https://doi.org/10.35945/gb.2024.18.003

FORECASTING CUSTOMER BANK BEHAVIOR USING WEKA AND CLASSIFICATION ALGORITHMS

Mokhtar Tahraoui
Ph.D. in Quantitative Economics, University Centre of Maghnia, Tlemcen, Algeria
tahraoui.mokhtar.enssea@gmail.com
https://orcid.org/0009-0006-6998-8418

Rachid Mansour
Ph.D. in Management Sciences, University Centre of Maghnia, Tlemcen, Algeria
mansour.rachid.ucm@gmail.com
https://orcid.org/0000-0002-8597-5979

Riadh Kadri
Ph.D. in Management Sciences Professor, University Centre of Maghnia, Tlemcen, Algeria
kadri.riadh@yahoo.fr
https://orcid.org/0009-0004-1896-5027

ABSTRACT. Banks derive significant profits from loans, necessitating careful customer selection to mitigate default risks. This study analyzes customer behavior at Société Générale Algeria Bank using three classification algorithms and using Weka software. A dataset comprising 300 customers was evaluated with three algorithms: Bayesian Network, Naïve Bayes, and the J48 Decision Tree. The goal of this study was to identify the most effective algorithm for classifying prospective customers as acceptable or not for loan approval. The performance of these algorithms was compared, focusing on accuracy and reliability. The results of this study indicated that the J48 Decision Tree algorithm outperformed the other methods, demonstrating superior classification accuracy. This suggests its potential as a robust tool for optimizing decision-making processes in the bank's loan system. By integrating the J48 Decision Tree into its operations, the bank could enhance its ability to identify suitable customers, minimize risks, and ensure sustainable profitability in its lending practices.

KEYWORDS: DATA MINING, WEKA, BANK, LOANS, CLASSIFICATION ALGORITHMS

1. INTRODUCTION

Banks and businesses alike are in a volatile, pressured world. They need to be more responsive to customers, helping them manage finances and access services quickly and easily on their channel of choice while at the same time delivering operational efficiencies. To accomplish that, these modern banks are not only employing management information systems but are also starting to introduce robust data mining practices more and more into their core business processes. Data mining, the process of identifying hidden patterns from the data that are relevant to the business, is considerably practical in the business area and yields successful results. The data mining methods are benefiting from ever-developing technology, and people are also intensifying their efforts. Over 60% of banks are currently using data mining methods to help them gain a better understanding of their customers' behaviors.

One of the best-known open-source tools for data mining practice is Weka. Weka contains tools for data pre-processing, classification, regression, clustering, association analysis, and visualization and allows for output model and inclusion func-

tion diagrams. The dataset used in the study contains financial transactions associated with people. The study has considered the prediction of customer banking behavior, and for this, several algorithms of the Weka tool are applied. The necessity of being able to predict the future possible application of the customer is needed to prevent possible negative results. The aim is to determine whether or not the account of the customer will be open or closed in the future. The methods applied are Naive Bayes Classifier, Random Tree Classifier, J48 Classifier, SMO Classifier, and Simple Logistic Classifier. The best success result has been achieved with the J48 Classifier. With the data mining methods, marketing campaigns for the customer can be applied, and business strategies can be developed and tailored to the target customers' behaviors.

2. LITERATURE REVIEW

A Classification-based model to assess customer behavior in the banking sector. This study used three classifiers, K-NN, decision tree, and artificial neural networks, for predicting customer behavior in the banking sector, and this study concluded that Artificial neural networks (ANNs) outperform both decision trees (DTs) and k-nearest neighbors (k-NNs).

A Case Study of Predicting Banking Customers' Behaviour by Using Data Mining. In this study, the proposed data mining framework manages the relations between banking organizations and their customers, the results indicate that the Neural Network model achieves better accuracy but takes a longer time to train the model.

An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour. The goal of this study is to predict the behavior of customers to enhance the decision-making processes for retaining valued customers, using two classification models, Naïve Bayes, and Neural networks, in the end, the results show that the accuracy of Neural Network is comparatively better.

A Prediction Model for Bank Loans Using Agglomerative Hierarchical Clustering with Classification Approach. The goal of this study is to leverage machine learning techniques to enhance decision-making processes in the banking sector, it reached the following findings: the data with decision tree obtained an accuracy of 84%, with the random forest obtained an accuracy of 85%

3. METHODOLOGY

Data mining is a scientific and computational process used to extract knowledge represented by hidden relationships in large datasets. It relies on various computational and statistical tools and techniques to enable machines to learn and deduce new knowledge. Below, we introduce the algorithms used in our data mining study and the Weka software utilized for different stages of the study.

3.1. Bayesian Network Algorithm (Bayes Net Classifier)

This probabilistic model addresses the problem of interdependent relationships among variables. It is used for the formal representation of variables and probabilistic relationships in a network structure called a Bayesian network, which consists of nodes (representing variables) and links (indicating relationships between variables). This network structure assumes that parent variables influence others, relying on mathematical and statistical principles such as conditional probability.

- **Conditional Probability: P(A/B)**

This represents the occurrence of event A, given that event B has occurred. It is expressed by the following phrase:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (1)$$

- **Chain Rule of Probability**

This rule allows us to express the joint probability of multiple variables using conditional probabilities. For example, given three variables A, B, and C, we can express it with the following formula:

$$P(A, B, C) = P((A|B, C) \times P(B|C) \times P(C) \quad (2)$$

- **Bayes' Rule or Bayesian Theorem**

Bayes' rule allows us to calculate the conditional probability of a hypothesis H given evidence E, using the likelihood of the evidence given the hypothesis and the prior probability of the hypothesis. It is expressed by the following formula:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \quad (3)$$

Where:

- P(H/E) is the **posterior probability** of the hypothesis H given the evidence E.
- P(E/H) is the **likelihood** of the evidence E given the hypothesis H.
- P(H) is the **prior probability** of the hypothesis H.
- P(E) is the **probability of the evidence** E.
- The Structure of the Bayesian Network

The Bayesian network is represented by a directed acyclic graph (DAG), where the nodes represent variables, and the directed edges represent probabilistic dependencies between the variables. The structure of the Bayesian network encodes the conditional probability distribution over all the variables.

3.2. Naïve Bayes Classifier

The Bayesian Theory, Named after Thomas Bayes, this theory is vital in probability studies. It relates the probability of event A given event B to the probability of event B given event A.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \quad (4)$$

Naïve Bayes classification organizes information into categories by making predictions based on a simplistic assumption that all variables are independent, though this is often not realistic. Despite this assumption, Naïve Bayes works effectively in fields such as healthcare.

The basic idea of the Naive Bayes classifier is to determine the probability that a person with certain characteristics (X) belongs to a particular group (class). It assumes that these characteristics are independent of each other and calculates the probabilities accordingly. Then, it selects the group with the highest probability as the most likely group for that person.

Let Ci be the individual's class, and X the characteristics. This probability is expressed by the following statement:

$$P(C_i/X) = \frac{P(C_i)P(X=x_1,.....x_p/C_i)}{P(X=x_1,.....x_p)} \quad (5)$$

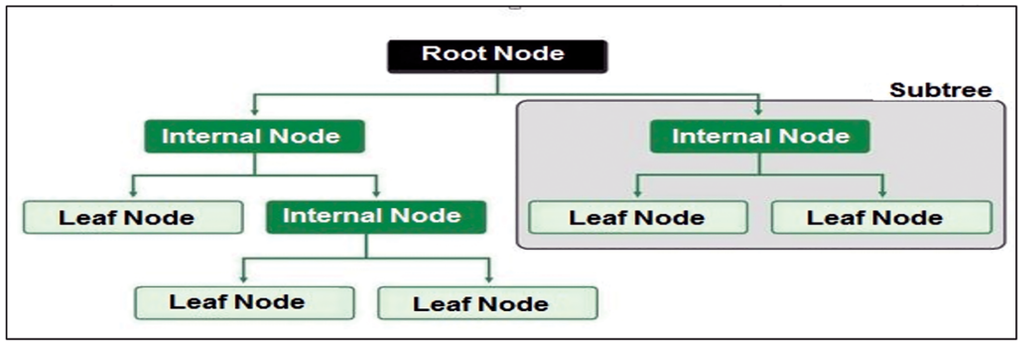
3.3. Decision Tree Algorithm (J48)

The J48 algorithm, an improved version of the C4.5 algorithm, classifies data by building a tree model that relies on sequential decisions. Each branch in the tree represents a decision, and the leaves represent final classifications. In data mining, decision trees help to describe, classify, and generalize specific datasets.

- **Tree Structure**

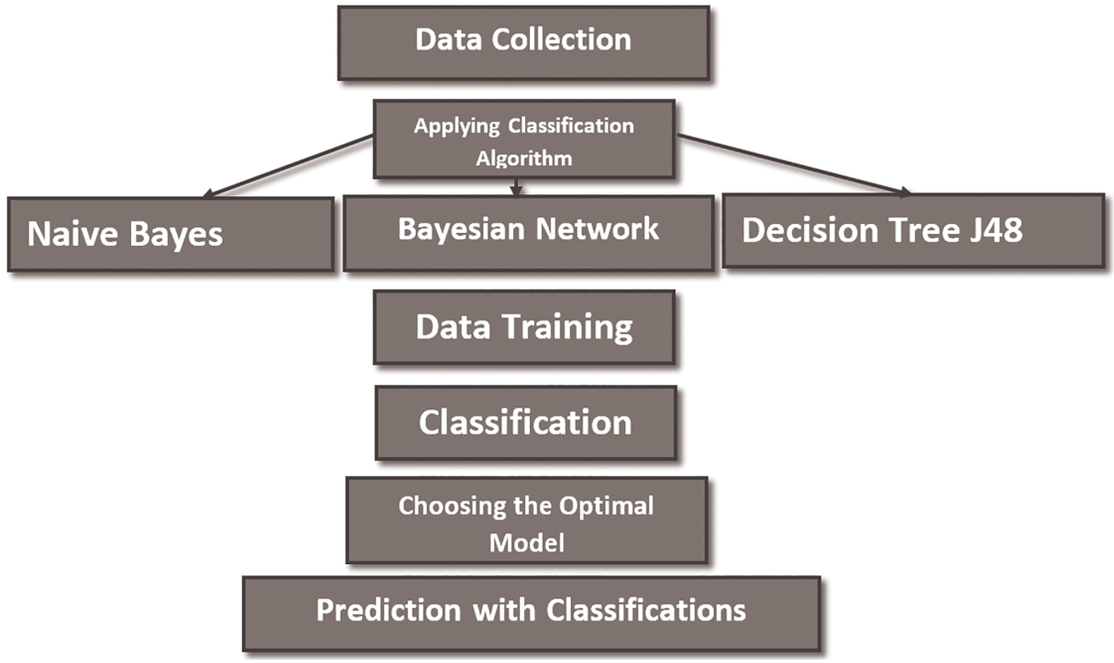
The internal non-terminal nodes are tests and are called decision nodes. The branches are the outcomes of these tests. The terminal leaf nodes represent decisions and are referred to as default classes. Each leaf represents a decision for a specific class based on all the tests performed from the root to that leaf (see Fig. 1).

FIG. 1: THE DECISION TREE DIAGRAM



Source: prepared by the researcher

FIG. 2: STUDY STAGES USING THE DATA MINING METHODOLOGY



Source: Prepared by the Researcher

3.4. Model Evaluation

- **Confusion Matrix**

Using the confusion matrix and performance indicators like accuracy, sensitivity, and error rate, we evaluated the model's performance, assessing correct and incorrect classifications.

- **Performance Metrics**

These indicators are calculated based on the confusion matrix (Adnan, Sarno, & Sungkono, 2019, 124) [15]:

We use the following symbols:

- VP: True Positive;
- FP: False Positive;
- VN: True Negative;
- FN: False Negative;
- P: Total number of positive observations;
- N: Total number of negative observations.

» Error Rate = (FP + FN) / (N + P), which represents the percentage of individuals or observations that the model classified incorrectly.

» Sensitivity = VP / (VP + FN), which measures the model's ability to detect true positives.

» Specificity = VN / (VN + FP), which measures the model's ability to correctly identify true negatives.

» Accuracy = (VP + VN) / (N + P), which represents the percentage of individuals that the model correctly predicted.

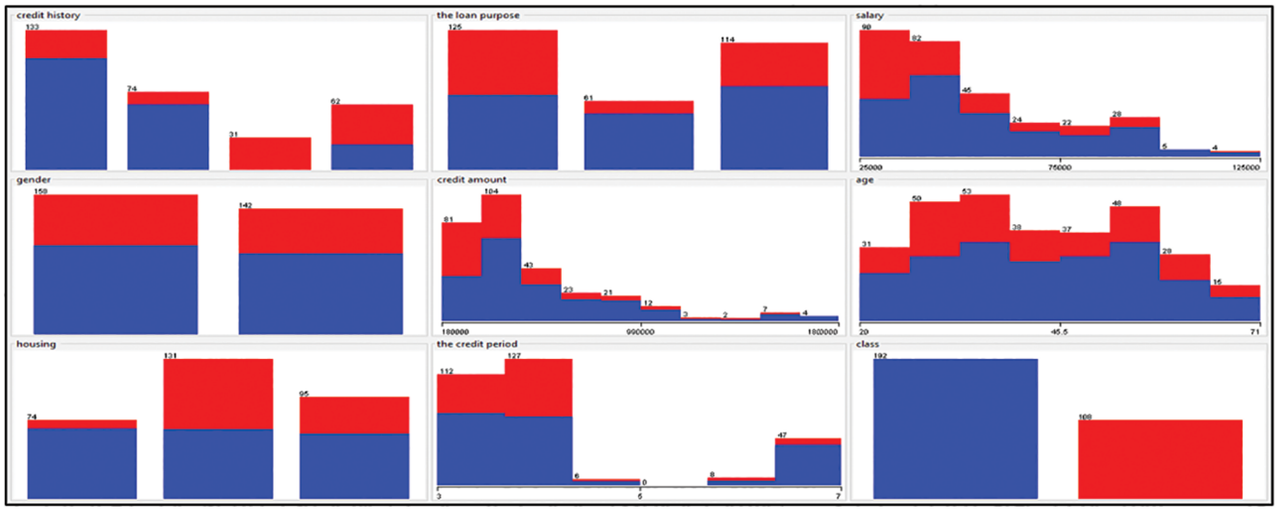
4. EMPIRICAL RESULT AND DISCUSSION

The data mining study using Weka followed these steps: data collection, training, application of classification algorithms, and selection of the optimal classification model to make predictions on new classifications (see Fig. 2).

- **Study Variables**

The study aimed to create a model aiding bank consultants in decision-making regarding loans. Variables considered include previous credit history, loan purpose, salary, gender, housing status, loan duration, and age. The sample comprised 300 customers.

FIG. 3: STATISTICS FOR ALL VARIABLES



Source: Outputs of the Weka Software

- **Descriptive Study of Study Variables**

The following section presents various statistics for each variable. The Weka software displays detailed information for each variable (see Fig.3).

For instance, regarding the customer's previous loans variable, 133 customers had no prior loans, 74 had paid off all their loans, 31 had delayed payments, and 62 still had outstanding balances. The loan purpose variable showed that 125 customers took loans for household appliances, 61 for home layouts, and 114 for purchasing a motorcycle. The income variable ranged between 25,000 DZD and 125,000 DZD. For the gender variable, 158 were male and 142 were female. The loan value variable ranged from 180,000 DZD to 1,800,000 DZD. The age variable spanned from 20 to 71 years. In terms of housing status, 74 customers owned their homes, 131 rented, and 95 lived in free accommodations. The loan term ranged between 3 and 7 years. The customer classification variable resulted in 192 classified as 'good' and 108 as 'bad'.

4.1. Study Results

The classification process involves grouping data based on shared variables in the form of categories. To achieve this, data mining offers numerous classification algorithms, including those used in this study:

» Naïve Bayes Algorithm;
» Bayesian Network Algorithm;

FIG. 4: RESULTS OF THE J48 DECISION TREE ALGORITHM

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         243               81      %
Incorrectly Classified Instances        57               19      %
Kappa statistic                          0.58
Mean absolute error                      0.2195
Root mean squared error                  0.3933
Relative absolute error                 47.5951 %
Root relative squared error             81.9316 %
Total Number of Instances              300
```

Source: Outputs of the Weka Software

FIG. 5: RESULTS OF THE BAYESIAN NETWORK ALGORITHM

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         241               80.3333 %
Incorrectly Classified Instances        59               19.6667 %
Kappa statistic                          0.558
Mean absolute error                      0.2743
Root mean squared error                  0.3651
Relative absolute error                 59.4814 %
Root relative squared error             76.0509 %
Total Number of Instances              300
```

Source: Outputs of the Weka Software

» J48 Decision Tree Algorithm.

These algorithms were applied to create a model to predict the behavior of new loan-seeking customers. The results were as follows:

- **Classification Results with J48 Decision Tree Algorithm**

After applying the J48 algorithm using Weka on the sample under study, the results were as shown in the table (see Fig. 4):

After applying the J48 algorithm using Weka on the sample data, the results are shown in the following table. We observe that the Kappa statistic has a value of 0.58, indicating a moderate agreement between predicted classifications and actual data categories. The model's classification accuracy was 81%, correctly classifying 243 customers, while 19% (57 customers) were misclassified.

- **Classification Results with Bayesian Network Algorithm**

After applying the Bayesian Network algorithm using Weka on the sample under study, the results were as shown in the table (see Fig. 5):

The Bayesian Network algorithm, when applied with Weka on the sample data, showed a Kappa statistic of 0.55, indicating a moderate level of agreement between predicted and actual classifications. The model's accuracy was 80.33%, correctly classifying 241 customers, while 19.66% (59 customers) were misclassified.

- **Classification Results with Naïve Bayes Algorithm**

After applying the Naive Bayes algorithm using the Weka program on the sample under study, the results are as shown in the table (see Fig. 6):

The Naïve Bayes algorithm produced a Kappa statistic of 0.36, indicating a lower agreement level. The model achieved a classification accuracy of 68.66%, correctly classifying 206 customers, while 31.33% (94 customers) were misclassified.

- **Selecting the Optimal Classification Algorithm**

The following table shows the performance

FIG. 6: RESULTS OF THE NAIVE BAYES ALGORITHM

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         206               68.6667 %
Incorrectly Classified Instances        94               31.3333 %
Kappa statistic                          0.3662
Mean absolute error                      0.3162
Root mean squared error                  0.4167
Relative absolute error                 68.5794 %
Root relative squared error             86.8041 %
Total Number of Instances              300
```

Source: Outputs of the Weka Software

TABLE 1: EVALUATION METRICS FOR THE THREE ALGORITHMS

INDICATOR	J48 ALGORITHM	BAYESIAN NETWORK ALGORITHM	NAIVE BAYES ALGORITHM
Sensitivity	0.810	0.803	0.687
False Positive Rate	0.241	0.265	0.294
Accuracy	0.808	0.801	0.719
F-Measure	0.808	0.799	0.693
MCC	0.581	0.563	0.378
ROC Area	0.842	0.863	0.819
PRC Area	0.825	0.875	0.832

Source: Prepared by the researcher based on Weka software.

metrics for the three algorithms, based on which we determined the optimal algorithm (see Table 1):

The J48 Decision Tree algorithm achieved the highest classification accuracy, making it the best choice for customer classification in this case.

- **Decision Tree Diagram**

Below is the decision tree diagram obtained using the J48 algorithm (see Fig. 7).

- **Predicting New Cases of Prospective Loan Customers**

To verify the quality of the obtained model, we tested it on five customer cases with known outcomes, and the results were as follows:

Figure 07 represents the decision tree obtained from the J48 classification algorithm, which will be used to predict the status of new loan applicants.

The figure illustrates the decision tree followed by decision-makers in the institution under study to determine or predict customer behavior. The tree considers all cases, and when applied to the cases shown in the following table, the results are as follows (see Table 2).

From the table, we have five out-of-sample cases. We predicted each case's behavior based on the J48 decision tree, and almost all results matched the actual classifications except for the first case. In that instance, the true classification categorized the customer as "Poor," while the de-

TABLE 2: PREDICTION RESULTS

PREVIOUS LOAN STATUS	LOAN PURPOSE	MONTHLY INCOME	GENDER	LOAN AMOUNT	AGE	HOUSING STATUS	LOAN PAYMENT TERM (YEARS)	ACTUAL CLASSIFICATION	PREDICTED CLASSIFICATION
Has outstanding payments	Motorcycle purchase	25,000	Male	340,000	28	Homeowner	3	Poor	Good
Payment delays	Home appliances	25,000	Female	280,000	30	Rented	4	Poor	Poor
Fully paid	Home appliances	60,000	Female	640,000	39	Rented	4	Good	Good
No previous loans	Home renovation	59,000	Male	890,000	60	Homeowner	5	Good	Good
Has outstanding payments	Home renovation	82,000	Male	600,000	54	Free	6	Poor	Poor

Source: Prepared by the researcher using J48 Decision Tree

cision tree classified them as “Good”.

4.2. Analysis of Study Results

Three different classification algorithms were applied: the J48 decision tree algorithm, the Naive Bayes algorithm, and the Bayesian Network algorithm. These were implemented using Weka software, with a sample of 300 customers for whom all relevant variables were available and whose loan behavior was known. Classification was based on whether the customer was “Good” or “Bad” for the loan. The model was trained using cross-validation, and the results were as follows:

Using the correct classification rate as the evaluation criterion, J48 Decision Tree had the highest accuracy at 81%. To evaluate this model’s performance, the indicators were as follows:

Kappa Statistic: The kappa coefficient was 0.58, indicating moderate agreement between predicted and actual classifications, suggesting a reliable classification model, though some error is present.

Correct Classification Rate: The model’s accuracy reached 81%, equating to 243 correctly classified customers, as indicated by the sensitivity metric, which reflects the model’s capability to detect correct classifications.

Incorrect Classification Rate: The error rate was 19%, corresponding to 57 misclassified customers.

False Positive Rate (FP Rate): The FP rate was 0.24, meaning that 24% of negative cases were incorrectly classified as positive.

Precision: The model’s precision reached 0.808, as reflected in the F-measure, indicating that 80.8% of cases predicted as positive were indeed positive, while 19.2% were incorrectly classified as positive.

ROC Area: The ROC area was 0.842, indicating a reasonable ability of the model to distinguish between positive and negative cases, with an 84.2% probability that a randomly selected positive case would be correctly classified over a negative one.

PRC Area: With a PRC area of 0.825, the model demonstrated a good balance between precision and recall for distinguishing positive and negative cases.

The final form of the decision tree comprised 36 nodes with 21 leaves. The root node was represented by the “Credit History” variable, indicating its primary influence on loan eligibility decisions.

We used this model to predict the behavior of loan-seeking customers with a sample of 5 customers outside the previously used dataset. The results were satisfactory, as 4 out of the five customers were classified correctly.

CONCLUSION

This methodology contributes to the early prediction of the behavior of loan-seeking customers in the banking institution, providing decision-makers, such as customer advisors, with a preliminary understanding of the individual’s behavior. This, in turn, enhances and rationalizes decision-making. It allows decision-makers to reject cases where indications suggest that the customer might be unsuitable for the loan, thereby reducing the risk of dealing with clients who are unable to repay loans and improving the overall quality of the bank’s clientele.

BIBLIOGRAPHY:

1. Adedipe, T., Shafiee, M., Zio, E. (2020). Bayesian network modeling for the wind energy industry: An overview. *Reliability Engineering & System Safety, 202*, 107053. <https://doi.org/10.1016/j.ress.2020.107053>.
2. Adnan, M., Sarno, R., Sungkono, K. R. (2019, September). Sentiment analysis of restaurant reviews with a classification approach in the decision tree-J48 algorithm. In *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*. IEEE. <https://doi.org/10.1109/ISEMANTIC.2019.8884282>.
3. Arowolo, M. O., Adeniyi, O. F., Adebiyi, M. O., Ogundokun, R. O. (2022). A prediction model for bank loans using agglomerative hierarchical clustering with a classification approach. *Covenant Journal of Informatics and Communication Technology, 10*(2).

4. Bahari, T. F., & Elayidom, M. S. (2015). An efficient CRM-data mining framework for the prediction of customer behavior. *Procedia Computer Science, 46*. <https://doi.org/10.1016/j.procs.2015.02.136>.
5. Bhargava, N., Sharma, G., Bhargava, R., Mathuria, M. (2013). Decision tree analysis on J48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, 3*(6).
6. Jain, A., Somwanshi, D., Joshi, K., Bhatt, S. S. (2022, April). A review: Data mining classification techniques. In *2022, the 3rd International Conference on Intelligent Engineering and Management (ICIEM)*. IEEE. <https://doi.org/10.1109/ICIEM54221.2022.9853036>.
7. Kaur, G., Oberai, E. N. (2014). A review article on Naive Bayes classifier with various smoothing techniques. *International Journal of Computer Science and Mobile Computing, 3*(10).
8. Kulkarni, E. G., Kulkarni, R. B. (2016). Weka: Powerful tool in data mining. *International Journal of Computer Applications, 975*(8887).
9. Larranaga, P., Karshenas, H., Bielza, C., Santana, R. (2013). A review on evolutionary algorithms in Bayesian network learning and inference tasks. *Information Sciences*. <https://doi.org/10.1016/j.ins.2012.12.051>.
10. Lee, C. S., Cheang, P. Y. S., Moslehpour, M. (2022). Predictive analytics in business analytics: Decision tree. *Advances in Decision Sciences, 26*(1).
11. Rahman, A., Khan, M. N. A. (2018). A classification-based model to assess customer behavior in the banking sector. *Engineering, Technology & Applied Science Research, 8*(3).
12. Thuraisingham, B., Maning, D. (1999). Technologies, techniques, tools, and trends. CRC Press.
13. Wickramasinghe, I., Kalutarage, H. (2021). Naive Bayes: Applications, variations, and vulnerabilities: A review of the literature with code snippets for implementation. *Soft Computing, 25*(3). <https://doi.org/10.1007/s00500-020-05297-6>.
14. Zhou, X., Bargshady, G., Abdar, M., Tao, X., Gururajan, R., Chan, K. C. (2019, October). A case study of predicting banking customers' behavior by using data mining. In *2019, the 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC)*. IEEE. <https://doi.org/10.1109/BESC48373.2019.8963436>.