International Conference on Informatics Engineering, Science & Technology

Data Mining Approach to Predict Bank Saving Decisions Employing Classification Method

Faisal Hanif Arifin*, Sri Supatmi

Master of Information Systems, Universitas Komputer Indonesia, Bandung, Indonesia

*Corresponding E-mail: faisal.75124014@mahasiswa.unikom.ac.id

Abstract. This study aims to predict which individuals would opt for saving their money in a bank by looking at matters concerning their age, their current occupation, gender, and how much income they provided each month. The authors employ the Naïve Bayes classification method, which is a frequently observed yet occasionally straightforward enough approach to mine data. The very initial stages were a questionnaire via the web that got 109 responses. After a rigorous data clean-up procedure that included dropping features and grouping continuous variables, the dataset was split up. About 70% of it was used to train the model, and its remaining 30% was used to test to determine how well it performed in circumstances it hadn't experienced previously. The outcomes pointed out that the Naïve Bayes classifier exceeded the threshold of 81% accuracy, with precision and recall around 85%. These were attained because the model indicated a strong ability to identify the majority of positive cases (high recall), while also showing a tendency to predict positives slightly more often than necessary, leading to some false positives, as indicated by the confusion matrix. This suggests that the model prioritizes minimizing missed detections over avoiding errors. These findings contribute to banks that use this type of data mining classification method may be able to improve their marketing strategies, prioritize specific leads, make decisions more quickly while still obtaining results that are thought to be fairly accurate, or optimize services to make hesitant customers feel more likely to open accounts. But this relies substantially on how properly the model indicates the way individuals conduct themselves in their actual lives.

Keywords: data mining, classification, naïve bayes, saving, bank, decision making, prediction,financial.

1. Introduction

Data mining appears to be potentially beneficial among industries and has been elevated with the intention of supporting companies to determine better decisions. It's transformed how businesses look at how the market changes and how to make daily operations better in customer segmentation (Dharmayanti et al., 2024; Mydyti et al., 2023; Ray et al., 2021; Sajid & Amin, 2023). This method is becoming more and more used in banking to guess who will save and why. By recognising which elements affect ensuring financial decisions, companies can spot patterns before they happen. Customers are more inclined to trust banks and put their money there when they see smart ads and understand what the banks are offering (Mulfi, 2022). The term "data mining" is broad. Statistics, machine learning, and predictive models are all employed collectively to deliver companies an improved grasp about the way consumers respond so they could develop mutually beneficial relationships rather than just making one-time sales (Kaderye et al., 2024).

Several financial, social, and demographic factors have influenced the decision to save in a bank, as consistently demonstrated by studies. Age, education, and profession contribute to demographic aspects. Younger, highly educated, and salaried individuals frequently claim that they could save more (Jumena et al., 2022; Kamil et al., 2023). Data mining is a single approach to predict savings decisions in banks because the customer information is highly valuable, complex to obtain, and dynamic. Conventional analysis methods may struggle to identify and capture important trends in large datasets comprising demographic, socioeconomic, behavioural, and transactional data (Cuomo et al., 2023; Vashi et al., 2024). Consumers who are likely to respond to or save funds on a specific banking product can be identified by effectively exploring hidden connections and patterns using data mining approaches, such as classification (Akkaya & Turgay, 2024; Alexandra & Sinaga, 2021; Vashi et al., 2024).

Researchers looking towards data-driven banking decisions, similar to figuring out who would establish a savings account or default on a loan, have usually employed classification methods like the Naïve Bayes algorithm in the past few years. It seems that one of the aspects that makes it desirable is how simple it is: it's conducted by inspecting the observed frequencies of various data amalgamations and figuring out the probability of certain outcomes. Even if it occasionally fails to deliver quite remarkable outcomes, its easily understood setup and unexpectedly trustworthy baseline accuracy allow it to be an attractive option, especially in banking circumstances (Setiani et al., 2021; Zhang et al., 2021, 2022; Zhang & Jiang, 2022). That said, the algorithm seems to punch above its weight: even though it rests on some fairly strict assumptions and keeps things relatively simple under the hood, it tends to train faster than many other classifiers, a pattern that's been observed repeatedly in practice. (Bamidele Awotunde et al., 2023).

The objectives of this study are to recognize patterns in individual savings behaviour to ensure that banks might employ classification data mining, primarily the Naïve Bayes algorithm, to pinpoint the traits of consumers who are more willing to save based on demographic parameters. Banks may swiftly and precisely improve their data-driven marketing campaigns by using this method.

2. Literature Review

2.1 Data Mining

Semantic data mining is another recent advancement that uses ontologies and domain knowledge to conceptually interpret data, yielding more in-depth and meaningful results. Applications such as fraud detection, customer segmentation, and predictive analytics are just a few of the many industries that utilize data mining, which continues to evolve with new tools and algorithms to meet new opportunities and challenges (Papakyriakou & Barbounakis, 2022; Sirichanya & Kraisak, 2021). Many people consider data mining as a synonym for another popular term, knowledge discovery from data, while others view data mining as just an important step in the overall knowledge discovery process (Han et al., 2012).

Leveraging data mining techniques on bank savings data offers substantial benefits, including improved risk assessment, enhanced fraud detection, and more effective decision-making by uncovering hidden patterns and trends within vast data sets. These techniques can automate the detection of questionable transactions, reduce manual audit costs, and help banks more effectively understand customer behaviour, ultimately resulting in better personalized services and stronger financial controls. Furthermore, it may be challenging for algorithms to comprehend, ensure data quality, and handle the fact that financial data can be complex and constantly changing. To overcome these previous issues, banks must focus on enhancing their data handling capabilities, utilizing various data mining methods, and implementing innovative solutions that make models more transparent and accessible to consumers (Sushkov et al., 2023; Yang, 2025).

Lots of essential demographic factors, notably higher income, higher educational attainment, and more earning family members, are regularly found to be linked to higher saving rates and more proactive financial planning in studies involving data mining (Hiremath & Afza, 2022; Yue et al., 2013). Saving behaviours and intentions might also be affected by gender and professional identity, though these impacts may be mitigated by contextual and cultural factors (Ahmadi & Farhoodi, 2023; Stancu et al., 2024; Yue et al., 2013). The overall knowledge discovery process is shown in Figure 1.

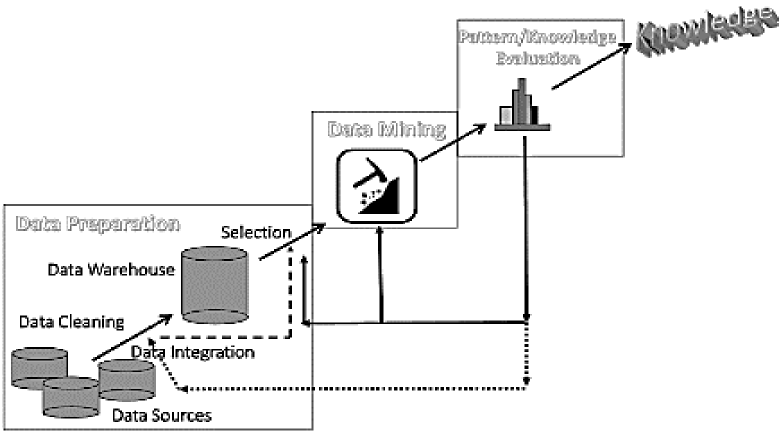

Figure 1. Data Mining: An essential step in the process of knowledge discovery (Han et al., 2012)

2.2 Classification Method

Classification is a standard data mining method that organizes data into separate categories based on predefined guidelines. Categories lead to the potential for discovering important patterns in large datasets. A few frequently observed algorithms for classification include decision trees (such as C4.5), naïve bayes, genetic algorithms, random forests, and advanced ensemble methods like XGBoost. However, all of these superior performance models regularly require higher-level calculations and are more challenging to understand than simpler models, which remain quite effective at generating predictions and are less complex to follow. When deciding on a classification method, one needs to consider its accuracy, the computing capacity it requires, and whether simplicity is necessary for the model, depending on the application and domain in question (Baghel et al., 2025; Sharma et al., 2024).

2.3 Naïve Bayes Algorithm

The Naïve Bayes classifier demonstrates effectiveness across various scenarios, despite its simplicity and foundational assumptions. The Naïve Bayes technique primarily aims to identify potential categories within a text document. Employing the aggregated probabilities of the terms and classes. The concept of sovereignty is foundational to its philosophy (Bamidele Awotunde et al., 2023). Naïve Bayes is a highly effective and efficient algorithm for inductive learning in the fields of machine learning and data mining. Although Naïve Bayes assumes attribute independence (i.e., no association between attributes), its performance in the classification process is relatively competitive. The assumption of attribute independence in real-world data is infrequent; however, despite violations of this assumption, Naïve Bayes demonstrates considerable classification performance. Various empirical studies have demonstrated this (Romli et al., 2021).

Naïve Bayes relies on the assumption that, given the output values, the attribute values are conditionally independent. The collective observed probability is the product of the individual probabilities, given the output values. The equation for Bayes' theorem is as follows in number (1) (Setiani et al., 2021):

$$P(H|X) = \frac{P(X|H)\ x\ P(H)}{P(X)} \quad (1)$$

Where:

- X: Data with an unknown class
- H: The data hypothesis is a specific class.
- P(H|X): The probability of hypothesis H under condition X (posterior probability)
- P(H): Probability hypothesis H (prior probability)
- P(X|H): Probability X based on the conditions in the hypothesis H
- P(X): Probability X

3. Method

The research process was carried out through several key, systematic stages. The initial stage involved data collection, which formed the foundation for the subsequent analysis. Following this, data pre-processing was conducted to ensure the quality and readiness of the data before its application in modelling. Once the data was prepared, the modelling process commenced, employing the Naïve Bayes algorithm as the primary method for classifying the data. After the model was developed, a testing phase was initiated to evaluate its performance. The authors utilized metrics suitable for predicting bank savings decisions to analyze the test results and assess the model's accuracy and effectiveness. Figure 2 illustrates the research progression method.

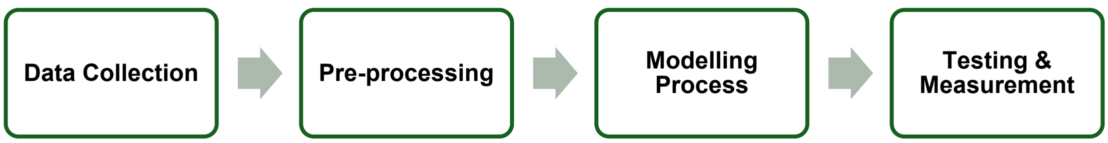

Figure 2. Research Method Progression

3.1 Data Collection

This research uses a dataset taken from an online questionnaire created by the authors in July 2019 using Google Forms and distributed to the public, particularly within the Indonesian Computer University environment. The minimum number of respondents expected for this research is 60-80 to achieve a high replication rate of 80% (Manyara et al., 2024). Online questionnaire recipients will be asked to fill out several questions containing data prepared by the authors, based on financial, social, and demographic aspects, including name, age, gender, employment status, average monthly income, and savings decisions (Jumena et al., 2022). Name was chosen to see the differences between respondents who completed the survey, although it will not be used in the modelling; age was chosen to see the age range to be measured; gender was chosen to see the gender comparison of respondents; status was chosen to see what job each respondent has; income was chosen to see the range of income earned each month; and savings decisions were chosen as the point of this dataset because it is related to previous data that will be classified.

3.2 Pre-processing

Pre-processing is considered a crucial phase in data mining, as it transforms raw data into understandable information for further analysis. Pre-processing aims to reduce the size of the data and transform it into a form suitable for further mining. Numerous data pre-processing techniques, including data transformation, data cleaning, data integration, data optimization, and data conversion, are employed to transform raw data into high-quality data (Joshi & Patel, 2021; Shetty et al., 2024). In performing this pre-processing, the authors can use cleaning alone, in combination with other techniques, or a single technique. However, at this stage, the techniques used are feature selection and discretization. Data cleaning will not be performed because the online questionnaire distributed requires all data to be completed, resulting in no incomplete rows.

3.3 Modelling Process

At this stage, 70% of the total dataset will be taken for training data and then read into pre-processed form (Fitri et al., 2020; Manyara et al., 2024). Next, the prior probability calculation for each class and attribute is performed using the Naïve Bayes algorithm. Once each probability is obtained, the data mining process can be performed to predict bank savings decisions.

3.4 Testing and Measurement

At this stage, testing is conducted on the training data, using a 70:30 ratio from the total dataset for testing purposes. Training data is used to train the model, and test data is used to evaluate the model's ability to categorize savings decisions accurately according to their attributes. Furthermore, the results of the classification algorithm test are measured using a confusion matrix table as shown in Table 1, which can be easier to understand (Fitri et al., 2020).

Table 1. Confusion Matrix

	Predicted Class	
Actual Class	TP	FP
	FN	TN

Note:
- TP: Correct Positive: Infarction is predicted, and there is an infarction
- FP: False Positive: Infarction is predicted, and there is no infarction
- FN: False Negative: There is no predictable infarction, and infarction exists
- TN: True Negative: There is no predictable infarction, and no infarction

Based on the confusion matrix table in Table 1, the performance of using the Naïve Bayes classification method can be measured by calculating the accuracy, precision, and recall values. Accuracy measures how well the model classifies a person's decision to save at a bank as a whole, precision measures how well the model classifies a person's decision to save at the correct bank as a person's decision to save at the correct bank, and recall measures how well the model finds all decisions of a person to save at the correct bank (Wahyuningsih et al., 2024).

4. Results and Discussion

4.1 Data Collection

The distributed online questionnaire yielded 109 original datasets from respondents, exceeding the minimum sample size of 60-80 for the research. Most responders were college students, young, and between the ages of 19 and 21. It makes sense that the majority of their monthly income was less than 1.5 million rupiah. This suggests that students with comparatively little purchasing power make up the majority of the response group. The original dataset can be accessed at this link. Table 2 displays a subset of the original dataset consisting of 7 columns, although not all of it is included.

Table 2. Original Dataset

Timestamp	Name	Age	Gender	Status	Average Monthly Income	Saving Decisions
7/26/2019 10:41:14	Reza	21	Male	College Student	Rp. 1,500,000 - Rp. 3,000,000 (Between 1.5 Million and 3 Million Rupiah)	Yes
7/26/2019 12:38:47	Alma Grace Sianturi	21	Female	College Student	< Rp 1,500,000 (Less than 1.5 Million Rupiah)	No
7/26/2019 14:10:53	Andi	22	Male	College Student	Rp. 1,500,000 - Rp. 3,000,000 (Between 1.5 Million and 3 Million Rupiah)	Yes
7/26/2019 14:40:12	Esti Utami	20	Female	College Student	< Rp 1,500,000 (Less than 1.5 Million Rupiah)	Yes
			....			
7/27/2019 7:49:42	sintYes	21	Female	College Student	< Rp 1,500,000 (Less than 1.5 Million Rupiah)	Yes

4.2 Pre-processing

The initial cleaning method employed involved basic feature selection by performing feature correlation analysis and statistical significance testing using chi-square, as this test assesses whether there is a statistically significant relationship between two categorical attributes. In addition to the name attribute, which is not used because it is not a feature but rather a meta attribute, the gender and timestamp attributes are also not used because they do not show a significant relationship with the target, where adding them would add noise. The results of the chi-square test are shown in Table 3.

Table 3. Results of The Chi-Square Test

Attribute	Type	Chi-Square Value (X^2)
Timestamp	Datetime	0.058
Age	Numeric	8.946
Gender	Categorical	0.103
Status	Categorical	11.935
Average Monthly Income	Categorical	5.595

Subsequently, for the discretization method, the authors will initially determine the number of categories to use and then calculate the mapping of continuous attribute values to categorical equivalents, as illustrated in Tables 4 and 5 below.

Table 4. Discretization of Age Attribute

Age Attribute	Category
< 17 (Under 17 Years)	Teenager
17 – 25 (Between 17 and 25 years old)	Young Adult
> 25 (Over 25 Years)	Adult

Table 5. Discretization and Labelling of Average Monthly Income Attribute

Average Monthly Income Attribute	Category
< Rp 1,500,000 (Less than 1.5 Million Rupiah)	Low
Rp. 1,500,000 - Rp. 3,000,000 (Between 1.5 Million and 3 Million Rupiah)	Medium
> Rp. 3.000.000 (More than 3 Million Rupiah)	High

That author obtained data on responses to the decision questionnaire to save at the bank, which has undergone cleaning and discretization techniques in pre-processing, as follows in Table 6:

Table 6. Pre-processed Data Result

Age	Status	Average Monthly Income	Saving Decisions
Young Adult	College Student	Low	Yes
Young Adult	College Student	Low	Yes
Young Adult	College Student	Medium	Yes
Young Adult	College Student	Low	Yes
	………		
Young Adult	College Student	Low	Yes

4.3 Modelling Process

The next step is to read the training data from the pre-processed data; in this case, the authors discuss a person's decision to save at a bank, which has been pre-processed from the original data. The training data used consists of 77 items and can be seen in Table 7:

Table 7. Training Data for Modelling Process

No.	Age	Status	Average Monthly Income	Saving Decisions
1	Young Adult	College Student	Low	Yes
2	Young Adult	College Student	Low	Yes
3	Young Adult	College Student	Low	Yes
4	Young Adult	School Student	Low	No
		………		
77	Young Adult	College Student	Low	No

The next step is to calculate the prior probability for each class. Starting from the first class, C1, which has a saving decisions value of "Yes", P(C1) = 50/77 = 0.649, because there are 50 data samples with a saving decisions value of "Yes". Meanwhile, the prior probability for the first class, C2, which is the saving decisions value of "No," is P(C2) = 27/77 = 0.350 because there are 27 data samples for the saving decisions value of "No". Thus, it will be found that:

- P(C1) = P (Saving Decisions = “Yes”) = 50/77 = 0.649
- P(C2) = P (Saving Decisions = “No”) = 27/77 = 0.350

Alternatively, based on data from 77 respondents, the authors observe that 50 people save in a bank, or with a probability of 0.649, compared to 27 people who do not save at a bank, or with a probability of 0.350. However, this estimate has not been adjusted for data from other attributes/features that the authors want to analyse, namely age, gender, status, and average monthly income. Then, to determine the probability of each required attribute, calculate the conditional probability for each class P(X1, X2, Xn | Ci) for i = 1.2, and for each attribute in the input data sample shown in Table 8.

Table 8. Probability of Each Attribute

Attribute	Value	C1	C2	P(X|C1)	P(X|C2)
Age	Young Adult	45	23	0.9	0.851
Age	Adult	2	1	0.04	0.037
Age	Teenager	3	3	0.06	0.111
Status	College Student	39	15	0.78	0.555
Status	School Student	6	7	0.12	0.259
Status	Employee	5	5	0.1	0.185
Average Monthly Income	Low	34	23	0.68	0.851
Average Monthly Income	High	12	2	0.24	0.074
Average Monthly Income	Medium	4	2	0.08	0.074

From Table 8, the probability of each attribute is obtained based on the decision to save in a bank. Information has been presented indicating that young adults are more likely to save in a bank compared to teenagers or older individuals. Furthermore, individuals with a low average monthly income are also more likely to save in a bank. However, in this research, the focus will be on data whose class is unknown, specifically whether individuals decide to save in a bank or not, based on the attributes of age, status, and income.

After that, perform a probability calculation based on the X data whose class is being analyzed to evaluate the experiment's results. For example, consider the data point X (the class is unknown), where X = (Age = Young Adult; Status = School Student; Average Monthly Income = Low).

- P(X | Saving Decisions = “Yes”) = (0.9 x 0.12 x 0.68) x 0.649 = 0.047
- P(X | Saving Decisions = “No”) = (0.851 x 0.259 x 0.851) x 0.350 = 0.065

Considering that P(X | Saving Decisions = “No”) = 0.065 is greater than P(X | Saving Decisions = “Yes”) = 0.047, the Naïve Bayesian classifier classifies a person with age value “Young Adult”, status value “Student”, and average monthly income value “Low”; then it will enter the saving decisions value of the “No” class, or it can be concluded that the new data above falls into the “No” category or does not save in a bank.

4.4 Testing and Measurement

The Naïve Bayes classification method was tested using a confusion matrix, where 30% of the dataset of individuals' decisions to save money in a bank was used as the test data. In this case, there is an actual class, which is the "Saving Decisions" class generated from the original data, and a predicted class, which is the "Saving Decisions" class generated from the Naïve Bayes classification process. The classification process using the Naïve Bayes method on the 32 test data records produced the following results in Table 9.

Table 9. Classification Results on Test Data

Age	Status	Average Monthly Income	*Actual Class*	P (X | Saving Decisions = "Yes")	P (X | Saving Decisions = "No")	*Predicted Class*
Young Adult	College Student	Low	Yes	0.669279	0.330721	Yes
Young Adult	College Student	Low	No	0.606588	0.393412	Yes
Young Adult	College Student	Low	Yes	0.69814	0.30186	Yes
Young Adult	College Student	Low	Yes	0.69814	0.30186	Yes
			....			
Young Adult	College Student	Low	Yes	0.669279	0.330721	Yes

From the results of the classification process presented in Table 9, it can be converted into a confusion matrix to show the number of calculations of the suitability of the Actual Class or class that matches the original data with the Predicted Class or class resulting from the Naïve Bayes classification process. The confusion matrix table is shown in Table 10.

Table 10. Confusion Matrix Results

	Predicted Class: Saving Decisions = "Yes"	***Predicted Class:*** Saving Decisions = "No"	∑
Actual Class: Saving Decisions = "Yes"	TP = 18	FP = 3	**21**
Actual Class: Saving Decisions = "No"	FN = 3	TN = 8	**11**
∑	**21**	**11**	**32**

Based on Table 10, the performance of the Naïve Bayes classification method can be measured by calculating the accuracy, precision, and recall values.

- Accuracy: (TP+TN) / (TP+TN+FP+FN): (18+8) / 32 = 81.25%
- Precision: TP / (TP+FP): 18 / 21 = 85.71%
- Recall: TP / (TP+FN): 18 / 21 = 85.71%

The Naïve Bayes algorithm in this research demonstrated solid performance with 81.25% accuracy, 85.71% precision, and 85.71% recall, particularly in identifying positive cases, as equivalent study has shown that Naïve Bayes models might achieve adequate or better performance metrics, such 85% accuracy and high precision and recall, which makes them appropriate for real-world use in a variety of fields, like as marketing (Polinar, 2020). The performance demonstrates the model's strength in identifying potential depositors, which is crucial for minimizing false negatives in banking campaigns (Maheswari et al., 2021). These results are similar to previous studies that reported that Naïve Bayes can perform an adequate task of predicting who will indicate their consent for bank deposits (around 90.8%) or spotting fraud (about 80.4%) (Gupta et al., 2021; Safarkhani & Moro, 2021). Despite this, it is pertinent to emphasise that alternative algorithms, like decision trees, are known for performing exceedingly well in similar circumstances, as previous studies have revealed that models like J48 could perform surpassing Naïve Bayes, with an accuracy rate exceeding as high as 94.39% (Safarkhani & Moro, 2021).

5. Conclusion

The present study employs a Naïve Bayes classifier to probe towards the possibility that could be applied to predict the individuals who may decide to save their money in a bank by considering matters concerning their age, gender, occupation, and monthly income. On paper, the model performed reasonably well, with an accuracy of about 81.25%, a precision of 85.71%, and a recall of 85.71%. It also indicates that they have been quite adept at figuring out potential consumers even before they get there. Because of its lightweight, easy to implement, and scales without much fuss, this solution is especially appealing in dynamic contexts like focused marketing or gaining new customers. Banks and other financial companies could employ this opportunity to enhance their marketing tactics, prioritise certain leads, or optimise services to encourage uncertain consumers to feel more inclined to open accounts.

The research limitations include the narrow scope of data collection limited to the Universitas Komputer Indonesia environment, which may not generalize broadly, and the reliance on self-reported data, which could introduce response biases. Future research could employ more advanced machine learning techniques, such as ensemble methods or deep learning, for improved prediction accuracy and to incorporate diverse data sources, like transactional or social media data, to enrich the understanding of consumer savings behaviour.

References

Ahmadi, K. A., & Farhoodi, G. (2023). Demographic factors and savings in agricultural house. *International Journal of Agriculture Extension and Social Development*, *6*(2), 35–37. https://doi.org/10.33545/26180723.2023.v6.i2a.479

Akkaya, E., & Turgay, S. (2024). Unveiling the Power: A Comparative Analysis of Data Mining Tools through Decision Tree Classification on the Bank Marketing Dataset. *WSEAS TRANSACTIONS ON COMPUTERS*, *23*, 95–105. https://doi.org/10.37394/23205.2024.23.9

Alexandra, J., & Sinaga, K. P. (2021). Machine Learning Approaches for Marketing Campaign in Portuguese Banks. *2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS)*, 1–6. https://doi.org/10.1109/ICORIS52787.2021.9649623

Baghel, K., Bhati, R., & Budhiraja, A. (2025). Classification Techniques in Data Mining: A Comprehensive Study. *2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, 605–611. https://doi.org/10.1109/CICTN64563.2025.10932482

Bamidele Awotunde, J., Misra, S., Katta, V., & Charles Adebayo, O. (2023). An Ensemble-Based Hotel Reviews System Using Naive Bayes Classifier. *Computer Modeling in Engineering & Sciences*, *137*(1), 131–154. https://doi.org/10.32604/cmes.2023.026812

Cuomo, M. T., Tortora, D., Colosimo, I., Ricciardi Celsi, L., Genovino, C., Festa, G., & La Rocca, M. (2023). Segmenting with big data analytics and Python: A quantitative exploratory analysis of household savings. *Technological Forecasting and Social Change*, *191*, 122431. https://doi.org/10.1016/j.techfore.2023.122431

Dharmayanti, D., Akma, A. O., Soegoto, E. S., & Warlina, L. (2024). APPLICATION OF DATA MINING FOR PREDICTING HORTICULTURAL COMMODITIES PRICE. *Journal of Engineering Science and Technology*, *19*(1), 163–175.

Fitri, S., Selsi, R., Rustam, Z., & Pandelaki, J. (2020). Naïve bayes classifier models for cerebral infarction classification. *Journal of Physics: Conference Series*, *1490*(1), 012019. https://doi.org/10.1088/1742-6596/1490/1/012019

Gupta, A., Lohani, M. C., & Manchanda, M. (2021). Financial fraud detection using naive bayes algorithm in highly imbalance data set. *Journal of Discrete Mathematical Sciences and Cryptography*, *24*(5), 1559–1572. https://doi.org/10.1080/09720529.2021.1969733

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques (Third Edition)*. Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00017-4

Hiremath, K., & Afza, N. (2022). Demographic Factors and Saving Behaviour: Its Relevance in Financial Planning for Retirement. *The Journal of Management and Entrepreneurship*, 89–97. https://doi.org/10.70906/20221601089097

Joshi, A. P., & Patel, B. V. (2021). Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process. *Oriental Journal of Computer Science and Technology*, *13*(0203), 78–81. https://doi.org/10.13005/ojcst13.0203.03

Jumena, B. B., Siaila, S., & Widokarti, J. R. (2022). Saving Behaviour: Factors That Affect Saving Decisions (Systematic Literature Review Approach). *Jurnal Economic Resource*, *5*(2), 217–235. https://doi.org/10.57178/jer.v5i2.365

Kaderye, G., Arif, A., & Kundu, R. (2024). Data Mining in Different Fields: A Study. *International Journal of Innovative Science and Research Technology (IJISRT)*, *9*(3), 1828–1838. https://doi.org/10.38124/ijisrt/IJISRT24MAR1384

Kamil, M. A. A., Wiliasih, R., & Irfany, M. I. (2023). Determinants of Saving Decisions at Indonesian Islamic Banks During the COVID-19 Pandemic. *Annals of Management and Organization Research*, *4*(1), 1–19. https://doi.org/10.35912/amor.v4i1.1387

Maheswari, B., Aswini, J., & Anita, M. (2021). Hybrid Feature Selection Approach for Naive Bayes to Improve Consumer Behavior Analysis. *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 1200–1204. https://doi.org/10.1109/ICICV50876.2021.9388439

Manyara, A. M., Purvis, A., Ciani, O., Collins, G. S., & Taylor, R. S. (2024). Sample size in multistakeholder Delphi surveys: at what minimum sample size do replicability of results stabilize? *Journal of Clinical Epidemiology*, *174*, 111485. https://doi.org/10.1016/j.jclinepi.2024.111485

Mulfi, P. I. (2022). Analysis of Factors Affecting Customer Decision to Saving at Bank Sharia Indonesia, Palopo City. *Journal Of Institution And Sharia Finance*, *5*(1), 42–50. https://doi.org/10.24256/joins.v5i1.3359

Mydyti, H., Kadriu, A., & Pejic Bach, M. (2023). Using Data Mining to Improve Decision-Making: Case Study of A Recommendation System Development. *Organizacija*, *56*(2), 138–154. https://doi.org/10.2478/orga-2023-0010

Papakyriakou, D., & Barbounakis, I. S. (2022). Data Mining Methods: A Review. *International Journal of Computer Applications*, *183*(48), 5–19. https://doi.org/10.5120/ijca2022921884

Polinar, E. L. (2020). Students Performance in Board Examination Analysis using Naïve Bayes and C4.5 Algorithms. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(1), 753–758. https://doi.org/10.30534/ijatcse/2020/107912020

Ray, M., Ray, M., Muduli, K., Banaitis, A., & Kumar, A. (2021). INTEGRATED APPROACH OF FUZZY MULTI-ATTRIBUTE DECISION MAKING AND DATA MINING FOR CUSTOMER SEGMENTATION. *E+M Ekonomie a Management*, *24*(4), 174–188. https://doi.org/10.15240/tul/001/2021-4-011

Romli, I., Pardamean, T., Butsianto, S., Wiyatno, T. N., & Mohamad, E. bin. (2021). Naive Bayes Algorithm Implementation Based on Particle Swarm Optimization in Analyzing the Defect Product. *Journal of Physics: Conference Series*, *1845*(1), 012020. https://doi.org/10.1088/1742-6596/1845/1/012020

Safarkhani, F., & Moro, S. (2021). Improving the Accuracy of Predicting Bank Depositor's Behavior Using a Decision Tree. *Applied Sciences*, *11*(19), 9016. https://doi.org/10.3390/app11199016

Sajid, A., & Amin, B. (2023). Data Mining Techniques in Decision Making. *NUML International Journal of Engineering and Computing*, *2*(1), 65–73. https://doi.org/10.52015/nijec.v2i1.39

Setiani, I., Tentua, M. N., & Oyama, S. (2021). Prediction of Banking Stock Prices Using Naïve Bayes Method. *Journal of Physics: Conference Series*, *1823*(1), 012059. https://doi.org/10.1088/1742-6596/1823/1/012059

Sharma, V., Srivastava, D., & Avasthi, V. (2024). Methods for Categorizing Data Mining Outputs: A Review. *2024 International Conference on Healthcare Innovations, Software and Engineering Technologies (HISET)*, 132–134. https://doi.org/10.1109/HISET61796.2024.00049
Shetty, R., M., G., Dinesh Acharya, U., & G., S. (2024). Enhancing Ovarian Tumor Dataset Analysis Through Data Mining Preprocessing Techniques. *IEEE Access*, *12*, 122300–122312. https://doi.org/10.1109/ACCESS.2024.3450520
Sirichanya, C., & Kraisak, K. (2021). Semantic data mining in the information age: A systematic review. *International Journal of Intelligent Systems*, *36*(8), 3880–3916. https://doi.org/10.1002/int.22443
Stancu, S., Hristea, A. M., Kailani, C., Bala, D. E., & Pernici, A. (2024). Socio-demographic factors' influence on the energy-saving behaviour of residential consumers: Evidence from Romania. *E+M Ekonomie a Management*, *27*(3), 120–134. https://doi.org/10.15240/tul/001/2024-5-012
Sushkov, V. M., Leonov, P. Y., Nadezhina, O. S., & Blagova, I. Y. (2023). Integrating Data Mining Techniques for Fraud Detection in Financial Control Processes. *International Journal of Technology*, *14*(8), 1675. https://doi.org/10.14716/ijtech.v14i8.6830
Vashi, H., Yadav, J., & Varde, A. S. (2024). Predictive Analysis of Bank Marketing for Financial Decision Support and Smart Economy. In K. Arai (Ed.), *Intelligent Systems and Applications* (pp. 474–490). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47715-7_33
Wahyuningsih, T., Manongga, D., Sembiring, I., & Wijono, S. (2024). Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments. *Procedia Computer Science*, *234*, 349–356. https://doi.org/10.1016/j.procs.2024.03.014
Yang, Y. (2025). Application of Data Mining Algorithms in Bank Credit Risk Assessment: Review and Prospect. *Applied and Computational Engineering*, *150*(1), 41–46. https://doi.org/10.54254/2755-2721/2025.22400
Yue, T., Long, R., & Chen, H. (2013). Factors influencing energy-saving behavior of urban households in Jiangsu Province. *Energy Policy*, *62*, 665–675. https://doi.org/10.1016/j.enpol.2013.07.051
Zhang, H., & Jiang, L. (2022). Fine tuning attribute weighted naive Bayes. *Neurocomputing*, *488*, 402–411. https://doi.org/10.1016/j.neucom.2022.03.020
Zhang, H., Jiang, L., & Yu, L. (2021). Attribute and instance weighted naive Bayes. *Pattern Recognition*, *111*, 107674. https://doi.org/10.1016/j.patcog.2020.107674
Zhang, H., Jiang, L., Zhang, W., & Li, C. (2022). Multi-view Attribute Weighted Naive Bayes. *IEEE Transactions on Knowledge and Data Engineering*, *35*, 1–1. https://doi.org/10.1109/TKDE.2022.3177634