a learning model capable of recognizing patterns and generating optimal predictions for new data. The essence of machine learning lies in its ability to build models that represent hidden patterns in data sets [7].

2.5. Confusion Matrix

A confusion matrix is an evaluation table used to describe the performance of a classification model by showing the number of correct and incorrect predictions against test data. This matrix presents a comparison between the model's predictions and the actual values (ground truth), making it easier to assess classification accuracy and errors. For binary classification cases, the confusion matrix structure typically consists of four main components: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) [8], as shown in Table 1.

Table 1: Confusion Matrix

Correct Classification	Classification	
	Positive	Negative
Positive	TP	TN
Negative	FP	FN

3. Research Methodology

This study uses an inductive approach by utilizing customer data obtained from bank datasets. The data contains information such as age, occupation, marital status, education, savings balance, and deposit opening decisions. A total of 30 customer data were randomly selected for analysis using the C4.5 algorithm in the deposit opening decision classification process.

3.1. *Collecting Data*

This research was conducted using secondary data collection methods, sourced from public datasets related to banking. The data was obtained online and reflects customer characteristics such as age, occupation, marital status, education, and decisions regarding opening deposits. The research did not use direct field methods, but focused on structured data analysis using classification algorithms.

3.2. *Cleansing Data*

At this stage, data cleansing is performed, which involves removing attributes that are irrelevant to the classification objective, such as the day and month columns, and ensuring that there is no missing data in the dataset. The data is also reviewed to ensure consistency of values and readiness for the next preprocessing stage.

3.3. *Data Integration*

After the data cleansing stage is complete, the next step is to select and combine relevant attributes. This process aims to determine the important variables that will be used in the prediction modeling stage, so that only features that contribute to the classification results are retained in the analysis.

3.4. *Data Transformation*

At this stage, data transformation is carried out, which involves grouping numerical attribute values into simpler and more informative categorical forms. One of the attributes that is transformed is age. The numerical values in this attribute are classified into three categories: Young for ages under 30, Adult for ages 30 to 59, and Elderly for ages 60 and above. This transformation aims to improve data readability and facilitate the C4.5 algorithm in building decision trees based on more representative age categories.

3.5. Python Programming Implementation

This research process focuses on the application of classification algorithms using the Python programming language as the main tool. Before the model is built, several supporting libraries are imported first to support the data analysis and prediction process. The libraries used in this research include pandas for data manipulation, numpy for numerical computation, scikit-learn for preprocessing and model evaluation, and chefboost as a tool for building a C4.5-based decision tree classification model.

The entire research process, from raw data processing to the final evaluation of the classification model, is visualized in the form of a research workflow illustration in Figure 1 to provide a comprehensive overview of the process carried out.

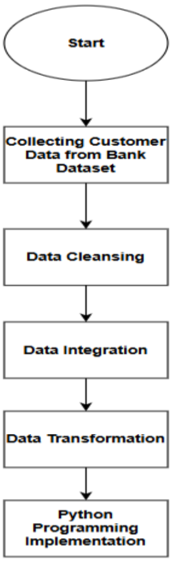

Start
Collecting Customer Data from Bank Dataset
Data Cleansing
Data Integration
Data Transformation
Python Programming Implementation

Fig. 1: Research Process

4. Result and Discussion

The dataset used in this study was derived from banking customer data consisting of 30 randomly selected data entries. The attributes used in the study included information such as customer age, occupation, marital status, education level, balance amount, and deposit opening decisions.

A summary of the initial data is presented in Table 2, which provides an overview of the structure and characteristics of the data analyzed.

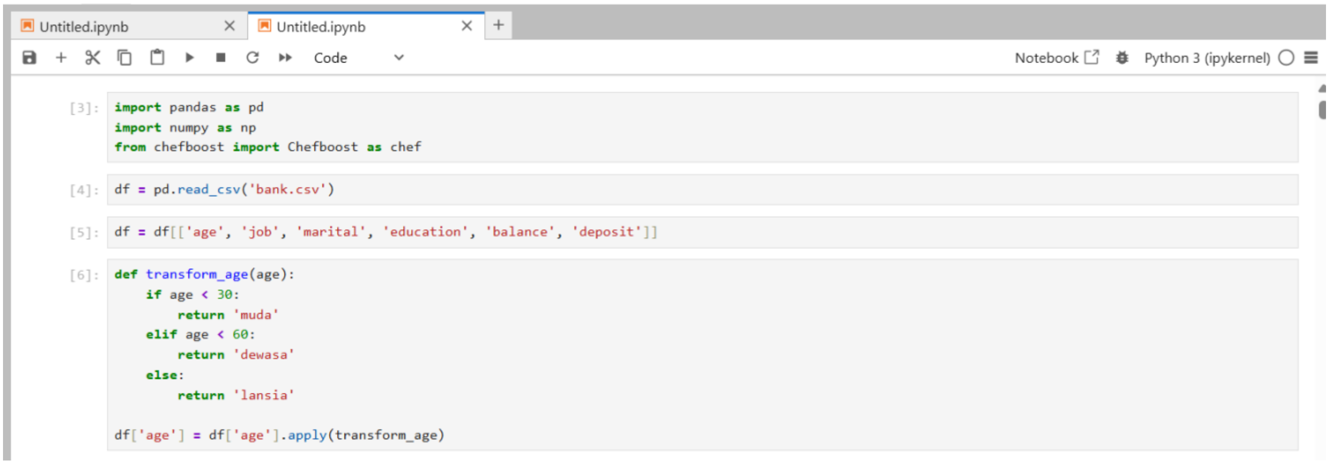
Table 2: Summary of customer dataset used for deposit classification

No	Age	Job	Marital	Education	Balance	Deposit
1	59	Admin	Married	Secondary	2243	Yes
2	56	Admin	Married	Secondary	45	Yes
3	41	Technician	Married	Secondary	1270	Yes

4	55	Services	Mariied	Secondary	2476	Yes
5	54	Admin	Married	Tertiary	184	Yes
6	42	Management	Single	Tertiary	0	Yes
7	56	Management	Maried	Tertiary	830	Yes
8	60	Retired	Divorced	Secondary	546	Yes
9	37	Technician	Married	Secondary	1	Yes
10	28	Services	Single	Secondary	5090	Yes
…	…	…	.. .	…	…	..
1162	34	Technician	Married	Secondary	0	No

The collected dataset was then applied in the form of an application using the Python programming language. The initial data stored in CSV (Comma-Separated Values) format was directly entered into Jupyter Notebook, which serves as an interactive work environment for analyzing data. The implementation was carried out using Python version 3.10.6, supported by various important libraries such as pandas for data processing, numpy for numerical calculations, sklearn for model preparation and evaluation, and chefboost as the library used to build a classification model based on the C4.5 decision tree algorithm.

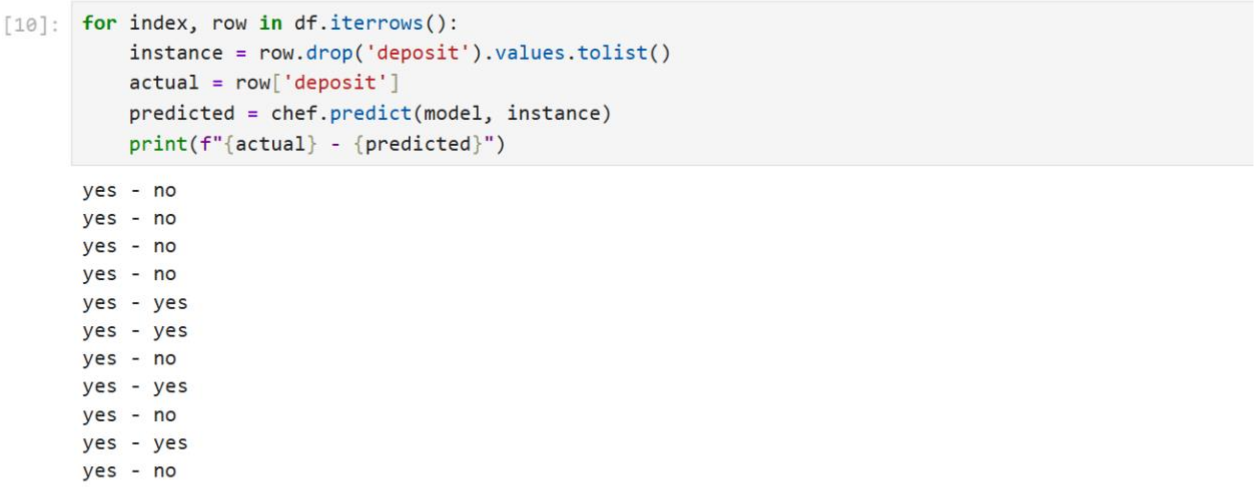
The entire programming process, including the data preparation phase, model training, and evaluation, is shown in Figure 2, which illustrates the overall code implementation.



```
[3]: import pandas as pd
     import numpy as np
     from chefboost import Chefboost as chef

[4]: df = pd.read_csv('bank.csv')

[5]: df = df[['age', 'job', 'marital', 'education', 'balance', 'deposit']]

[6]: def transform_age(age):
         if age < 30:
             return 'muda'
         elif age < 60:
             return 'dewasa'
         else:
             return 'lansia'

     df['age'] = df['age'].apply(transform_age)
```

Fig. 2: Python implementation

Once the dataset has been imported into the Jupyter Framework, the next step is to input the equations from the C4.5 Algorithm and execute the program. The outcome of the prediction statement using Python may be seen in Figure 3.



```
[10]: for index, row in df.iterrows():
          instance = row.drop('deposit').values.tolist()
          actual = row['deposit']
          predicted = chef.predict(model, instance)
          print(f"{actual} - {predicted}")

      yes - no
      yes - no
      yes - no
      yes - no
      yes - yes
      yes - yes
      yes - no
      yes - yes
      yes - no
      yes - yes
      yes - no
```

Fig. 3: Python prediction results

A comparison between the actual results from the dataset and the predictions generated by Python using the C4.5 algorithm can be seen in Table 3. The dataset used in this study includes 11,162 customer data with various attributes, including age, occupation, marital status, education level, balance, and decisions regarding opening a deposit account.

Table 3: Comparison of actual deposit data and C4.5 prediction results using Python

No	Result Description Dataset	Result Description Python Prediction

1	Yes	No
2	Yes	No
3	Yes	No
4	Yes	No
5	Yes	Yes
6	Yes	Yes
7	Yes	Yes
8	Yes	Yes
9	Yes	No
10	Yes	Yes

Confusion Matrix
A confusion matrix is used to perform the relative accuracy of what was learned from description of dataset results with description of results Utilize Python for predictions. A confusing confusion matrix as seen in Table 4

Table 4: Confusion matrix of deposit classification using C4.5 algorithm

Correct Classification	Classification	
	Positive	Negative
Positive	2248	3041
Negative	1793	4080

Explanation:
1. Positive Classification with Positive = 2248 because the number of positive data was correctly classified by the system.
2. Negative Classification with Positive = 1793 because the number of negative data was classified as positive by the system.
3. Positive Classification with Negative = 3041 because the number of positive data was classified as negative by the system.
4. Negative Classification with Negative = 4080 because the number of negative data correctly classified by the system.
5. The test results obtained are as follows:
Accuracy = (TP + TN) / (TP + TN + FP + FN)
= (2,248 + 4,080) / 11,162 × 100% = 56.69%
Precision = TP / (TP + FP)
= 2,248 / (2,248 + 1,793) × 100% = 55.63%
Recall = TP / (TP + FN)
= 2,248 / (2,248 + 3,041) × 100% = 42.50%

5. Conclusion

This study successfully created a classification model to predict customer decisions regarding opening deposits using the C4.5 decision tree algorithm implemented through the Python programming language. The data used consisted of 11,162 customer records, which included information such as age, occupation, marital status, education level, and balance. The entire research process followed the KDD (Knowledge Discovery in Databases) method, which included data cleaning, integration of relevant attributes, conversion of age values into categories of young, adult, and elderly, and application of the classification model using the chefboost library in Jupyter Notebook. The evaluation results for the model showed that the C4.5 algorithm was able to make fairly accurate predictions, although it was not yet optimal. Based on the confusion matrix results, the True Positive (TP) value was 2,248, True Negative (TN) was 4,080, False Positive (FP) was 1,793, and False Negative (FN) was 3,041. From these values, the accuracy is 56.69%, precision is 55.63%, and recall is 42.50%. This data indicates that the model is able to identify most customers according to the actual decision, but there is still a lot of misclassified data, especially in the positive (Yes) category. A comparison between the original data in the dataset and the prediction results obtained from Python shows discrepancies in some data, possibly due to class imbalance, lack of important attributes, or customer behavior patterns that cannot be fully captured by the existing attributes. Nevertheless, the use of Python has proven effective in supporting this classification process, from data pre-processing to evaluating prediction results. This demonstrates that the use of data mining technologies such as the C4.5 algorithm can make a significant contribution to facilitating data-driven decision-making, particularly in the banking industry.

6. Acknowledgment

The author would like to thank Mr. Elkin Rilvani, S.Kom., M.Kom., as the lecturer of Data Mining, who has provided valuable knowledge, guidance, and motivation so that the author can write this article well. Thanks also go to all those who have provided support during the research and writing process of this journal.

References

[1] D. H. Nuryuliani, “Analisis Prediksi Nasabah yang Berpotensi Membuka Deposito pada Bank Umum di Bekasi Menggunakan Algoritma C4.5 dan Naive Bayes,” *J. Ilm. Komputasi*, vol. 19, no. 3, pp. 317–326, Sep. 2020, doi: 10.32409/jikstik.19.3.66.

[2] M. R. Takakobi and K. D. Hartomo, “Analisis Metode Klasifikasi Nasabah Potensial dalam Membuka Deposito Jangka Panjang Melalui Telemarketing Menggunakan Metode Gradient Boosting Classifier”.

[3] N. Nia, “Klasifikasi Data Mining untuk Prediksi Potensi Nasabah dalam Membuat Deposito Berjangka,” *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 3, no. 1, pp. 65–75, May 2021, doi: 10.46772/intech.

[4] W.-T. Wu *et al.*, “Data mining in clinical big data: the frequently used databases, steps, and methodological models,” *Mil. Med. Res.*, vol. 8, no. 1, Aug. 2021, doi: 10.1186/s40779-021-00338-z.

[5] Y. F. S. Sugiarto, “Algoritma C4.5 sebagai Penerapan Decision Tree-Based Classification Model untuk Mengklasifikasikan Tingkat Omzet UMKM Berdasarkan Profil Bisnis,” *J. Ekon. Manaj. Akunt. Dan Perbank. Syari’ah*, vol. 13, no. 2, pp. 188–197, Sep. 2024.

[6] Alfarizi *et al.*, “Penggunaan Python sebagai Bahasa Pemrograman untuk Machine Learning dan Deep Learning,” *J. Karimah Tauhid*, vol. 2, no. 1, pp. 1–6, 2023.

[7] A. D. Sidik and A. Ansawarman, “Prediksi Jumlah Kendaraan Bermotor Menggunakan Machine Learning,” *Formosa J. Multidiscip. Res.*, vol. 1, no. 3, pp. 559–568, Jul. 2022, doi: 10.55927/fjmr.v1i3.745.

[8] D. Normawati and S. A. Prayogi, “Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter,” *J. Sains Komput. Inform. J-SAKTI*, vol. 5, no. 2, pp. 697–711, 2021.